

# Vícerozměrné statistické metody

Podobnosti a vzdálenosti ve vícerozměrném prostoru, asociační matice I

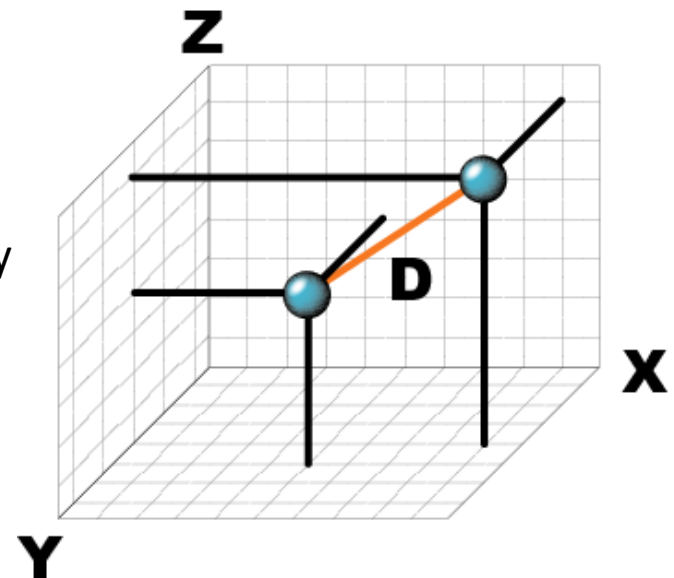
Jiří Jarkovský, Simona Littnerová

# Vícerozměrné statistické metody

Princip využití vzdáleností ve vícerozměrném prostoru

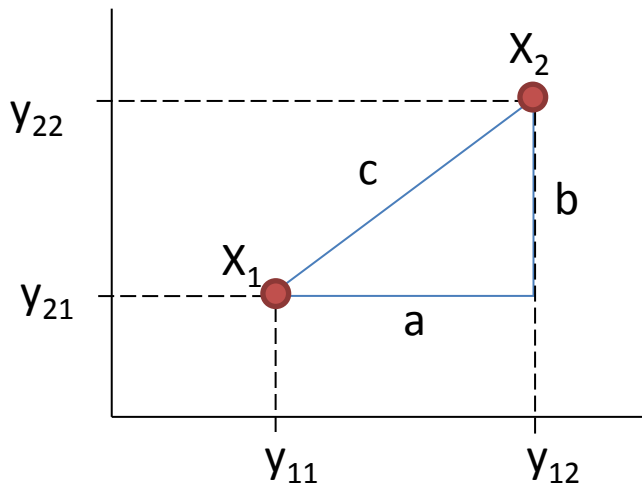
# Vzdálenosti nebo podobnosti objektů ve vícerozměrném prostoru

- Vícerozměrný popis objektů představuje jejich pozici ve vícerozměrném prostoru
- Vztahy mezi objekty lze vyjádřit pomocí jejich vzdálenosti v prostoru
- Existuje celá řada způsobů měření vzdálenosti v prostoru pro různé typy dat (binární, kategoriální, spojitá)
- Výběr metriky vzdálenosti nebo podobnosti silně ovlivňuje výsledky analýzy, protože definuje jakým způsobem vztah mezi objekty interpretujeme
- Výběr metriky je dán dvěma pohledy:
  - Typ dat – s různými typy dat jsou spjaty různé metriky
  - Předpoklady výpočtu metriky – obdobně jako klasické statistické metody ani metriky nelze použít ve všech situacích a v některých by dokonce díky jejich předpokladům šlo o hrubou chybu
  - Expertní interpretace vztahů objektů



# Euklidovská vzdálenost jako princip výpočtu vícerozměrných analýz

- Nejsnáze představitelným měřítkem vztahu dvou objektů ve vícerozměrném prostoru je jejich vzdálenost
- Nejjednodušším typem této vzdálenosti (bohužel s omezeným použitím na data společenstev) je Euklidovská vzdálenost vycházející z Pythagorovy věty



$$D_1(x_1, x_2) = \sqrt{\sum_{j=1}^p (y_{1j} - y_{2j})^2}$$

# Různé přístupy k měření vzdálenosti

Jednou na Manhattanu .....



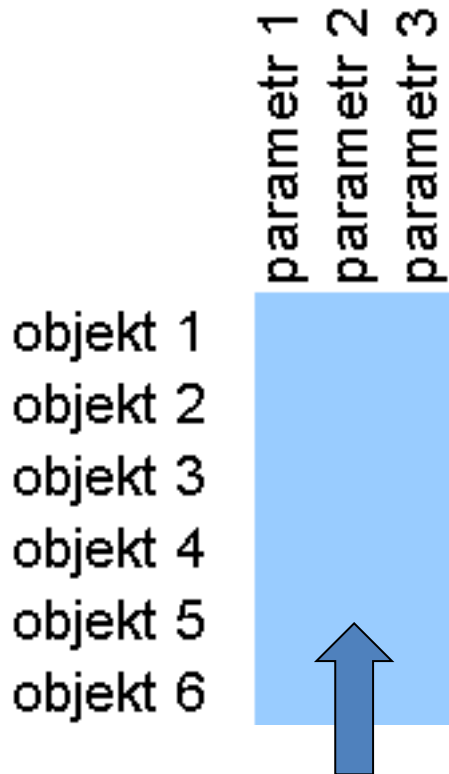
A

B



# Asociační matice

## NxP MATICE

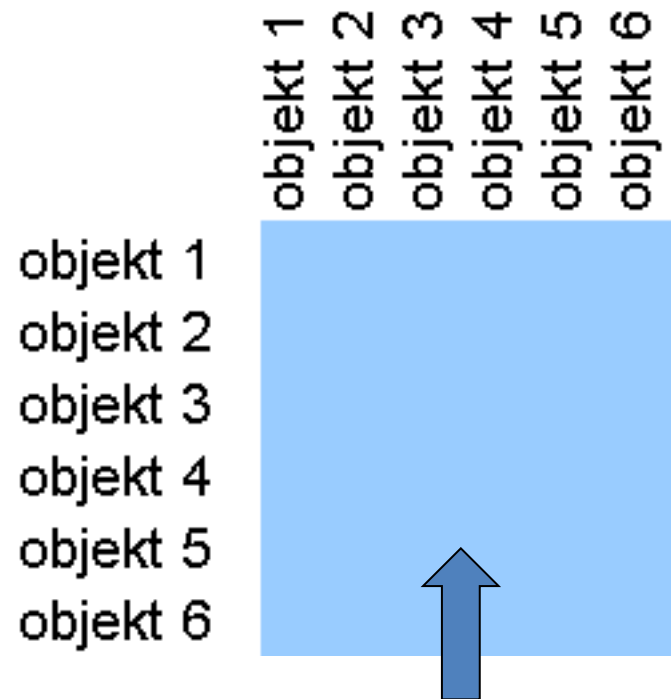


Hodnoty parametrů pro jednotlivé objekty

Výpočet metriky  
podobností/  
vzdáleností

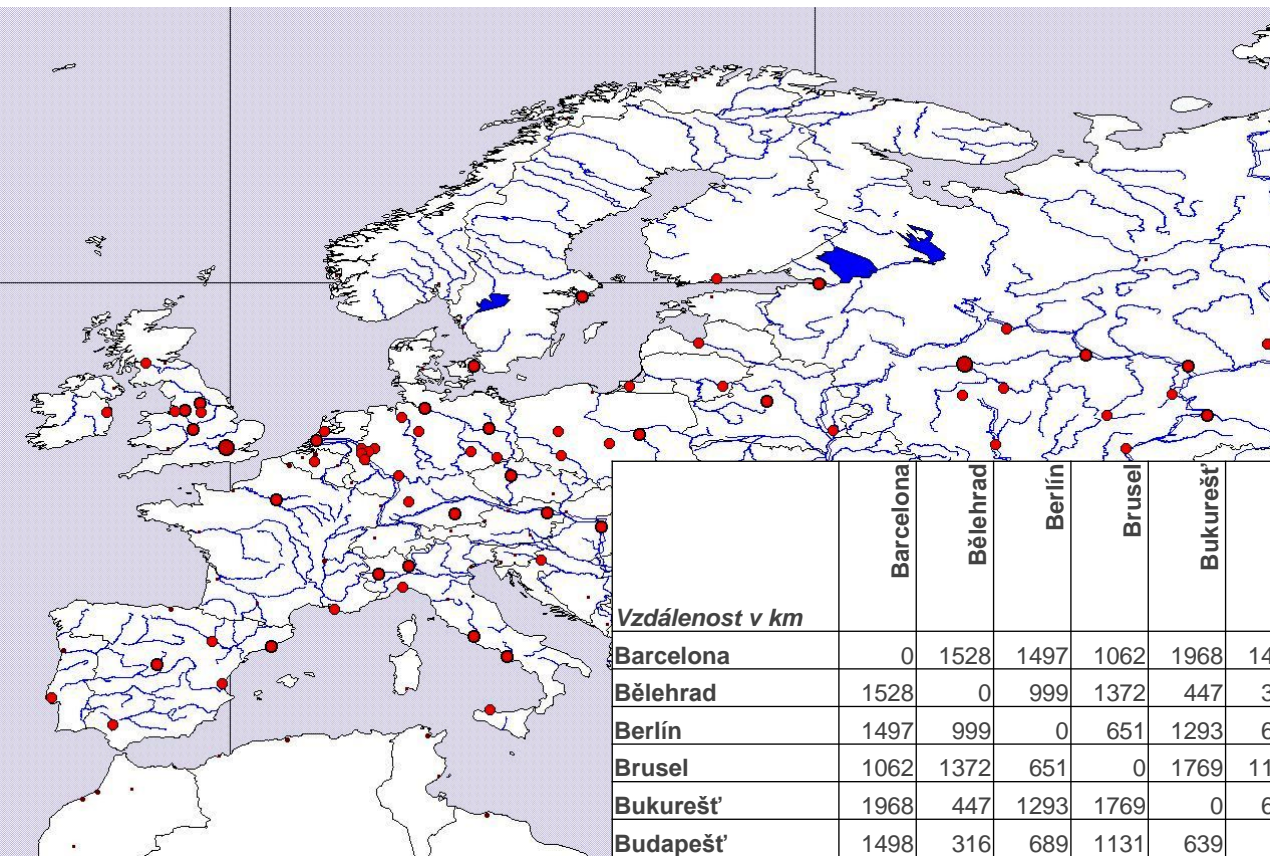


## ASOCIAČNÍ MATICE



Korelace, kovariance, vzdálenost, podobnost

# Mapa prostoru



Vzdálenost měst v mapě není ničím jiným než maticí vzdálenosti v 2D prostoru

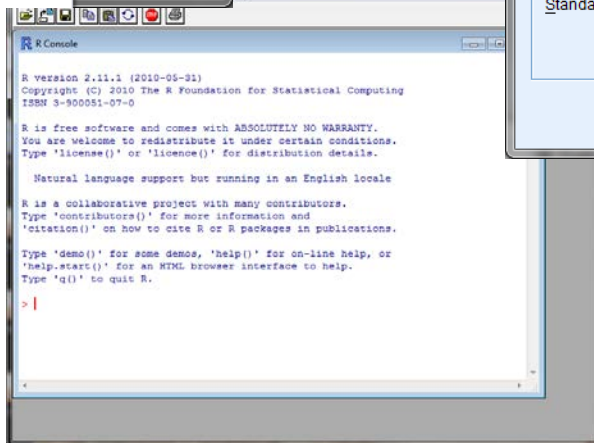
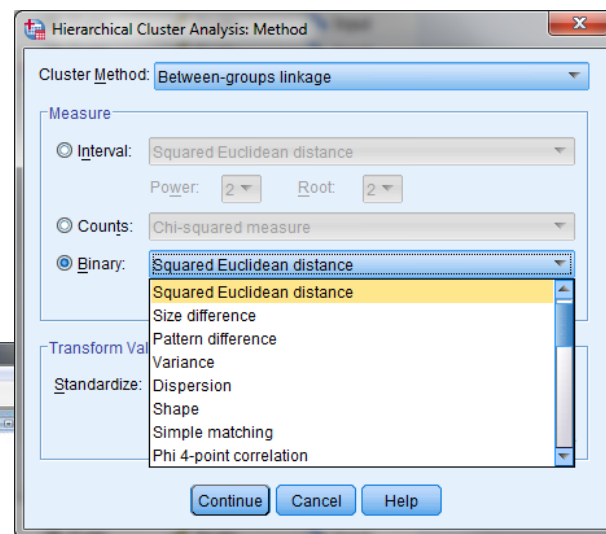
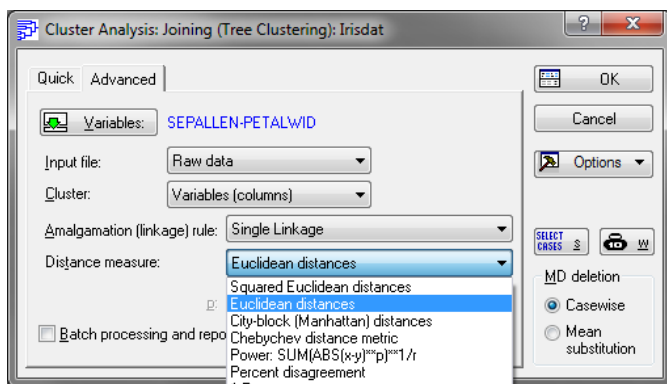
	Barcelona	Bělehrad	Berlín	Brusel	Bukurešť	Budapešť	Kodaň	Dublin	Hamburg	Istanbul	Kiev	Londýn	Madrid
<i>Vzdálenost v km</i>													
Barcelona	0	1528	1497	1062	1968	1498	1757	1469	1471	2230	2391	1137	504
Bělehrad	1528	0	999	1372	447	316	1327	2145	1229	809	976	1688	2026
Berlín	1497	999	0	651	1293	689	354	1315	254	1735	1204	929	1867
Brusel	1062	1372	651	0	1769	1131	766	773	489	2178	1836	318	1314
Bukurešť	1968	447	1293	1769	0	639	1571	2534	1544	445	744	2088	2469
Budapešť	1498	316	689	1131	639	0	1011	1894	927	1064	894	1450	1975
Kodaň	1757	1327	354	766	1571	1011	0	1238	287	2017	1326	955	2071
Dublin	1469	2145	1315	773	2534	1894	1238	0	1073	2950	2513	462	1449
Hamburg	1471	1229	254	489	1544	927	287	1073	0	1983	1440	720	1785
Istanbul	2230	809	1735	2178	445	1064	2017	2950	1983	0	1052	2496	2734
Kiev	2391	976	1204	1836	744	894	1326	2513	1440	1052	0	2131	2859
Londýn	1137	1688	929	318	2088	1450	955	462	720	2496	2131	0	1263
Madrid	504	2026	1867	1314	2469	1975	2071	1449	1785	2734	2859	1263	0

# Metrika vzdálenosti/podobnosti jako klíčový bod vícerozměrné analýzy

- Výběr metriky vzdálenosti/podobnosti je klíčovým bodem každé vícerozměrné analýzy:
  - Některé metody umožňují úplnou volnost ve výběru metriky podobnosti (hierarchická aglomerativní shluková analýza, multidimensional scaling)
  - Některé metody jsou přímo spjaté s konkrétní metrikou (PCA, CA, k-means clustering)
- Chybný výběr metriky může vést k chybným závěrům analýzy (stejně jako v klasické statistické analýze výběr nevhodného testu nebo popisné statistiky)
- Metriky podobností nebo vzdáleností kromě vícerozměrných statistických metod mohou vstupovat i do klasických statistických výpočtů:
  - Popisná statistika a vizualizace metrik
  - Analogie t-testů a ANOVA pro asociační matice
  - Korelace asociačních matic
  - Regrese asociačních matic

# Software pro výpočet metrik podobnosti/vzdálenosti

- Různé SW obsahují různé typy metrik
  - Statistica – velmi omezený seznam
  - SPSS – velké množství metrik
  - R – jakékoliv metriky, potřeba nainstalování knihoven



# Vícerozměrné statistické metody

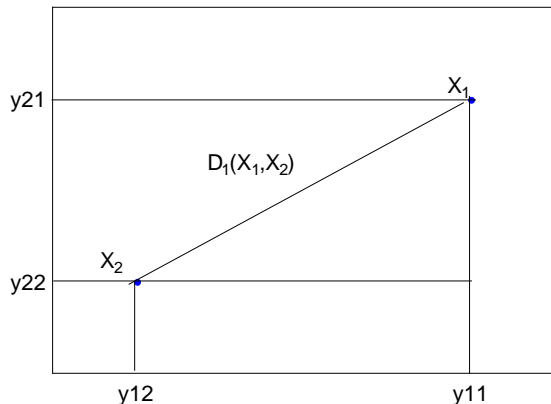
Kvantitativní metriky vzdáleností a podobností

# Euklidovská vzdálenost

- Jde o základní metrické měřítko vzdálenosti a počítá vzdálenost objektů obdobně jako Pythagorova věta počítá přeponu pravoúhlého trojúhelníku. Metoda je citlivá na rozdílný rozsah hodnot vstupujících proměnných (vhodným řešením může být standardizace) a double zero problém. Nemá horní hranici hodnot.

$$D_1(x_1, x_2) = \sqrt{\sum_{j=1}^p (y_{1j} - y_{2j})^2}$$

- Jako další měřítko se používá také čtverec této vzdálenosti. Jeho nevýhodou jsou semimetrické vlastnosti.



$$D_1^2(x_1, x_2) = \sum_{j=1}^p (y_{1j} - y_{2j})^2$$

# Průměrná vzdálenost

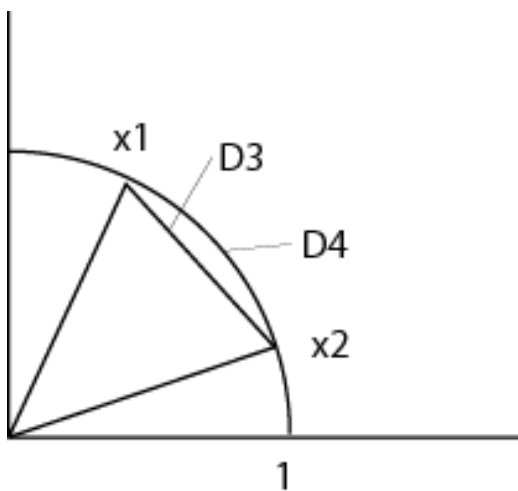
- Euklidovská vzdálenost je přepočítána na počet parametrů (druhů v případě vzdálenosti společenstev odběrů).

$$D_2^2(x_1, x_2) = \frac{1}{p} \sum_{j=1}^p (y_{1j} - y_{2j})^2$$

$$D_2(x_1, x_2) = \sqrt{D_2^2}$$

# Chord distance (Orlóci, 1967)

- Odstraňuje double zero problém a vliv rozdílného počtu jedinců druhů ve vzorcích při výpočtu Euklidovské vzdálenosti. Její maximální hodnota je druhá odmocnina ze dvou a minimum 0. Při výpočtu počítá pouze s poměry druhů v rámci jednotlivých vzorků. Jde vlastně o Euklidovskou vzdálenost počítanou pro vektory vzorků standardizované na délku 1, nebo je možný přímý výpočet už zahrnující standardizaci. Vnitřní část výpočtu je vlastně cosinus úhlu svíraného vektory, zápis vzorce je možný i v této formě.

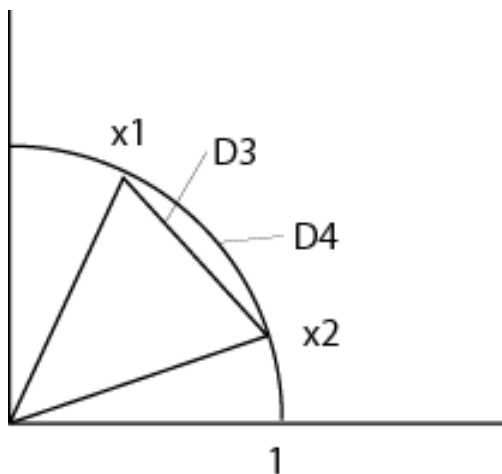


$$D_3(x_1, x_2) = \sqrt{2 \left( 1 - \frac{\sum_{j=1}^p y_{1j} y_{2j}}{\sqrt{\sum_{j=1}^p y_{1j}^2 \sum_{j=1}^p y_{2j}^2}} \right)}$$

$$D_3 = \sqrt{2(1 - \cos \theta)}$$

# Geodetická metrika

- Počítá délku výseče jednotkové kružnice mezi normalizovanými vektory (viz. Chord distance).



$$D_4(x_1, x_2) = \arccos \left[ 1 - \frac{D_3^2(x_1, x_2)}{2} \right]$$

# Mahalanobisova vzdálenost (Mahalanobis 1936)

- Jde o obecné měřítko vzdálenosti beroucí v úvahu korelaci mezi parametry a je nezávislá na rozsahu hodnot parametrů. Počítá vzdálenost mezi objekty v systému souřadnic jehož osy nemusí být na sebe kolmé. V praxi se používá pro zjištění vzdálenosti mezi skupinami objektů. Jsou dány dvě skupiny objektů  $w_1$  a  $w_2$  o  $n_1$  a  $n_2$  počtu objektů a popsané  $p$  parametry:

$$D_5^2(w_1, w_2) = \overline{d}_{12} V^{-1} \overline{d}_{12}$$

- Kde  $\overline{d}_{12}$  je vektor o délce  $p$  rozdílů mezi průměry  $p$  parametrů v obou skupinách.  $V$  je vážená disperzní matice (matice kovariancí parametrů) uvnitř skupin objektů.

$$V = \frac{1}{n_1 + n_2 - 2} [(n_1 - 1)S_1 + (n_2 - 1)S_2]$$

- kde  $S_1$  a  $S_2$  jsou disperzní matice jednotlivých skupin. Vektor  $\overline{d}_{12}$  měří rozdíl mezi  $p$ -rozměrnými průměry skupin a  $V$  vkládá do rovnice kovarianci mezi parametry.

# Minkowskeho metrika

- Je obecnou formou výpočtu vzdálenosti – podle zadaného koeficientu může odpovídat např. Euklidovské nebo Manhattanské metrice. Se stoupající koeficientem umocňování stoupá významnost větších rozdílů. Existuje ještě obecnější forma, kdy koeficient umocňování a odmocňování je zadáván zvlášť.

$$D_r(x_1, x_2) = \left[ \sum_{j=1}^p |y_{1j} - y_{2j}|^r \right]^{1/r}$$

# Manhattanská vzdálenost

- Jde vlastně o součet rozdílů jednotlivých parametrů popisujících objekty

$$D_7(x_1, x_2) = \sum_{j=1}^p |y_{1j} - y_{2j}|$$

# Mean character difference (Czekanowski 1909)

- Manhattanská vzdálenost přepočítaná na počet parametrů.

$$D_8(x_1, x_2) = \frac{1}{p} \sum_{j=1}^p |y_{1j} - y_{2j}|$$

# Whittakerův asociační index (Whittaker 1952)

- Je dobře použitelný pro data abundancí, každý druh je nejprve transformován ve svůj podíl ve společenstvu, následující výpočet je opět obdobou Manhattanské vzdálenosti.

$$D_9(x_1, x_2) = \frac{1}{2} \sum_{j=1}^p \left| \frac{y_{1j}}{\sum_{j=1}^p y_{ij}} - \frac{y_{2j}}{\sum_{j=1}^p y_{2j}} \right|$$

- Jeho hodnota je 0 v případě identických proporcí druhů. Stejný výsledek lze získat i jako součet nejmenších podílů v rámci obou vzorků.

$$D_9(x_1, x_2) = \left[ 1 - \min \left( \frac{y_j}{\sum_{j=1}^p y_j} \right) \right]$$

# Canberra metric (Lance & Williams 1966)

- Varianta Manhattanské vzdálenosti (před výpočtem musí být odstraněny double zero a není jím tedy ovlivněna). Stejný rozdíl mezi početnými druhy ovlivňuje vzdálenost méně než mezi druhy vzácnějšími.

$$D_{10}(x_1, x_2) = \sum_{j=1}^p \left[ \frac{|y_{1j} - y_{2j}|}{(y_{1j} + y_{2j})} \right]$$

- Stephenson et al. (1972) a Moreau & Legendre (1979) použili tuto metriku jako součást koeficientu podobnosti

$$S(x_1, x_2) = 1 - \frac{1}{p} D_{10}$$

# Koeficient divergence

- Obdobná metrika jako D10 ale založená na Euklidovské vzdálenosti a vztažená na počet parametrů.

$$D_{11}(x_1, x_2) = \sqrt{\frac{1}{p} \sum_{j=1}^p \left( \frac{y_{1j} - y_{2j}}{y_{1j} + y_{2j}} \right)^2}$$

# Coefficient of racial likeness (Pearson 1926)

- Umožňuje srovnávat skupiny objektů podobně jako Mahalanobisova vzdálenost, ale na rozdíl od ní neeliminuje vliv korelace parametrů. Dvě skupiny objektů  $w_1$  a  $w_2$  jsou charakterizovány  $\bar{y}_{ij}$  (průměr parametrů ve skupinách) a  $s_{ij}^2$  (rozptyl parametrů ve skupinách).

$$D_{12}(w_1, w_2) = \sqrt{\frac{1}{p} \sum_{j=1}^p \frac{(\bar{y}_{1j} - \bar{y}_{2j})^2}{\left(\frac{s_{1j}^2}{n_1}\right) + \left(\frac{s_{2j}^2}{n_2}\right)}} - \frac{2}{p}$$

# $\chi^2$ metrika (Roux & Reyssac 1975)

- První ze skupiny metrik založených na  $\chi^2$  pro výpočet vzdáleností odběrů založených na abundancích druhů nebo jiných frekvenčních datech (nejsou přípustné žádné záporné hodnoty). Data původní matice abundancí/frekvencí  $Y$  jsou nejprve přepočítána do matice poměrných frekvencí (součty frekvencí v řádcích (odběry) jsou rovny 1). Jako dodatečné charakteristiky uplatňované při výpočtu jsou spočteny součty řádků  $y_{i+}$  a sloupců  $y_{+j}$  celé! matice  $n(i)$  odběrů  $\times$   $p(j)$  druhů.

$$Y = \begin{matrix} & \begin{bmatrix} y_{ij} \\ \vdots \\ y_{ij} \end{bmatrix} \\ \begin{bmatrix} y_{+j} \\ \vdots \\ y_{+j} \end{bmatrix} & \end{matrix} \rightarrow \begin{bmatrix} y_{ij} / y_{i+} \\ \vdots \\ y_{ij} / y_{i+} \end{bmatrix}$$

$$D(x_1, x_2) = \sqrt{\sum_{j=1}^p \left( \frac{y_{1j}}{y_{1+}} - \frac{y_{2j}}{y_{2+}} \right)^2}$$

- Výpočet odstraňuje problém double zero. Nejjednodušším výpočtem je obdoba Euklidovské vzdálenosti
- která je dále vážena součty jednotlivých druhů

$$D_{15}(x_1, x_2) = \sqrt{\sum_{j=1}^p \frac{1}{y_{+j}} \left( \frac{y_{1j}}{y_{1+}} - \frac{y_{2j}}{y_{2+}} \right)^2}$$

# $\chi^2$ vzdálenost (Lébart & Fénelon 1971)

- Výpočet je podobný  $\chi^2$  metrice, ale vážení je prováděno relativní četností řádku v matici místo jeho absolutního součtu, při výpočtu se užívá parametr  $y_{++}$  (celkový součet matice). Je využívána také při výpočtu vztahů řádků a sloupců kontingenční tabulky.

$$D_{16}(x_1, x_2) = \sqrt{\sum_{j=1}^p \frac{1}{y_{+j} / y_{++}} \left( \frac{y_{1j}}{y_{1+}} - \frac{y_{2j}}{y_{2+}} \right)^2} = \sqrt{y_{++}} \sqrt{\sum_{j=1}^p \frac{1}{y_{+j}} \left( \frac{y_{1j}}{y_{1+}} - \frac{y_{2j}}{y_{2+}} \right)^2}$$

# Hellingerova vzdálenost (Rao 1995)

- Koeficient související s D15 a D16.

$$D_{17}(x_1, x_2) = \sqrt{\sum_{j=1}^p \left[ \sqrt{\frac{y_{1j}}{y_{1+}}} - \sqrt{\frac{y_{2j}}{y_{2+}}} \right]^2}$$

# Vícerozměrné statistické metody

Symetrické binární koeficienty podobnosti

# Koeficienty podobosti (indexy podobnosti)

- Ve vícerozměrné analýze se využívá řada indexů podobnosti založených buď na přítomnosti/nepřítomnosti kategorií objektů

## Binární koeficienty podobnosti

		Společenstvo 1	
		1	0
Společenstvo 2	1	a	b
	0	c	d

a, b, c, d = počet případů, kdy souhlasí binární charakteristika společenstev 1 a 2  
 $a+b+c+d=p$

**Symetrické binární koeficienty** - není rozdíl mezi případem 1-1 a 0-0

**Asymetrické binární koeficienty** - rozdíl mezi případem 1-1 a 0-0

Více informací a další měření vzdáleností a podobností najdete v knize **LEGENDRE, P. & LEGENDRE, L. (1998). *Numerical ecology*. Elsevier Science BV, Amsterdam.**

# Simple matching coefficient (Sokal & Michener, 1958)

- Obvyklou metodou pro výpočet podobnosti mezi dvěma objekty je podíl počtu deskriptorů, které kódují objekt stejně, a celkového počtu deskriptorů. Při použití tohoto koeficientu předpokládáme, že není rozdíl mezi nastáním 0 a 1 u deskriptorů.

$$S_1(x_1, x_2) = \frac{a + d}{p}$$

# Rogers & Tanimoto koeficient (1960)

- Dává větší váhu rozdílům než podobnostem.

$$S_2(x_1, x_2) = \frac{a + d}{a + 2b + 2c + d}$$

# Sokal & Sneath (1963)

- Další čtyři navržené koeficienty obsahují double-zero, ale jsou navrženy tak, aby se snížil vliv double-zero:

$$S_3(x_1, x_2) = \frac{2a + 2d}{2a + b + c + 2d}$$

- tento koeficient dává dvakrát větší váhu shodným deskriptorům než rozdílným;

$$S_4(x_1, x_2) = \frac{a + d}{b + c}$$

- porovnává shody a rozdíly prostým podílem v měřítku jdoucím od 0 do nekonečna;

$$S_5(x_1, x_2) = \frac{1}{4} \left[ \frac{a}{a+b} + \frac{a}{a+c} + \frac{d}{b+d} + \frac{d}{c+d} \right]$$

- porovnává shodné deskriptory se součty okrajů tabulky;

$$S_6(x_1, x_2) = \frac{a}{\sqrt{(a+b)(a+c)}} \frac{d}{\sqrt{(b+d)(c+d)}}$$

- je vytvořen z geometrických průměrů členů vztahujících se k  $a$  a  $d$ , podle koeficientu  $S_5$ .

# Hammannův koeficient

$$S = \frac{a + d - b - c}{p}$$

## Yuleho koeficient

$$S = \frac{ad - bc}{ad + bc}$$

## Pearsonovo $\Phi$ (phi)

$$\phi = \frac{ad - bc}{\sqrt{(a + b)(c + d)(a + c)(b + d)}}$$

# Vícerozměrné statistické metody

Kvantitativní asymetrické metriky podobnosti a vzdálenosti

# „Klasické“ indexy podobnosti

- Sørensenův kvantitativní koeficient, kde  $aN$  a  $bN$  jsou celkové počty jedinců v společenstvech A a B,  $jN$  je pak suma abundancí pokud se druh nachází v obou společenstvech, je počítána vždy z nižší abundance daného druhu ve společenstvu

$$C_N = \frac{2jN}{(aN + bN)}$$

- Morisita-Horn index, kde  $aN$  je celkový počet jedinců ve společenstvu A a  $an_i$  počet jedinců druhu  $i$  ve společenstvu A (obdobně platí pro společenstvo B)

$$C_{mH} = \frac{2\sum (an_i bn_i)}{(da + db).aN.bN}$$

$$da = \frac{\sum an_i^2}{aN^2}$$

# Jednoduchý srovnávací koeficient (Sokal & Michener, 1958)

- modifikovaný simple matching coefficient může být použit pro multistavové deskriptory - číselník obsahuje počet deskriptorů, pro které jsou dva objekty ve stejném stavu – např. je-li dvojice objektů popsána následujícími deseti multistavovými deskriptory: hodnota  $S_1$ , vypočítaná pro 10 multistavových deskriptorů bude  $S_1(x_1, x_2) = 4 \text{ agreements} / 10 \text{ descriptors} = 0.4$
- Podobným způsobem je možné rozšířit všechny binární koeficienty pro multistavové deskriptory.

$$S_1(x_1, x_2) = \frac{\text{agreements}}{p}$$

	Deskriptors										$\Sigma$
Object $x_1$	9	3	7	3	4	9	5	4	0	6	
Object $x_2$	2	3	2	1	2	9	3	2	0	6	
Agreements	0	+	+	+	+	+	+	+	+	+	4
		1	0	0	0	1	0	0	1	1	

# Gowerův obecný koeficient podobnosti (1971) I.

- Gower navrhl obecný koeficient podobnosti, který může kombinovat různé typy deskriptorů. Podobnost mezi dvěma objekty je vypočítána jako průměr podobností, vypočítaných pro všechny deskriptory. Pro každý deskriptor  $j$  je hodnota parciální podobnosti  $s_{12j}$  mezi objekty  $x_1$  a  $x_2$  vypočítána následovně:

$$S_{15}(x_1, x_2) = \frac{1}{p} \sum_{j=1}^p s_{12j}$$

- ✓ Pro binární deskriptory  $s_j=1$  (shoda) nebo 0 (neshoda). Gower navrhl dvě formy tohoto koeficientu. Následující forma je symetrická, dává  $s_j=1$  double-zero. Druhá forma, Gowerův asymetrický koeficient  $S_{19}$  dává pro double-zero  $s_j=0$
- ✓ Kvalitativní a semikvantitativní deskriptory jsou upraveny podle jednoduchého zaměřovacího pravidla,  $s_j=1$  při souhlasu a  $s_j = 0$  při nesouhlasu deskriptorů. Double zero jsou ošetřeny stejně jako v předchozím odstavci.
- ✓ Kvantitativní deskriptory (reálná čísla) jsou zpracovány následovně: pro každý deskriptor se nejprve vypočte rozdíl mezi stavy obou objektů který je poté vydělen největším rozdílem ( $R_j$ ), nalezeným pro daný deskriptor mezi všemi objekty ve studii (nebo v referenční populaci – doporučuje se vypočítat největší diferenci  $R_j$  každého deskriptoru  $j$  pro celou populaci, aby byla zajištěna konzistence výsledků pro všechny parciální studie).

# Gowerův obecný koeficient podobnosti (1971) II.

- normalizovaná vzdálenost může být odečtena od 1 aby byla transformována na podobnost:

$$s_{12j} = 1 - \left[ \frac{|y_{1j} - y_{2j}|}{R_j} \right]$$

- Gowerův koeficient může být nastaven tak, aby zahrnoval přídatný flexibilní prvek: žádné porovnání není vypočítáno u deskriptorů, u nichž chybí informace buď u jednoho, nebo u druhého objektu. Toto zajišťuje člen  $w_j$ , nazývaný Kroneckerovo delta, popisující přítomnost/nepřítomnost informace v obou objektech: je-li informace o deskriptoru  $y_j$  přítomna u obou objektů ( $w_j=1$ ), jinak ( $w_j=0$ ), tento koeficient nabývá hodnot podobnosti mezi 0 a 1 (největší podobnost objektů). Další možností je vážení různých deskriptorů prostým přiřazením čísla v rozsahu 0-1  $w_j$ .

$$S_{15}(x_1, x_2) = \frac{\sum_{j=1}^p w_{12j} s_{12j}}{\sum_{j=1}^p w_{12j}}$$

# Vícerozměrné statistické metody

Asymetrické binární koeficienty

# Jaccardův koeficient (1900, 1901, 1908)

- Všechny členy mají stejnou váhu

$$S_7(x_1, x_2) = \frac{a}{a + b + c}$$

# Sørensenův koeficient (1948) (Coincidence index, Dice(1945))

- varianta předchozího koeficientu dává dvojnásobnou váhu dvojitým prezencím , protože se může zdát, že přítomnost druhů je více informativní než jejich absence, která může být způsobena různými faktory a nemusí nutně odrážet rozdílnost prostředí. Prezence druhu na obou lokalitách je silným ukazatelem jejich podobnosti. S7 je monotónní k S8, proto podobnost pro dvě dvojice objektů vypočítaná podle S7 bude podobná stejnému výpočtu S8. Oba koeficienty se liší pouze v měřítku. Tento index byl poprvé použit Dicem v R-mode studii asociací druhů. Jiná varianta tohoto koeficientu dává duplicitním prezencím trojnásobnou váhu.

$$S_8(x_1, x_2) = \frac{2a}{2a + b + c}$$

$$S_8(x_1, x_2) = \frac{3a}{3a + b + c}$$

# Sokal & Sneath (1963)

- navržen jako doplněk Rogers & Tanimotova koeficientu (S2), dává dvojnásobnou váhu rozdílům ve jmenovateli.

$$S_{10}(x_1, x_2) = \frac{a + d}{a + 2b + 2c}$$

## Russel & Rao (1940)

- navržená míra umožňuje porovnání počtu duplicitních prezencí (v čitateli) proti celkovému počtu druhů, nalezených na všech lokalitách, zahrnujícím druhy, které chybějí ( $d$ ) na obou uvažovaných lokalitách.

$$S_{11}(x_1, x_2) = \frac{a}{p}$$

# Kulczynski (1928)

- koeficient porovnávající duplicitní prezence s diferencemi

$$S_{12}(x_1, x_2) = \frac{a}{b + c}$$

# Binární verze asymetrického kvantitativního Kulczynski koeficientu (1928)

- Mezi svými koeficienty pro presence/absence data zmiňují Sokal & Sneath (1963) tuto verzi kvantitativního koeficientu  $S_{18}$ , kde jsou duplicitní prezence srovnávány se součty okrajů tabulky  $(a+b)$  a  $(a+c)$ .

$$S_{13}(x_1, x_2) = \frac{1}{2} \left[ \frac{a}{a+b} + \frac{a}{a+c} \right]$$

# Ochiachi (1957)

- použil jako míru podobnosti geometrický průměr poměrů  $a$  k počtu druhů na každé lokalitě, tj. se součty okrajů tabulky  $(a+b)$  a  $(a+c)$ , tento koeficient je obdobou  $S_6$ , bez části, týkající se double-zero (d).

$$S_{14}(x_1, x_2) = \sqrt{\frac{a}{(a+b)} \frac{a}{(a+c)}} = \frac{a}{\sqrt{(a+b)(a+c)}}$$

# Faith (1983)

- V tomto koeficientu je neshoda (přítomnost na jedné a absence na druhé lokalitě) vážena proti duplicitní prezenci. Hodnota S26 klesá s růstem double-zero

$$S_{26}(x_1, x_2) = \frac{a + d/2}{p}$$

# Vícerozměrné statistické metody

Práce s asociační maticí

# Asociační matice

- Typická asociční matice je čtvercová matice
- Typická asociční matice je symetrická kolem diagonály
  - Ve speciálních případech existují i asymetrické asociční matice
- Diagonála obsahuje 0 (v případě vzdáleností) nebo identitu objektu se sebou samým (podobnosti, obvykle 1 nebo 100%)
- Asociční matice může být spočtena mezi objekty pomocí metrik podobnosti a vzdálenosti (Q mode analýza) nebo mezi proměnnými pomocí korelací a kovariancí (R mode analýza)
- Asociční matice mohou být jak vstupem do vícerozměrných analýz tak vstupem pro klasické jednorozměrné statistické výpočty, kdy základní jednotkou není jeden objekt, ale podobnost/vzdálenost dvojice objektů

# Příklad výpočtu asociační matice

STATISTICA - [Data: Irisdat\* (5v by 150c)]

File Edit View Insert Format Statistics Data Mining Graphs Tools Data Win

Arial 10 B I U

Fisher (1936) iris data: length & width of sepals and petals, 3 types of I

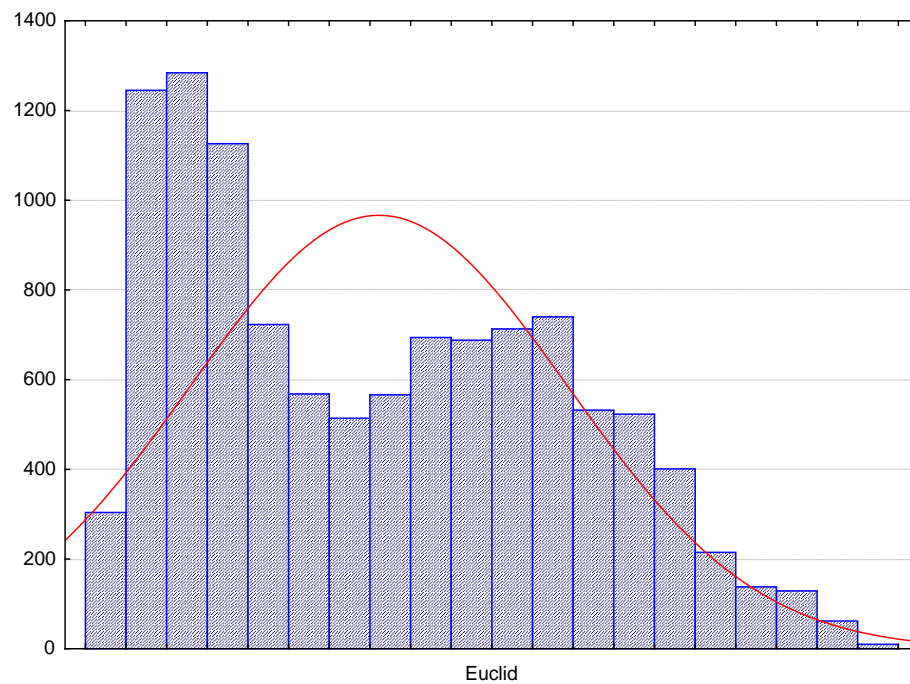
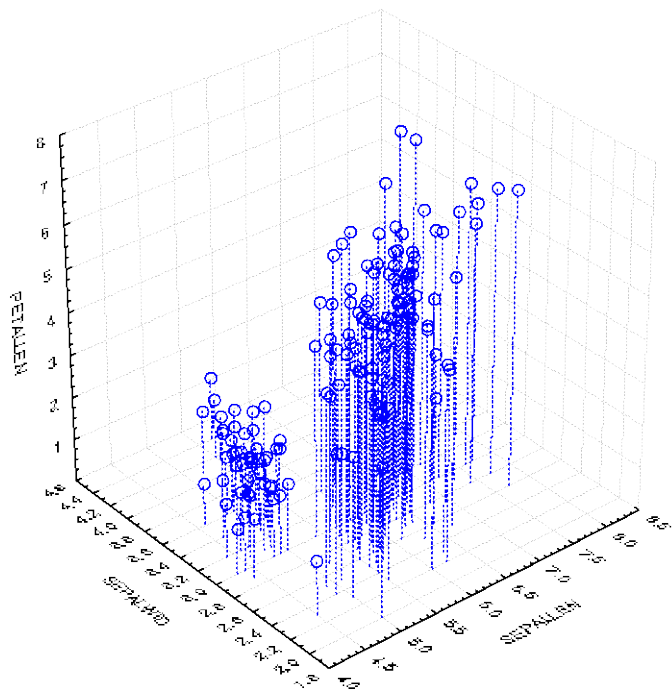
	1 SEPALLEN	2 SEPALWID	3 PETALLEN	4 PETALWID	5 IRISTYPE
1	5.0	3.3	1.4	0.2	SETOSA
2	6.4	2.8	5.6	2.2	VIRGINIC
3	6.5	2.8	4.6	1.5	VERSICO
4	6.7	3.1	5.6		
5	6.3	2.8	5.1		
6	4.6	3.4	1.4		
7	6.9	3.1	5.1	2.3	VIRGINIC
8	6.2	2.2	4.5	1.5	VERSICO
9	5.9	3.2	4.8	1.8	VERSICO
10	4.6	3.6	1.0	0.2	SETOSA
11	6.1	3.0	4.6	1.4	VERSICO
12	6.0	2.7	5.1	1.6	VERSICO
13	6.5	3.0	5.2	2.0	VIRGINIC
14	5.6	2.5	3.9	1.1	VERSICO
15	6.5	3.0	5.5	1.8	VIRGINIC
16	5.8	2.7	5.1	1.9	VIRGINIC
17	6.8	3.2	5.9	2.3	VIRGINIC
18	5.1	3.3	1.7	0.5	SETOSA
19	5.7	2.8	4.5	1.3	VERSICO
20	6.2	3.4	5.4	2.3	VIRGINIC
21	7.7	3.8	6.7	2.2	VIRGINIC
22	6.3	3.3	4.7	1.6	VERSICO
23	6.7	3.3	5.7	2.5	VIRGINIC
24	7.6	3.0	6.6	2.1	VIRGINIC
25	4.9	2.5	4.5	1.7	VIRGINIC
26	5.5	3.5	1.3	0.2	SETOSA
27	6.7	3.0	5.2	2.3	VIRGINIC
28	7.0	3.2	4.7	1.4	VERSICO
29	6.4	3.2	4.5	1.5	VERSICO
30	6.1	2.8	4.0	1.3	VERSICO
31	4.8	3.1	1.6	0.2	SETOSA
32	5.9	3.0	5.1	1.8	VIRGINIC
33	5.5	2.4	3.8	1.1	VERSICO
34	6.3	2.5	5.0	1.9	VIRGINIC

Euclidean distances (Irisdat)

Case No	C 1	C 2	C 3	C 4	C 5	C 6	C 7	C 8	C 9	C 10	C 11	C 12	C 13	C 14	C 15	C 16	C 17
C 1	0.00	4.88	3.80	5.04	4.16	0.42	4.66	3.73	3.87	0.64	3.60	4.12	4.47	2.84	4.66	4.19	5.28
C 2	4.88	0.00	1.22	0.47	0.87	4.98	0.77	1.45	1.10	5.39	1.33	0.88	0.50	2.20	0.47	0.84	0.65
C 3	3.80	1.22	0.00	1.39	0.54	3.96	1.07	0.68	0.81	4.35	0.46	0.72	0.81	1.24	0.97	0.95	1.61
C 4	5.04	0.47	1.39	0.00	1.14	5.15	0.55	1.75	1.28	5.54	1.54	1.24	0.61	2.48	0.65	1.21	0.35
C 5	4.16	0.87	0.54	1.14	0.00	4.29	1.04	0.85	0.71	4.69	0.58	0.33	0.58	1.48	0.57	0.65	1.30
C 6	0.42	4.98	3.96	5.15	4.29	0.00	4.80	3.88	3.94	0.46	3.72	4.22	4.59	2.95	4.78	4.26	5.40
C 7	4.66	0.77	1.07	0.55	1.04	4.80	0.00	1.52	1.16	5.17	1.31	1.21	0.52	2.22	0.76	1.24	0.81
C 8	3.73	1.45	0.68	1.75	0.85	3.88	1.52	0.00	1.13	4.30	0.82	0.81	1.21	0.98	1.35	0.96	1.99
C 9	3.87	1.10	0.81	1.28	0.71	3.94	1.16	1.13	0.00	4.34	0.53	0.62	0.77	1.37	0.94	0.60	1.51
C 10	0.64	5.39	4.35	5.54	4.69	0.46	5.17	4.30	4.34	0.00	4.12	4.64	4.98	3.38	5.17	4.69	5.78
C 11	3.60	1.33	0.46	1.54	0.58	3.72	1.31	0.82	0.53	4.12	0.00	0.62	0.94	1.04	1.06	0.82	1.74
C 12	4.12	0.88	0.72	1.24	0.33	4.22	1.21	0.81	0.62	4.64	0.62	0.00	0.71	1.37	0.73	0.36	1.42
C 13	4.47	0.50	0.81	0.61	0.58	4.59	0.52	1.21	0.77	4.98	0.94	0.71	0.00	1.89	0.36	0.77	0.84
C 14	2.84	2.20	1.24	2.48	1.48	2.95	2.22	0.98	1.37	3.38	1.04	1.37	1.89	0.00	2.03	1.47	2.71
C 15	4.66	0.47	0.97	0.65	0.57	4.78	0.76	1.35	0.94	5.17	1.06	0.73	0.36	2.03	0.00	0.87	0.73
C 16	4.19	0.84	0.95	1.21	0.65	4.26	1.24	0.96	0.60	4.69	0.82	0.36	0.77	1.47	0.87	0.00	1.43
C 17	5.28	0.65	1.61	0.35	1.30	5.40	0.81	1.99	1.51	5.78	1.74	1.42	0.84	2.71	0.73	1.43	0.00
C 18	0.44	4.48	3.41	4.63	3.77	0.62	4.25	3.36	3.46	0.96	3.21	3.73	4.07	2.47	4.26	3.79	4.88
C 19	3.40	1.58	0.83	1.87	0.87	3.49	1.70	0.81	0.73	3.91	0.47	0.74	1.29	0.71	1.39	0.86	2.08
C 20	4.68	0.67	1.32	0.62	1.05	4.75	0.82	1.70	0.86	5.14	1.27	1.05	0.62	2.20	0.71	0.95	0.81

Asociační matice euklidovských vzdáleností mezi rostlinami

# Histogram jako popis asociační matice



# Vztahy mezi různými metrikami vzdáleností

