VÍCEROZMĚRNÉ STATISTICKÉ METODY

Mgr. Martin Sebera

Fakulta sportovních studií
Masarykovy univerzity

20. 1. 2006

Sledování vzájemné závislosti dvou kvantitativních proměnných, které splňují předpoklad normality

U kvantitativních proměnných, které splňují předpoklad normality, lze zkoumat vzájemnou závislost pomocí korelační analýzy. Intenzita závislosti je posuzována pomocí Pearsonova *korelačního koeficientu*, který nabývá hodnot z intervalu <-1;1>, přičemž hodnota 0 znamená nezávislost. Testem o nezávislosti je tedy *test o nulovosti* korelačního koeficientu.

Sledování vzájemné závislosti dvou kategoriálních proměnných

Základním testem je *test chí-kvadrát o nezávislosti* dvou proměnných, založený na četnostech v kontingenční tabulce. Pro sledování intenzity závislosti jsou používány různé koeficienty, které obvykle nabývají hodnot z intervalu <0;1>, případně <-1;1>, přičemž hodnota 0 znamená nezávislost. Dalšími testy jsou tedy *testy o nulovosti* těchto koeficientů.

Pro speciální situace (např. tabulka 2 x 2 – tj. obě proměnné jsou dichotomické) existují míry, které v případě nezávislosti nabývají hodnoty 1. Tehdy testujeme, zda příslušný koeficient se rovná jedné. Jestliže není splněn předpoklad pro použití chí-kvadrát testu v kontingenční tabulce (týká se teoretických četností v políčkách tabulky), pak jsou používány tzv. *exaktní testy - Fisherův*.

Chí-kvadrát test o nezávislosti patří mezi neparametrické testy. Dalšími z těchto testů, které můžeme zařadit k postupům pro řešení prvního typu úloh, jsou *testy pro 2 závislé (párové) výběry - Wilcoxonův*. Pomocí nich testujeme, zda dvě ordinální proměnné jsou výběry ze stejného rozdělení. Existuje i speciální test pro dvě dichotomické proměnné, který je založen na četnostech v asociační tabulce (*Mc Nemarův*).

Koeficienty závislosti můžeme klasifikovat na základě

- typu tabulky (tj. podle počtu kategorií u sledovaných proměnných),
- typu proměnných (zda jde o proměnné nominální, ordinální či kvantitativní),
- způsobu výpočtu (zda je při výpočtu použita hodnota chí-kvadrát, zda je koeficient vypočítán na základě koeficientů vyjadřující míru jednostranné závislosti apod.).

Sledování jednostranné závislosti dvou kategoriálních proměnných

V tomto případě obvykle *testujeme nulovost koeficientů*, počítaných na základě četností v kontingenční tabulce. Jde vždy o dvojici asymetrických koeficientů, které posuzují míru závislosti jednak proměnné X na Y, jednak Y na X (z každé takové dvojice můžeme vypočítat koeficient symetrický).

Další testy řadíme k neparametrickým. Sledujeme při nich závislost ordinální proměnné na proměnné kategoriální, u níž nezáleží na typu. Tyto testy dělíme na případy, kdy

- vysvětlující proměnná nabývá dvou hodnot (testy pro 2 nezávislé výběry-Mann-Whitneův) a kdy
- vysvětlující proměnná nabývá tří či více hodnot (testy pro 3 a více nezávislých výběrů).

Sledování závislosti kvantitativní spojité proměnné na proměnných kategoriálních

Pokud jsou vysvětlujícími proměnnými pouze proměnné kategoriální, provádí se zjišťování závislosti pomocí *analýzy rozptylu*. Jestliže chceme též odhadovat hodnoty vysvětlované proměnné, provedeme *regresní analýzu* (lineární či nelineární), přičemž můžeme jako vysvětlující použít proměnné různých typů. Nominální proměnnou však musíme převést na pomocné proměnné, například na binární.

Sledování závislosti kvantitativní spojité proměnné na kvantitativních proměnných

Modelování těchto vztahů je předmětem *regresní analýzy*.

Sledování vzájemné závislosti tří a více kvantitativních spojitých proměnných

K tomuto účelu slouží *korelační analýza*, intenzita závislosti je posuzována pomocí celkového (vícenásobného) korelačního koeficientu.

Sledování vzájemné závislosti tří a více kategoriálních proměnných

Obecným přístupem je *rozšíření chí-kvadrát testu o nezávislosti* pro dvě proměnné přidáváním dalších rozměrů. Jestliže zamítneme hypotézu o nezávislosti, zjišťujeme dále, zda není nezávislost porušena pouze u některé skupiny proměnných.

Pro ordinální proměnné lze použít *neparametrické testy pro 3 a více závislých výběrů*. K nim patří *Friedmanův test*, který je založen na pořadí hodnot. Jsou porovnávána průměrná pořadí pro všechny proměnné. Základní idea je taková, že pokud není rozdíl mezi výběry, pak není rozdíl mezi průměrnými pořadími. Friedmanovo testové kritérium má při platnosti hypotézy H_0 přibližně chí-kvadrát rozdělení.

Ve výstupu z programového systému získáváme dvě tabulky. První obsahuje průměrná pořadí a druhá vlastní výsledek testu, který zahrnuje: počet pozorování, hodnotu testového kritéria chí-kvadrát, počet stupňů volnosti a minimální hladinu významnosti.

Sledování závislosti kategoriální proměnné na 2 a více kategoriálních proměnných

Tímto typem úloh se zabývá *loglineární analýza*.

Zkoumání podobnosti proměnných – shlukování (segmentace)

Kromě neparametrických testů pro závislé výběry, které jsou určeny pro ordinální proměnné a při nichž je nutno zadávat, podobnost kterých proměnných chceme zjišťovat, existují metody zaměřené na shlukování. Protože je současně zjišťována rozdílnost skupin proměnných, jsou v současné literatuře (zejména v souvislosti s temínem „data mining“) označovány tyto úlohy jako segmentace.

Shlukovou analýzu můžeme použít v případech, kdy jsou proměnné stejného typu. Speciální míry vzdálenosti (resp. podobnosti) existují pro diskrétní číselné proměnné a pro proměnné binární Při *hierarchické shlukové analýze* se počítá matice vzdálenosti, resp. podobnosti, nejprve pro všechny dvojice proměnných a poté se kombinují vzdálenosti jednak mezi jednotlivými proměnnými, jednak mezi vzniklými shluky proměnných. Uvedený postup je použitelný jak pro zjišťování podobnosti proměnných, tak pro zjišťování podobnosti objektů.

Speciálním přístupem pro zjišťování podobnosti kvantitativních proměnných je faktorová analýza. Jejím základem je analýza hlavních komponent, která slouží ke snížení rozměrů úlohy (místo mnoha proměnných je pro další výpočty určen malý počet hlavních komponent, které lze vyjádřit lineární kombinací původních proměnných). Pro případ, kdy nelze použít lineární kombinaci, je určena *nelineární faktorová analýza*.

Zkoumání podobnosti objektů (shlukování)

Pro tento typ úloh je určena *shluková analýza*, o níž bylo pojednáno v souvislosti s výše uvedeným typem úloh.

Zkoumání podobnosti kategorií (jedné či dvou proměnných)

Pro tento případ lze použít shlukovou analýzu založenou na četnostech v kontingenční tabulce, pro dvě proměnné dvourozměrnou shlukovou analýzu.

Zařazování objektů do skupin (klasifikace)

Jestliže je kategoriální proměnnou pouze proměnná vysvětlovaná a vysvětlující proměnné jsou kvantitativní spojité, používá se *diskriminační analýza*. Na základě analýzy vztahů mezi vysvětlujícími a vysvětlovanou proměnnou lze pro neznámý objekt se známými hodnotami vysvětlujících proměnných odhadnout zařazení tohoto objektu do definovaných skupin (tj. odhadnout hodnotu vysvětlované proměnné).

ANALÝZA HLAVNÍCH KOMPONENT

(Principal Component Analysis - PCA)

Cílem analýzy hlavních komponent je snížení dimenze dat. Poměrně často vykazují jednotlivé měřené veličiny silnou korelaci. V takovém případě je možné celou skupinu proměnných nahradit veličinou jedinou (nebo menším počtem veličin), které budou nést o datech téměř stejnou informaci, jako nesly veličiny původní. Jedná se vlastně o rotaci souřadnicového systému takovou, aby obrazy případů v nové souřadné soustavě vyhovovaly určitému kritériu.

Příklad: Struktura ekonomiky evropských zemí v roce 1979

1. Data

Data se nachází v souboru *Země.sta.* Proměnné představují procenta obyvatel zaměstnaných v různých odvětvích ekonomiky. Údaje jsou za evropské země v roce 1979. Zdrojem dat je Euromonitor (1979), European Marketing Data and Statistics, London: Euromonitor Publications, 76-77.

2. Cíl analýzy

Zajímá nás, zda lze devět údajů o každé zemi nějakým způsobem shrnout, tak aby bylo možno označit zemi např. jako průmyslově zaměřenou nebo orientovanou na zemědělství či služby. Rovněž bychom chtěli identifikovat země, které mají podobnou strukturu ekonomiky.

3. Standardizace dat

Analýzu můžeme provádět buď na základě korelační matice nebo na základě kovarianční matice. Pokud vycházíme z korelační matice, pak je standardizace zbytečná. Kovarianční matici používáme naopak v případě, kdy data standardizovat nechceme. Ve výsledcích analýzy se pak projeví rozdílná měřítka jednotlivých proměnných.

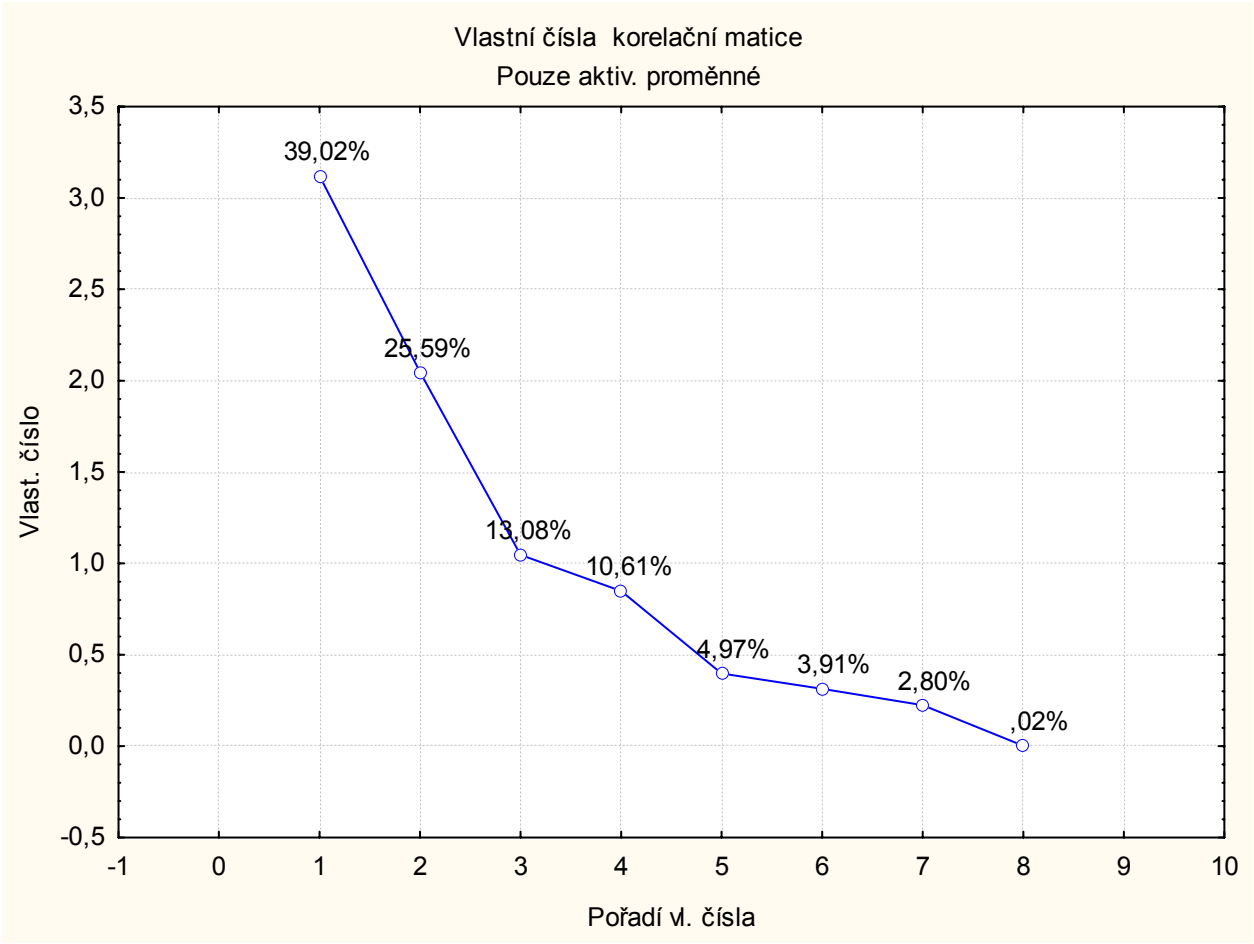
4. Spuštění analýzy a výběr proměnných

Statistika - Vícerozměrné průzkumné techniky • Hlavni komponenty & klasifikační analýza • tlačítko *Proměnné:* • v oddělení *Proměnné pro analýzu:* vybereme všechny kromě poslední (poslední proměnná představuje dopočet do 100% a je pro analýzu nadbytečná - *OK· OK.*

5. Vlastní čísla & Kolik potřebujeme hlavních komponent?

Bylo již řečeno, že cílem analýzy je „vměstnat co nejvíce informací“ do několika málo nových proměnných - hlavních komponent. První o co se tedy ve výsledcích budeme zajímat je, nakolik je taková „komprese“ možná a kolik bude třeba výsledných komponent.

Výpočet hlavních komponent vychází z vlastních čísel korelační matice (resp. kovarianční matice), přičemž každému vlastnímu číslu přísluší jedna hlavní komponenta (což je vlastní vektor). Vlastní čísla zobrazíme tímto postupem: karta *Základní výsledky* - tlačítko *Sutinový graf* (viz Obrázek 1).



Obrázek 1

Samotná vlastní čísla pro nás příliš důležitá nejsou. Mnohem důležitější informaci nám v tomto grafu přináší procenta nad vlastními čísly. Tato procenta udávají, jaká část rozptylu původních proměnných je vysvětlena komponentou, která přísluší danému vlastnímu číslu. Např. první, „nejdůležitější“, komponenta v našem příkladě vysvětluje zhruba 40 % rozptylu (přesněji 39,02 %). To lze chápat tak, že pokud bychom všech osm původních proměnných nahradili touto jednou proměnnou, uchováme zhruba 40 % informací, které byly v původních proměnných. Druhá komponenta vysvětluje dalších 26 %, čili společně tyto dvě hlavní komponenty již obsáhnou přes 60 % informace. Postupné „narůstání“ vysvětleného rozptylu lze vidět také v tabulce (viz Obrázek 2).

Vlastní čísla korelační matice a související statistiky (Zeme) Pouze aktiv. proměnné

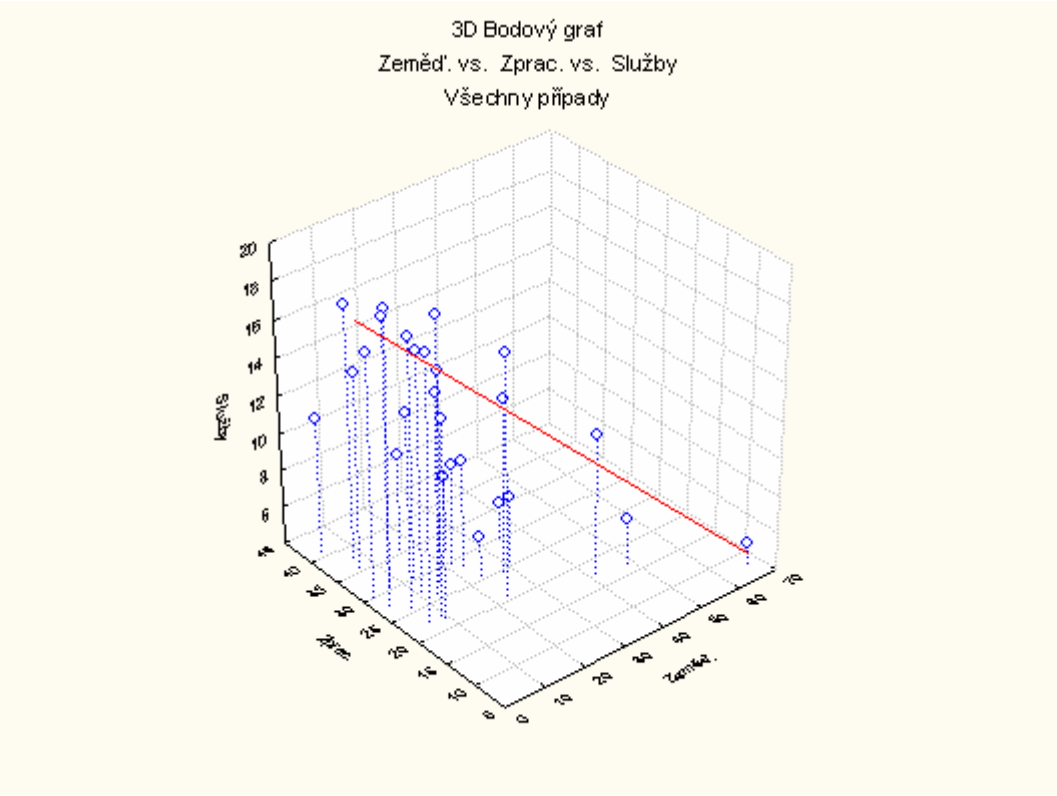
	vl. číslo	% celk.	Kumulativ.	Kumulativ.
1	3,121264	39,01579	3,121264	39,0158
2	2,046991	25,58739	5,168255	64,6032
3	1,046581	13,08227	6,214836	77,6855
4	0,848820	10,61025	7,063656	88,2957
5	0,397831	4,97288	7,461487	93,2686
6	0,313145	3,91431	7,774631	97,1829
7	0,223609	2,79511	7,998240	99,9780
8	0,001760	0,02200	8,000000	100,0000

Obrázek 2

Tabulku získáme na záložce *Základní výsledky* - tlačítko *Vlastní čísla.* Vidíme, že budeme-li přidávat komponenty, bude stoupat i procento reprodukované variability. Pokud bychom uvažovali všechny komponenty, dosáhneme 100 %. To však není cílem analýzy. Snažíme se docílit co nejvyšší procento, ovšem při co nejmenším počtu komponent. Protože tyto požadavky jdou „proti sobě“, je většinou třeba zvolit kompromis. Na to, kolik procent je dostačujících a kolik by celkem mělo být komponent, žádné obecné pravidlo není a závisí to na cílech analýzy a datech. Někdy se doporučuje použít ty komponenty, které odpovídají vlastním číslům větším než 1 (vycházíme-li z kovarianční matice, pak komponenty odpovídající kladným vlastním číslům). Zde by to byly první 3 komponenty, které společně vystihují 78 % rozptylu. Jiný přístup vychází ze sutinového grafu. Zde se vychází z analogie valících se kamenů v horách (odtud název tohoto grafu). Přidávání dalších komponent by se mělo zastavit v takovém místě, kde by se zastavily sesouvající se kameny, které by na grafu padaly od prvního vlastního čísla. To by v tomto případě znamenalo vzít 3 nebo 5 komponent.

6. Co jsou hlavní komponenty

Opusťme na chvíli náš příklad a představme si, že máme jako vstupní pouze 3 proměnné (např. zemědělství, zpracovatelský prům. a služby). Hodnoty těchto tří proměnných u každé země představují souřadnice bodu - země v třírozměrném prostoru. Když si tyto body zobrazíme v grafu, dostaneme Obrázek 3. První hlavní komponenta je v tomto grafu zakreslena červenou přímkou. Je to přímka, která nejlépe vystihuje data ve smyslu nejmenších čtverců. Druhá komponenta by byla opět přímka, vedená tak, aby byla kolmá na první komponentu a zároveň co nejvíce zlepšovala aproximaci dat. Máme-li tedy odpovědět na otázku, co jsou hlavní komponenty, můžeme říct, že jsou to vlastně osy nového souřadnicového systému, jehož dimenze je „rozumně“ malá. Analýza hlavních komponent tedy vytvoří nový souřadnicový systém a převede do něj původní proměnné. Zde by například stačila jedna dimenze daná první hlavní komponentou na to, aby vystihla podstatné chování tří původních proměnných. Každé zemí by pak příslušelo pouze jedno číslo - hodnota na nové červené přímce – namísto původních tří.



Obrázek 3

7. Interpretace komponent

Z matematického hlediska jsou hlavní komponenty souřadnými osami. Co ale tyto osy měří? To lze zjistit na základě jejich korelace s původními proměnnými. Tabulku s korelacemi (viz Obrázek 4) dostaneme na záložce *Proměnné* - tlačítko *Korelace faktorů & proměnných.*

Korelace faktorů a proměnných (faktor. zátěže) podle korelací (Zeme)

	Faktor 1	Faktor 2	Faktor 3	Faktor 4	Faktor 5	Faktor 6	Faktor 7	Faktor 8
Zeměď.	0,984816	-0,069479	0,015595	0,091869	0,124129	0,005292	-0,008305	-0,033497
Těžba	0,117003	0,909035	0,168922	-0,054168	-0,088986	-0,012531	0,347012	-0,002004
Zprac.	-0,611281	0,653176	-0,143788	0,136420	-0,298899	0,174293	-0,201168	-0,014987
Energie	-0,426550	0,461182	0,688541	0,024128	0,309063	-0,109848	-0,152006	-0,001951
Staveb.	-0,595711	0,170336	-0,537111	0,463779	0,222330	-0,243149	0,062939	-0,004656
Služby	-0,803198	-0,373862	0,038513	-0,018446	0,227085	0,369978	0,157304	-0,008978
Finance	-0,282613	-0,575013	0,482111	0,509713	-0,279237	-0,100651	0,095504	-0,005237
Sociál. s.	-0,708685	-0,277606	-0,025259	-0,585675	-0,103304	-0,253730	0,041486	-0,016602

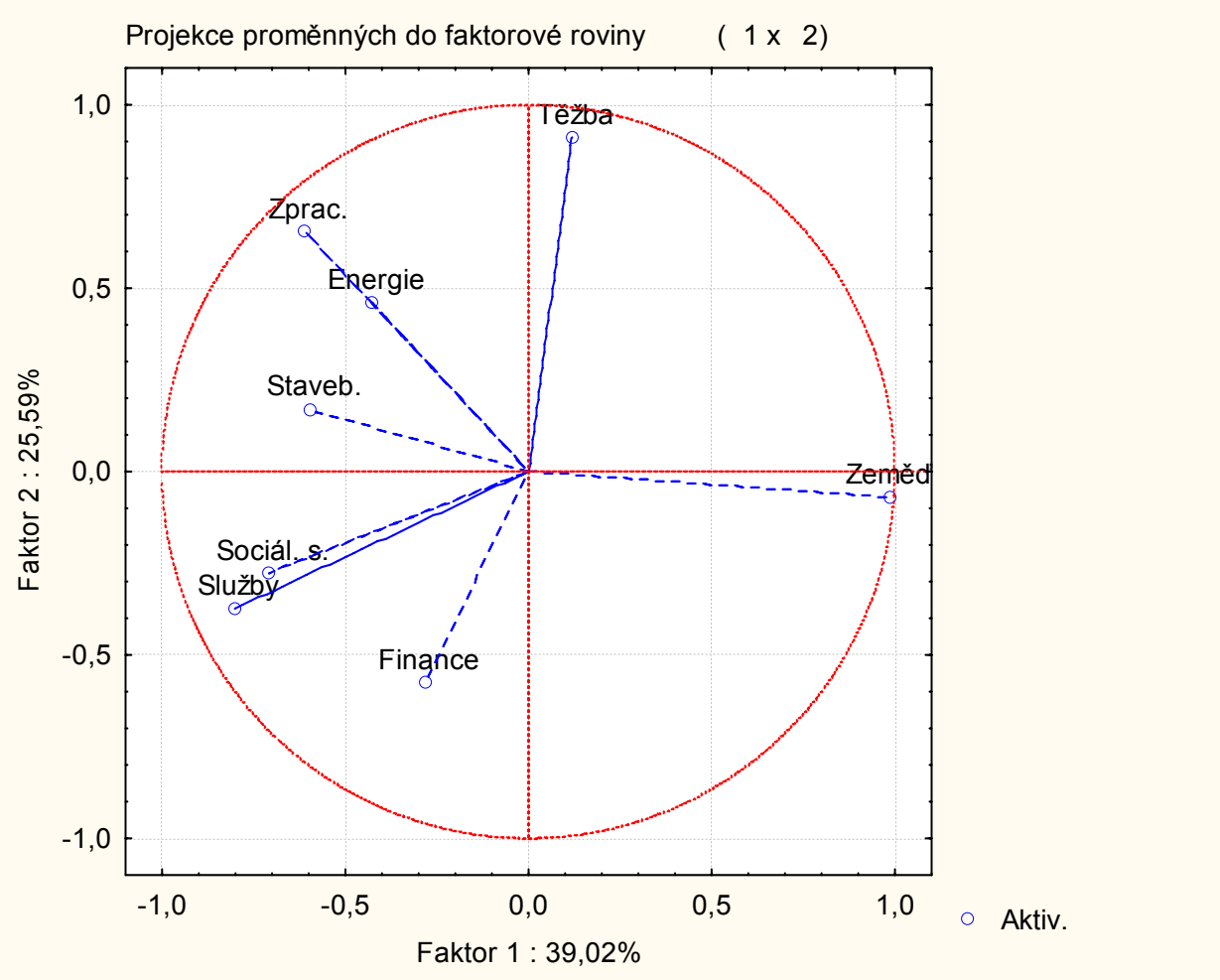
Obrázek 4

POZOR. Pojmenování Faktor 1 a Faktor 2 atd. může být trochu matoucí. Ve většině vícerozměrných metod v systému *STATISTICA* jsou takto obecně pojmenovány nově tvořené proměnné. Je třeba mít na paměti, že tyto proměnné obecně nemají stejný význam, jako faktory ve faktorové analýze!

Souvislost mezi vstupními a novými proměnnými lze také zobrazit graficky. Příslušný graf (viz Obrázek 5) vytvoříme následujícím postupem: záložka *Proměnné* - tlačítko *2D graf fakt. souřadnic prom.* - *Osa x:* Faktor I, *Osa y:* Faktor 2 - *OK*. Na ose x budou skóre vstupních proměnných vzhledem k první hlavní komponentě, na ose Y vzhledem ke druhé komponentě.

POZOR. Vycházíme-li při výpočtu komponent z korelační matice, jsou faktorové souřadnice proměnných normované a jsou zároveň i korelacemi s původními proměnnými (podobně jako v regresi jsou standardizované regresní koeficienty korelacemi s vysvětlujícími proměnnými). Toto neplatí, pokud vycházíme z kovarianční matice.

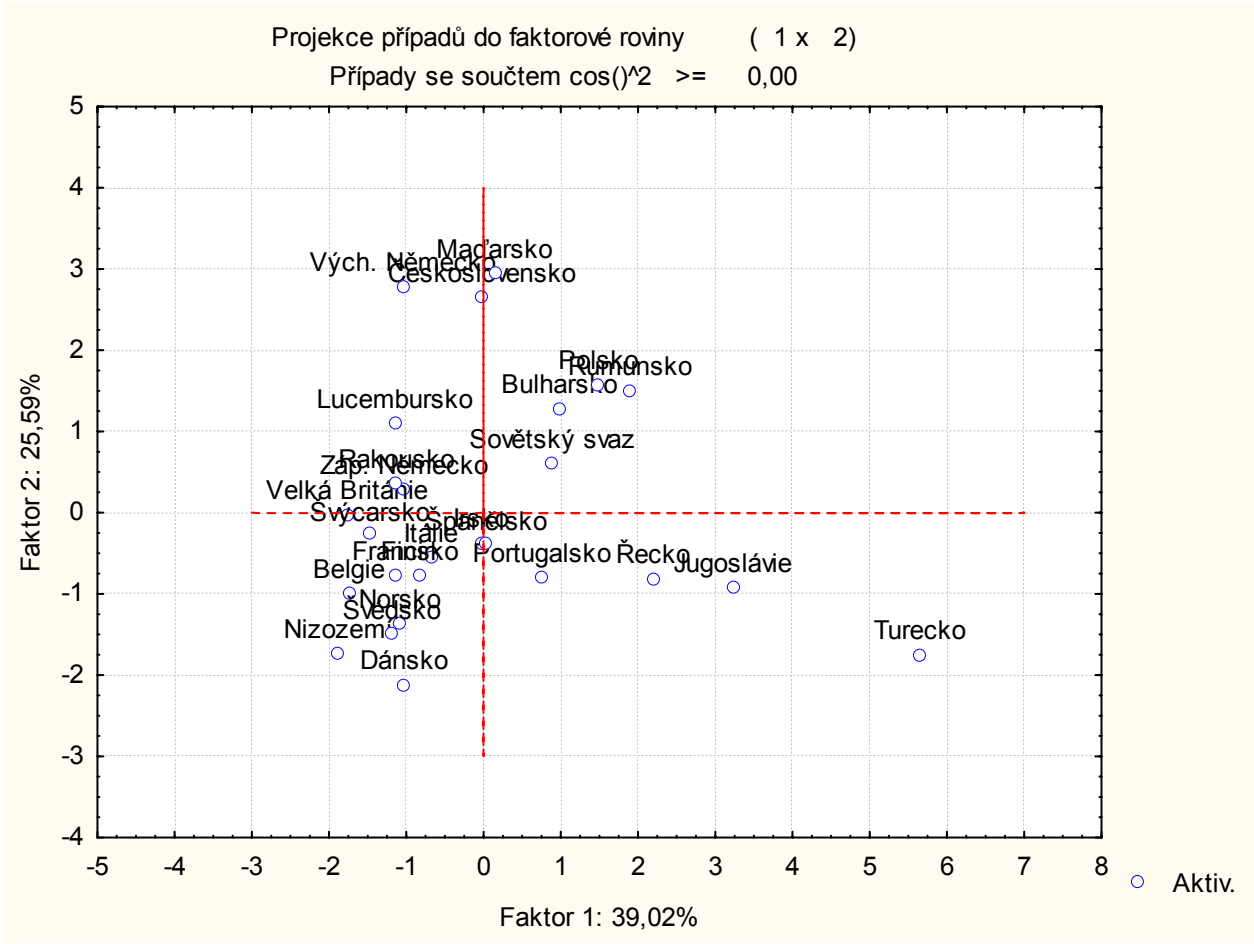
Z grafu i tabulky je zřetelné, že první komponenta je vysoce pozitivně korelována se zemědělstvím a naopak negativně se službami. Jelikož je podíl lidí v zemědělství a ve službách obecně považován za určité měřítko vyspělosti země, můžeme první komponentu interpretovat jako míru zaostalosti/vyspělosti. Druhá komponenta výrazně pozitivně koreluje s těžebním průmyslem, energetikou a zpracovatelským průmyslem. Negativně koreluje se službami a finanční sférou. Budeme ji proto interpretovat jako míru toho, nakolik se země orientuje na průmyslovou výrobu. (Podotkněme ještě, že ne vždy mají komponenty takto jasnou interpretaci. Jsou jen jistou matematickou transformací vstupních proměnných, která může a nemusí odrážet nějakou reálnou vlastnost objektů!). Podobně bychom si mohli nechat vytvořit grafy dalších komponent.



Obrázek 5

8. Co jsou projekce případů

Nyní tedy známe význam komponent. Podívejme se, jak jsou na tom jednotlivé země ve vztahu k těmto komponentám. Neboli, nakolik jsou jednotlivé státy zemědělsky zaměřené (což indikuje komponenta 1) a nakolik průmyslově zaměřené (komponenta 2). To nám ukazuje Obrázek 6, který dostaneme na záložce *Případy* - tlačítko *2D graf fakt. souřadnic příp.* - *Osa* x: Faktor I, *Osa y:* Faktor 2 - *OK*.



Obrázek 6

Státy napravo jsou státy s vysokým podílem zemědělství. Vyniká zde zejména Turecko a Jugoslávie. Všechny státy obvykle považované za ekonomicky vyspělé jsou naopak na levé straně. Jsou to státy, kde je nižší podíl osob zaměstnaných v zemědělství, zato vyšší podíl osob pracujících ve službách. Je zde také hezky vidět zaměření zemí tehdejšího socialistického bloku na průmyslovou výrobu - horní část grafu. A naopak severské státy a státy Beneluxu orientované na finanční a další služby v dolní části.

Pokud by nás zajímaly přesné číselné hodnoty, získáme je postupem: záložka *Případy* - tlačítko *Faktorové souřadnice případů*. „Ručně“ bychom tyto dvě nové souřadnice získali tak, že bychom u každého případu vynásobili vektor vstupních proměnných prvním a druhým vlastním vektorem. (přesněji bychom měli říci vektor standardizovaných hodnot X, pokud vycházíme z korelační matice, nebo vektor odchylek od průměrů, vycházíme-li z kovarianční matice.)

FAKTOROVÁ ANALÝZA
(Factor analysis)

Faktorová analýza má především dva cíle: zmenšit počet proměnných, které vstupují do dalších analýz, a zjistit strukturu vztahů mezi proměnnými. V mnohém se podobá analýze hlavních komponent, ale její podstata a interpretace výsledků je odlišná:

(a) Zatímco v analýze hlavních komponent jsou komponenty vytvořeny tak, aby maximalizovaly vysvětlený rozptyl, faktory ve faktorové analýze se snaží vystihnout korelační strukturu.

(b) Hlavní komponenty jsou lineární kombinace proměnných - vznikají z proměnných. Naproti tomu ve faktorové analýze to „funguje obráceně“ - z faktorů „vznikají“ proměnné. Každou proměnnou v ní chápeme jako výsledek souhry faktorů (= skrytých, často neměřitelných vlivů stojících v pozadí) a náhodné složky. Uveďme si ilustrační příklad. V jisté studii se zjistilo, že existuje silná korelace mezi velikostí bot dětí a výsledky v testech. Děti s většíma nohama měly lepší výsledky. Znamená to, že čím máme větší nohy, tím jsme inteligentnější? Bohužel ne. Autoři studie jaksi zapomněli zohlednit fakt, že obě veličiny závisí na věku a do studie zahrnuli děti v různých věkových kategoriích. Pochopitelně starší děti měly jak větší nohy, tak lepší výsledky testů, kdežto u mladších dětí to bylo obráceně. Když se změřila korelace nohou a testů u dětí téhož věku, zjistilo se, že veličiny spolu nekorelují. A jak s tím souvisí faktorová analýza? Faktorová analýza hledá právě takové veličiny/faktory, jako byl v našem příkladu věk. Něco, co skrytě ovlivňuje proměnné a způsobuje, že spolu korelují. Když se tento vliv eliminuje, korelace **„vymizí“.**

(c) Hlavní komponenty jsou umělé proměnné, jasně dané korelační / kovarianční maticí. Jsou čistě matematickou záležitostí a nemusí mít vždy věcnou interpretaci. Faktorová analýza oproti tomu nedává jednoznačné výsledky. Záleží na úsudku řešitele, které řešení zvolí. Důraz se klade především na to, aby faktory byly smysluplné. Aby měly věcné opodstatnění. Je zřejmé, že tato analýza se neobejde bez znalosti zkoumané problematiky.

Příklad*: *Výsledky desetiboje z olympiády v Athénách

1. Data

Soubor s daty se nazývá ***Desetiboj.sta.*** Proměnnými jsou výsledky desetiboje mužů na letních olympijských hrách v Athénách (2004).

2. Cíl analýzy

Zajímá nás, zda lze identifikovat faktory, na kterých závisí výsledky v jednotlivých disciplínách. A dále které faktory jsou nejdůležitější pro vítězství.

3. Úprava dat

Při pohledu na datový soubor zjistíme, že část závodníků neabsolvovala všechny disciplíny. Vyloučení těchto případů z analýzy zajistíme buď pomocí filtru nebo až bezprostředně v analýze pomocí správy chybějících dat (viz bod 5).

4. Standardizace dat

Není nutná. Faktorová analýza, jakožto metoda vycházející z korelační matice, není závislá na měřítku vstupních hodnot.

5. Spuštění analýzy a výběr proměnných

Statistika - *Vícerozměrné průzkumné techniky* - *Faktorová analýza* - tlačítko *Proměnné:* zvolíme proměnné 4-13 s výsledky disciplín (další možností by bylo použít body za disciplíny) - *OK* - v oddělení *Vstupní soubor:* ponecháme *Zdrojová data*· pokud není použit filtr, ujistíme se, že v oddělení *ChD vynechána* je zatržena volba *Celé případy* - *OK*.

6. Prozkoumání vstupních dat

K tomu, abychom mohli provést faktorovou analýzu, jsou prvním předpokladem nenulové korelace mezi proměnnými. Těžko bychom pochopitelně mohli modelovat korelační strukturu, když by mezi proměnnými žádné korelace nebyly. Nejprve si proto prohlédneme korelační matici. Na záložce Popisné statistiky - tlačítko Přehled korelací, průměru, směrodatných odchylek – v následujícím dialogu záložka Základní výsledky - tlačítko Korelace. Pro posouzení normality jsou zde k dispozici na záložce Detaily - tlačítko Histogramy a tlačítko Normální p-graf.

7. Metody odhadu faktorů

Nyní stojíme před první fází faktorové analýzy, v níž budou odhadnuty faktory. (Cílem druhé fáze pak bude rotace vzniklých faktorů tak, aby se usnadnila jejich interpretace). Zde je třeba zvolit, jakým způsobem budou faktory odhadnuty. Jednou z nejčastěji používaných metod je metoda hlavních komponent, se kterou jsme se již seznámili v předešlé kapitole. Vzniklé faktory jsou pak vlastně hlavními komponentami. V systému *STATISTICA* je tato metoda nastavena jako výchozí. Kromě ní je k dispozici řada dalších metod. V dialogu Metoda extrakce faktoru je nalezneme na záložce Detaily.

Obecně se doporučuje jako moudré vyzkoušet více různých metod. Je-li za daty skutečná faktorová struktura, pak výsledky různých metod bývají konzistentní.

8. Vlastní čísla & Kolik vytvořit faktorů

Je třeba také zvolit, kolik faktorů chceme vytvořit. Máme-li ze znalostí tématiky představu, kolik faktoru připadá v úvahu, pak toto nastavíme na záložce *Detaily* - volba *Max. počet faktorů:*. Pokud apriori nevíme, kolik faktoru vytvořit, je dobré pro začátek nastavit raději vyšší počet a ve volbě *Min. vlastni číslo:* snížit hranici pro minimální vlastní číslo. Nastavme počet faktorů na 10 a minimum na O. *OK*.

Stejně jako v předchozím příkladě si nejprve prohlédneme vlastní čísla a podíly vysvětleného rozptylu. Ve výsledkovém dialogu je získáme na záložce *Zákl. výsledky* - tlačítko *Vlastní čísla* (Obrázek 6). Vidíme, že vlastní čísla větší než 1 jsou tři a faktory/ komponenty jim příslušející vystihují zhruba 70% variability původních proměnných. Je na zvážení, zda použít ještě čtvrtý faktor, kterým bychom zvýšili procento vysvětleného rozptylu na 78%.

Vl. čísla (Desetiboj) Extrakce: Hlavní komponenty

	vl. číslo	% celk.	Kumulativ.	Kumulativ.
1	3,545628	35,45628	3,545628	35,45628
2	1,969494	19,69494	5,515122	55,15122
3	1,421791	14,21791	6,936913	69,36913
4	0,903646	9,03646	7,840559	78,40559
5	0,563241	5,63241	8,403800	84,03800
6	0,527759	5,27759	8,931559	89,31559

Obrázek 7

Prostřednictvím tlačítka *Storno* se vrátíme do předešlého dialogu *Metoda extrakce faktorů* a zde zadáme, že požadujeme odhadnout 3 faktory. Po *OK* se vrátíme opět do výsledkového dialogu. Tentokrát se odhadnou a zobrazí výsledky pouze pro tři faktory. Uvědomme si, že jelikož výpočet probíhá metodou hlavních komponent, jsou tyto tři faktory stejné jako byly první tři faktory v předcházejícím kroku. Proč jsme se tedy vraceli a zadávali, že chceme tři faktory?

Rozdíl nastane v následující fázi, kdy budeme faktory rotovat. Rotace tří faktoru dá jiné výsledky než rotace 5 faktoru a opět jiné pro 8 faktorů.

9. Faktorové zátěže & Rotace a interpretace faktorů

Dostáváme se do druhé fáze - máme prvotní faktory a chceme je transformovat tak, aby měly smysluplnou interpretaci. Tato fáze se nazývá rotace faktorů a jde v ní o to, že námi vytvořené faktorové souřadnice rotujeme tak, aby vznikla jasná korelační struktura. Snažíme se dosáhnout toho, aby byl každý faktor korelován pouze s určitou skupinou proměnných a korelace s ostatními proměnnými byly nulové. Cílem je najít smysluplné faktory.

Stejně jako u všech ostatních vícerozměrných metod odvozujeme interpretaci faktoru z jejich korelací se vstupními proměnnými. Tyto korelace se zde nazývají faktorové zátěže (factor loadings) a nalezneme je na záložce *Zákl. výsledky* resp. na záložce Zátěže - tlačítko *Shrnutí: Faktorové* zátěže. Obrázek 8 ukazuje korelace jednak před provedením rotace, jednak po rotaci metodou Varimax. Najít interpretaci faktoru před rotací je zde problematické. První faktor by se určitým způsobem interpretovat dal - koreluje se všemi disciplínami tak, že čím lepši výsledek, tím nižší hodnota faktoru. Byl by tedy ukazatelem celkového výkonu. Interpretace druhého faktoru je však záhadou.

Faktor. zátěže (Bez rot.) (Desetiboj) Extrakce: Hlavní komponenty (Označené zátěže jsou >,700000)

	Faktor	Faktor	Faktor
Běh 100 m	**0,795801**	-0,253471	0,252856
Skok do dálky	**-0,793590**	0,324855	-0,155018
Vrh koulí	-0,628538	-0,623456	-0,024067
Skok do výšky	-0,625730	-0,472605	0,011378
Běh 400 m	**0,734613**	-0,493797	-0,228731
110 m překážky	**0,708973**	-0,231895	0,044744
Hod diskem	-0,541686	-0,667541	-0,018709
Skok o tyči	-0,179397	0,326781	-0,623548
Hod oštěpem	-0,286747	-0,338819	0,522717
Běh 1500 m	0,214050	-0,472077	**-0,785112**

Faktor. zátěže (Varimax pr.) (Desetiboj) Extrakce: Hlavní komponenty (Označené zatěže jsou >,700000)

	Faktor	Faktor	Faktor
Běh 100 m	**-0,842660**	-0,217283	-0,064756
Skok do dálky	**0,853894**	0,167822	-0,045244
Vrh koulí	0,185353	**0,863229**	0,069407
Skok do výšky	0,253925	**0,741986**	0,001451
Běh 400 m	**-0,798454**	-0,036115	0,443821
110 m překážky	**-0,708282**	-0,207337	0,117338
Hod diskem	0,090322	**0,850633**	0,087463
Skok o tyči	0,475904	-0,230255	0,498277
Hod oštěpem	-0,079191	0,493663	-0,469341
Běh 1500 m	-0,224590	0,183246	**0,895017**

Obrázek 8

U rotovaných faktorů je situace jednodušší. První faktor jasně souvisí s výsledky krátkých běhů a skoku do dálky - čím lepší výsledek, tím vyšší hodnota faktoru. Mohli bychom ho charakterizovat jako faktor rychlosti. Nejsilnější korelace druhého faktoru jsou se všemi

„vrhačskými“ disciplínami a skokem do výšky. Nebýt skoku do výšky, interpretovali bychom ho jako sílu paží. Takto bychom zřejmě řekli, že souvisí se schopnosti zkoncentrovat sílu v jediném okamžiku a napřít ji požadovaným směrem - do hodu, či skoku. Poslední třetí faktor je jasně korelován s během na delší vzdálenost, z čehož je vidět, že tato disciplína se od ostatních nejvíce odlišuje. POZOR. Tentokrát je korelace opačná - čím lepší výsledek, tím nižší hodnota faktoru. U předchozích dvou faktoru lepší výsledky znamenaly vyšší hodnoty faktorů. Tato orientace je ovšem zvolena náhodně a nezáleží na ní.

SHLUKOVÁ ANALÝZA
(Cluster analysis)

Shluková analýza představuje typickou metodu pro učení bez učitele. Představme si situaci, kdy máme naměřená určitá data na nějakých objektech a chtěli bychom tyto objekty klasifikovat do několika tříd. Naše trénovací množina ovšem neobsahuje žádnou informaci o tom, který objekt do které třídy patří (proto učení bez učitele). Nejdříve se může zdát, že taková úloha je nesmyslná. Když si ale uvědomíme, že třídy objektů jsou obvykle definovány tak, že prvky z jedné třídy si jsou něčím podobné, jsme již na dobré cestě, jak náš problém, alespoň pro určité typy úloh, vyřešit.

Klíčovým pojmem je tu *podobnost* obrazů objektů. Pokud jsme schopní určit, „jak moc" jsou si dva objekty podobné, nebrání nám nic v tom, abychom se pokusili vytvořit z těchto objektů „trsy" či „shluky" vzájemně podobných obrazů. Mírou podobnosti bývá obvykle nějaká vzdálenost v prostoru naměřených hodnot.

Skutečnost, že dva objekty z téže třídy si jsou podobné, je v praxi velice častá. Shluková analýza již byla uplatněna v mnoha oborech lidské činnosti: první aplikace byly v biologii a týkali se taxonomie živočichů a rostlin, v medicíně se úspěšně shlukovaly např. symptomy různých nemocí (každý shluk pak označoval osoby trpící stejnou chorobou), podobně v psychiatrii byly vytvořeny shluky příznaků pro paranoiu, schizofrenií apod., v archeologii se zase vyskytly pokusy vytvořit taxonomie kamenných nástrojů, pohřebních předmětů atd.

Příklad*: *Rozdělení zemí

1. Data

Použijeme datový soubor Země.sta, se kterým jsme pracovali v kapitole věnované analýze hlavních komponent.

2. Cíl analýzy

Budeme se snažit určit skupiny zemí, které mají podobnou strukturu zaměstnanosti v jednotlivých odvětvích.

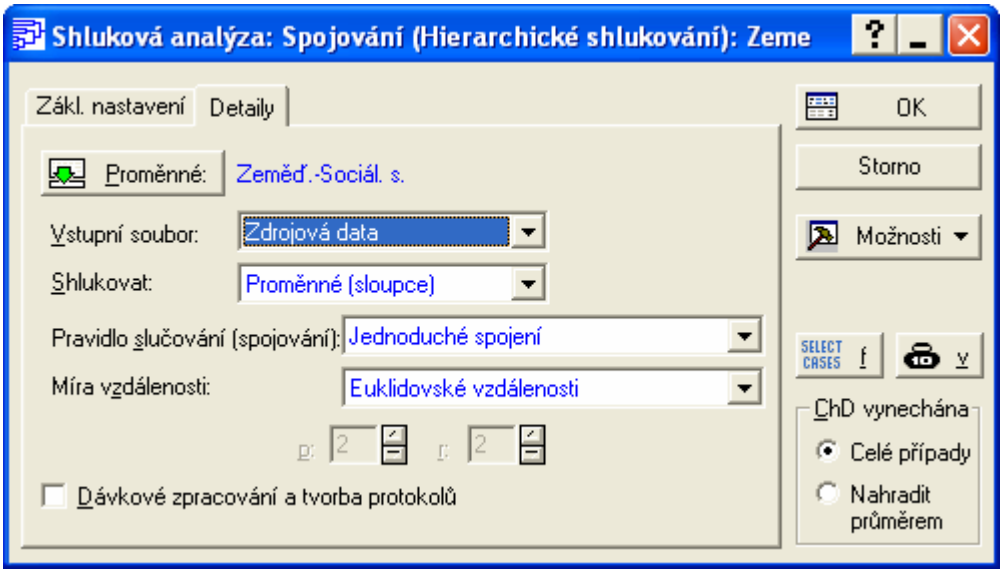
3. Standardizace dat

Všechny shlukovací algoritmy potřebují během své činnosti vyhodnotit vzdálenost mezi shluky nebo objekty. Vzdálenost samotná je velice závislá na měřítkách jednotlivých veličin. V tomto případě jsou všechny veličiny měřeny v procentech, proto není třeba data standardizovat. Často se ale v praxi setkáváme se situací, že měřítka proměnných jsou značně nesourodá (statisíce korun, počet sekund, …). V takových případech je úprava nezbytná, protože jinak by celá analýza závisela nejvíce na proměnné s největším rozsahem. Jindy ovšem může nastat situace, kdy data standardizovat nechceme.

4. Spuštění analýzy a volba proměnných

Statistika - Vícerozměrné průzkumné techniky - Shluková analýza - Spojováni (hierarchické shlukováni) - OK - záložka ***Zákl. nastaveni*** - tlačítko ***Proměnné*** - zvolit vše bez poslední proměnné **Zbytek. *OK.***

Nyní je třeba zvolit, jakým způsobem má být shlukování provedeno. K tomu slouží záložka ***Detaily*** (viz Obrázek 9).



Obrázek 9

První volba *Vstupní soubor: Zdrojová data* ponecháme, protože skutečně vycházíme z datové tabulky, nikoli z korelační matice.

Druhá volba *Shlukovat:* - nastavíme *Případy (řádky),* protože chceme zjišťovat podobnost mezi jednotlivými zeměmi-případy.

5. Jak zvolit vzdálenost objektů

Postup při shlukové analýze probíhá v zásadě ve dvou krocích. V prvním kroku se vypočtou vzdálenosti objektů (proměnných nebo případů) - každého od každého - a uloží se do matice vzdáleností. Ve druhém kroku se na základě této matice objekty postupně sdružují do clusterů. Podívejme se nejdřív podrobněji na první fázi. V ní hraje klíčovou úlohu to, jakým způsobem definujeme vzdálenosti mezi objekty. Výsledky analýzy se mohou podstatně lišit v důsledku různých použitých měr

V programu je možné zvolit z několika různých způsobů ve volbě *Míra vzdáleností:*. Podívejme se, jaké to jsou:

- Euklidovské vzdálenosti - $d(x,y) = \{\Sigma_i (x_i - y_i)\}^{1/2}$, - klasická míra vzdálenosti, která pro dva body v prostoru určuje délku „nejkratší cesty“ z jednoho bodu do druhého
- Blokové vzdálenosti (Manhattan) - $d(x,y) = \Sigma_i |x_i - y_i|$ - suma vzdáleností v jednotlivých dimenzích. Název i výpočet je inspirován vzdáleností, kterou na Manhattanu člověk urazí při cestě z jednoho bodu - nákupního střediska/baru/apod. :-) do druhého. Nelze jít po spojnici, musí se jít po kolmých ulicích
- Čebyševovy vzdálenosti - $d(x,y) = \text{Max} |x_i - y_i|$ - maximum ze vzdáleností v jednotlivých dimenzích
- Mocninné vzdálenosti - $d(x,y) = (\Sigma_i |x_i - y_i|^p)^{1/r}$ - uživatelem definovaná míra vzdáleností. Čím vyšší parametr p, tím vyšší váha se přikládá větším vzdálenostem v jednotlivých dimenzích a snižuje se význam malých vzdáleností. Vysoké p nejvíce „propaguje“ body hodně vzdálené ve všech dimenzích. Parametr r působí opačným směrem, čím vyšší r, tím menší váha se přikládá větším vzdálenostem. r ovšem působí celkově bez ohledu na dimenze
- Procentuální neshoda - $d(x,y) = (\text{počet } x_i \neq y_i) / i$ - je vhodná pouze pro kategorické proměnné. Pro dva objekty se spočte jako podíl počtu dimenzí, v nichž se jejich hodnota liší, ku celkovému počtu dimenzí
- 1- Pearsonův r - $d(x,y) = 1 - r(x,y)$ - míra založená na korelaci. Největší vzdálenost přiřazuje negativně korelovaným objektům, nejmenší naopak pozitivně korelovaným objektům. Nevhodná pro malý počet dimenzí.

Kterou míru vzdálenosti vybrat v našem konkrétním příkladu? Předem můžeme vyloučit Procentuální neshodu, protože ta je určena pro kategorické proměnné. Dále si uvědomme, že jsme neprováděli standardizaci dat. Některé proměnné mají tudíž větší rozptyl (např. zemědělství), jiné menší. Pokud bychom zvolili Čebyševovu míru vzdálenosti, bude o zařazení do clusterů rozhodovat právě zemědělství s největším rozptylem a vliv ostatních proměnných bude zanedbatelný. Zřejmě proto ani tato míra nebude nejvhodnější. Všechny ostatní míry jsou přijatelné a lze je postupně vyzkoušet. Pro začátek ponechejme Euklidovské vzdálenosti. U všech měr bude pochopitelně vliv zemědělství také podstatný. To je ovšem namístě, protože se země liší nejvíce právě díky němu.

POZOR. Obecně ale nemusí být proměnná s největším rozptylem skutečně ta, která nejvíce odlišuje objekty!

6. Pravidla slučování

Zde se dostáváme k druhé fázi výpočtu. Volbou ***Pravidlo slučováni*** *(spojováni):* určíme, jak se budou na základě matice vzdáleností objekty sdružovat do shluků. Obecně všechny shlukovací algoritmy pracují tak, že v matici vzdáleností naleznou minimum a objekty, jimž tato vzdálenost přísluší spojí do clusteru. Pak následuje přepočet matice vzdáleností, během něhož se vypočtou vzdálenosti nového clusteru od ostatních objektů a clusterů. Poté se celý cyklus opakuje nalezne se minimum, objekty nebo jejich dříve vytvořené shluky se sdruží do nového clusteru a opět se spočtou jeho vzdálenosti - dokud nejsou všechny vstupní jednotky sdruženy do jednoho velkého clusteru.

To, v čem se odlišují jednotlivé shlukovací algoritmy je, jak počítají vzdálenost mezi dvěma shluky.

- Jednoduché spojení - vzdálenost dvou shluků se určí jako vzdálenost dvou nejbližších objektů (případů/proměnných). Rozumí se dvou nejbližších objektů z různých clusterů! Tento algoritmus má tendenci spojovat objekty do dlouhých „řetízků“
- Úplné spojení - vzdálenost shluků je naopak dána vzdáleností těch dvou objektů, které jsou nejdále od sebe. Algoritmus je vhodný pro případy, kdy jsou objekty přirozeně rozdělené do určitých skupin. Má tendenci spíše tvořit skupiny s podobným počtem objektů
- Nevážený průměr skupin dvojic - vzdálenost shluků je prostým průměrem vzdáleností všech párů objektů, které lze vytvořit tak, že z každého shluku vezmeme jeden objekt. Tato varianta algoritmu pracuje lépe v případech, kdy vstupní objekty mají spíš charakter oddělených skupin. Lze ale použít i pro objekty mající „řetízkovou“ strukturu
- Vážený průměr skupin dvojic - obdoba předchozího algoritmu. Při výpočtu průměru se navíc berou jako váhy počty objektů v jednotlivých clusterech
- Nevážený centroid skupin dvojic - vzdálenost shluků je určí jako vzdálenost mezi centroidy shluků. (Centroid je bodem definovaným průměry v jednotlivých dimenzích)
- Vážený centroid skupin dvojic (medián) - vážená varianta předchozího algoritmu
- Wardova metoda - metoda založená na principu analýzy rozptylu. Metoda je považována za velmi efektivní. Má ale tendenci vytvářet spíše malé shluky.

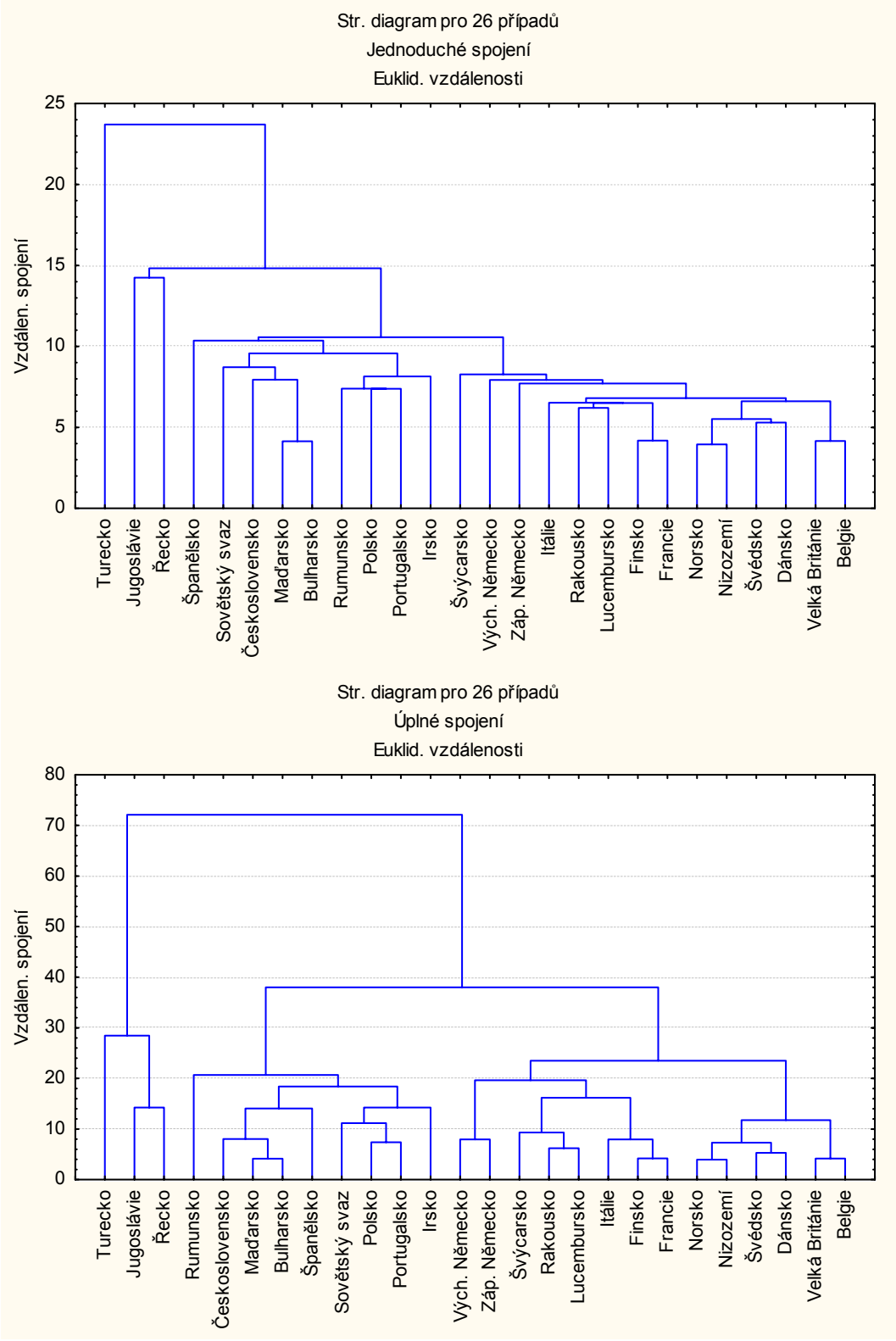
Když se opět vrátíme k našemu příkladu, víme, že jednotlivé státy by měly být klasifikovatelné do skupin minimálně na základě tehdejší příslušnosti k východnímu či západnímu bloku. Na druhou stranu ale nelze předem vyloučit ani možnost, že země budou tvořit „řetízkovou“ strukturu. Nemáme tedy jasné vodítko, které z pravidel slučování použít. V analýze proto vyzkoušíme Jednoduché spojení a Úplné spojení, jakožto dvě krajní a nejpoužívanější alternativy.

7. Jak interpretovat dendrogram

Po výběru míry vzdálenosti a pravidla slučování klikněte na ***OK.*** Ve výsledkovém dialogu, který se otevře, se nabízí dvě varianty výsledného grafu, tzv. dendrogramu horizontální vertikální. tlačítko. Zvolme ***Vertikální*** „***třásňový“ graf.*** Výsledky ukazuje Obrázek 10.

Interpretace grafů je velmi jednoduchá. Na ose X jsou jednotlivé objekty.Objekty spojené „vidličkou“ jsou součástí clusteru. Výška, v jaké jsou spojeny vodorovnou čarou, označuje, jak jsou od sebe objekty příp. clustery vzdálené. Jinými slovy, čím delší je svislá čára u daného objektu, tím menší je jeho podobnost s nejbližším objektem.

Grafy hezky demonstrují odlišnost obou postupů. Zřetězení zemí v prvním grafu a proti tomu tři jasně odlišené skupiny v druhém grafu. Všimněte si také odlišných měřítek Y-ové osy.



Obrázek 10

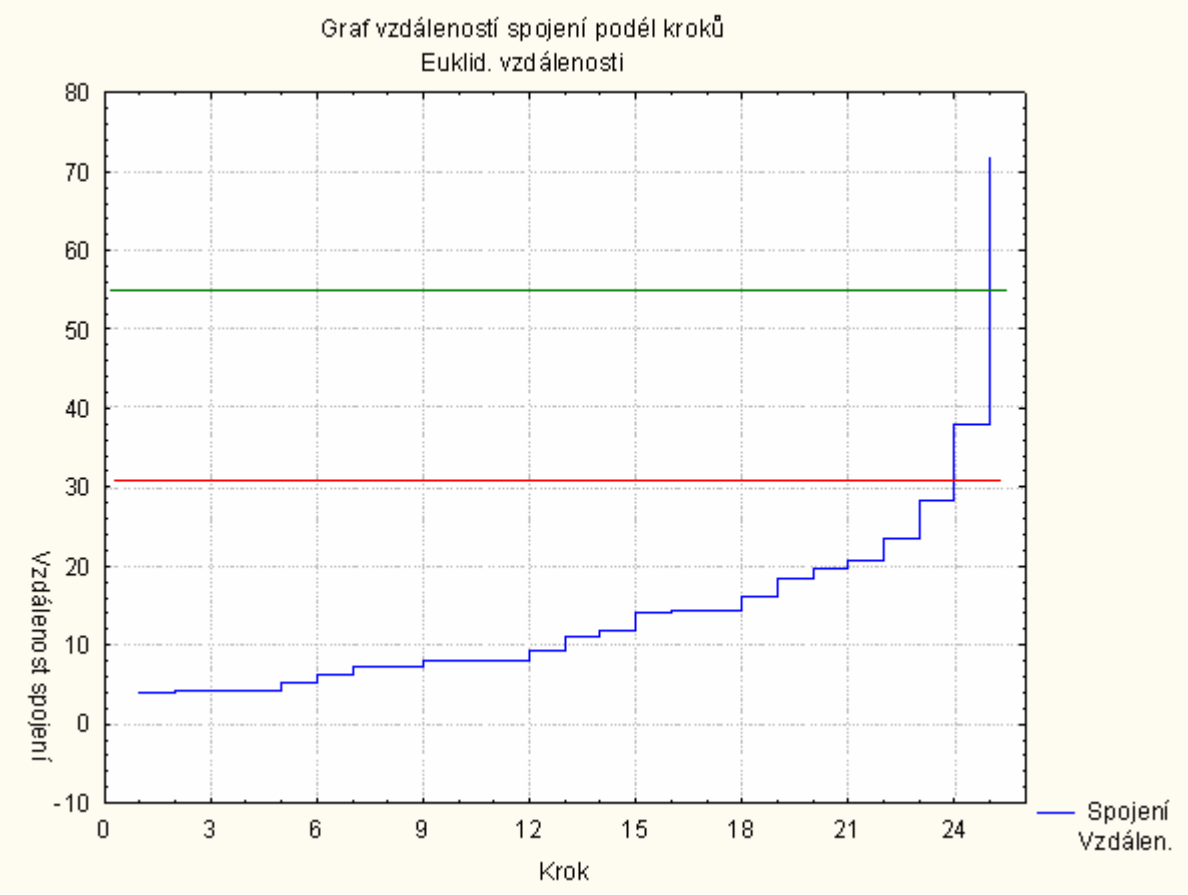
Celkové závěry jsou ale jednoznačné. Tři země v levé části - Turecko, Jugoslávie a Řecko - stojí zcela mimo hlavní proud. (Jak jsme zjistili v části 3, je to díky vysokému podílu zemědělství). Dále jsou si podobné země tehdejšího východního bloku - Československo, Maďarsko, Bulharsko... K nim se rovněž zařadily „chudší“ státy ze západu - Irsko, Španělsko a Portugalsko. Ostatní západní státy se shromáždily ve shluku v pravé části obou grafů. Výjimku tvoří pouze Východní Německo, které na základě společného historického vývoje bylo blíže Západnímu Německu než ostatním zemím východního bloku, kam patřilo.

8. Rozvrh shlukování

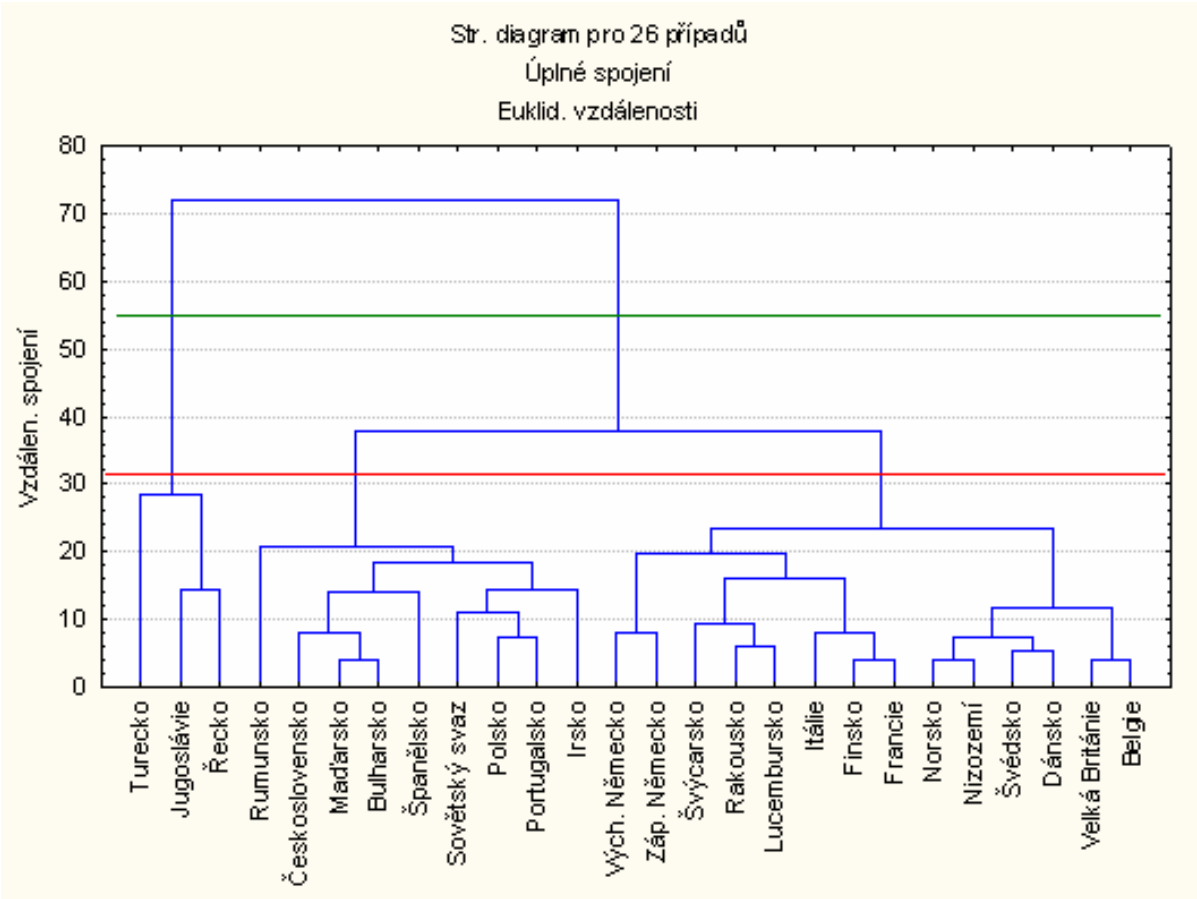
Podrobnější výsledky lze získat na záložce Detaily. Je zde jednak matice vzdáleností, o které jsme již hovořili, a jednak rozvrh shlukování.

Prostřednictvím záložka *Detaily* - tlačítko *Rozvrh shlukování* lze sledovat, jak byly postupně vytvářeny jednotlivé shluky. Graf rozvrhu shlukování nám zase může napovědět, zda lze objekty rozdělit do vzájemně odlišných skupin a kde určit hranici mezí skupinami.

Graf získáme pomocí tlačítka Graf rozvrhu shlukování (viz Obrázek 11). Na ose X je pořadí kroku, osa Y označuje vzdálenost objektů/shluků spojených v daném kroku. Čím plošší je křivka, tím více objektů je shlukováno zhruba na stejné úrovní vzdálenosti a dá se očekávat, že patří k téže skupině. Schody v grafu naopak naznačuj í větší odlišnost mezi shlukovanými objekty/shluky. V těchto místech je rozumné vést pomyslnou hranici, která oddělí jednotlivé skupiny jednotek. Dvě možná dělení jsou naznačena linkami.



Obrázek 11



Obrázek 12

LITERATURA

Blahuš P. *Základní pojmy statistické teorie psychologických testů*. Československá psychologie 33, 1989: 233-241.

Cyhelský, Kahounová, Hidls. *Elementární statistická analýza*. Praha: Management Press, 1996.

Hendl Jan. *Přehled statistických metod zpracování dat*. Praha 2004. ISBN: 80-7178-820-1

Kubíček J., Dufek J. *Statistika*. Brno: Vysoká škola zemědělská, 1978. 55-913a-78

McDonald R. P. *Faktorová analýza a příbuzné metody v psychologii*. Academia, Praha 1992, 252 s.

Meloun, Militký. *Statistické zpracování experimentálních dat*. Praha: PLUS, 1994

Seger, Hindls. *Statistické metody v tržním hospodářství*