

Vícerozměrné statistické metody

Podobnosti a vzdálenosti ve vícerozměrném prostoru, asociační matice II

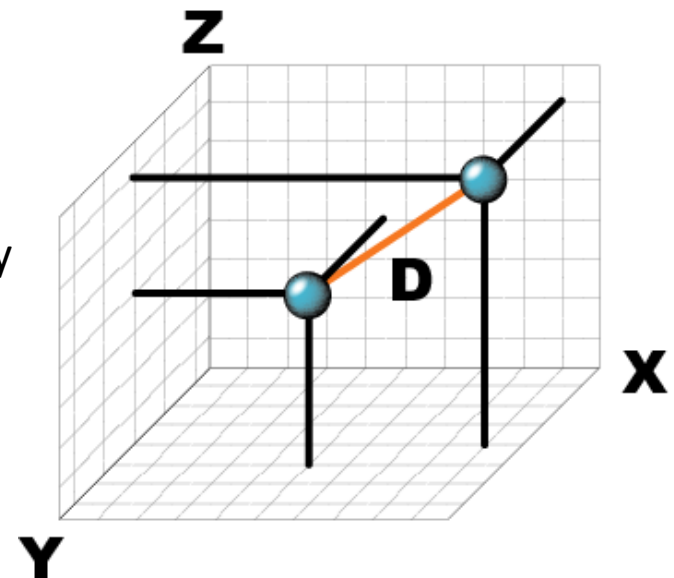
Jiří Jarkovský, Simona Littnerová

Vícerozměrné statistické metody

Práce s asociační maticí

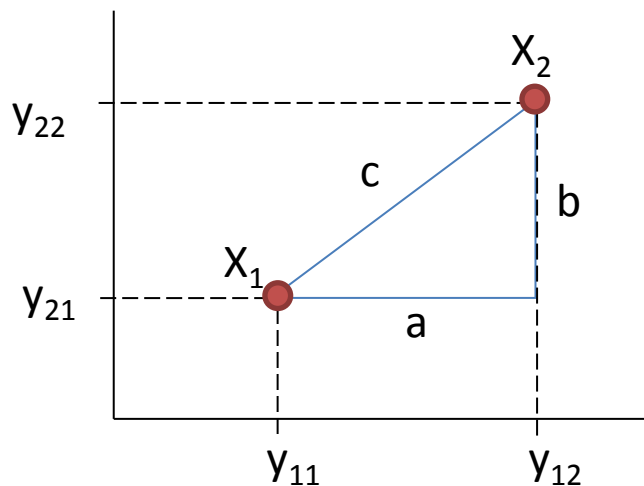
Vzdálenosti nebo podobnosti objektů ve vícerozměrném prostoru

- 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

- Typická asociační matice je čtvercová matice
- Typická asociační matice je symetrická kolem diagonály
 - Ve speciálních případech existují i asymetrické asociační matice
- Diagonála obsahuje 0 (v případě vzdáleností) nebo identitu objektu se sebou samým (podobnosti, obvykle 1 nebo 100%)
- Asociační matice může být spočtena mezi objekty pomocí metrik podobnosti a vzdálenosti (Q mode analýza) nebo mezi proměnnými pomocí korelací a kovariancí (R mode analýza)
- Asociační 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

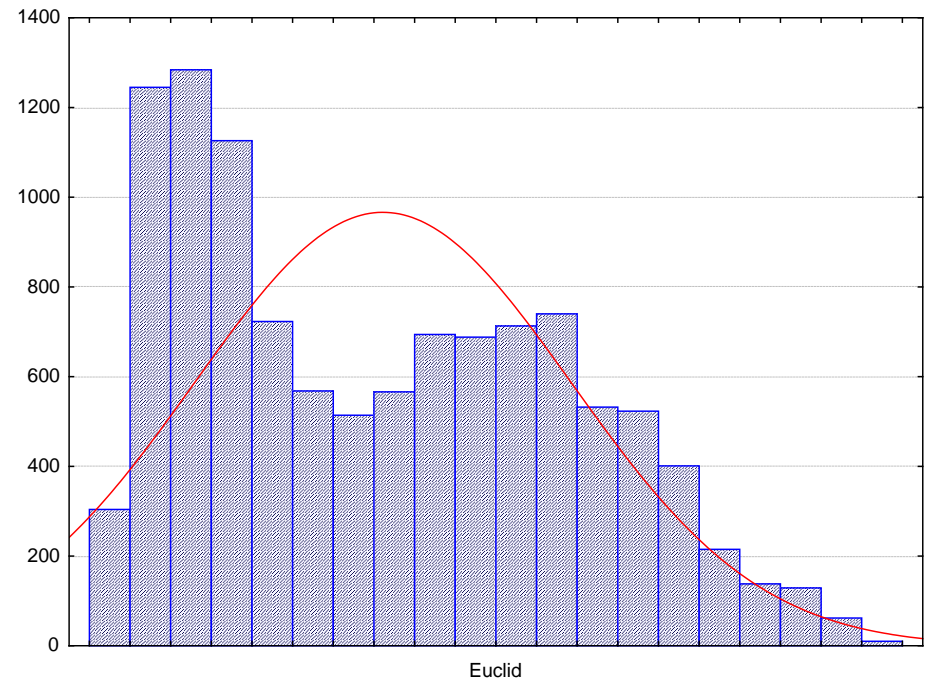
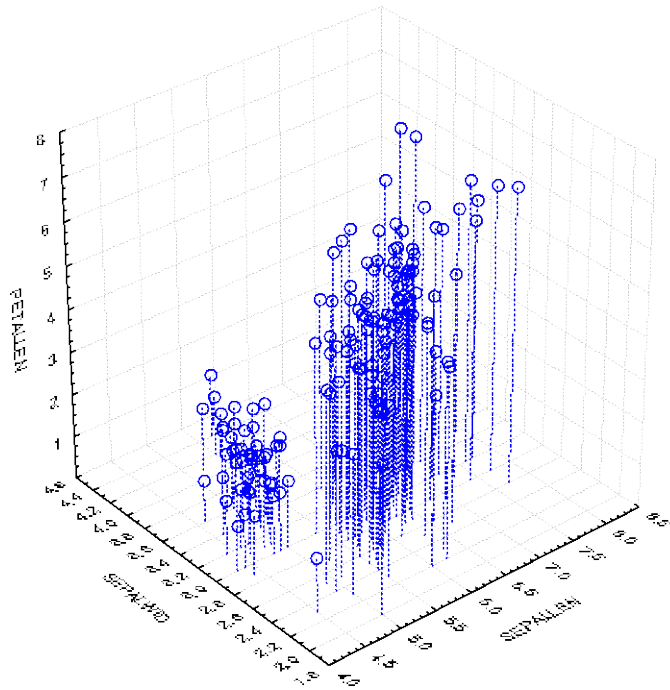
	1	2	3	4	5
	SEPALLEN	SEPALWID	PETALLEN	PETALWID	IRISTYPE
1	5.0	3.3	1.4	0.2	SETOSA
2	6.4	2.8	5.6	2.2	VIRGINIC
3	6.5	2.8	4.6	1.5	VERSICO
4	6.7	3.1	5.6		
5	6.3	2.8	5.1		
6	4.6	3.4	1.4		
7	6.9	3.1	5.1	2.3	VIRGINIC
8	6.2	2.2	4.5	1.5	VERSICO
9	5.9	3.2	4.8	1.8	VERSICO
10	4.6	3.6	1.0	0.2	SETOSA
11	6.1	3.0	4.6	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)

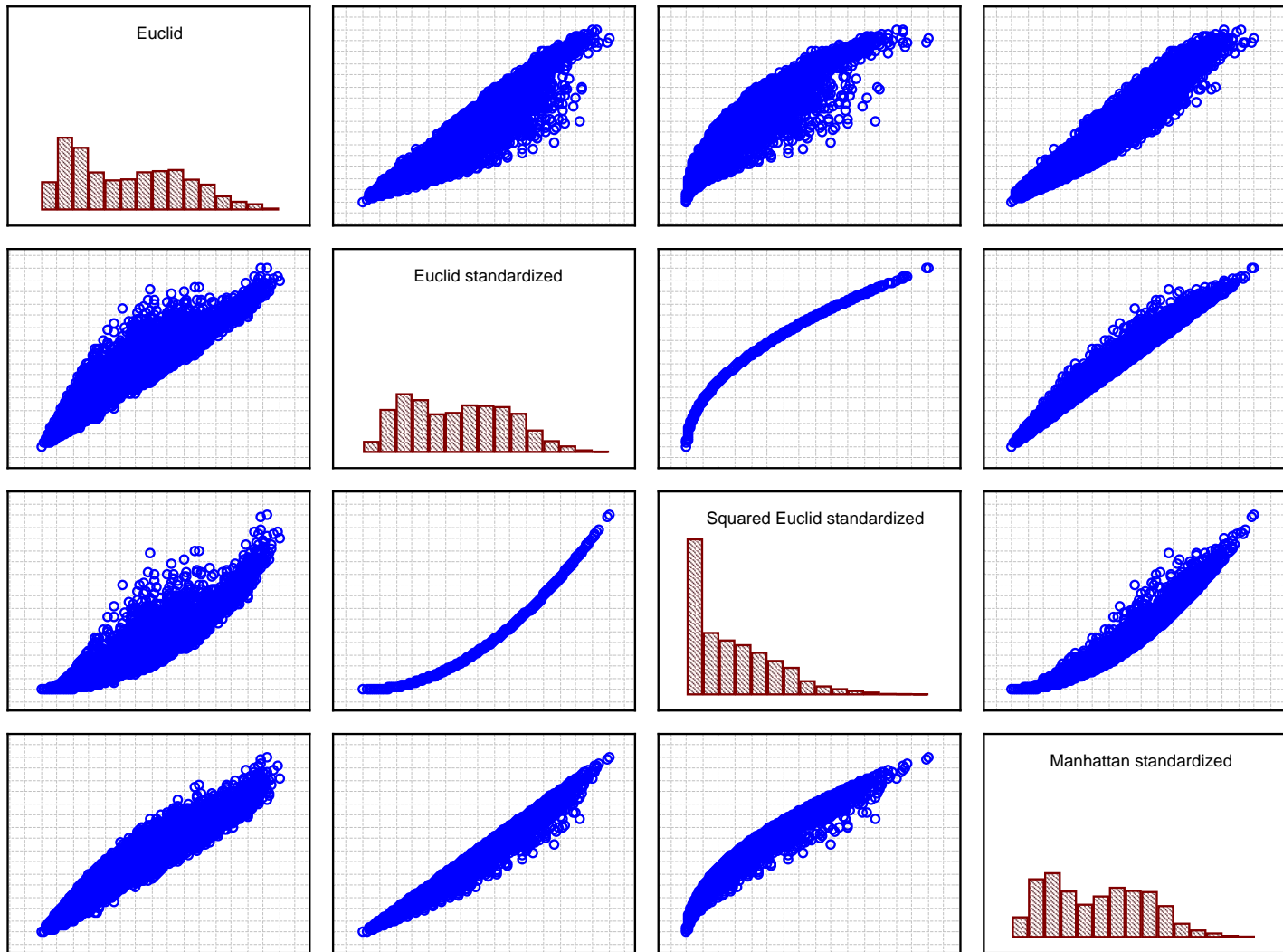
Case No.	C 1	C 2	C 3	C 4	C 5	C 6	C 7	C 8	C 9	C 10	C 11	C 12	C 13	C 14	C 15	C 16	C 17
C 1	0.00	4.88	3.80	5.04	4.16	0.42	4.66	3.73	3.87	0.64	3.60	4.12	4.47	2.84	4.66	4.19	5.28
C 2	4.88	0.00	1.22	0.47	0.87	4.98	0.77	1.45	1.10	5.39	1.33	0.88	0.50	2.20	0.47	0.84	0.65
C 3	3.80	1.22	0.00	1.39	0.54	3.96	1.07	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



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

Vícerozměrné statistické metody

Analogie klasických statistických metod s využitím asociačních matic

Klasické statistické metody na asociační matici

- Na datech asociačních koeficientů je možné počítat libovolné jednorozměrné statistické metody
- Je nezbytné zohlednit
 - 1 hodnota není jeden objekt, jde o vztah dvou objektů !!!
 - Hodnoty nejsou nezávislé !!!
 - Díky nesouladu mezi N hodnot a počtem stupňů volnosti není možné klasické statistické testování, ale je nezbytný permutační přístup
- Pro vizualizaci i výpočet statistik je možné použít klasické statistické SW
- Pro výpočet statistické významnosti a intervalů spolehlivosti je nezbytné použít specializovaný SW

Konverze asociační matice pro jednorozměrné analýzy

Similarita		Jaccard index																
	PL-VIS	GE-RHI	PL-SLE	CZ-ELO	CZ-ELV	CZ-KYJ	CZ-MOR	SK-DAN	IT-RMO	BG-DAN	FR-DUR	BG-ISK	BG-STR	GR-NES	TU-ESK	TU-BAL	TU-MAS	TU-KUR
PL-VIS	0.389	0.333	0.190	0.227	0.286	0.333	0.350	0.190	0.182	0.130	0.154	0.333	0.200	0.208	0.292	0.435	0.370	
GE-RHI	0.389	0.357	0.333	0.500	0.200	0.364	0.385	0.333	0.214	0.214	0.167	0.267	0.250	0.250	0.222	0.500	0.400	
PL-SLE	0.333	0.357	0.333	0.500	0.313	0.286	0.500	0.267	0.250	0.176	0.200	0.222	0.125	0.211	0.190	0.350	0.292	
CZ-ELO	0.190	0.333	0.357	0.800	0.125	0.250	0.385	0.231	0.214	0.133	0.105	0.188	0.154	0.111	0.158	0.263	0.217	
CZ-ELV	0.227	0.500	0.500	0.800	0.176	0.308	0.429	0.286	0.267	0.188	0.150	0.235	0.133	0.158	0.200	0.368	0.304	
CZ-KYJ	0.286	0.200	0.313	0.125	0.176	0.308	0.429	0.200	0.267	0.267	0.211	0.235	0.214	0.158	0.200	0.182	0.200	
CZ-MOR	0.333	0.364	0.286	0.250	0.308	0.308	0.308	0.154	0.143	0.143	0.111	0.200	0.167	0.118	0.167	0.211	0.174	
SK-DAN	0.350	0.385	0.500	0.385	0.429	0.429	0.308	0.385	0.188	0.357	0.278	0.235	0.214	0.222	0.263	0.300	0.300	
IT-RMO	0.190	0.333	0.267	0.231	0.286	0.200	0.154	0.385	0.133	0.417	0.313	0.357	0.364	0.250	0.294	0.263	0.263	
BG-DAN	0.182	0.214	0.250	0.214	0.267	0.267	0.143	0.188	0.133	0.200	0.375	0.250	0.143	0.400	0.353	0.316	0.316	
FR-DUR	0.130	0.214	0.176	0.133	0.188	0.267	0.143	0.357	0.417	0.200	0.294	0.176	0.231	0.167	0.353	0.190	0.208	
BG-ISK	0.154	0.167	0.200	0.105	0.150	0.211	0.111	0.278	0.313	0.375	0.294	0.500	0.176	0.471	0.421	0.261	0.320	
BG-STR	0.333	0.267	0.222	0.188	0.235	0.235	0.200	0.235	0.357	0.250	0.176	0.500	0.286	0.278	0.316	0.350	0.348	
GR-NES	0.200	0.250	0.125	0.154	0.133	0.214	0.167	0.214	0.143	0.231	0.176	0.286	0.267	0.313	0.211	0.174	0.174	
TU-ESK	0.208	0.250	0.211	0.111	0.158	0.158	0.118	0.222	0.250	0.400	0.167	0.471	0.278	0.267	0.444	0.400	0.333	
TU-BAL	0.292	0.222	0.190	0.158	0.200	0.200	0.167	0.263	0.294	0.353	0.353	0.421	0.316	0.313	0.444	0.364	0.360	
TU-MAS	0.435	0.500	0.350	0.263	0.368	0.182	0.211	0.300	0.263	0.316	0.190	0.261	0.350	0.211	0.400	0.364	0.565	
TU-KUR	0.370	0.400	0.292	0.217	0.304	0.200	0.174	0.304	0.273	0.318	0.208	0.320	0.348	0.174	0.333	0.360	0.565	

Jaccard	row	column
0.389	PL-VIS	GE-RHI
0.333	PL-VIS	PL-SLE
0.357	GE-RHI	PL-SLE
0.190	PL-VIS	CZ-ELO
0.333	GE-RHI	CZ-ELO
0.357	PL-SLE	CZ-ELO
0.227	PL-VIS	CZ-ELV
0.500	GE-RHI	CZ-ELV
0.500	PL-SLE	CZ-ELV
0.800	CZ-ELO	CZ-ELV
0.286	PL-VIS	CZ-KYJ
0.200	GE-RHI	CZ-KYJ
0.313	PL-SLE	CZ-KYJ
0.5	CZ-ELO	CZ-KYJ
0.176	CZ-ELV	CZ-KYJ
0.333	PL-VIS	CZ-MOR
0.364	GE-RHI	CZ-MOR
0.286	PL-SLE	CZ-MOR
0.250	CZ-ELO	CZ-MOR
0.308	CZ-ELV	CZ-MOR
0.308	CZ-KYJ	CZ-MOR
0.350	PL-VIS	SK-DAN
0.385	GE-RHI	SK-DAN
0.500	PL-SLE	SK-DAN
0.385	CZ-ELO	SK-DAN
0.429	CZ-ELV	SK-DAN
0.429	CZ-KYJ	SK-DAN
0.308	CZ-MOR	SK-DAN
0.190	PL-VIS	IT-RMO
0.333	GE-RHI	IT-RMO
0.267	PL-SLE	IT-RMO
0.231	CZ-ELO	IT-RMO
0.286	CZ-ELV	IT-RMO
0.200	CZ-KYJ	IT-RMO
0.154	CZ-MOR	IT-RMO
0.385	SK-DAN	IT-RMO
0.182	PL-VIS	BG-DAN
0.214	GE-RHI	BG-DAN
0.250	PL-SLE	BG-DAN
0.214	CZ-ELO	BG-DAN
0.267	CZ-ELV	BG-DAN
0.267	CZ-KYJ	BG-DAN
0.143	CZ-MOR	BG-DAN
0.188	SK-DAN	BG-DAN
0.133	IT-RMO	BG-DAN
0.130	PL-VIS	FR-DUR
0.214	GE-RHI	FR-DUR
0.176	PL-SLE	FR-DUR
0.133	CZ-ELO	FR-DUR
0.188	CZ-ELV	FR-DUR
0.267	CZ-KYJ	FR-DUR
0.143	CZ-MOR	FR-DUR
0.357	SK-DAN	FR-DUR



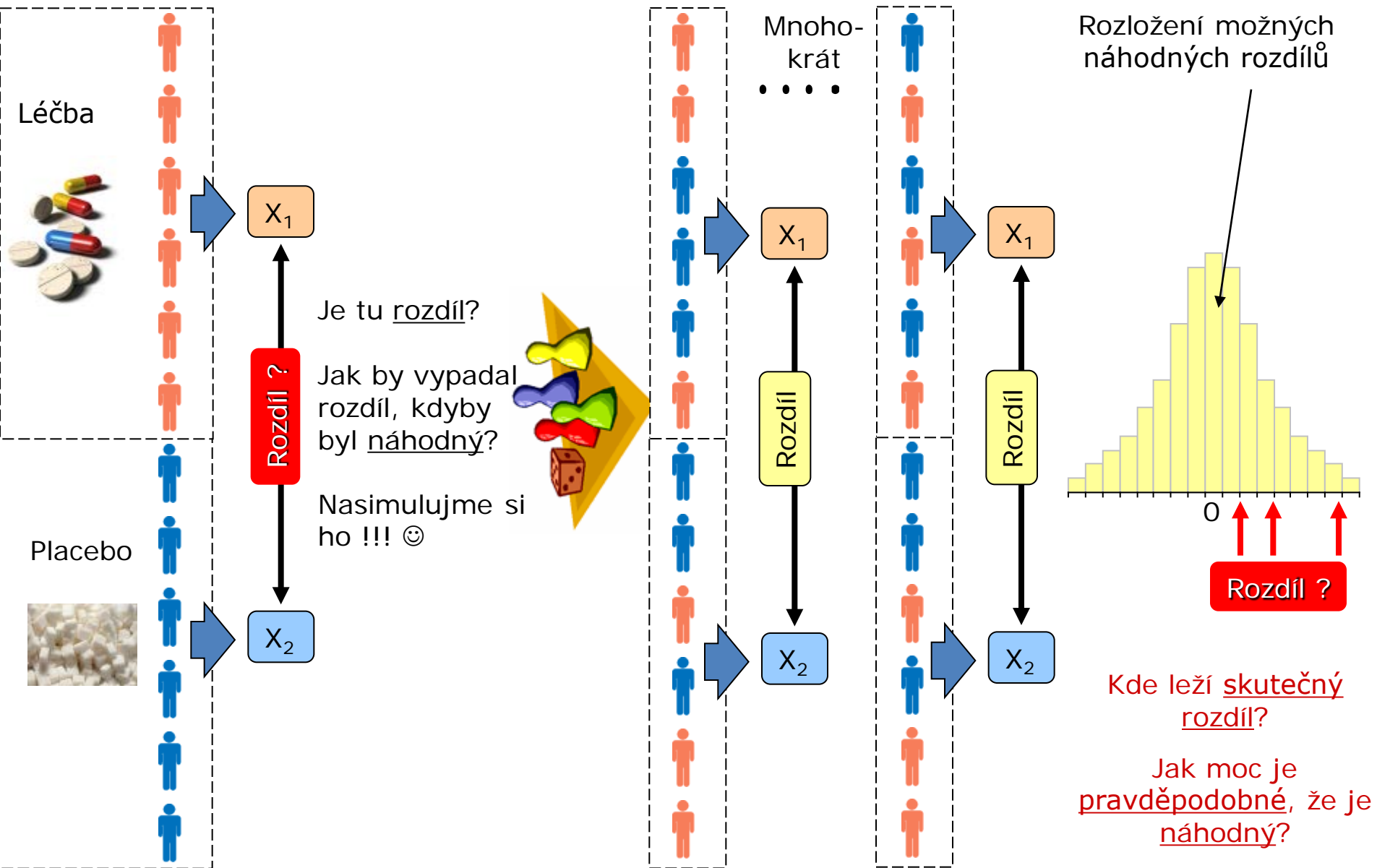
Konverzí horní trojúhelníkové matice získáme sloupec hodnot = míry asociace řádků a sloupců tabulky

Tabulku je možné dále libovolně rozšiřovat o zařazení objektů do skupin nebo o asociace objektů pomocí jiných proměnných

Příklad složitého souboru pro analýzu vztahů asociačních matic

row	column	Jaccard index	Geographical distance	Phylogenetic distance	Temperature	No fishes	No microsatellites	Y JTSK	X JTSK	No taxa	No parasites	Shannon index	Shannon evenness	Berger Parker index
PL-VIS	GE-RHI	0.389	907	0.658	5.5	26	7	906437	56332	9	385	0.303	0.182	0.001
PL-VIS	PL-SLE	0.333	246	0.100	3.7	30	7	190920	156350	6	6	0.432	0.001	0.226
GE-RHI	PL-SLE	0.357	746	0.555	1.8	4	0	715518	212681	3	391	0.129	0.181	0.226
PL-VIS	CZ-ELO	0.190	433	0.330	0.5	5	17	377143	214481	9	783	1.307	0.136	0.543
GE-RHI	CZ-ELO	0.333	594	0.281	5	21	24	529294	270812	0	1168	1.004	0.318	0.542
PL-SLE	CZ-ELO	0.357	195	0.209	3.2	25	24	186223	58131	3	777	0.875	0.137	0.317
PL-VIS	CZ-ELV	0.227	393	0.661	3.715	41	11	300529	254241	7	432	0.665	0.040	0.302
GE-RHI	CZ-ELV	0.500	680	0.345	1.785	15	4	605908	310573	2	47	0.362	0.222	0.302
PL-SLE	CZ-ELV	0.500	147	0.539	1.50E-02	11	4	109609	97892	1	438	0.233	0.042	0.076
CZ-ELO	CZ-ELV	0.800	86	0.156	3.215	36	28	76614	39761	2	1215	0.642	0.095	0.241
PL-VIS	CZ-KYJ	0.286	457	1.218	1.8	2	16	215024	403783	7	719	0.284	0.089	0.060
GE-RHI	CZ-KYJ	0.200	830	0.359	3.7	28	23	691413	460115	2	334	0.019	0.093	0.060
PL-SLE	CZ-KYJ	0.313	248	0.821	1.9	32	23	24104	247433	1	725	0.148	0.088	0.166
CZ-ELO	CZ-KYJ	0.125	249	0.220	1.3	7	1	162119	189302	2	1502	1.023	0.225	0.483
CZ-ELV	CZ-KYJ	0.176	172	0.171	1.915	43	27	85505	149542	0	287	0.381	0.130	0.242
PL-VIS	CZ-MOR	0.333	467	1.218	3.4	5	5	218534	413948	10	764	0.577	0.116	0.106
GE-RHI	CZ-MOR	0.364	833	0.356	2.1	31	12	687903	470280	1	379	0.274	0.066	0.106
PL-SLE	CZ-MOR	0.286	259	0.847	0.3	35	12	27615	257599	4	770	0.145	0.115	0.120
CZ-ELO	CZ-MOR	0.250	255	0.224	2.9	10	12	158609	199468	1	1547	0.730	0.252	0.437
CZ-ELV	CZ-MOR	0.308	180	0.184	0.315	46	16	81995	159707	3	332	0.088	0.156	0.196
CZ-KYJ	CZ-MOR	0.308	11	0.005	1.6	3	11	3510	10165	3	45	0.293	0.027	0.046
PL-VIS	SK-DAN	0.350	540	1.118	8	6	14	190569	506010	7	647	0.278	0.092	0.102
GE-RHI	SK-DAN	0.385	909	0.318	2.5	20	7	715869	562342	2	262	0.025	0.090	0.102
PL-SLE	SK-DAN	0.500	349	0.807	4.3	24	7	351	349661	1	653	0.154	0.091	0.124
CZ-ELO	SK-DAN	0.385	346	0.198	7.5	1	31	186574	291530	2	1430	1.029	0.228	0.441
CZ-ELV	SK-DAN	0.429	275	0.199	4.285	35	3	109960	251769	0	215	0.387	0.132	0.200
CZ-KYJ	SK-DAN	0.429	105	0.040	6.2	8	30	24455	102227	0	72	0.006	0.003	0.042
CZ-MOR	SK-DAN	0.308	96	0.056	4.6	11	19	27966	92062	3	117	0.299	0.024	0.004
PL-VIS	IT-RMO	0.190	1120	1.416	8	30	7	894871	676397	9	756	0.009	0.353	0.043
GE-RHI	IT-RMO	0.333	731	0.142	2.5	4	0	11567	732728	0	371	0.294	0.171	0.044
PL-SLE	IT-RMO	0.267	874	1.143	4.3	0	0	703951	520047	3	762	0.423	0.352	0.269
CZ-ELO	IT-RMO	0.231	692	0.736	7.5	25	24	517728	461916	0	1539	1.298	0.489	0.586
CZ-ELV	IT-RMO	0.286	728	0.711	4.285	11	4	594342	422156	2	324	0.656	0.393	0.346
CZ-KYJ	IT-RMO	0.200	731	0.565	6.2	32	23	679847	272614	2	37	0.275	0.264	0.104
CZ-MOR	IT-RMO	0.154	724	0.587	4.6	35	12	676336	262449	1	8	0.568	0.237	0.149
SK-DAN	IT-RMO	0.385	723	0.483	0	24	7	704302	170387	2	109	0.269	0.261	0.146
PL-VIS	BG-DAN	0.182	1002	1.079	7.5	28	1	203173	982589	8	92	0.257	0.147	0.042

Permutační testování



Meansim – analogie k ANOVA

The screenshot shows a Mozilla Firefox browser window displaying the Western Ecology Division website. The address bar shows the URL: <http://www.epa.gov/wed/pages/models/dendro/meansim6.htm>. The page header includes the U.S. Environmental Protection Agency logo and the text "U.S. ENVIRONMENTAL PROTECTION AGENCY". The main heading is "Western Ecology Division" with a search bar and navigation links. The page content is titled "Documentation for MEANSIM, Version 6.0" and describes a set of programs for Mean Similarity Analysis. It lists contact information for John Van Sickle and provides a detailed description of the software's purpose and usage. The page also includes a section for "1. MEAN SIMILARITY DENDROGRAMS" which references a journal article by Van Sickle, J., 1997.

Western Ecology Division | US EPA - Mozilla Firefox

Soubor Úpravy Zobrazení Historie Záložky Nástroje nápověda

US EPA <http://www.epa.gov/wed/pages/models/dendro/meansim6.htm> Google

Nejnávštěvovanější Jak začít Přehled zpráv Report Manager - Logi... Úřad průmyslového vl... 6. letní škola Matemati... FEKT VUT - Časový plá... LogMeIn - Remote Ac... IBA sportovní aktivity

Převést

US EPA Western Ecology Division | US EPA

U.S. ENVIRONMENTAL PROTECTION AGENCY

Western Ecology Division

Contact Us Search: All EPA This Area Go

You are here: EPA Home » Western Ecology Division (WED) » About WED » Models, Software, Data Sets

Documentation for MEANSIM, Version 6.0

A set of programs for Mean Similarity Analysis 10/29/98

John Van Sickle
ph. 541-754-4314
fax 541-754-4338
email: vansickle.john@epa.gov

This distribution contains software for Mean Similarity Analysis, a method for evaluating the strength of a classification of many objects (sites) into a relatively small number of groups. The classification strength of a grouping is evaluated by the extent to which objects within the same group are more similar to each other, on average, than they are to objects in different groups.

The enclosed programs run under Windows 95/98/NT. If you need a DOS or Windows 3.1 version, please contact the author.

The analysis is based on a matrix of pairwise similarities (or dissimilarities) for all possible pairs of objects. The mean similarity analysis outputs consist of a small matrix of average similarities within and between the chosen groups, and a statistical test of whether average within-group similarities are greater than between-group similarities.

This software is provided free of charge with the understanding that it will not be used for any commercial purposes. It is reasonably reliable, but has not been exhaustively tested and must be applied at the user's own risk. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.
The software was developed partly under US Environmental Protection Agency Contract 68-C5-0005 to Dynamac, Inc.

1. MEAN SIMILARITY DENDROGRAMS

This documentation closely follows the language and notation of the article: Van Sickle, J., 1997, Using Mean Similarity Dendrograms to Evaluate Classifications, *Journal of Agricultural, Biological and Environmental Statistics* 2, 370-388. A copy of this article may be downloaded from the same source as this software. For a paper reprint of this article, contact the author at the above address.

The article describes a convenient dendrogram format for displaying mean similarities, but it also describes the mean similarity approach and provides numerous references. This software distribution does not include software for actually drawing mean similarity dendrograms. However, users can easily draw such graphs with any presentation graphics package, once the mean similarity matrix has been computed using the programs discussed below.

Hotovo

Meansim – analogie k ANOVA

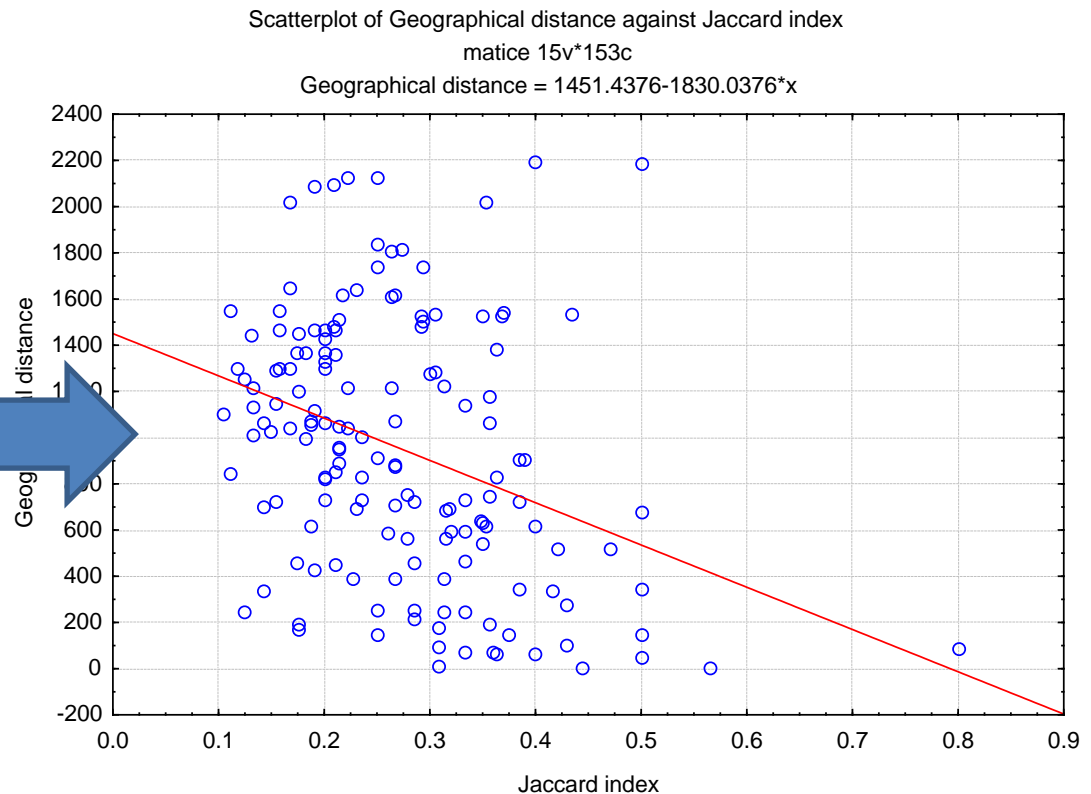
	A	B	C
A			
B			
C			

- Meansim pracuje s pojmy průměrná vnitroshluková vzdálenost a průměrná mezishluková vzdálenost
- Ty mají obdobný význam jako variabilita uvnitř a mezi skupinami v klasické ANOVA
- Rozdíl oproti ANOVA je ve výpočtu statistické významnosti:
 - Objekty (v řádcích a sloupcích) jsou náhodně zpřeházeny mezi skupinami
 - Je spočten poměr mezishlukové a vnitroshlukové variability
 - Postup je opakován x krát až získáme rozdělení náhodného vztahu asociace objektů ke kategoriím
 - Výsledek testu porovnán se simulovaným rozdělením náhodného vztahu asociace objektů ke kategoriím

Mantel test – analogie ke korelaci

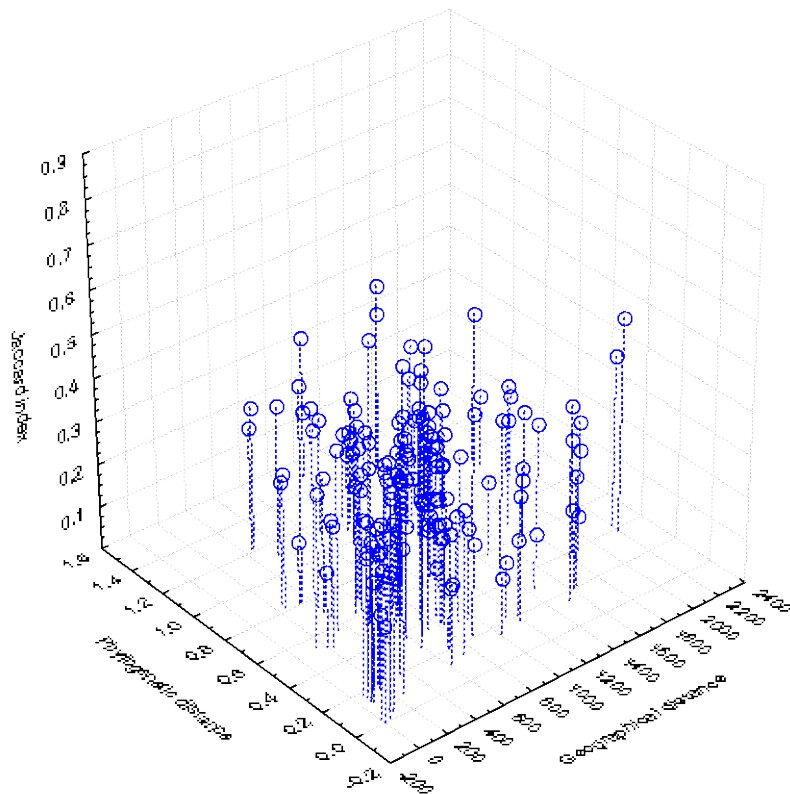
- Počítán pomocí Pearsonovy nebo Spearmanovy korelace, lze použít libovolný korelační koeficient
- Rozdíl je opět ve výpočtu statistické významnosti, která je počítána permutačně

3	4
Jaccard index	Geographical distance F
0.38889	907
0.33333	246
0.35714	746
0.19048	433
0.33333	594
0.35714	195
0.22727	393
0.5	680
0.5	147
0.8	86
0.28571	4
0.2	8
0.3125	248
0.125	249
0.17647	172
0.33333	467
0.36364	833
0.28571	259
0.25	255
0.30769	180
0.30769	11



Regrese na asociačních maticích

- Obdobná výpočtu klasické regrese, ale na maticích vzdáleností



```
RGui
File Edit View Misc Packages Windows Help
R Console
151 0.33333 73 0.166 0.600 10 1 66852.4894 29299.4435 8 798 0.4030 0.0290 0.1493
152 0.36000 74 0.129 0.900 10 2 68213.2231 28783.5358 6 1452 0.0530 0.1293 0.0430
153 0.56522 6 0.003 0.300 0 2 5047.6618 994.6586 4 907 0.3380 0.0279 0.1759
> mode(graZe)
[1] "list"
> mode(mx)
[1] "list"
> pokus <-MRM(mx$Jaccard-mx$Geographical + mx$Phylogenetic,data=mx, nperm=100)
Error in model.frame.default(formula = mx$Jaccard ~ mx$Geographical + :
invalid type (NULL) for variable 'mx$Jaccard'
> help(MRM)
> help(MRM)
> data(graZe)
> LOAR10.mrm <- MRM(dist(LOAR10) ~ dist(sitelocation) + dist(forestpct), data=graZe, nperm=100)
> loaer10.mrm
Error: object 'loaer10.mrm' not found
> LOAR10.mrm
Scoef
          dist(LOAR10) pval
Int          6.9372046 0.95
dist(sitelocation) -0.4840631 0.33
dist(forestpct)    0.1456083 0.03

Sr.squared
      R2      pval
0.04927212 0.06000000

SF.test
      F      F.pval
31.66549 0.06000
```