Statistické metody a zpracování dat

Faktorová a komponentní analýza
(Úvod do vícerozměrných metod)

Petr Dobrovolný

Úvod do vícerozměrných metod

O řadě jevů či procesů máme k dispozici ne jeden statistický znak ale znaků několik.

Př. Struktura obyvatelstva, vlastnosti povodí, klimatické poměry místa, ...

Vstupní data: Statistické jednotky (např. městské obvody) a k nim několik charakteristik (např. demografická data).

Cíle prezentovaných metod:

1. **redukovat počet proměnných**
2. **detekovat strukturu vztahů mezi proměnnými (klasifikovat, vytvořit typologii dat)**

Faktorová analýza (Factor Analysis – **FA**)

Analýza hlavních komponent (Principal Component Analysis – **PCA**)

Literatura:

Heřmanová, E. (1991): Vybrané vícerozměrné statistické metody v geografii. SPN, Praha, 133 s.

http://www.statsoft.cz/textbook/stathome.html

Ilustrativní příklad – vstupní data

Podíl zaměstnaných v devíti odvětvích ve 26 evropských zemích (údaje z konce 70. let 20. století)

1. AGR = agriculture
2. MIN = mining
3. MAN = manufacturing
4. PS = power suplies
5. CON = construction
6. SER = service industries
7. FIN = finance,
8. SPS = social and personal services
9. TC = transport and communications

Vstupní matice: 9 řádků (proměnných – odvětví) a 26 sloupců (případy – státy)

Cíl: Redukce počtu proměnných a odhalení typických znaků v zaměstnanosti jednotlivých států

Příklad – typický výstup PCA I.

No.	Eigenvalue	Individual Percent	Cumulative Percent	Scree Plot
1	3.487151	38.75	38.75	||||||||
2	2.130173	23.67	62.41	|||||
3	1.098958	12.21	74.63	|||
4	0.994483	11.05	85.68	|||
5	0.543218	6.04	91.71	||
6	0.383428	4.26	95.97	|
7	0.225754	2.51	98.48	|
8	0.136790	1.52	100.00	|
9	0.000046	0.00	100.00	|

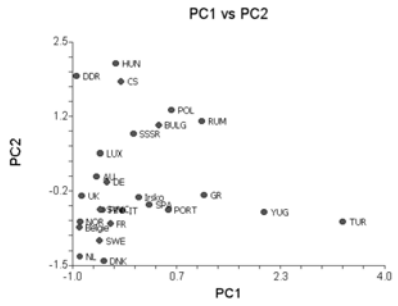
- pořadové číslo nové proměnné (PC - **hlavní komponenty**)
- tzv. **vlastní číslo** – část z celkového rozptylu původních dat vysvětlená každou z nových komponent
- procentuální vyjádření množství **rozptylu vysvětleného** komponentou
- **kumulativní** hodnota procentuálního podílu vysvětleného příslušnými komponentani (např. první 4 komponenty vysvětlují 85,68 % celkové variability původních dat)
- tzv. **sutinový graf** sloužící k určení počtu významných komponent

Příklad – typický výstup PCA II.

Variables	Factor1	Factor2	Factor3	Factor4
AGR	0.523791	0.053594	0.048674	0.028793
MIN	0.001323	0.617807	-0.201100	0.064085
MAN	-0.347495	0.355054	-0.150463	-0.346088
PS	-0.255716	0.261096	-0.561083	0.393309
CON	-0.325179	0.051288	0.153321	-0.668324
SER	-0.378920	-0.350172	-0.115096	-0.050157
FIN	-0.074374	-0.453698	-0.587361	-0.051567
SPS	-0.387409	-0.221521	0.311904	0.412230
TC	-0.366823	0.202592	0.375106	0.314372

Tzv. **zátěže** (loadings) - představují míru korelace mezi původními a novými proměnnými

Příklad – typický výstup PCA



Struktura zaměstnanosti jednotlivých zemí vyjádřená polohou v grafu hodnot prvních dvou (nejvýznamnějších) hlavních komponent.

Princip FA a PCA

Charakteristiky, které na jednotkách měříme, jsou jen určitou formou projevu tzv. **skrytých veličin**, které přímo měřit nemůžeme.

Řada měřených charakteristik spolu do značné míry **souvisí** – vypovídá o stejné vlastnosti, **koreluje** spolu (mezi proměnnými existují „překryvy“).

Cílem obou metod je **eliminování duplicit**, **zhuštění informace** obsažené v původních proměnných do menšího počtu vzájemně nekorelovaných proměnných.

Tyto nové proměnné (**faktory, hlavní komponenty**) popisují soubor jednotek syntetičtěji a úsporněji.

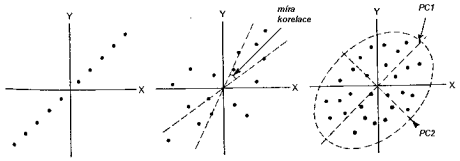
Základní východiska

Princip redukce dat a „skryté“proměnné (interpretace následujícího obrázku)

Máme-li pro soubor znaků dvě proměnné a ty spolu vzájemně korelují – potom vypovídají z velké části o tomtéž – jsou redundantní.

Pokud takového dvě (korelované) proměnné vyneseme do grafu a proložíme rovnicí přímky – potom tuto přímku můžeme považovat za osu, na niž jsou vyneseny hodnoty **nové proměnné**, která ponese podstatnou informaci z obou proměnných původních.

Základní východiska



Základní východiska

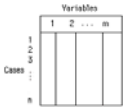
Tedy – dvě původní proměnné redukujeme do jedné nové proměnné – do tzv. faktoru (FA) či hlavní komponenty (PC).

Faktor či hlavní komponenta je lineární kombinací původních proměnných.

Uvedený princip lze zobecnit na větší počet proměnných a je podstatou metod FA a PCA.

Tyto metody se používají k analýze vztahů závislosti ve vícerozměrném (obecně r-rozměrném) ortogonálním (pravoúhlém) prostoru.

Vstupní datová matice

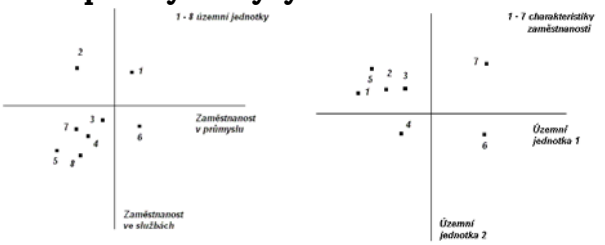


Vstupní data představuje matice, která obsahuje n případů pro m proměnných. V běžném případě představují proměnné sloupce datové matice a případy její řádky.

Charakteristiky vstupují do analýzy obvykle ve standardizovaném tvaru (ve formě směrodatných proměnných.

$$t_i = \frac{x_i - \mu}{\sigma}$$

Dva způsoby analýzy



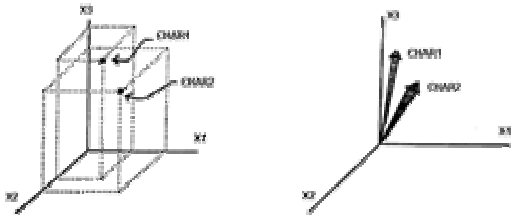
Analýza podobnosti jednotek – dimenze r-rozměrného prostoru jsou charakteristiky (proměnné). Cílem analýzy je redukovat sloupce datové matice

Analýza podobnosti proměnných - dimenze r-rozměrného prostoru jsou jednotky (případy). Cílem analýzy je redukovat dimensionalitu řádků.

Geometrický model

Dvojice charakteristik může být vyjádřena dvěma vektory se společným počátkem. Orientace a těsnost jejich vztahu je určena velikostí sevřeného úhlu.

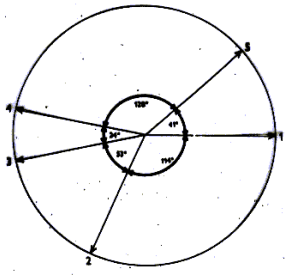
Příklad pro tři proměnné a dva případy



Geometrický model

Úhly mohou nabývat hodnot od 0 do 180 stupňů a *cos* úhlu odpovídá hodnotě korelačního koeficientu:

Grafické znázornění korelací mezi pěti proměnnými



	1	2	3	4	5
1	1	-0,41	-0,97	-0,98	0,75
2		1	0,60	0,22	-0,91
3			1	0,91	-0,88
4				1	-0,62
5					1

cos 0 =1, r_{xy} = 1

cos 90 =0, r_{xy} = 0

cos 180 =-1, r_{xy} = -1

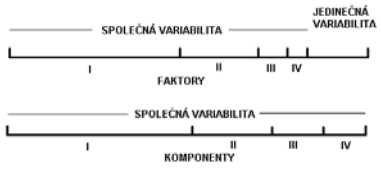
Rozdíly mezi FA a PCA

Obě metody lze považovat za dva modely založené na stejném principu.

PCA – uzavřený systém, ve kterém veškerá variabilita v hodnotách proměnných je vysvětlena proměnnými samotnými. Nepředpokládáme žádnou strukturu a jde nám jen o redukci počtu proměnných

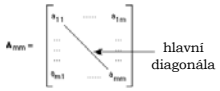
FA – model, který předpokládá, že nemáme k dispozici všechny proměnné, které popisují daný problém. S souboru existuje i variabilita, která není vysvětlena jednotlivými faktory a přísluší reziduální složce (neznámé či chybové). Jen část celkové variability je vysvětlena použitými proměnnými.

Rozdíly mezi FA a PCA



Za jistých podmínek oba modely dávají podobné výsledky – např. v případě, že korelace mezi původními proměnnými jsou vysoké.

Komunalita



FA používá k výpočtu tzv. komunality . Hodnoty komunality se nacházejí na hlavní diagonále korelační matice. U PCA se na hlavní diagonále nacházejí hodnoty 1.

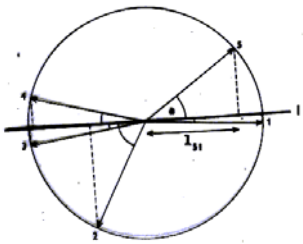
Jedničky na hlavní diagonále korelační matice vyjadřují předpoklad, že celková variabilita daného souboru je vysvětlena vybranými proměnnými.

Komunalita se značí h^2 a a lze ji interpretovat jako část rozptylu připadajícího na společné faktory

Obecný algoritmus výpočtu komponentní a faktorové analýzy

1. **Sestavení matice standardizovaných charakteristik typu n,m**
2. **Výpočet korelační matice typu m,m**
3. **Pro FA odhad komunalit, kterými jsou nahrazeny jedničky na hlavní diagonále korelační matice.**
4. **Výpočet r ortogonálních proměnných (faktorů či hlavních komponent) z příslušných korelačních matic**
5. **Rotace faktorů či komponent**
6. **Interpretace výsledků**

Extrahování PC1 či FA1 - geometrický model



Extrahování faktorů - geometrický model

Cílem extrakce je nalezení průmětu použitých vektorů se společným počátkem do prostoru o menším počtu dimenzí tak, aby zůstala zachována co možná největší délka jednotlivých vektorů („ostnů“ ježka) (délka ostnů = variabilita)

Soustavou vektorů se společným počátkem se postupně prokládají osy definující nový prostor – jsou na sebe kolmé a jsou prokládány tak, aby každá osa vystihovala maximální variabilitu – geometricky – aby průměty vektorů – původních proměnných – na novou osu byly co nejdelší.

Faktorové zátěže

Délka projekce vektorů se označuje *l* a odpovídá hodnotě korelačního koeficientu mezi původní a extrahovanou proměnnou. Hodnota *l* je definována jako váha (zátěž – loading).

Druhá nová osa je proložena tak aby vystihovala maximum ze zbývající variability atd.

Vektory proměnných u PCA mají jednotkovou délku.

U FA je délka vektoru rovna odmocnině z příslušné komunality.

Výpočetní model pro první faktor (hlavní komponentu)

Faktorové zátěže: suma korelací každé proměnné /druhá odmocnina z celkové sumy koeficientů.

Proměnná	X1	X2	X3
X1	1,0	0,6	0,7
X2	0,6	1,0	0,7
X3	0,7	0,8	1,0
Σ korelací	**2,3**	**2,4**	**2,5**

Celková suma koeficientů v matici: 7,2

Druhá odmocnina z celkové sumy koeficientů (tj. společná variabilita): 2,68

Výpočet zátěží l_1, l_2, l_3 pro první faktor

l_1 = 2,3/2,68 = 0,86

l_2 = 2,4/2,68 = 0,90

l_3 = 2,5/2,68 = 0,93

Zátěže představují míru korelace mezi původními proměnnými a novým faktorem – tedy korelační koeficient.

Z toho tedy plyne, že druhá mocnina zátěže vyjadřuje část rozptylu původní proměnné, která je vysvětlena novým faktorem (analog. koeficientu determinance).

Výpočet velikosti korelace reprodukované poslední extrahovanou komponentou (faktorem)

Proměnná	Zátěž (l)	l^2
X1	0,86	0,72
X2	0,90	0,81
X3	0,93	0,86
Vlastní číslo		2,39

Vlastní číslo (eigenvalue) vypočteme jako sumu druhých mocnin zátěží jednotlivých proměnných.

Vlastní číslo představuje hodnotu rozptylu vysvětleného faktorem či komponentou

Významnost extrahovaného faktoru

Rozptyl nového faktoru můžeme vztáhnout k celkovému rozptylu obsaženému v korelační matici původních proměnných:

Procento rozptylu vysvětlené faktorem = Vlastní číslo faktoru / počet původních proměnných * 100

V našem případě tedy část variability vysvětlená prvním faktorem činí 79 % (2,39/3*100)

Vlastnosti první hlavní komponenty

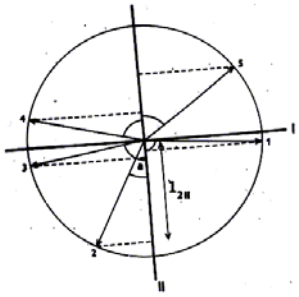
PC1 je lineární kombinací vstupních proměnných

PC1 vystihuje 79 % variability původních dat

První hlavní komponenta tedy nepostihuje veškerou variabilitu.

Proto v následném kroku tedy extrahujeme druhou hlavní komponentu (či faktor), která by objasňovala zbývající proměnlivost původních proměnných.

Extrahování PC2 či FA2 - geometrický model



Celý proces se opakuje výpočtem PC2, PC3, ... následovně:

1. Sestavíme matici, která vyjadřuje variabilitu vysvětlenou první komponentou.
2. Tuto matici odečteme od korelační matice původních proměnných.
3. Dostaneme tzv. matici reziduálních (zbytkových) korelací.
4. Vypočteme váhy (zátěže) a procento variability reprodukované dalšími PC či FA
5. Celý výpočet se opakuje pro tolik komponent, kolik bylo vstupních proměnných

Určení matice vyjadřující variabilitu vysvětlenou první komponentou

Zátěže mezi první PC či FA a původními proměnnými:

X1 X2 X3

0,86 0,90 0,93 např. 0,86 * 0,90 = 0,77

Potom matice, která vyjadřuje velikost korelace reprodukované právě extrahovanou komponentou bude:

	X1	X2	X3
X1	0,74	0,77	0,80
X2	0,77	0,81	0,84
X3	0,80	0,84	0,86

Tuto matici odečteme od původní korelační matice a dostaneme matici reziduálních (zbytkových) korelací

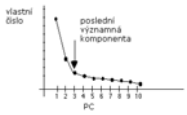
Shrnutí vlastností vypočtených faktorů (komponent)

- **Druhá (a každá následující) PC či FA postihuje rozptyl, který nesouvisí s PC či FA první (předchozí)**
- **Jednotlivé faktory jsou vzájemně nekorelované (ortogonální)**
- **Postupně obsahují (či vysvětlují) menší část variability původních dat.**

Rozhodování o počtu interpretovatelných nových faktorů

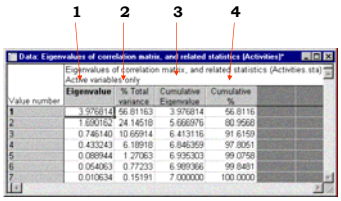
Dvě základní kritéria:

- **Je-li hodnota vlastního čísla větší než 1, potom daný faktor vysvětluje více celkového rozptylu než jedna původní proměnná.**



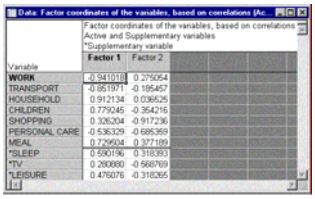
- **„Scree"-graf – hledá se zřejmý zlom ve sklonu křivky, která prezentuje spojnici hodnot celkového rozptylu vysvětleného jednotlivými faktory.**

Typický výstup FA či PCA



Tabulka ve sloupcích obsahuje pro sedm extrahovaných faktorů (hl. komponent) hodnotu vlastního čísla (1), dále procento variability vysvětlené daným faktorem (2), kumulovanou hodnotu vlastních čísel (3) a kumulovanou hodnotu vysvětlené variability (4)

Typický výstup FA či PCA



Váhy (zátěže) pro první dva faktory, které informují o těsnosti korelace mezi určitým faktorem a každou ze vstupních proměnných.

Interpretace výsledků I.

Zátěže informují o tom, které proměnné nejvíce „zatěžují" jednotlivé nové faktory (které v nich mají největší zastoupení).

Pro identifikaci struktury v datech jsou důležité absolutní hodnoty zátěží.

Strukturu lze odhalit i na základě zkušenosti.

Cílem je dát vypočteným faktorům konkrétní význam, název, označení,...

K lepší interpretaci výsledků PCA lze provést jejich rotaci

Příklad

Vstupní data: výsledky dosažené ve výběru 220 žáků v šesti předmětech:

1. gaelština
2. angličtina
3. dějepis
4. aritmetika
5. algebra
6. geometrie

Korelační matice vstupních dat

1.00					
0.44	1.00				
0.41	0.35	1.00			
0.29	0.35	0.16	1.00		
0.33	0.32	0.19	0.59	1.00	
0.25	0.33	0.18	0.47	0.46	1.00

Příklad – výstup: vlastní čísla a zátěže

Eigenvalues

No.	Eigenvalue	Individual Percent	Cumulative Percent	Scree Plot
1	2.728683	45.48	45.48	||||||||||
2	1.128792	18.81	64.29	||||
3	0.615291	10.25	74.55	|||
4	0.602809	10.05	84.59	|||
5	0.522514	8.71	93.30	||
6	0.401910	6.70	100.00	||

Factor Loadings

Variables	Factor1	Factor2
Gaelic	-0.660803	-0.444475
English	-0.688465	-0.289771
History	-0.516356	-0.639552
Arithmetic	-0.735620	0.417018
Algebra	-0.741868	0.372759
Geometry	-0.678168	0.354100

Příklad – výstup: vlastní čísla a zátěže (výsledek po provedení rotace)

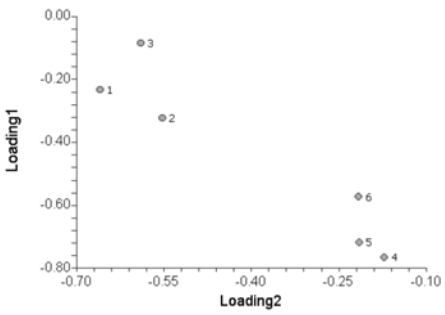
Eigenvalues after Varimax Rotation

No.	Eigenvalue	Individual Percent	Cumulative Percent	Scree Plot
1	1.596863	56.94	56.94	||||||||||||
2	1.207981	43.08	100.02	||||||||
3	0.050820	1.81	101.83	|
4	0.011910	0.42	102.26	|
5	-0.008657	-0.31	101.95	|
6	-0.054642	-1.95	100.00	|

Factor Loadings after Varimax Rotation

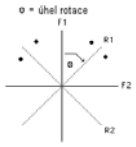
Variables	Factor1	Factor2
Gaelic	-0.233132	-0.659253
English	-0.322810	-0.552071
History	-0.084713	-0.589192
Arithmetic	-0.765986	-0.170657
Algebra	-0.718105	-0.214689
Geometry	-0.573340	-0.214994

Příklad



Korelační strukturu pozorovaných dat lze vysvětlit dvěma faktory. První faktor vyjadřuje matematickou dispozici žáka, druhý dispozici jazykově-humanitní.

Rotace faktorů

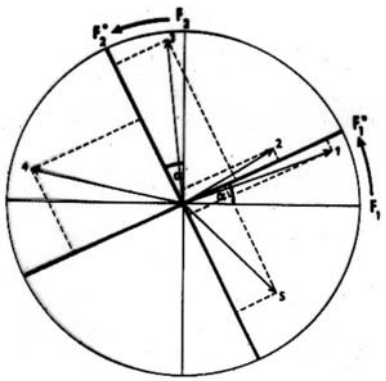


Cíl PCA či FA – nalézt nové proměnné, které by zřetelněji a úsporněji popisovaly vstupní datový soubor.

Hledá se „**jednoduchá struktura**“ – tedy výsledek, kdy každá původní proměnná „hodně zatěžuje“ jeden faktor a málo jiný. Ve většině případů prvotní analýza tuto jednoduchou strukturu neposkytuje a odvozené faktory nejasně (ve smyslu obtížné interpretace) popisují původní proměnné.

Možným řešením je tzv. **rotace faktorů**. Smyslem rotace je nalezení stejně výstižného, ale z hlediska věcné interpretace podstatně výhodnějšího řešení.

Geometrické vyjádření rotace



Rotace faktorů

Cílem rotace je zvýraznit shluky proměnných beze změny jejich relativní polohy ve vícerozměrném prostoru.

Jedná se vlastně o pootočení souřadné soustavy faktorů kolem počátku.

Podstata rotace – otočení systému os o určitý úhel tak, aby se co nejvíce přiblížily vektorům proměnných.

Změní se vztah mezi osami a proměnnými a tedy **změní se i struktura zátěží**. Vzájemné **vztahy mezi vektory proměnných se nezmění**.

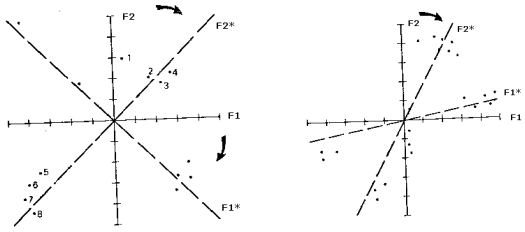
	F1	F2
1	0,83	0,33
2	0,50	0,33
3	-0,10	0,95
4	-0,86	0,20
5	0,53	-0,52

Matice nerotovaných faktorových vah (zátěží)

	F1*	F2*
1	0,90	-0,07
2	0,60	0,08
3	0,33	0,90
4	-0,68	0,55
5	0,25	-0,70

Matice rotovaných faktorových vah

Rotace ortogonální a neortogonální



F – nerotované faktory (komponenty)

F* - rotované faktory (komponenty)

Rotace ortogonální a neortogonální

Neortogonální rotace se hůře iterpretuje.

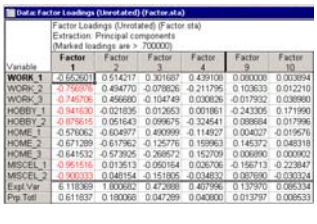
Ortogonální rotace – ideální je případ, kdy každá proměnná má zátěž jednoho faktoru rovnu jedné a zátěže ostatních faktorů jsou nulové.

Existuje několik metod rotace - nejpoužívanější je metoda VARIMAX - rotace ve směru maximálního rozptylu.

Kritérium jednoduché struktury:

- **V rotované matici vah má být co nejvíce nulových zátěží (-0,1 ; 0,1)**
- **Každá proměnná má být významně obsažena v co nejmenším počtu faktorů**
- **Každý faktor má být představován kombinací jen několika málo proměnných**

Typický výstup FA či PCA


Data: Factor Loadings (Unrotated) (Factor.sta)

Factor Loadings (Unrotated) (Factor.sta)
Extraction: Principal components
(Marked loadings are > ,700000)

Variable	Factor 1	Factor 2	Factor 3	Factor 4	Factor 9	Factor 10
WORK_1	-0.653501	0.514217	0.301687	0.439108	0.000008	0.003894
WORK_2	-0.756976	0.494770	0.078826	-0.211795	0.103633	0.012210
WORK_3	-0.745706	0.456680	-0.104749	0.030626	-0.017932	0.038980
HOBBY_1	-0.941630	-0.021835	0.012653	0.001861	-0.243305	0.171990
HOBBY_2	-0.875615	0.051643	0.099675	-0.324541	0.088684	0.017996
HOME_1	-0.576062	-0.604977	0.490999	-0.114927	0.004027	-0.019576
HOME_2	-0.671289	-0.617962	-0.125776	0.159963	0.145372	0.048318
HOME_3	-0.641532	-0.573925	-0.268572	0.152709	0.006990	0.000902
MISCEL_1	-0.961516	0.013513	-0.050164	0.026706	-0.156713	-0.223847
MISCEL_2	-0.900333	0.048154	-0.151805	-0.034832	0.087690	-0.030324
Expl.Var	6.118369	1.800682	0.472688	0.407996	0.137970	0.085334
Prp.Tot	0.611837	0.180068	0.047269	0.040800	0.013797	0.008533

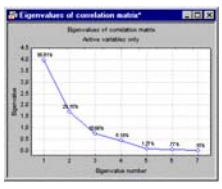

Data: Factor Loadings (Varimax raw) (Factor.st...

Factor Loadings (Varimax raw) (Factor.sta)
Extraction: Principal components
(Marked loadings are > ,700000)

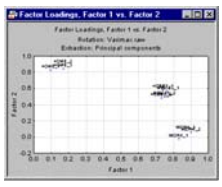
Variable	Factor 1	Factor 2
WORK_1	0.830623	-0.019320
WORK_2	0.902408	0.058905
WORK_3	0.870524	0.082595
HOBBY_1	0.736667	0.582885
HOBBY_2	0.731191	0.484489
HOME_1	0.097371	0.829676
HOME_2	0.165722	0.897242
HOME_3	0.168370	0.844159
MISCEL_1	0.768988	0.560555
MISCEL_2	0.748861	0.502121
Expl.Var	4.561544	3.357507
Prp.Tot	0.456154	0.335751

Nerotované a rotované hodnoty zátěží („korelačních koeficientů“) pro jednotlivé extrahované faktory. Rotovaný výsledek má „jednoduchou strukturu“

Typický výstup FA či PCA



Graf umožňující odhadnout počet interpretovatelných faktorů“c



Projekci původních proměnných do 2-D prostoru definovaného prvními dvěma (nejvýznamnějšími) vypočtenými faktory

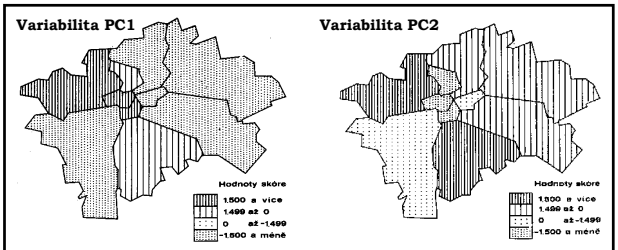
Využití faktorových skóre

Matice faktorových skóre je jedním z důležitých výsledků FA.

Je důležitá pro interpretaci výsledků v geografii při analýze prostorových struktur (uspořádání).

Ukazuje do jaké míry je konkrétní pozorování zastoupeno v nových faktorech (poskytuje míru vztahu mezi každým pozorováním (případem) a novými faktory).

Jestliže určitý (případ) má vysokou hodnotu v určité proměnné a ta má vysokou zátěž v daném faktoru, potom také tento případ bude mít vysokou hodnotu skóre u tohoto faktoru.



Faktorová skóre mohou sloužit k vynášení do mapy – k jednotlivým prostorovým objektům - k vytváření typologií a klasifikaci.

Každý případ (např. okres, povodí, ...) může být přiřazen k určitému faktoru podle hodnoty faktorového skóre. Tedy statisticky podobné jednotky budou patřit ke stejnému faktoru. Pro každý faktor můžeme vytvořit mapu.

Faktorová skóre mohou být dále využita pro vytváření grafů ve vícerozměrném prostoru definovaném nově extrahovaným faktory.