INSTITUT PRO KRIMINOLOGII A SOCIÁLNÍ PREVENCI

VYBRANÉ METODY VÍCEROZMĚRNÉ STATISTIKY

(se zvláštním zaměřením na kriminologický výzkum)

Jaroslav Košťál

Vybrané metody kriminologického výzkumu

Svazek 4

Praha 2013

Autor:
JUDr. Jaroslav Košťál, CSc.

Recenzenti:
PhDr. Martina Klicperová, CSc.
PhDr. Ing. Petr Soukup

Technická spolupráce:
Lucie Černá

Řada Vybrané metody kriminologického výzkumu
Svazek 4
Odborný garant: PhDr. Martin Cejp, CSc.

Tento text neprošel jazykovou korekturou.

ISBN 978-80-7338-128-8

Obsah

Předmluva		4
Úvod		5
1.	Obecná charakteristika vícerozměrné statistiky	6
2.	Faktorová analýza (Factor analysis, Principal component analysis)	15
3.	Shluková analýza (Cluster analysis, klastrovací analýza, shlukovací analýza)	46
4.	Korespondenční analýza	80
5.	Mnohorozměrné škálování (multidimensional scaling)	88
Závěr		97
Souhrn		98
Summary		103
Seznam literatury		108

Předmluva

Na kriminologickém bádání se podílejí badatelé z různých oborů: právníci, psychologové, sociologové, ekonomové, politologové, statistici. Vzhledem k tomu, že každý z nich má specifické zkušenosti s výzkumnými postupy daných disciplín, chceme v rámci ediční řady, zaměřené na aplikaci výzkumných postupů v kriminologii, postupně pojednat o těch výzkumných metodách a technikách, které jsou v oblasti kriminologie aktuálně využívány. Pokoušíme se o zhodnocení používaných výzkumných metod a technik z hlediska jejich spolehlivosti a platnosti, přehodnocujeme, jak jsou nebo nejsou schopny zachytit skutečnost, do jaké míry je jejich použití při zkoumání kriminality funkční, jaké mohou obsahovat nepřesnosti a případná zkreslení. V jednotlivých studiích uvádíme četné příklady, na nichž chceme ukázat, jak a s jakým efektem byly v posledních letech různé výzkumné postupy použity. Zároveň se pokoušíme hledat inovace, které by mohly vést k větší plastičnosti, přesnosti a spolehlivosti, a tím i k využitelnosti zjištěných výsledků.

Na možnosti využití novějších, nebo alespoň v kriminologii určitě dosud nevyužívaných postupů, se podstatným způsobem zaměřuje čtvrtá studie ediční řady věnovaná využitelnosti vícerozměrných statistik. Ve studii je na příkladech prezentována jejich použitelnost v kriminologickém výzkumu, což by mohlo být prvním krokem k jejich soustavnějšímu využívání. V řadě zemí jsou vícerozměrné analýzy již desítky let používány, protože se ukazuje, že při aplikaci jednorozměrných a dvourozměrných statistik – které dosud zcela převažovalo - nejen že není při interpretaci zjištěných dat využit všechen potenciál, ale zjištěné souvislosti mezi proměnnými mohou být i ne zcela správně pochopeny.

Martin Cejp
Odborný garant ediční řady
Vybrané metody kriminologického výzkumu

Úvod

Následující text byl vytvořen na základě přednášek, které autor měl v roce 2011 v IKSP. Nabízí vhled jen do několika technik vícerozměrné statistiky, mnohé pomíjí (např. regresní analýzy, které zde nejsou zahrnuty, by si zasloužily zvláštní pojednání; podobně strukturální modelování (SEM-zmíněné jen letmo). Hlavním záměrem je probudit o vícerozměrné analýzy a jejich tvůrce zájem a ukázat jejich použitelnost na příkladech z praxe českého kriminologického výzkumu. Jde o techniky, které podle jednoho z recenzentů mají dvojí hlavní přínos:

a) výsledky založené na vícerozměrných technikách nemohou být klamné (zavádějící), protože se zohledňuje komplexně vliv mnoha faktorů
b) roste naděje, že výsledky české kriminologie lze uplatnit v kvalitních zahraničních časopisech (citováno z recenzního posudku P. Soukupa).

Opírám se dále nejčastěji o ukázky z práce se statistickým balíčkem SPSS, který v posledních desetiletích u nás i ve světě téměř zevšedněl a je čtenářům znám z vlastní praxe nebo aspoň z doslechu od kolegů na vlastním pracovišti.

1. Obecná charakteristika vícerozměrné statistiky

Nejpřístupnější a také nejčastěji používanou metodou analýzy dat je zkoumání jednorozměrného a dvourozměrného vztahu. Například se konstatuje, že v roce 2010 podíl recidivistů na policí stíhaných osobách činil 47,5 %. Nebo se hledá závislost počtu výrobců a držitelů psychotropních látek na jejich věku a pohlaví (a vyjadřuje se pomocí procent, korelací, kontingenčních koeficientů). Ovšem tyto vztahy se nezkoumají najednou, nýbrž postupně, tj. pokaždé jenom mezi dvěma proměnnými: zpravidla jednou závislou a jednou nezávislou. Trendové studie IKSP spadají do téže kategorie, protože sledují vývoj nějaké kriminální charakteristiky nebo složení pachatelů a obětí v čase zachyceném řadou na sebe navazujících následných let.

Zopakujeme si stručně základy statistického testování. Ve výběrových šetřeních se na vybraných případech snažíme zjistit, jaký je stav zkoumaného jevu v širším základním souboru („populaci“ nebo „universu“). Zpravidla přitom vycházíme ze dvou předpokladů či hypotéz o zkoumaném jevu: a) tzv. H0, nulové hypotézy, podle níž mezi proměnnými zkoumaného jevu nejsou žádné rozdíly (neliší se průměrem, procenty ap.) nebo že na sobě nezávisí (jedna proměnná na druhou nemá žádný vliv nebo zkoumaná dvojice proměnných nemá nic společného); b) tzv. H1, alternativní hypotézy, která naopak tvrdí, že mezi zkoumanými proměnnými existuje vztah nebo rozdíl. Statistické testování probíhá tak, že z dat výběrového souboru vypočteme statistiku a porovnáme ji s jejím rozdělením (testovacím kritériem). Konstatujeme pak pro danou hladinu významnosti (ve společenských vědách zpravidla alfa=.05 a .01), že např. pokud je námi vypočtená statistika nižší než testovací kritérion, je H0 vyvrácena a H1 potvrzena.

Vraťme se ale k dvourozměrnému přístupu k datům. Například můžeme snadno za určitých podmínek dopočíst očekávaná procenta a porovnat je pak (za pomoci statistických testů) s pozorovanými procenty. Nebo jestliže pracujeme s průměry, jak často musíme při zhodnocování výsledků rozsáhlejších testů, můžeme kromě rozdílů v průměrech a hodnotách

jemněji posoudit kolísání hodnot pomocí analýzy rozptylu. Použití sofistikovanějších statistických analýz může zpravidla přinést lepší vysvětlení faktů, nalézt nové souvislosti, vygenerovat nové podněty k zamyšlení.

Sledování dvourozměrných vztahů se zdá být bez problémů, jasné: zjišťujeme například, že procento osob, které byly stíhány za hospodářskou kriminalitu ve věku 18-20 let (2,7 %) je podstatně nižší než procento 40-60 ti letých (35 %) a vyvozujeme z toho přímo závěry o vlivu věku na tento typ zločinnosti. Nebo sledujeme klesající procento evidovaných majetkových trestných činů a konstatujeme rozdíl za posledních deset let v procentních bodech (-3,1 %). Vzhledem ke každoročně evidovanému počtu trestných činů zjišťujeme, že dochází v období 2000-2010 k poklesu evidovaných majetkových trestných činů významně teprve od roku 2005. Jakmile tedy chceme vědět z údajů něco víc, neobejdeme se bez statistické analýzy. Ta nám potvrdí nebo vyvrátí domněnku, že ve věkové skupině 30-39 let je srovnatelný podíl osob s hospodářskou kriminalitou jako ve skupině 40-59 ti letých. Objektivizuje naše úvahy nad daty, podpoří s vysokou pravděpodobností naše domněnky, které, jak sami cítíme, se jinak pohybují dost ve vakuu.

Testování dvourozměrných (párových) asociací, o kterých je zde řeč, ovšem není bez problémů. Někdy výsledky takových testů mohou zastírat skutečný stav věcí, „neodrážet“ skutečné vlivy proměnných. Testování nezávislosti párových asociací může snadno vést k vyvrácení nulové hypotézy, tedy k nesprávnému závěru, že např. zkoumané průměry nebo procenta si nejsou rovny nebo že mezi zkoumanými jevy neexistuje žádný vztah (tzv. chyby I. druhu). Naše testování spadne do intervalu, který hladina významnosti určuje jako možnost omylu (např. mezi 5 % selhání podle alfa=.05). Toto riziko se dá dokonce odhadnout podle vztahu $\alpha_{EW}=1-(1-\alpha)^c$, kde α je hladina významnosti a c je počet položek v testu, např. při 19 položkách baterie jako ve výzkumu, kde veřejnost hodnotila práci orgánů činných v trestním řízení (Zeman et al., 2011b), $\alpha=0,05$ je toto riziko 62 %, protože $\alpha_{EW} = 1-(1-0,05)^{19}=0,62$ (postup dle Kline, 2004:39). Zvláště často tedy k tomu může docházet při sériovém testování dvourozměrných asociací v rozsáhlých bateriích, např. při vyhodnocování významných vzájemných korelací mezi položkami otázky na kriminální citlivost veřejnosti. Nebo naopak, můžeme např. z testování alternativní hypotézy (H1) vyvodit závěr, že tu nějaká interakce existuje, ačkoliv ve skutečnosti žádná není, jde o artefakt (tzv. chyby II. druhu). Např. se to stane, pokud není náš výběrový soubor dostatečně velký a test tak ztrácí na síle.

Příklad 1: Povědomí o trestnosti ve vztahu k příjmu a vzdělání. Příklad 1 byl s malou úpravou převzat z výzkumu kriminality (Zeman et al., 2011b). Informovanost veřejnosti o trestnosti činů a druzích trestu za různé aktivity (zjištěná testy s minimem 0 a maximálním skórem 100) silně souvisí s příjmem (r=.356). Čím vyšší příjem, tím vyšší znalosti. Když ale přibereme souvislost se vzděláním, korelace skoro vymizí (r=.095).

Můžeme to provést rozkouskováním dvourozměrných asociací do několika kroků, podsouborů. Postupně rozporcujeme, parcializujeme korelace výše příjmu se znalostí podle 5 kategorií vzdělání. Základní vzdělání: výší příjmu se sice respondenti liší (průměr 12.000 Kč ±2.000), ale jejich znalosti jsou skoro shodné, zřídkakdy přesahují 48 bodů. Souvislost mezi příjmem 10-14.000 Kč lidí se základním vzděláním a jejich znalostí je mizivá (r = -,052, se stoupajícím příjmem se zde znalost nezvyšuje, téměř se nemění).

Obdobná situace je v dalších vzdělanostních kategoriích, kde sice průměrná znalost a průměrný příjem jsou vždy o něco vyšší, ale korelace k příjmu je také pokaždé nízká. Znalosti tedy nerostou se zvyšujícím se příjmem (ačkoliv to dvourozměrné korelace zpočátku naznačily), nýbrž se zvyšujícím se vzděláním. Tj. od nejnižšího vzdělání se 48 body až po nejvyšší se 63 body. Vliv příjmu na znalosti je tedy namnoze jenom vedlejším produktem vlivu vzdělání a je dán tím, že příjem se vzděláním stoupá. (Pozn. ve skutečnosti korelace nebyly tak vysoké: r=.264 pro vzdělání a r=.173 pro příjem, takže pokles parciální korelace mezi příjmem a znalostmi na 0,095 při kontrole vzdělání není tak dramatický jako v příkladu shora).

Tabulka 1: Průměrný skor znalostí o trestnosti a druzích trestů podle příjmu a vzdělání

	Průměrná znalost (ze 100)	Průměrný příjem (Kč)	Korelace k příjmu	N
1 ZŠ	48	12.000	-0,052	98
2 Vyučen	55	16.000	0,06	532
3 SŠ bez maturity	56	18.000	0,053	146
4 SŠ s maturitou	57	19.000	0,106	106
5 VŠ	63	23.000	0,015	148
Celkem/průměr	56	17.000	0,095	1.030

V případě, kdy místo dvojice proměnných statisticky zkoumáme a analyzujeme vztahy více proměnných zároveň a hledáme hlouběji vztahy mezi proměnnými, jde o **vícerozměrné metody.** Např. k věku pachatelů hospodářské kriminality přidáme jejich pohlaví, druhy a opakovanost udělených trestů. Vedle vzdělání a příjmu vložíme do vícerozměrného modelu věk a pohlaví a zkoumáme, které z nich mají nebo nemají vliv na informovanost o trestnosti. Pokud vliv mají, tak se snažíme zjistit, jak silný v porovnání s vlivem dalších proměnných, přičemž se zbavujeme falešně zkorelovaných položek (kdy vztah mezi A a B je zprostředkován ve skutečnosti proměnnou C) a snažíme se odhalit vlivy latentní.

Matematici v této souvislosti hovoří o mnohorozměrných statistických metodách (anglicky multidimensional), mnozí další odborníci o vícerozměrných, multivariátních, multivariačních nebo multivaričních metodách (z anglického multivariate techniques nebo multivariate statistical analysis). Přidržíme se recenzenty doporučené terminologie a budeme dále používat výrazy vícerozměrná statistika, techniky vícerozměrné statistiky apod. na rozdíl od jednorozměrných, popř. dvourozměrných rozdělení a analytických postupů práce s daty. Celkový přehled o těchto metodách poskytují také někteří čeští autoři (např. Hendl, 2004; Hebák et al., 2005; Meloun & Militký, 1994).

Často, i když rozhodně ne výlučně, je vícerozměrná statistika používána k analýzám dat získaných bateriemi otázek, tj. dotazy, v nichž se respondentovi nabízí série položek (tvrzení nebo dílčích podotázek) k posouzení, odpovědi nebo k odsouhlasení.

Vícerozměrné měření zde operuje s odpověďmi na několikanásobné otázky, na baterii otázek: analyzuje tyto odpovědi najednou, zároveň a nikoliv pouze v oddělených sekvencích po dvojicích jako při párových (dvourozměrných) analýzách.

Příklad 2: Hodnocení orgánů činných v trestním řízení. Ukázka dotazové baterie (Zeman et al., 2011b). 6 proměnných ve shora již citovaném výzkumu, bylo spolu s hodnocením dalších orgánů činných v trestním řízení podrobeno vícerozměrným statistickým postupům: shlukové a faktorové analýze.

24) Oznámkujte jako ve škole od 1 do 5 (1 = nejlepší hodnocení, 5 = nejhorší hodnocení), nakolik se podle vašeho názoru daří Policii ČR v boji s kriminalitou plnit následující úkoly:

	1	2	3	4	5	nevím
Odhalovat pachatele trestných činů	1	2	3	4	5	9
Dodržovat práva osob podezřelých ze spáchání trestného činu	1	2	3	4	5	9
Citlivě přistupovat k obětem trestné činnosti	1	2	3	4	5	9
Plnit její heslo „pomáhat a chránit“	1	2	3	4	5	9
Dodržovat práva poškozených	1	2	3	4	5	9
Nenechat se ovlivnit korupcí, politickými ani jinými nepřípustnými vlivy	1	2	3	4	5	9

Někdy při vícerozměrném měření nemá vůbec smysl hovořit o závislých a nezávislých proměnných, jindy jde o vztahy mezi množinami závislých a nezávislých proměnných.

Institucionalizace, rozšíření a komponenty

Ačkoliv některé vícerozměrné techniky mají více než stoletou tradici, k jisté institucionalizaci, mnohostrannému rozvoji a značnému rozšíření došlo až v 60. - 70. letech 20. století. Jsou obvykle kalkulačně náročné, a tak jejich rozmach se dostavil teprve s rozvojem výpočetní techniky. Výuka vícerozměrných technik statistické analýzy se stala zároveň akademickým předmětem a zejména na amerických universitách začaly vznikat související učené společnosti a časopisy jako např. Society of Multivariate Experimental Psychology založená 1960 a spravující od r. 1966 časopis Multivariate Behavioral Research, Journal of Multivariate Analysis Kalifornské university v Los Angeles – UCLA od r. 1971 nebo Multivariate Behavioral Research (Oklahomská universita 1966). O popularitě těchto metod svědčí také to, že dodnes vznikají také na evropském kontinentu specializované mezinárodní projekty, např. Society for Multivariate Analysis in the Behavioral Sciences (založeno ve Frankfurtu nad Mohanem 2004- vydává časopis Methodology).

Šíře aplikace je neuvěřitelná. Kromě psychologie a sociologie se tyto metody uplatňují mimo jiné v archeologii, potravinářství, analytické chemii a farmacii až po genetiku, zoologii, biologii, medicínu, ekologii, fyziku a samozřejmě také kriminologii.

Dříve než přikročím k podrobnějšímu popisu možností, které kriminologickému výzkumu nabízejí některé vícerozměrné statistické techniky, stručná globální charakteristika těchto metod. Zaprvé, jedná se o různé **výpočetní postupy**, techniky. Zaměříme se dále hlavně na faktorovou, korespondenční a shlukovací analýzu, mnohorozměrné škálování a okrajově také na rozhodovací stromy. Nejvíce se výklad soustředí na postupy z SPSS, protože tento statistický balíček se stal za desítky let užívání v ČR (i ve světě) jakýmsi standardem pro většinu společenskovědních statistických analýz. Klíčovými osobnostmi při rozvoji zmiňovaných metod byli hlavně R.B.Cattell, G.E.P.Box, M.Bénzecri, C.H.Coombs, R.A.Fisher, M.Greenacre, L.Guttman, H.Hotelling, M.G.Kendall, P.F.Lazarsfeld, C.E.Osgood, C.Spearman, W.S.Thorgerson a L.R.Tucker (Mardia et al., 1979; Everitt a Dunn, 1991; Hebák et al., 2005; aj.).

Výběr zmiňovaných technik se řídí zadáním IKPS. Ponecháváme mnohé druhy vícerozměrné analýzy stranou, zejména celou rodinu různých typů velmi zajímavých a v kriminologické praxi jistě využitelných regresních analýz (včetně lineární, hierarchické, logistické, Coxovy nebo kategoriální).

Za druhé, testování vypočtených výsledků je zpravidla umožněno tím, že hodnoty se porovnávají s náhodnými veličinami a vylučuje se tak možnost, že by konstelace našich čísel vznikla čistě samovolně bez vnitřní souvislosti. ***Rozdělení náhodných veličin*** tak bývá kritériem k odhadu parametrů základního souboru včetně chybovosti tohoto odhadu (chyba bývá volena zpravidla na hladině významnosti 0,05 nebo 0,01). Jen vysvětlující poznámka k parametrům základního souboru: ve výběrových šetřeních (a to se týká i statistik kriminality z nějakého po sobě jdoucími léty vymezeného/vybraného období) odhadujeme z poměrů ve výběrovém souboru s určitou pravděpodobností procenta, poměry, průměry a podobné parametry panující v širším základním souboru, v celé populaci. Další specifikou této statistiky testování hypotéz je, že nikdy vlastně přímo neměříme, zda vztah mezi proměnnými opravdu existuje nebo neexistuje. Pravdivost nebo nepravdivost tvrzení o vztahu statisticky nezkoumáme. Vyvracíme nebo potvrzujeme pouze s určitou pravděpodobností, že vztah neexistuje. Otcem tohoto trendu byl William S.Gosset (1908). Jeho průkopnická práce

inspirovala vznik dnešního testování významnosti výsledků vícerozměrných statistických technik, tj. tři typy rozdělení: vícerozměrné t-rozdělení, vícerozměrné normální rozdělení a Wishartovo rozdělení (zobecnění gamma a CHI^2 rozdělení na vícerozměrné).

V tomto pojednání se budeme zabývat statistickými analýzami, které zpravidla nepatří k testovacím procedurám. Jejich významnost většinou nelze přímo testovat. Faktorovou analýzu lze testovat buď jen sporadicky (například při extrakci metodou maximální věrohodnosti) nebo nepřímo, s využitím odlišných výpočetních statistických postupů (hlavně strukturálního modelování - za pomoci tzv. konfirmační faktorové analýzy). Na relevanci korespondenční a shlukové analýzy lze usuzovat až odvozeně z následných testů významnosti asociací nebo rozptylu sdíleného s dalšími proměnnými. Pouze další dvě techniky, jimiž se zde budeme zabývat, jsou zároveň výpočetním postupem i testovací procedurou: rozhodovací stromy (jmenovitě metoda růstu CHAID nebo QUEST) a mnohorozměrné škálování.

Uvedené skutečnosti vedou nutně k závěru, že vícerozměrné statistiky, jejichž výsledky nejsou testovány na statistickou významnost, je nutné ověřovat: hledat stabilní řešení (opakovat analýzy na různých souborech a podsouborech dat) a nasazovat pokud možno různé výpočetní postupy.

Po testovacích procedurách a rozdělení náhodných veličin jsou třetím důležitým komponentem měřící jednotky. Vícerozměrné statistické metody kalkulují s prostorem a tedy s ***určitými vzdálenostmi*** mezi případy nebo proměnnými (u většiny těchto metod lze výsledky zobrazovat také graficky s projekcí do dvourozměrného prostoru, aby se umožnilo jejich zobrazení). Měříme tedy, ***jak těsně souvisí proměnné (nebo skory respondentů)***, čili jak moc jsou si proměnné nebo případy navzájem vzdáleny. Variant vzdáleností využívaných v různých statistických testech je velké množství. Jejich charakter a použití jsou dány typem zkoumaných proměnných.

S malou odbočkou si jen zopakujeme ***typy proměnných/dat*** podle vztahu mezi jejich hodnotami: A) kategoriální data (také kvalitativní) – 1) nominální -nabývají číselných hodnot jen čistě podle nějaké konvence, čísly se nahrazují jména, např. odliší pohlaví podle 1= muž a 2= žena; nemůžeme s nimi provádět aritmetické operace, např. je sčítat, odčítat, násobit nebo dělit- pouze odlišují hodnoty podle toho, jestli jsou stejné nebo různé (výjimkou jsou

data typu ano/ne, dichotomická, z nichž lze např. počítat průměry) 2) ordinální (pořadové), např. 5-místná stupnice spokojenosti od 1=velmi nespokojen, přes 2=spíše nespokojen atd. až 5=po velmi spokojen. Zde již víme, že spíše spokojen je méně než velmi spokojen, jen nedokážeme říci o kolik. B) metrická data (také kvantitativní, numerická, kardinální), k nimž patří: 3) intervalová nebo také rozdílová proměnná. Nabývá číselných hodnot a známe velikost rozdílů (intervaly) mezi jednotlivými hodnotami – stupni, např. měsíční příjem domácnosti (kdy můžeme zjistit, o kolik větší nebo menší příjem má rodina A oproti rodině B); 4) poměrová (podílová) proměnná- její míra má nějaký počátek, podobně jako teploměr má nulu, a navíc oproti intervalové se tak u poměrových dat dá zjistit, kolikrát je jedna hodnota větší nebo menší než druhá hodnota. Např. dvoučlenná domácnost je 2x menší než čtyřčlenná. Ve společenskovědní praxi se mj. o poměrové měření výzkumníci pokoušejí např. tím, že nabídnou respondentům odpovědi na otázky v dotazníku tak, že si vybírají z hodnot na graficky zobrazené stupnici se stejnými intervaly od 0 do 9. Metrické proměnné mohou být také členěny na diskrétní (nabývají hodnot celých čísel, např. počet členů domácnosti) a kontinuální, spojité (někdy se jenom jim v užším slova smyslu říká metrické) nabývající v daném rozmezí libovolných hodnot jako např. věk respondentů.

Vrátíme se nyní k pojetí vzdáleností ve vícerozměrné statistice. Jsou odvozeny z několika základních typů. Pro metrická, zejména pro poměrová data lze konstruovat vzdálenost eukleidovskou, což je odmocněný čtverec rozdílu mezi „body“, tj. např. mezi dosaženými skory nebo četnostmi výskytu jevu či chování. Pro kategoriální i metrická data lze hovořit o vzdálenostech založených na různých standardních statistikách, např. na rozdílech v procentech, počtech shodných a neshodných párů asociace (CHI^2) nebo na korelacích (u metrických dat). Jako zvláštní případ statistických vzdáleností lze uvést vzdálenost Mahalanobisovu (je založena na kovariancích a rozdílech bodů).

Vícerozměrné statistické procedury lze také členit podle účelu na a) apriorní, vycházející z určité dané struktury dat (diskriminační analýza, konfirmační faktorová analýza) a b) aposteriorní, hledající v datech (data mining), které strukturu určují a odvozují z dat (shluková, explorační faktorová a korespondenční analýza).

Při používání apriorních metod máme nějakou teorii o tom, že na námi studovanou množinu proměnných působí nebo že je ovlivňuje určitý činitel a vycházíme z této teorie. Vytváříme podle ní model, týkající se například vztahu latentních (faktorů) a pozorovaných

proměnných (jednotlivých položek dotazované baterie). Posléze tento model můžeme testovat pomocí modelování strukturálními rovnicemi (SEM). Nebo analýzu založíme na tom, že data mohou mít nějakou společnou podmiňující charakteristiku. Tou může být vzdělání či profese respondentů. Nebo skory orgánů činných v trestním řízení, získané hodnocením od veřejnosti, rozdělíme diskriminační analýzou. Zvolíme přitom vzdělání a testujeme pak podle apriorního třídícího kritéria, zda se lidé se základním vzděláním liší od středoškolsky a vysokoškolsky vzdělaných pokud jde o (baterii) hodnocení práce policie. Zjišťujeme, zda skory různě vzdělaných lidí mají odlišnou konstelaci, charakteristický nezaměnitelný vzorec přes celou baterii. (V tomto, dále citovaném příkladu 1, se hodnotilo, jak se daří odhalovat pachatele, respektovat práva podezřelých, jak citlivý má policie přístup k obětem trestné činnosti, naplňuje heslo pomáhat a chránit, respektuje práva poškozených a odolává korupci nebo nežádoucímu vnějšímu ovlivňování vyšetřování.)

V druhém případě, kdy jde o aposteriorní metody, se teprve nějaká struktura, která není napohled vůbec zřejmá, v datech hledá. Počet proměnných se například účelně sníží na minimum: větší počet pozorovaných proměnných se přiřadí k nějakému společnému skrytému, přímo nepozorovanému základu. Tedy k tomu, co je pro ně společné (faktorová analýza). Nebo se objekty či respondenti rozdělí na skupiny podle toho, co jedny sbližuje a odděluje od druhých (shluková analýza). Případně se graficky názorně zobrazí, které objekty jsou si blízko a které jsou si vzdálené (korespondenční analýza, multidimenzionální škálování).

2. Faktorová analýza (Factor analysis, Principal component analysis)

Pozadí

Faktorová analýza (FA) a analýza hlavních komponent (PCA) jsou statistické metody, kterými se vysvětluje nebo popisuje rozptyl zjevných nebo též měřených (anglicky manifest či measured) proměnných. Děje se tak za pomoci menšího počtu latentních, konstruovaných (unobserved, latent, constructs) proměnných, tj. faktorů.

Pozorované proměnné zde vysvětlujeme jako lineární kombinaci faktorů plus chybu, tj. nevysvětlenou část rozptylu nebo nepřesnost v měření. Před 100 lety tuto metodu vynalezl armádní důstojník, později filosof (v Lipsku) a ještě později psycholog Charles Spearman (PCA: seminární práce na téma faktorová analýza inteligence z r. 1904). Charles Spearman (1863-1945) hledal obecnou vlastnost myšlení, inteligenci pomocí testů

- *matematických schopností,*
- *schopnosti se slovně vyjádřit*
- *logicky uvažovat*
- *uměleckých skonů.*

Podle něho existuje jeden společný, obecný jmenovatel či faktor „G“ (general) inteligence. Je to obecná intelektová schopnost, která způsobuje, že jsou výsledky všech zmiňovaných testů zkorelovány. (Kromě toho má samozřejmě každá oblast svou specifiku – faktor „s“.)

Další významný podnět k rozvoji FA dala psychometrika, především R. B. Cattell (1905-1998). Vystudoval chemii v Cambridgi r. 1926. Měl politické a sociální zájmy a ty ho vedly k tomu, že obrátil pozornost k psychologii. V r. 1929 dokončil studia psychologie. Položky (skory) jeho testů inteligence byly rozlišeny na tři faktory: verbální, matematické a logické schopnosti.

V krizových 30. letech se snažil pochopit a řešit ekonomické, morální problémy pomocí objektivního psychologického poznání morální stránky člověka.

Hledal adjektiva popisující osobnost. Použil: a) L-data (týkající se života, life record-chování ve společnosti, např. soudní záznamy); b) Q-data (sebehodnocení subjektů z dotazníku), c) T-data (testové situace, kdy si subjekt neuvědomuje, že je mu měřena nějaká vlastnost).

Jeho výběrový soubor přesahoval 1.000 osob, které byly různého věku a pocházely z různých zemí (U. S., Británie, Austrálie, Nový Zéland, Francie, Italie, Německo, Mexiko, Brazílie, Argentina, Indie a Japonsko). Pomocí faktorové analýzy našel 16 osobnostních rysů (Cattell, 1943). Ty později jiní psychologové (Fiske 1949, Tupes a Christal 1961, Tucker, 2009, John 1999) zjednodušili na pět: živost, přívětivost, svědomitost, emocionální stabilita a otevřenost vůči zkušenostem.

Rozvoj faktorové analýzy nastal hlavně v 60. a 70. létech 20. století podobně jako všech dalších vícerozměrných technik. FA a PCA pronikly do dalších oborů: sociologie, medicíny, výzkumu trhu a samozřejmě také do kriminologie. Také někteří naši autoři se už v 70. letech faktorovou analýzou aktivně zabývali (např. Überla, 1974).

Rozvoj metody měl i stinnou stránku: vznikla nepřehledná situace, množství odborné terminologie se stalo nezvladatelným. Ke konsolidaci postupů došlo až v současnosti pod vlivem prací a odkazu takových historických osobností jako byl Thurstone (1934) a statistiků jako Lawley, Hotelling, Bartlett aj. (viz např. Garson, 2011b).

Lze **hlouběji porozumět příčinám názorů, jednání**? Faktorová analýza na základě korelací mezi větším množstvím (zjevných, manifestních) proměnných statistickým způsobem určuje, zda jsou si některé blízké, patří k sobě, tj. zda za nimi stojí jeden společný faktor (latentní proměnná) nebo zda patří k jinému společnému faktoru.

Počet faktorů by měl být co nejmenší a nalezené závislosti by měly být vysvětleny co nejjednodušeji. Pokud faktory zjistíme a pojmenujeme, můžeme pak s nimi pracovat dál-např. vytvoříme na základě jednotlivých faktorů souhrnné indexy ze skorů jednotlivých položek (hodnocení soudního procesu, hodnocení policie apod.). Ty potom dál třídíme, analyzujeme a testujeme. Nebo pracujeme s tzv. faktorovými skory jednotlivců, které si lze při analýze zadat: vytvoří se jako nové proměnné a jsou tak přímo součástí procedury a dat.

Východiska a požadavky na data

První přiblížení k faktorové analýze nabízí korelační matice. Běžná korelace je statisticky vyjádřený vzestupný nebo sestupný (lineární) vztah mezi dvěma proměnnými (např. věkem recidivisty a počtem trestných činů: čím vyšší věk pachatele, tím vyšší počet stíhání pro trestné činy). Nabývá hodnot od -1 do +1. Výchozí pro faktorovou analýzu je **korelační nebo kovariační matice.**

Vrátíme se k **Příkladu 2: hodnocení orgánů činných v trestním řízení** (Zeman et al., 2011b). Respondenti byli požádáni, aby zhodnotili práci policie, státních zástupců, soudů a pracovníků vězeňské správy. K posouzení na školské stupnici jim bylo předloženo 19 položek týkajících se práce těchto orgánů (=odtud se vytvořily manifestní, měřené proměnné). Ty jsme pak pomocí faktorové analýzy vyjádřili jako silnější nebo slabší důsledek působení čtyř (nedotazovaných, latentních) faktorů.

Tabulka 2 níže zahrnuje korelace skorů (známek na školské stupnici) : proměnné r26a r26b a r26c **mají vysoké vzájemné korelace,** které mohou (pokud se splní další podmínky) být důsledkem působení stejného faktoru (hodnocení soudů). Proměnné r27a r27b r27c s nimi příliš nesouvisejí a mohou spadat pod jiný faktor, např. hodnocení vězeňství.

Tabulka 2: Korelační matice hodnocení soudů a vězeňství veřejností ČR

	t26a	t26b	t26c	t26d	t26e	t26f	t27a	t27b	t27c	t27d
t26a	1									
t26b	**,487**	1								
t26c	**,351**	**,559**	1							
t26d	**,442**	**,555**	**,587**	1						
t26e	**,327**	**,443**	**,448**	**,451**	1					
t26f	**,484**	**,527**	**,499**	**,509**	**,426**	1				
t27a	,136	,214	,257	,195	,228	,196	1			
t27b	-,006	,211	,251	,175	,207	,136	**,467**	1		
t27c	,317	,338	,319	,304	,279	,324	**,323**	**,276**	1	
t27d	,281	,268	,233	,238	,241	,258	**,296**	**,279**	**,684**	1

Legenda k tab. 2: Nakolik se podle Vašeho názoru daří (soudům/vězeňství) plnit úkol? t26a rozhodovat bez zbytečných průtahů; t26b ke všem obviněným přistupovat stejně; t26c trestat skutečné pachatele a osoby neprávem obviněné osvobozovat; t26d ukládat spravedlivé tresty;t26e citlivě přistupovat k obětem trestné činnosti; t26f nenechat se ovlivnit korupcí, politickými ani jinými nepřípustnými vlivy; t27a znemožnit útěky odsouzených; t27b dodržovat práva odsouzených; t27c přispívat k nápravě odsouzených; t27d připravovat odsouzené na návrat z vězení do společnosti

Dále v tabulce 3 z téhož výzkumu příkladu č. 2 jsou vysoce zkorelovány poskytování a využívání sexuálních služeb a držba s užíváním nelegálních drog. Také tyto dvojice proměnných se octnou v totožném faktoru.

V tomto **příkladu 3 Možný postih různých skutků jako trestný čin** (Zeman et al., 2011b-tentýž pramen jako příklad 2) otázka zněla odpovídajícím způsobem: zda by vyjmenované činy měly být postihovány jako trestný čin. Respondenti mohli k 13 vyjmenovaným skutkům odpovídat ano, ne nebo nevím, tedy stupnice odpovědí byla kategoriální a nevyhovuje předpokladům faktorové analýzy. Faktory by měly zachytit oblasti,

v nichž se veřejnost názorově rozchází. Ale při kategoriálních datech, dichotomiích, kdy odpovědi na otázky nemusí být dost protichůdné, by se naopak faktory mohly rozdělit na základě shodných odpovědí: podle marginálních četností tak, že některé by měly vysoké souhrnné průměry (převaha kladné odpovědi) a jiné naopak nízké (převaha záporných odpovědí).

Tabulka 3: Část matice tetrachorických koeficientů* z výzkumu mínění veřejnosti o kvalifikaci skutků jako trestný čin

Proměnná	r18_1	r18_2	r18_3	r18_4	r18_5	r18_6	r18_7
r18_1	1.000						
r18_2	0.830	1.000					
r18_3	0.389	0.415	1.000				
r18_4	0.252	0.306	0.388	1.000			
r18_5	0.322	0.378	0.379	0.399	1.000		
r18_6	0.233	0.331	0.396	0.455	0.497	1.000	
r18_7	0.411	0.393	0.395	0.364	0.432	0.481	1.000
r18_8	0.391	0.342	0.335	0.251	0.396	0.355	0.911
r18_9	0.218	0.156	0.163	0.357	0.181	0.317	0.206
r18_10	0.159	0.182	0.180	0.364	0.051	0.186	0.257
r18_11	0.215	0.212	0.083	0.366	0.150	0.346	0.290
r18_12	0.224	0.126	0.163	0.280	0.246	0.341	0.255
r18_13	0.216	0.196	0.201	0.205	0.132	0.408	0.299

*tetrachorické koeficienty jsou korelace vzniklé metrizací výchozích kategoriálních hodnot (dichotomie typu ano/ne)

Legenda k Tab. 3: r18_1 užívání nelegálních drog; r18_2 držení nelegálních drog pro vlastní potřebu; r18_3 dobrovolný pohlavní styk s osobou ve věku 14 let; r18_4 šíření urážlivých nepravdivých skutečností o jiné osobě; r18_5 usmrcení trpící nevyléčitelně nemocné osoby na její žádost; r18_6 nelegální kopírování hudby, filmů, počítačových programů apod.; r18_7 poskytování sexuálních služeb za úplatu; r18_8 využívání sexuálních služeb za úplatu; r18_9 závažné poškozování životního prostředí; r18_10 poskytování půjček na neúměrně vysoký úrok; r18_11 vyhýbání se splácení dluhů; r18_12 veřejné projevování sympatií k rasismu; r18_13 šizení na daních

Dáme-li důraz na kladnou odpověď, získáme data tzv. nepravé dichotomie, tj. ano/nikoliv ano (=ne+nevím). Další variantou je zahrnout do faktorové analýzy pouze odpovědi typu ano-ne (pravou dichotomii) a vynechat nepoužitelné, neskorovatelné „nevím“. V obou případech odborná literatura doporučuje data nejdříve metrizovat, tj. převést skory 1-0 nebo 1-2 na odstupňované poměrové kontinuální (spojité) skory a vypočíst vzájemné

vztahy pomocí tzv. tetrachorických koeficientů. SPSS sice přímo tuto transformaci neumožňuje, ale dá se nahradit makrem nebo výpočtem v jiném programu. Práce s pravou nebo nepravou dichotomií má své výhody i nevýhody. V našem příkladu č. 3, kdy odpovědí „nevím“ bylo pokaždé hodně (přes 100 při N=1692) jsme dali přednost nepravé dichotomii.

Další přiblížení nabízí matice parciálních korelací. **Parciální korelace** je lineární vztah, korelace mezi dvěma proměnnými, pokud odhlédneme od jejich vztahu ke třetí proměnné- ten jakoby se odečte nebo vyruší („třetí proměnné jsou konstantní“). Čím je parciální korelace vyšší, tím méně je vztah dvojice proměnných ovlivněn, zprostředkován něčím třetím (zde: faktorem) a naopak. Princip parciální korelace je popsán shora v příkladu 2.

V matici níže (Tabulka 4) jsou uvedeny parciální korelace mezi faktory a proměnnými hodnocení orgánů činných v trestním řízení z téhož příkladu. Jsou na diagonále blízké 1, kdežto mezi sebou, po „odečtení“ vlivu společných faktorů jsou blízké nule (≤+-0,3). Pohled na příklad níže nám dokládá, že na každou proměnnou z našeho kriminologického výzkumu působí společné faktory, protože korelační koeficienty mimo diagonálu jsou opravdu nízké.

K výchozím podmínkám faktorové analýzy patří: a) vysoké korelace většího počtu proměnných a b) nízké parciální korelace. Matice není vždy snadné zkoumat, zvláště při velkém počtu proměnných. Navíc osobní posouzení matice nemusí být vždy korektní. Zda korelační a parciální korelační matice splňují předpoklady je proto testováno různými koeficienty a indexy, zpravidla KMO a Bartlettovým testem sféricity. Např. vysoké hodnoty KMO (blízké 1, minimálně 0,6) a významnost Bartlettova testu ($p<.001$) ukazují, že zmiňované výchozí předpoklady faktorové analýzy jsou v pořádku.

Kaiser-Meyer-Olkinova míra (KMO), čili míra adekvátnosti výběru. Je to průměrová charakteristika vyjadřující poměr naměřené a ideální, tj. maximálně možné hodnoty. Aby bylo možné provést FA, musí být vztahy mezi dvěma proměnnými skutečné, těsné a nejenom zprostředkované vlivem nějaké třetí sousední proměnné v baterii otázek výzkumu. KMO nabývá hodnot mezi 0 a 1. Čím vyšší, tím lépe (jde o součet mocnin korelačních koeficientů dělený součtem mocnin korelačních a parciálních koeficientů, čili o poměry koeficientů determinace – ukazatelů společné variance). Doporučuje se minimálně hodnota KMO=0,6. **Bartlettův test sféricity-**testuje se, jestli nejsou proměnné pouze autokorelované, jestli tedy mají kromě korelací samých se sebou nějakou další sféru vztahů

k dalším proměnným. (Vyvrací se nulová hypotéza, že proměnné mají korelace jen k sobě samým, že matice se skládá z jedniček na diagonále a nul mimo ni.) Bartlettův test ověřuje, že nejde o tzv. jednotkovou (identitní) korelační matici. Tento test už není tak dobrým vodítkem, protože velké výběrové soubory jím projdou, i když jsou interkorelace různých proměnných velmi nízké. Při testu se porovnávají skutečně naměřené hodnoty s náhodně vygenerovanými a rozdíl se vyjadřuje CHI^2 a jeho významností.

Data by měla být v ideálním případě spojitá (metrická), povlovně se zvětšující nebo zmenšující, s možností lineárně je kombinovat, např. věk nebo teplota (od 0 po 100 a víc), hmotnost tělesa (od 0 po n kilogramů), nebo vzdálenost v kilometrech. V odborné literatuře se zpravidla v současnosti připouští změkčení těchto požadavků. V praxi ve společenských vědách totiž taková data bývají zřídkakdy k dispozici. Musíme obvykle vystačit i s kategoriálními nebo metrickými **diskrétními daty,** tj. s hrubým **odstupňováním nebo bez přesně změřené nebo měřitelné vzdálenosti** mezi etapami a bez přirozeného počátku cesty.

Tabulka 4: Parciální korelace mezi hodnocením orgánů činných v trestním řízení a faktory

	t26a	t26b	t26c	t27a	t27b	t27c
t26a	**0,929**					
t26b	-0,204	**0,948**				
t26c	0,041	-0,214	**0,938**			
t27a	-0,096	-0,076	-0,121	**0,932**		
t27b	-0,105	-0,149	-0,080	-0,156	**0,916**	
t27c	0,019	-0,036	0,029	-0,024	-0,035	**0,866**

Legenda k Tab. 4: Nakolik se podle Vašeho názoru daří 1) soudům - t26a rozhodovat bez zbytečných průtahů; t26b ke všem obviněným přistupovat stejně; t26c trestat skutečné pachatele a osoby neprávem obviněné osvobozovat; 2) vězeňství - t27a znemožnit útěky odsouzených; t27b dodržovat práva odsouzených; t27c přispívat k nápravě odsouzených

Literatura (např. Garson, 2012a: 47-48) uvádí, že tak se s ordinálními daty zachází jako s intervalovými, což je „forma chyby měření přinášející oslabení korelace... (a následně) se faktorové zátěže hůře interpretují“.

Je to jako bychom říkali o vzdálenosti: místo A je blízko, místo B je dál a C ještě dál od nás. Podobně je to se známkami školské stupnice: při zhoršení prospěchu z 1 na 5 předstíráme, že je to totéž jako pokles o 4 stupně z 5 na 1 stupeň na spojité stupnici Celsia.

Nebo když jde o pokles spokojenosti o 4 stupně od naprosto spokojen – spíše spokojen – ani tak, ani tak – spíše nespokojen – až k naprosto nespokojen: předstíráme, že je to srovnatelné např. s poklesem o 4 decibely na 0 na hlukoměru měřícím spojitě hluk potlesku.

Existuje dokonce zdarma přístupný program, který tvrdí, že umí počítat faktorovou analýzu z takových pořadových, méně přesných dat (VISTA- viz např. Ledesma & Mora, 2007; Ledesma & Molina, 2009). Navíc řada odborníků, zejména psychologických statistiků, ordinální data pro faktorovou analýzu nezakazuje, pokud jde o alespoň pětistupňové škály.

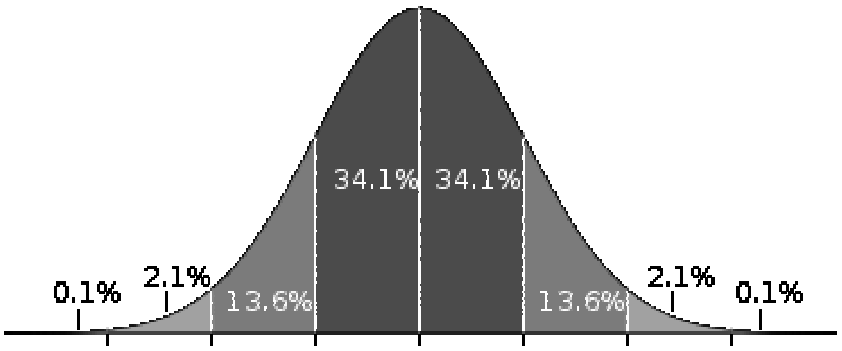
Zcela nevhodná k analýze jsou data kvalitativního, nominálního charakteru (např. čísla na dresech hráčů, označení 1=muž/2=žena, kraj 1 až 14, sledování média X). Ovšem i kategoriální data lze podrobit analýze latentních tříd, jisté obdobě faktorové analýzy vhodné pro tento účel (viz blíže např. Hagenaars & McCutcheon, 2002).

Dále je **nevhodný příliš malý počet případů,** tj. malé výběrové soubory respondentů. Faktorová analýza vychází z korelací a ty se stabilizují teprve na poměrně velkých výběrových souborech. Odborná literatura uvádí: 50 respondentů je velmi špatné, 100 je špatné, 200 ujde, 300 je dobré, 500 je velmi dobré a 1000 a víc je skvělé. Někteří autoři uvádějí, že respondentů (případů) musí být minimálně 5x víc než proměnných pro FA (Costello, Osborne, 2005, 4 a 7). Není tím samozřejmě řečeno, že čím je výběrový soubor větší, tím je faktorová analýza přesnější a kvalitnější. Záleží také na kvalitě sebraných dat a jejich tzv. „síle“. „Silná“ data, schopná dobře vysvětlovat rozptyl proměnných, vážící každou proměnnou jenom k jednomu faktoru, očištěná od odlehlých zkreslujících pozorování, nevyžadují příliš velké výběry (Mulaik, 1990; Widaman, 1993).

Standardizace, extrakce a analýza hlavních komponentů

Faktorová analýza výchozí hodnoty respondentů nejdříve **standardizuje**. (Od naměřených hodnot respondenta se odečte celkový průměr a rozdíl se pak vydělí směrodatnou odchylkou. Touto úpravou nabudou proměnné statisticky výhodnějších hodnot, kdy průměr=0 a rozptyl=1)

Např. soudní orgány i státní zástupci byli ve výzkumu příkladu 3 známkováni jako ve škole podle toho, jestli jsou podle respondentů spravedliví, neúplatní, důslední, nestranní k obviněným a citliví k obětem apod. Hodnotící výroky a korelace obsahuje tabulka 11 a 13 na předchozích stranách. Při posuzování jednotlivých známek respondentů přihlédneme k tomu, jak moc se liší od celkového známkového průměru. Zjistíme u každého rozdíl od tohoto průměru a tento rozdíl (vydělený směrodatnou odchylkou) pak převedeme na jednu společnou stupnici. Jde o počet směrodatných odchylek - je jich u normálního rozdělení celkem 6 a v jejich dosahu leží přibližně 99,7 % všech pozorování ve výběrovém souboru, takže už můžeme na této „normální“ stupnici od 0 do 6 či výše porovnávat hodnoty.



Obr. 1: Normální rozdělení

Soustavou lineárních rovnic se pak pro každou (standardní) hodnotu vypočte vztah k různým (pro více proměnných společným) faktorům. Např. proměnná „neúplatnost soudce“ z příkladu 2 měří faktor A a mnohem méně také další faktory B, C a D.

Naše proměnná totiž neměří určitý faktor sama, ale spolu s dalšími proměnnými (např. vedle neúplatnosti je faktor A měřen také spravedlností, důsledností a nestranností soudů). Jako jediná ze všech proměnných použité dotazové baterie měří pouze ona určitou

vlastnost (neúplatnost soudce). Ovšem jedinečnost (parciální vztahy) v tomto případě nebereme v úvahu. Každý faktor se pak vypočte postupně, jeden po druhém, po vyčerpání možností a tedy nezávisle na jiném faktoru.

Extrakce faktorů je způsob, jak z množiny proměnných vytvořit, vybrat faktory (Garson, 2012b: 5 ad.). Existuje řada metod, z nichž zmíníme 3 nejčastěji používané: PC (hlavní součásti), PAF (faktorování podle hlavních os) a ML (maximální věrohodnost).

Metoda hlavních součástí či hlavních komponent (principal components, PC) produkuje na sobě nezávislé (nekorelované) faktory uspořádané tak, že první na sebe strhává nejvíc rozptylu (má největší varianci) a poslední faktor má nejmenší varianci. Do analýzy se zahrne veškerý rozptyl, tj. včetně jedinečnosti proměnné. Původní komunality (tj. podíl zastoupení proměnné ve faktoru) jsou tedy rovny 1. Hlavní komponenty/ faktory jsou určeny jednoznačně, takže když zvolíme větší počet faktorů (nebo u PC komponentů), ty původně zjištěné se nezmění.

První faktor (hl. komponenta) se vypočte lineárními rovnicemi jako ta kombinace původních proměnných (manifestních), která na sebe strhuje největší rozptyl. ($F1_{ij}$=V1*V2 + V1*V3 + V2*V3 +V_i*V_j = max var; přičemž součet mocnin korelací mezi všemi standardizovanými proměnnými V se musí celkem rovnat 1, čili 100 %; ovšem faktor vyčerpá 100 %, když je faktorová analýza celkově jednodimenzionální). Další faktor se vypočte obdobně, ale nesmí být k tomu prvnímu korelován atd.

Vysoce zkorelované proměnné mohou mít jen jeden faktor. Ve výzkumu kriminality (Zeman et al., 2011b) to byl případ deklarovaného dostatku informací o kriminalitě, trestních zákonech a řízení a průběhu výkonu trestů. Vztahy byly tak těsné, že FA nalezla pouze jednodimenzionální řešení. Toto zjištění nám umožnilo sestavit spolehlivý součtový index deklarovaného dostatku informací. Když mezi proměnnými není těsný vztah (korelace se blíží 0), může být počet faktorů rovný počtu proměnných. Tím by se samozřejmě FA stala zbytečnou.

V příkladu 3 (Možný postih různých skutků jako trestný čin**)** si nejdříve ujasníme jeden z výstupů analýzy hlavních komponentů (i faktorové analýzy): nerotovanou faktorovou matici. Po extrakci zjistíme jednak počet dimenzí (faktorů nebo u PC komponentů: zde 2)

a také korelace či „zátěže“ (loadings) jednotlivých proměnných. Zátěže nabývají hodnot od -1 do +1. Např. nestrannost soudů v jednom z našich předchozích příkladů má k faktoru 1 (spravedlivého soudního procesu) korelaci 0,790. Abychom spolehlivě přiřadili proměnnou k faktoru (nebo ke komponentu) vyžaduje se zpravidla, aby daná proměnná měla vyšší zátěž než 0,3.

Představíme-li si každou proměnnou jako bod v prostoru, zde dvourozměrném - tabulka 5, pak například držba nelegálních drog je vzdálena od osy C1 („komponent 1“) -0.631 a od osy C2 („komponent 2“) -0.423. Bod má extrémní polohu v záporném pásmu.

Naproti tomu vyhýbání se splácení dluhů leží vysoko nad průsečíkem os v rozměru C2 (0.570) a mírně nad v rozměru C1 (-0.571). Tato konstelace naznačuje, že zde existuje významný protiklad.

Poslední sloupeček našeho příkladu zobrazuje tzv. komunality. Společnému rozptylu (součtu mocnin všech těchto koeficientů bez jedinečné variance) říkáme **komunalita.** Dosahuje maximálně 1. Vyjadřuje, na kolik procent je daná proměnná vysvětlena faktory (např. 64 % u r18_7 poskytování sexuálních služeb za úplatu nebo 65 % u r18_11 vyhýbání se splácení dluhů. Toto číslo vzniká umocněním zátěží položky např. r18_7 faktorové matice= $(-0,571)^2$+ $(0,571)^2$=0,652).

Komunalita se původně pro každou proměnnou rovná 1 (včetně její jedinečnosti), až po extrakci se mění. Výše komunality také informuje o tom, zda je rozptyl příslušné položky dostatečně faktorem vysvětlen. Např. veřejné projevování sympatií k rasismu (r18_12) má komunalitu velmi nízkou ($-.631^2$=0.363).

Tabulka 5: Nerotovaná faktorová struktura (mínění veřejnosti o trestním postihu za skutky)

Proměnná	C1	C2	Komu-nalita
r18_1	-0.628	-0.388	0.545
r18_2	-0.631	-0.423	0.577
r18_3	-0.577	-0.322	0.436
r18_4	-0.632	0.100	0.410
r18_5	-0.593	-0.331	0.461
r18_6	-0.691	0.015	0.477
r18_7	-0.755	-0.271	0.643
r18_8	-0.680	-0.300	0.552
r18_9	-0.509	0.422	0.437
r18_10	-0.500	0.522	0.522
r18_11	-0.571	0.570	0.652
r18_12	-0.513	0.315	0.363
r18_13	-0.567	0.466	0.538

Legenda k Tab. 5: r18_1 užívání nelegálních drog; r18_2 držení nelegálních drog pro vlastní potřebu; r18_3 dobrovolný pohlavní styk s osobou ve věku 14 let; r18_4 šíření urážlivých nepravdivých skutečností o jiné osobě; r18_5 usmrcení trpící nevyléčitelně nemocné osoby na její žádost; r18_6 nelegální kopírování hudby, filmů, počítačových programů apod.; r18_7 poskytování sexuálních služeb za úplatu; r18_8 využívání sexuálních služeb za úplatu; r18_9 závažné poškozování životního prostředí; r18_10 poskytování půjček na neúměrně vysoký úrok; r18_11 vyhýbání se splácení dluhů; r18_12 veřejné projevování sympatií k rasismu; r18_13 šizení na daních

Ke každému výpočtu faktorových zátěží existuje množství alternativních řešení. Nejlepší řešení jsou ta, při nichž se faktorová zátěž u jednoho faktoru (korelace s faktorem či komponentem A) blíží 1 a s jinými faktory se blíží 0, tj. proměnná „patří“ do faktoru A nepatří do faktorů B až M. Docílíme toho tzv. rotací.

V našem příkladu pokračujeme dalším charakteristickým výstupem, rotovanou faktorovou maticí (maticí dřívějších hlavních komponent C1 a C2, tabulka 6). Osy pootočíme kolem jejich průsečíku tak, že např. r8_11 (vyhýbání se splácení dluhů) se jednoznačně přiblíží k faktoru 1 (0.801) a vzdálí od faktoru 2 (0.102) a zároveň držba nelegálních drog pro vlastní potřebu, r18_2, se vzdálí od faktoru 1 (0.052) a přiblíží k faktoru 2 (0.758). V důsledku rotace se tedy změní faktorové zátěže, avšak vzájemné uspořádání jednotlivých bodů (proměnných) a jejich vzdálenosti se nemění, tj. výchozí korelace zůstávají stejné.

Po rotaci (Tabulka 6) jsme získali faktor činů poškozujících integritu společnosti (poškození finanční, životního prostředí a sociálních skupin nebo menšin) a činů poškozujících osobní integritu.

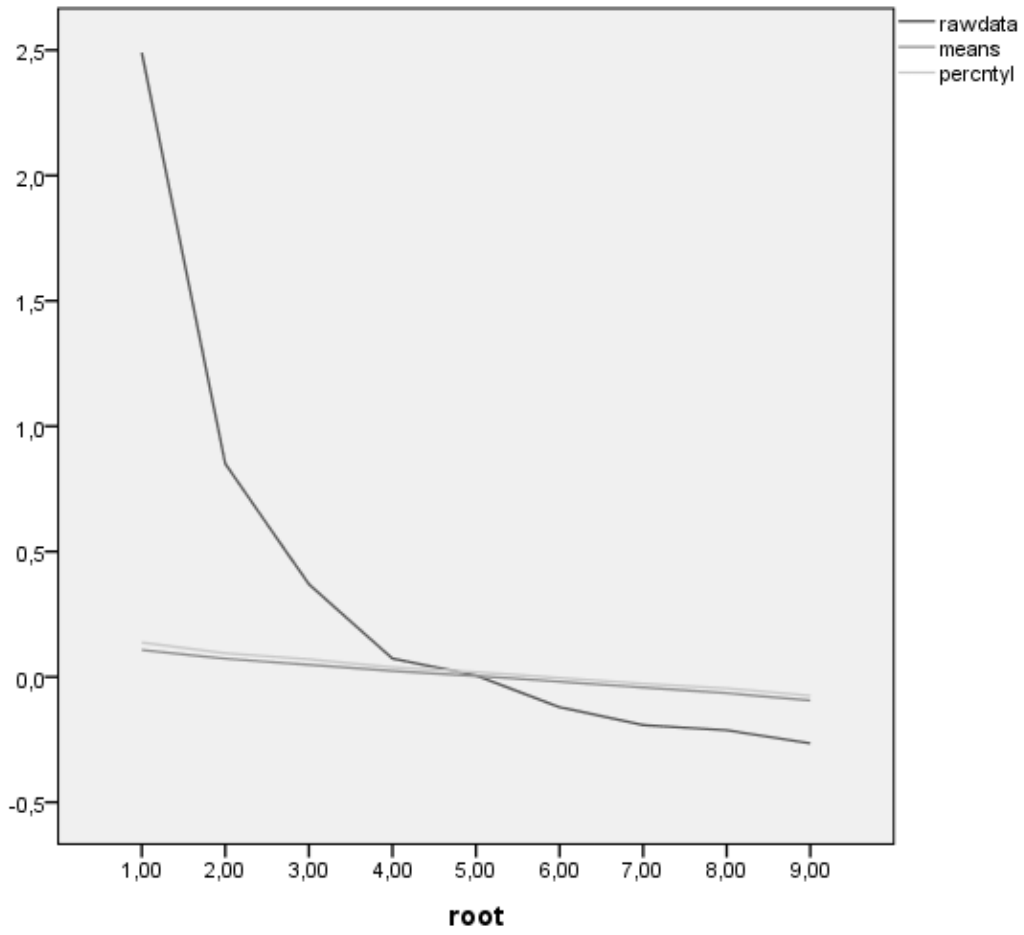
Počet faktorů byl určen paralelní analýzou např. podle tzv. sutinového grafu (Obr. 2) nebo podle tabulkového výstupu (Tabulka 7).

Tabulka 6: Rotovaná faktorová struktura (mínění veřejnosti o trestním postihu za skutky)

Proměnná	C1	C2
r18_1	**0.801**	
r18_2	**0.718**	
r18_3	**0.715**	
r18_4	**0.645**	
r18_5	**0.563**	
r18_6	**0.466**	0.438
r18_7		**0.763**
r18_8		**0.758**
r18_9		**0.734**
r18_10		**0.721**
r18_11		**0.671**
r18_12		**0.653**
r18_13	0.435	**0.537**

Legenda k Tab. 6: r18_1 užívání nelegálních drog; r18_2 držení nelegálních drog pro vlastní potřebu; r18_3 dobrovolný pohlavní styk s osobou ve věku 14 let; r18_4 šíření urážlivých nepravdivých skutečností o jiné osobě; r18_5 usmrcení trpící nevyléčitelně nemocné osoby na její žádost; r18_6 nelegální kopírování hudby, filmů, počítačových programů apod.; r18_7 poskytování sexuálních služeb za úplatu; r18_8 využívání sexuálních služeb za úplatu; r18_9 závažné poškozování životního prostředí; r18_10 poskytování půjček na neúměrně vysoký úrok; r18_11 vyhýbání se splácení dluhů; r18_12 veřejné projevování sympatií k rasismu; r18_13 šizení na daních

Korelační matice: tetrachorické koeficienty; metoda extrakce: Principal components; metoda rotace: VARIMAX; počet faktorů: podle Hornovy paralelní analýzy (Horn, 1965). Faktory vysvětlují 67,7 % celkového rozptylu, z toho C1 36,9 %. Spolehlivost podle Mislevy & Bock (1990) je vysoká. 786 a .829



Obr. 2: Sutinový graf (po Hornově paralelní analýze)

Obr. 2 ukazuje, že na ose X (=root) lze sledovat modely počtů faktorů (z 13 možných= 13 vstupních proměnných) a na ose Y odpovídající vstupní data. Průměry a percentily vymezují úsek rozhodující pro určení počtu faktorů. Nejlepší řešení jsou 2 faktory (velký je rozdíl mezi 1. a 2.fakorem, mezi 2. a 3. už rozdíl není velký – jenom 0,4). Je to patrné i z tabulky 16: rozdíl mezi 1. a 2. faktorem je na hrubých datech 3,36-1,6=1,76, kdežto mezi 2. a 3. faktorem jenom 1,6-1,2=0,4 atd.

Tabulka 7: Výstup z paralelní analýzy

root	rawdata	means	percentyl
1	3,36	1,18	1,22
2	1,6	1,14	1,17
3	1,2	1,1	1,12
4	1,08	1,08	1,09
5	0,87	1,05	1,07
6	0,82	1,02	1,04
7	0,8	1	1,02
8	0,79	0,97	0,99
9	0,64	0,95	0,96
10	0,62	0,92	0,95
11	0,58	0,89	0,92
12	0,4	0,86	0,89
13	0,25	0,83	0,85

Rozhodující je tedy způsob, jakým určíme počet faktorů. Je možné si přímo určitý počet zadat, což je vhodné v případě, že máme nějakou teoretickou představu, kolik faktorů se za pozorovanými daty skrývá. Nejběžnější je ovšem zvolit nějaký automatizmus, který rozhodne za nás. Nejčastěji se používá tzv. Kaiserovo pravidlo (Kaiser, 1960). Podle tohoto pravidla se při Eigenvalue ≥1 vyberou ty faktory, které mají diferenciační sílu jednotlivých proměnných v hodnotě nejméně jedné proměnné a faktory s nižší hodnotou se vypustí. (To alespoň 1 zde tedy znamená, že v daném faktoru je nejméně jedna silně diferencující proměnná). Někteří autoři tuto praxi kritizují. Např. Velicer (Velicer & Jackson, 1990, 99) uvádějí: „Většina statistických balíčků implicite podrží všechny faktory s eigenvalue větším než 1.0. V literatuře panuje široký konsensus, že toto je jedna z nejméně přesných metod výběru faktorů“.

Ke Kaiserovu pravidlu se ještě nezřídka pro kontrolu využívá tzv. sutinový graf (screeplot). K dalším metodám patří Velicerův koeficient (Velicer&Jackson, 1990, Velicer&Fava, 1998). V Hornově metodě, považované momentálně za nejlepší, se porovnávají pozorované hodnoty s náhodně vygenerovanými očekávanými. V SPSS je k dispozici volba podle Kaiserova pravidla, screeplot a pokud jde o ostatní metody, lze je řešit makrem.

Rozptyl faktorů se určuje jako vlastní číslo (Eigenvalue). Součet všech rozptylů všech faktorů je roven počtu proměnných, např. v našem výzkumu měla baterie 13 proměnných. Každý faktor s vlastním číslem (Eigenvalue) > 1 jde podle **Kaiserova pravidla** do FA.

Velicerův test MAP (Minimum Average Partial) pracuje s analýzou hlavních komponent a s následně vzniklými maticemi parciálních korelačních koeficientů. Hledá, které komponenty jsou společné (tedy nejlepší faktorové řešení) a podle toho určuje počet faktorů. Podceňuje faktory s menším počtem sytících položek nebo s nízkými zátěžemi (Ledesma, Valero-Mora, 2007). **Hornova metoda paralelní analýzy:** z korelační matice se extrahují vlastní hodnoty (Eigenvalues). Simulací (Monte Carlo) se vytvoří souběžné soubory o stejném N a se stejným počtem proměnných. Generují se pro ně srovnávací Eigenvalues. Takových porovnávání přitom může být 100 i více. Pokud je vlastní číslo vyšší než průměr (nebo 95 % percentil) náhodně vygenerovaných Eigenvalues, faktor se použije.

Výsledky faktorové analýzy mohou být z větší nebo menší části odlišné v případě, kdy zvolíme jinou výchozí matici koeficientů, např. místo tetrachorických koeficientů pearsonovské korelace (srovnejte Tabulku 6 a 8).

K následující tabulce 8: jde o rozptylovou vstupní matici pearsonovských korelací; dále o metodu extrakce Hlavních komponentů (Principal components) a metodu rotace VARIMAX. Počet faktorů byl určen na základě Kaiserova pravidla. Čtyři faktory vysvětlují 55,3 % celkového rozptylu, z toho C1 25,3 %. Spolehlivost (Cronbachova alfa) prvých dvou faktorů není vysoká: C1 .593; C2: .616. Naproti tomu vyhovují C3: .847 a C4: .737.

Analýza hlavních komponentů v Tabulce 8 má poněkud odlišná východiska (místo tetrachorické rozptylové matice jako v tab. 6 jsou zde použity Pearsonovy korelace, místo paralelní analýzy Kaiserovo pravidlo pro určení počtu faktorů). Přestože už víme, že data (dichotomie) docela nemusí vyhovovat předpokladům analýzy, pokusíme se je interpretovat, abychom názorně viděli rozdíl. Získali jsme komponent C1 hodnocení deliktů proti integritě společnosti (finančních, environmentálních a rasistických přestupků) a C2 hodnocení přestupků proti integritě osob, tedy dost podobné faktorům z tabulky 6. Navíc ovšem C3 hodnocení prostituce a C4 hodnocení užívání/držby drog jako zvláštní faktory (hlavní komponenty).

Tabulka 8: Rotovaná faktorová struktura (mínění veřejnosti o trestním postihu za skutky)

Rotated Component Matrix				
	Component			
	1	2	3	4
r18_11	**,678**			
r18_13	**,650**			
r18_10	**,633**			
r18_9	**,535**			
r18_12	**,499**			
r18_6		**,700**		
r18_5		**,683**		
r18_4		**,640**		
r18_3		**,520**		,360
r18_8			**,904**	
r18_7			**,869**	
r18_1				**,867**
r18_2				**,845**

Legenda k Tab. 8: r18_11 vyhýbání se splácení dluhů; r18_13 šizení na daních; r18_10 poskytování půjček na neúměrně vysoký úrok; r18_9 závažné poškozování životního prostředí; r18_12 veřejné projevování sympatií k rasismu; r18_6 nelegální kopírování hudby, filmů, počítačových programů apod.; r18_5 usmrcení trpící nevyléčitelně nemocné osoby na její žádost; r18_4 šíření urážlivých nepravdivých skutečností o jiné osobě; r18_3 dobrovolný pohlavní styk s osobou ve věku 14 let; r18_8 využívání sexuálních služeb za úplatu; r18_7 poskytování sexuálních služeb za úplatu; r18_1 užívání nelegálních drog; r18_2 držení nelegálních drog pro vlastní potřebu

Některé dvouznačné proměnné z tabulky 6 (r18_4: šíření urážlivých a nepravdivých skutečností o jiné osobě jsou dvouznačné, sytí skoro stejně dva faktory) jsou zde přiřazeny do působnosti převážně jiného faktoru. Nebo se zde naopak faktorová přináležitost rozvolní (r18_3: dobrovolný pohlavní styk s osobou mladší 15 let zde sytí dost silně dva faktory), popř. zpevní (r18_6: nelegální kopírování hudby, filmů, počítačových programů apod. – sytí velmi silně jen jeden faktor).

V příkladu 4 – Zkušenost české veřejnosti s vyjmenovanými psychotropními látkami (Zeman et al., 2011a) - jsou respondenti mj. v otázce 17 dotazováni, zda už v životě vyzkoušeli nějakou vyjmenovanou psychotropní látku a kdy to bylo naposledy. Odpovědi

byly nabídnuty na 4-místné stupnici od ano: v posledních 30 dnech, 12 měsících, před více než 12 měsíci až po nikdy. Paralelní analýzou se stanovil počet faktorů pro skory otázky 17 na 2-3.

Dvoufaktorové řešení se zdá být spolehlivější s výraznějšími zátěžemi u vícepoložkových faktorů a vysvětluje více celkového rozptylu. Čtyř faktorové řešení přináší rovněž zajímavé podněty k zamyšlení, ale 2 faktory jsou zastoupeny vždy jen malým počtem položek (C3 a C4 mají po 2 proměnných) a prvé dva faktory mají slabší spolehlivost. Z těchto důvodů bychom měli dát přednost dvou faktorovému řešení na bázi tetrachorických koeficientů.

Shrnutí a rozvedení problematiky rotace, extrakce a hledání jednoduché struktury

Analýza hlavních komponentů je nejčastěji používanou metodou extrakce, jak ukazují nedávné přehledy mnoha stovek výzkumných projektů (Osborne & Costello, 2009). Ve skutečnosti nejde o faktorovou analýzu: komponenty se vypočítávají z veškeré variance pozorovaných proměnných a společný rozptyl není oddělován od individuálního. Mezi použitými metodami skoro není rozdíl, pokud se vychází z velkých matic, např. z velkého počtu položek v dotazové baterii. Extrakce hlavních komponentů obvykle přináší vyšší zátěže, což může vést někdy k chybným závěrům.

Vcelku lze metodu hlavních komponent doporučit hlavně při položkové analýze nebo tehdy, když chceme snížit počet otázek v dotazníku tak, že vyloučíme překrývající se otázky.

Faktorovou analýzu (FA) používáme hlavně k potvrzení, že se za množinou měřených proměnných skrývá nějaká latentní struktura (přímo nepozorované faktory). FA modeluje vztahy mezi pozorovanými proměnnými, faktory a chybou. Východiskem je tzv. redukovaná korelační matice, kde na diagonále jsou komunality. Nejčastěji se používají metody faktorování podle hlavních os, zřídkakdy maximální pravděpodobnosti.

Metoda faktorování podle hlavních os (principal axis factoring, PAF): tím, že pracuje se společným rozptylem (položka A se na něm podílí stejně jako položka B) se PAF soustřeďuje hlavně na vztahy proměnných, na jejich korelace. Používá se, pokud hodnoty jedné nebo více proměnných významně postrádají normální rozložení.

Metoda maximální věrohodnosti (Maximum Likelihood, ML) se používá na extrakci z dat, která jsou relativně normálně rozložena. Statistici tuto metodu (a také např. metodu Obecných nejmenších čtverců: Generalized Least Squares, GLS) rádi doporučují, protože se při ní testuje statistická významnost faktorových zátěží (testem dobré shody CHI^2) a vypočítávají také korelace mezi faktory (podobně jako u šikmých rotací). Ovšem ohledně spolehlivosti testu shody modelu s daty existují v literatuře značné rozpory, přičemž značná část, ne-li většina praktiků a odborníků se přiklání k názoru vyjádřenému např. jednou z našich odbornic: omezuje jenom na malé datové soubory, protože „s rostoucí velikostí analyzovaného souboru se i malé neshody stávají statisticky významnými a test zamítá i jinak správný model. Test je tedy třeba používat pouze jako doplněk analýzy a jeho použití se omezuje na menší soubory.“ (Škaloudová, 2012).

Zmiňovali jsme rotaci jako metodu, která umožňuje v rámci stejných korelačních parametrů našich proměnných zlepšit vztahy k jednotlivým faktorům, vidět je jednoznačněji. V nerotované verzi máme proměnné promítnuté do prostoru v různých překryvech a přesazích. V rotovaném pohledu je od sebe lépe rozlišíme. Cílem rotace je nalézt tzv. **jednoduchou strukturu** (simple structure). Je to stav, kdy každá proměnná koreluje jen s jedním faktorem a u ostatních má nízké zátěže.

Přesněji: a) v každém řádku faktorové matice je aspoň jedna zátěž blízká 0, b) v každém sloupci je alespoň tolik malých zátěží, kolik je faktorů, c) pro každou dvojici sloupců platí, že obsahuje alespoň tolik dvojic proměnných s malou zátěží u jednoho a velkou zátěží u druhého faktoru, kolik je faktorů.

V praxi za malou považujeme zátěž $\leq 0{,}32$ nebo ≥ -0.32 (Tabachnick and Fidell, 2001). Za velkou zátěž od ±0,50 výš a uspokojivě velkou od ±0,7 výš. Zopakujeme si: faktorová zátěž 1 nebo -1 (=100 %) znamená, že variance proměnné je faktorem zcela vyčerpána a 0 naopak že faktor se vůbec proměnné nedotýká.

Rotací, způsobů, jak nalézt jednoduchou strukturu, je celá řada. Většina se opírá o tzv. simplicitní funkci (Bentlerova faktorová zátěž: Garson, 2011b). Ta nabývá mezních hodnot (vysokých či naopak nízkých), když se k jednoduché struktuře přiblížíme.

Nejdříve se rozhodneme mezi rotací pravoúhlou (ortogonální) nebo šikmou (oblique). V prvním případě hledáme na sobě **nezávislé faktory**, které jsou jakoby vůči sobě v prostoru pravoúhle uspořádány. Je to např. nejčastěji užívaná metoda VARIMAX a patří sem i QUARTIMAX. K šikmým rotacím počítáme PROMAX nebo OBLIMIN a další. V tomto případě jsou faktory na sobě závislé. Toto rozhodnutí by měl výzkumník uskutečnit na základě teoreticky zdůvodněných očekávání. Je třeba mít na paměti, že pokud faktory považujeme za nezávislé (při volbě pravoúhlého řešení), neměly by být silně zkorelované. To může potvrdit šikmá rotace, která sleduje zkorelovanost faktorů.

Extrakce provedené jak a) faktorovou analýzou tak i b) analýzou hlavních komponent, pokud je řešení stabilní, přinášejí stejné výsledky, tj. stejné faktory. Mohou se ovšem lišit jejich pořadím a tím i podílem vysvětlené variance.

Tabulka 9: Popisné statistiky k otázce 7 výzkumu mladých lidí

Proměnná	N	Minimum	Maximum	Mean	Std. Deviation	Skewness		Kurtosis	
						Statistic	Std. Error	Statistic	Std. Error
q7_1	1477	1	6	3,90	1,571	-,358	,064	-,910	,127
q7_2	1464	1	6	2,62	1,479	,595	,064	-,648	,128
q7_3	1467	1	6	3,78	1,534	-,218	,064	-,962	,128
q7_4	1467	1	6	2,62	1,482	,653	,064	-,528	,128
q7_5	1471	1	6	3,99	1,608	-,364	,064	-,987	,128
q7_6	1457	1	6	3,46	1,551	,012	,064	-,980	,128
q7_7	1437	1	6	2,57	1,386	,584	,065	-,484	,129
q7_8	1479	1	6	2,73	1,389	,502	,064	-,559	,127

Legenda k tab. 9: q7_1 Oko za oko, zub za zub; q7_2 Kdo nekrade, okrádá svoji rodinu; q7_3 Kdo nepracuje, ať nejí; q7_4 Žít na dluh se nevyplácí; q7_5 Poctivou prací nelze zbohatnout; q7_6 Uzavírání manželství je zbytečné, svatba je přežitek; q7_7 Většina starých lidí jen ekonomicky zatěžuje náš stát; q7_8 Většina mladých lidí jsou flákači a příživníci

Příklad č. 5 (Večerka et al., 2011) **Souhlas mladých lidí v ČR s vybranými životními zásadami a názory**. V tabulce 9 je zachycena část matice vytvořené z baterie 24 položek. Hodnoty rozdělení odpovědí ve sloupcích zkosení (skewness) a špičatost (kurtosis) nepřekračují 1, ale v poměru ke směrodatné odchylce nevyhovují. Data proto nemohou být

považována předběžně za (jednorozměrně) normálně rozložená. Odpovědi měly minimum 1=určitě souhlasím až maximum 6=určitě nesouhlasím.

Abychom z ordinálních dat získali data velmi blízká metrickým, požadovaným pro faktorovou analýzu, přetransformovali jsme běžné (pearsonovské) korelace na tzv. polychorické. Data sice nevyhovují podmínkám kladeným na metodu maximální věrohodnosti (ML), ale mohou být faktorována pomocí metody faktorování podle hlavních os. Podle Hornovy PA by měl optimální počet faktorů být 3 (2 položky s velmi nízkou komunalitou jsme z analýzy vypustili).

Zjištěné faktory podle následující tabulky 10 jsou 1. mediální a politicko-policejní stabilita; 2. xenofobie a preference vlády pevné ruky; 3. morálka pouličních gangů a vrstevnického snobismu. Hodnoty v tabulce, které nejsou vyznačené tučně, ukazují výši podvojných zátěží. Jde v těchto případech o položky, které se odchylují od jednoduché struktury, např. položka „policie je při omezování kriminality neschopná“ sytí jak faktor xenofobie, tak i (přestože v menší míře) faktor morálky pouličních gangů.

Tabulka 10: Rotovaná faktorová struktura skorů z odpovědí na otázku 7 výzkumu mladých lidí

Rotated Factor Matrix[a]				
	Factor			Cronba-chova alfa
	1	2	3	
q7_20	**,843**			0,732
q7_21	**,830**			
q7_18	**,802**			
q7_19	**,800**			
q7_24	**,346**			
q7_17	**,336**			
q7_12		**.759**		0,739
q7_13		**,696**		
q7_11		**,621**		
q7_14		**,523**		
q7_16		**,422**	,388	
q7_23		**,417**		
q7_22		**,352**		
q7_6			**,564**	0,74
q7_10			**,563**	
q7_7			**,547**	
q7_15			**,545**	
q7_4			**,492**	
q7_5			**,463**	
q7_9		,307	**,409**	
q7_1		,323	**,348**	
Extraction Method: Principal Axis				
a. Rotation converged in 5 iterations.				

Legenda k Tab. 10: q7_20 Zprávy, které se vysílají v komerčních televizích, jsou pravdivé; q7_21 Zprávy, které se dočítáte v denním tisku, jsou pravdivé; q7_18 Zprávy, které se vysílají v České televizi, jsou pravdivé; q7_19 Zprávy, které se získávají z Internetu, jsou pravdivé; q7_24 V současné době se zlepšuje celkový přístup policie ČR k občanům; q7_17 Politická situace v ČR se dobře vyvíjí, důležité věci jdou správným směrem; q7_12 V ČR je v současnosti moc divných cizinců; q7_13 V ČR je moc bezdomovců a žebráků; q7_11 Romové jsou nebezpeční lidé; q7_14 Za nejtěžší zločiny by měl být udělován trest smrti; q7_16 Policie je při omezování kriminality neschopná; q7_23 Nejlepší je neplést se do ničeho, co se mne netýká; q7_22 Chování občanů je málo kontrolováno, takže si každý dělá, co chce; q7_6 Uzavírání manželství je zbytečné, svatba je přežitek; q7_10 Než hlásit kriminalitu na policii, je lepší si to s pachatelem vyřídit důrazně sám; q7_7 Většina starých lidí jen ekonomicky zatěžuje náš stát; q7_15 Kdo nemá peníze, nic neznamená; q7_4 Žít na dluh se nevyplácí; q7_5 Poctivou prací nelze zbohatnout; q7_9 Jestliže nemáš to, co je v módě a nejsi „in“, nepřijmou tě mladí lidé mezi sebe; q7_1 Oko za oko, zub za zub

Většina autorů připouští, že ve společenskovědním výzkumu se užívá metoda faktorace podle hlavních os (PAF) častěji než metoda maximální věrohodnosti (ML) a zvláště pak při vážném narušení normality rozdělení proměnných. Ale mnozí věnují ML (u nás např. již shora citovaná Škaloudová, 2012) velkou pozornost. Zopakujeme si: Bývá to hlavně proto, že výstup procedury je doprovázen testem významnosti, lze vypočítat různé koeficienty dobré shody, korelaci mezi faktory a konfidenční intervaly. Avšak někteří renomovaní odborníci, např. Osborne & Costello (2009: 143) nejenom preferují faktorovou analýzu před PC jako mnozí jiní, ale doporučují jako nejvhodnější metodu extrakce ML: „..optimální výsledky (tj. výsledky, které lze zobecňovat i na další výběrové soubory a které odrážejí povahu základního souboru) budou dosaženy extrakční metodou skutečné faktorové analýzy (dáváme přednost maximální věrohodnosti)“.

Proč je ve společenskovědních výzkumech tak málo analýz s extrakcí ML je zřejmě dáno přísnými požadavky na normalitu vstupních dat, jimž nelze v praxi vždy vyhovět. Jednou z možných příčin malé úspěšnosti ML v českém kriminologickém výzkumu může být také fakt, že se při dotazování nejčastěji využívají čtyř až šestimístné lickertovské stupnice (např. v European Value Survey se v bateriích často vyskytují lépe ML metodou zpracovatelné stupnice skorované 0-9 nebo 1-10, viz dotazník 2008, EVS). A tak se také pomocí ML dosahují významné výsledky jen výjimečně (ale dobře výzkumně argumentované názory citované shora jsou podnětem k tomu, abychom ML nezavrhovali). Uvádím dále jeden takový poměrně vzácný a poměrně úspěšný případ (mj. také kvůli zdařilému zpracování otázek): příklad č. 6. Test shody faktorového modelu s daty zde neukázal významný rozdíl. Správně by tomu mělo být pokaždé, když chceme dokázat, že náš model se neliší od dat (nevyvrátila se nulová hypotéza shody). Ale výsledek analýzy bohužel nevyhovuje z jiných důvodů, popsaných dále.

Tabulka 11: Rotovaná faktorová struktura skorů z odpovědí na otázku 6 výzkumu mínění mladých lidí o kriminalitě

Rotovaná faktorová matice[a]			
	Faktor		
Proměnná	1	2	3
q6f	**,759**		
q6e	**,561**		
q6h	**,518**	*x(0,373)*	
q6b	**,514**		
q6d	*x(0,322)*	**,677**	
q6c	*x(0,373)*	**,631**	
q6g		**,493**	
q6a			**,576**
Metoda extrakce: Maximum Likelihood.			
a. Rotace konvergovala v 5 iteracích.			

Legenda k Tab. 11: q6f „Občan vidí, že se někdo pokouší o sebevraždu“; q6e „Občan vidí, že někdo prodává před základní školou dětem drogy“; q6h „Občan slyší ve svém bytě, že někdo v blízkém parku zoufale volá o pomoc; q6b „Občan se stal svědkem závažné dopravní nehody se zraněním“; q6d „Občan se stal svědkem pokusu o vykradení cizího auta na ulici“; q6c „Občan se stal svědkem přepadení na ulici“; q6g „Občan vidí, že policie pronásleduje pachatele trestného činu“; q6a „Nález peněženky s větší množstvím peněz a s adresou a jménem majitele“

Goodness-of-fit Test		
Chi-Square	df	Sig.
10,169	7	,179

Obr. 3: Test dobré shody faktorového modelu ML s daty ve výzkumu mínění mladých lidí o kriminalitě (Otázka 6)

Příklad č. 6 (Večerka et a., 2011): **Názory mladých lidí na chování svědků ve vybraných situacích souvisejících s trestným činem.** Otázka 6 se zaměřila na zjištění projektivně konstruovaných odhadů respondentů, do jaké míry by se ostatní lidé zachovali morálně nebo nemorálně, kdyby byli svědky různých situací včetně vážné dopravní nehody doprovázené zraněním, přepadení na ulici, vykrádání cizího auta na ulici, prodeje drog dětem před školou, pokusu o sebevraždu, pronásledování pachatele policií, volání o pomoc nebo kdyby nalezli peněženku s větším obnosem a adresou majitele.

V příkladu 6 je velmi pěkně pořadová stupnice kvalifikována: 1) téměř nikdo by..., 2) více lidí by ne... 3) zhruba polovina by... 4) více lidí by (ano) a 5) téměř všichni by... To zřejmě přispělo k přesnějším odhadům, větší vyváženosti celé baterie bez obvyklých „šumů". Nalezli jsme pomocí ML extrakce tyto 3 faktory (Tabulka 20): 1) obětí násilí, drog, 2) aktérů krádeže a přepadení a 3) přisvojení si opuštěné, ztracené, ale neanonymní věci.

Nicméně třetí faktor (Tabulka 10) je zastoupen jen jednou položkou, což se všeobecně nepovažuje za dobré stabilní řešení (přestože faktorový model včetně 3 latentních proměnných dobře odpovídá vstupním datům- viz obr. 3). Tento výsledek je dobrou ukázkou toho, že na statistické testy nelze spoléhat mechanicky. 3. faktor by se možná stabilizoval, kdyby baterie obsahovala ještě další položky podobného typu.

V již citovaném **příkladu č. 4** (Zeman et al., 2011a), týkajícím se osobní zkušenosti české veřejnosti s užíváním drog, většina položek baterie nebyla normálně rozložena. Nemohli bychom se tedy rozhodnout správně, kdybychom zvolili při faktorové analýze jako metodu extrakce ML, která normalitu vyžaduje. Častěji v takovéto situaci zvolíme např. faktorování podle hlavní osy (PAF). Alternativní přístup zahrnuje možnosti transformace dat na logity nebo probity, které už mohou předpoklady normality splňovat.

Ortogonální metoda rotace přináší jednodušší a jednoznačnější, přehlednější výsledky. Předpokládá se přitom, že zkoumané jevy spolu moc nesouvisejí. Šikmá rotace však bývá bližší realitě. Psychologové např. uvádějí, že logické, matematické a verbální schopnosti jsou zkorelované, tyto faktory k sobě mají blízko. Reflektují jednu společnou vlastnost-inteligenci. Šikmá rotace se hůře interpretuje. Jako výstup jsou k dispozici dvě matice: faktorová struktura a faktorový vzor (factor pattern). Velmi výstižně tyto matice charakterizuje Garson (2012b: 32): faktorová struktura je „maticí faktorových zátěží jako při ortogonální rotaci", kdy rozptyl měřených proměnných je vysvětlován faktorem celkově, jak na „základě jedinečného tak i společného příspěvku". Naproti tomu matice faktorového vzoru „obsahuje koeficienty, které představují pouze jedinečné příspěvky. Čím je faktorů více, tím jsou koeficienty faktorového vzoru nižší, protože se dostaví více společných příspěvků k vysvětlenému rozptylu".

Při interpretaci se hlavně opíráme o matici faktorové struktury, která je uvedena dále (viz tabulka 12), ale matici s faktorovým vzorem používáme nezřídka jako klíč k té dříve zmíněné.

V příkladu 4 a v tabulce 12 jsou použity jako metody extrakce jak hlavní komponenty tak i faktorová analýza. Rotovali jsme pravoúhle (VARIMAX) i šikmo (OBLIMIN) a dospěli ke stabilním výsledkům, které se tedy navzájem potvrzují. Odlišily se podle různého stupně užívanosti (v pořadí od nejméně k nejvíce užívaným): tvrdé a měkké drogy a běžné (dostupné nebo dostupnější) psychotropní látky.

Šikmá řešení prozradila, že některé drogy jsou vnímány v nejednoznačných kontextech (marihuana, LSD a houby). Především marihuanu (r=.468) část respondentů vnímá ve stejném kontextu jako alkohol a tabák, tedy jako běžně dostupné a sociálně tolerovanější psychotropní látky. Podobně (r=.298) LSD je zčásti vnímána/užívána také jako měkká droga, přestože je silněji respondenty řazena mezi tvrdé drogy. Nicméně faktory jsou poměrně nezávislé, málo zkorelované, čímž se zhruba potvrzuje správnost orthogonálního řešení (pouze faktor 1 a 3 má o něco vyšší závislost: zhruba 22 % společné variance).

Tabulka 12: Rotovaná faktorová struktura otázky 17 z výzkumu drogové problematiky

Extrakce:		Hlavní komponenty		Faktorace podle	
Rotace		Pravoúhlá	Šikmá	Pravoúhlá	Šikmá
% vysvětleného		63,6	x	47,4	x
	Vyzkoušel/a jste někdy v životě … kdy to bylo				
Faktor 1: měkké drogy	q17_4	**,850**	**,859**	**,819**	**,831**
	q17_3	**,732**	**,765**	**,606**	**,665**
	q17_5	**,728**	**,741**	**,564**	**,600**
	q17_9	**,515**	**,589**	**,433**	**,515**
	q17_8	x(.315)	x(0,437)	x(0,322)	x(.444)
	q17_6				x(.305)
	q17_7				x(.331))
Faktor 2: tvrdé drogy	q17_6	**,875**	**,873**	**,890**	**,885**
	q17_7	**,783**	**,793**	**,592**	**,622**
	q17_8	**,666**	**,715**	**,528**	**,589**
	q17_9	x(.396)	x(0,489)	x(0.330)	x(0,425)
	q17_5		x(0,332)		x(0,357)
	q17_4				x(0.327)
Faktor 3: běžné látky	q17_2	**,845**	**,840**	**,592**	**,592**
	q17_1	**,796**	**,812**	**,656**	**,674**
	q17_3	x(.326)	x(0,404)	x(.336)	x(.426)
Korelace faktorů:					
	F1 s F2		,209		,298
	F1 s F3		,376		,468
	F2 s F3		,032		.050

Legenda k Tab. 12: q17_4 extázi; q17_3 marihuanu či hašiš; q17_5 pervitin, amfetaminy; q17_9 halucinogenní houby; q17_8 LSD „krystal, trip, papír“; q17_6 kokain; q17_7 heroin; q17_6 kokain; q17_7 heroin; q17_8 LSD „krystal, trip, papír“; q17_9 halucinogenní houby; q17_5 pervitin, amfetaminy; q17_4 extázi; q17_2 alkohol; q17_1 tabák; q17_3 marihuanu či hašiš

Nejsilnější zátěže jsou vyznačeny tučně. Jako x a číslem v závorce jsou označeny podružné zátěže vyšší než .300

Závěr k faktorové analýze, explorační a konfirmační formy

Výhodou FA je, že vychází ze skutečně naměřených souvztažností mezi pozorovanými jevy a neseskupuje je dohromady na základě nějaké vnější podobnosti, jakou např. může být vzdálenost od naměřeného středu určité zkoumané množiny hodnot.

Dále, i když při dotazování obrátíme stupnice a známkujeme např. místo od 1 do 5 obráceně od 5 do 1, faktory a jejich výklad se nijak nezmění.

Nevýhodou může být dopad různých „divokých" pozorování (např. provokativní odpovědi na konci dlouhého dotazníku, kontra konformistické reakce mladých mužských respondentů na otázky týkající se kriminalizace prostituce a drog). Odtud odvozené korelace mohou způsobit, že faktorová analýza extrahuje singulární, zamlžující faktor, s nímž se pak výzkumník trápí. Tomuto riziku se čelí analýzou tzv. odlehlých pozorování (outliers), která se odstraňují buď vyloučením těchto odlehlých pozorování z datového souboru nebo transformací dat (změnou vzdáleností mezi stupni použité škály při zachování jejich sledu).

Objektivita a spolehlivost faktorové analýzy je dána dokonalým a nezkresleným sběrem dat (k němuž pravděpodobně nikdy nemůže dojít).

Dále, interpretace faktorů není absolutní: je závislá na představivosti, heuristice. Je možné se setkat s více interpretacemi dat shodně faktorovaných. Nevýhody FA lze efektivně snížit jednak pečlivou předběžnou analýzou dat a jednak týmovou prací na interpretaci faktorů.

Abychom doplnili ještě nezmíněnou typologii metod faktorové analýzy, je potřeba zmínit, že podle teoretických východisek nebo jejich absence jde o 1) explorativní či heuristickou, 2) konfirmační či absolutní faktorovou analýzu.

Explorační faktorovou analýzou zjišťujeme, odhalujeme povahu zkoumaného jevu: ovlivňuje něco (a co?) celou řadu názorů, projevů chování apod.? Někdy se nemusíme shodnout v počtu faktorů. Usilujeme o hypotetické a nikoliv o absolutní vysvětlení. Chceme

podstatě jevů porozumět relativně, jako když např. říkáme, že slunce vychází/zapadá a rozumíme si (ačkoliv je vědecky i prakticky doloženo, že vychází/zapadá země).

Konfirmační faktorovou analýzou testujeme již existující teorii nebo výzkumem získané výsledky (např. z explorativní faktorové analýzy). Působí tyto proměnné na různé reakce (odpovědi) podle naší předpovědi (teorie)? Nebo podle našich předešlých výsledků explorativní analýzy? Opírá se o teorii nebo zobecněný empirický výzkum a testuje výchozí hypotézy. Začala se používat cca kolem roku 1979 (Thompson, 2004). Používáme při ní např. modelování strukturálními rovnicemi (SEM, Structural Equation Modelling).

Na obr. 3 uvádíme výsledky jedné takové konfirmační faktorové analýzy. Jde o příklad (č. 6) zaměřený na názory mladých lidí na chování lidí ve vybraných situacích souvisejících s trestným činem (Večerka et al., 2011). Obr. 3 představuje cestičkový diagram (path diagram) získaný zpracováním dat v programu AMOS (metodou ML). Předchozí explorační faktorová analýza, provedená metodou PAF, naznačila existenci 2-3 faktorů (podle Hornovy paralelní analýzy). Data byla přitom blízká normálnímu rozdělení, tedy vhodná pro použití metody ML. Nalezenou faktorovou strukturu jsme proto využili ke konstrukci modelu, který by měl být potvrzen konfirmační faktorovou analýzou. Ta by tedy měla zodpovědět otázku, jestli mladí lidé skutečně při setkání s kriminalitou nahlížejí hodně odlišně na situace, jimiž mohou být svědky a v nichž se objevuje v popředí zájmu a) oběť nebo b) pachatel.

Malá odbočka, než přistoupíme k samotné interpretaci obr. 4. Česká pracoviště kriminologie, pokud jsem informován, zpravidla příslušné softwary na zpracování takových dat nemají (jde o Lisrel, AMOS, EQS aj.). SEM (modelování pomocí strukturálních rovnic: Structural Equation Modeling), který tyto softwary implementují, představuje poměrně složitou problematiku. Její popularita v kriminologii v posledních letech roste (viz např. poslední ročníky Journal of Criminal Justice Education, Criminal Justice and Behavior nebo monografii Choi, 2008).

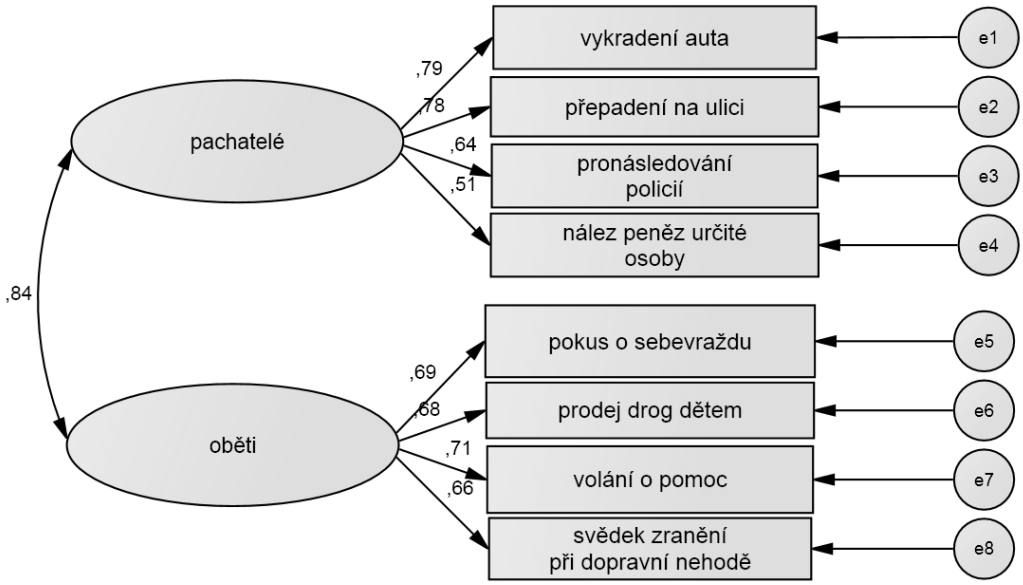
Ani AMOS (ale i další softwary) není tak snadný na zvládnutí, jak někteří autoři díky pohodlnému grafickému uživatelskému rozhraní tvrdí (Arbuckle, 2011: 2). Je pravda, že příkazy se poměrně komfortně prostředkují kreslením okének a oválů, dále jednoduchým ukládáním proměnných do těchto obrazců a spojováním obrazců šipkami, které naznačují

cestičku (směr) vlivu jedné proměnné na druhou nebo vzájemné závislosti. Avšak je zapotřebí prostudovat rozsáhlou literaturu a vyzkoušet četné postupy, než program začne přinášet očekávané výsledky. Hlubší pochopení vyžaduje prostudovat problematiku kovariačních matic a regresí, na nichž je SEM založeno. Dále seznámit se s charakterem a výší četných koeficientů, které se při analýze využívají k poměřování shody modelu s daty a z nichž dále uvádíme ukázku.

Zpět k obrázku 4: Analýza ukázala, že veškeré koeficienty splňují požadavky kladené na úspěšně postavený model (jde o výši koeficientů IFI a CFI, které mají být vysoké alespoň 0,9, TLI aspoň 0,90 a RMSEA má ležet mezi 0,0 a 0,08). Faktorové zátěže všech pozorovaných proměnných jsou vysoké, nejméně 0,511 a všechny statisticky významné (podle kritického poměru chyb a zátěže- CR- p=.001). Respondenti podle naší předchozí teorie, založené mj. na výsledcích explorační faktorové analýzy, odlišují situace, kdy člověk a) pozoruje pachatele určitého přestupku či trestného činu a kdy je b) svědkem ohrožení někoho dalšího na životě nebo zdraví. Avšak toto rozlišení se díky SEM potvrdilo jen jako velmi jemné: oba faktory jsou vysoce zkorelované (r=.837), řešení připomíná šikmou rotaci (ortogonální řešení by naopak ukázalo nízkou zkorelovanost obou faktorů).

Jeden pohled, jedno hodnocení (na pachatele v akci nebo na útěku před policií) snadno přechází v druhé hodnocení (s pohledem na újmu a oběti trestného činu) a SEM proto spíše naši výchozí teorii vyvrací.

Analýza s využitím SEM pochopitelně může dále postihnout případné další rozdíly, např. generační, vzdělanostní, jejich kombinace, nebo rozdíly podle potencionální (kriminální) problémovosti, kriminální citlivosti, zkušeností s užíváním drog. Může se pak ukázat, že některé skupiny populace (např. ženy v určitém věku) ostře oba aspekty odlišují a jiným do značné míry splývají. V možnostech práce s různými soubory a podsoubory dat je SEM (modelování pomocí strukturálních rovnic) velmi pružný.



Konfirmační faktorová analýza (Kriminalita mládeže)
CHI-kvadrát=76,102 df=19 (p=,000)
CFI=,986 TLI=,979 RMSEA=,045 (Pclose=,748) SRMR=.0241

Obr. 4: Cestičkový diagram z baterie otázky 6 výzkumu mínění mladých lidí o kriminalitě

3. Shluková analýza (Cluster analysis, klastrovací analýza, shlukovací analýza)

Pozadí a východiska

Roku 1935 použil pojem „cluster analysis“ poprvé genetik a psycholog Robert Choate Tryon (1901-1967), proslavený svými pokusy o testování inteligence krys pomocí nástrah v bludišti. Napsal také první monografii o shlukové analýze (Tryon, 1939). Povšiml si tzv. korelačního profilu, tj. skutečnosti, že některé proměnné mají shodný nebo velmi podobný průběh či konstelaci vztahů k ostatním proměnným.

Tryon vyvinul později počítačový program BC TRY, propracoval v 50. letech 20. století metodu hlouběji. Definoval ji takto: „Shluková analýza je obecný logický postup formulovaný jako procedura, pomocí níž seskupujeme objektivně jedince do skupin na základě jejich podobnosti a rozdílnosti“. (Tryon, 1939, 3) Metoda se brzy rozšířila do různých vědních oborů.

Další osobností, která významně zasáhla do vývoje této metody, byl Clyde Hamilton Coombs (1912-1988). Byl matematickým psychologem, studoval na Kalifornské universitě, kde ho ovlivnil Tryon. Pod vlivem Paula Lazarsfelda (1901-1976) a Thurstonových (1887-1955) seminářů i jeho proslaveného vystoupení Vectors of Mind (1934) se zaměřil na měření v psychologii. V jedné ze zásadních prací (Theory of Data, 1964) Coombs zformuloval myšlenku, že veškeré psychologické měření lze pojmout jako prostorové zobrazení vztahů mezi body. Preference podnětu jednotlivcem představuje kratší vzdálenost k ideálu (jednotlivci) než nepreferovaný podnět. „Buď jde o **vztahy nadřazenosti, dominance** (např. jedno světlo je jasnější než druhé, student A má jasně lepší matematické schopnosti než student B) nebo **příbuznosti, blízkosti** (např. dva výroky vyjadřují podobnější postoj než jiné dva výroky, nebo častý případ v dotaznících - jedno tvrzení vystihuje názor někoho lépe než jiné tvrzení). Kromě toho body můžeme rozlišovat tak, že patří do stejné sady (např. stejně

jasná světla) nebo do různých sad (např. učební předmět a schopnosti ho zvládat). (Dawes & Tversky, 1989, 1415)

Coombsovy práce někteří současníci označili jako „nemetrickou revoluci“ (Dawes & Tversky, 1989, 1416). Podle Coombsovy koncepce se pak se rozvíjely **hierarchické** (tj. se vztahy nadřazenosti a podřazenosti) **a nehierarchické či rozdělující** (se vztahy příbuznosti/nepříbuznosti, vzdálenosti/blízkosti) **metody** shlukování.

K rozvoji metody shlukové analýzy přispěla celá řada dalších přírodních vědců, sociologů, psychologů a matematiků: P.Lazarsfeld, M.Lorr, L.L.McQuitty, J.McQueen, Bulhaři D.Vandev, Y.Tzvetanova. Snad nejvíce ji poznamenal rakousko-americký biologický statistik a entomolog Robert R.Sokal (1926-2012) myšlenkou, že shluky jsou objektivně existující, „přirozené“ skupiny. V článcích a v zásadní práci Principles of Numerical Taxonomy (spolu s P.H.E.Sneathem, 1963) charakterizoval shlukovou analýzu jako „matematický výzkum metodou určování **přirozených skupin** mezi třídami jsoucen“(Sneath&Sokal,1973). Podle něho jde navíc o základní vlastnost žijících organismů, že takovéto klasifikace vytváří, že tedy hodnotí podobnosti a rozdíly a zařazují si je za účelem, aby se přizpůsobily prostředí.

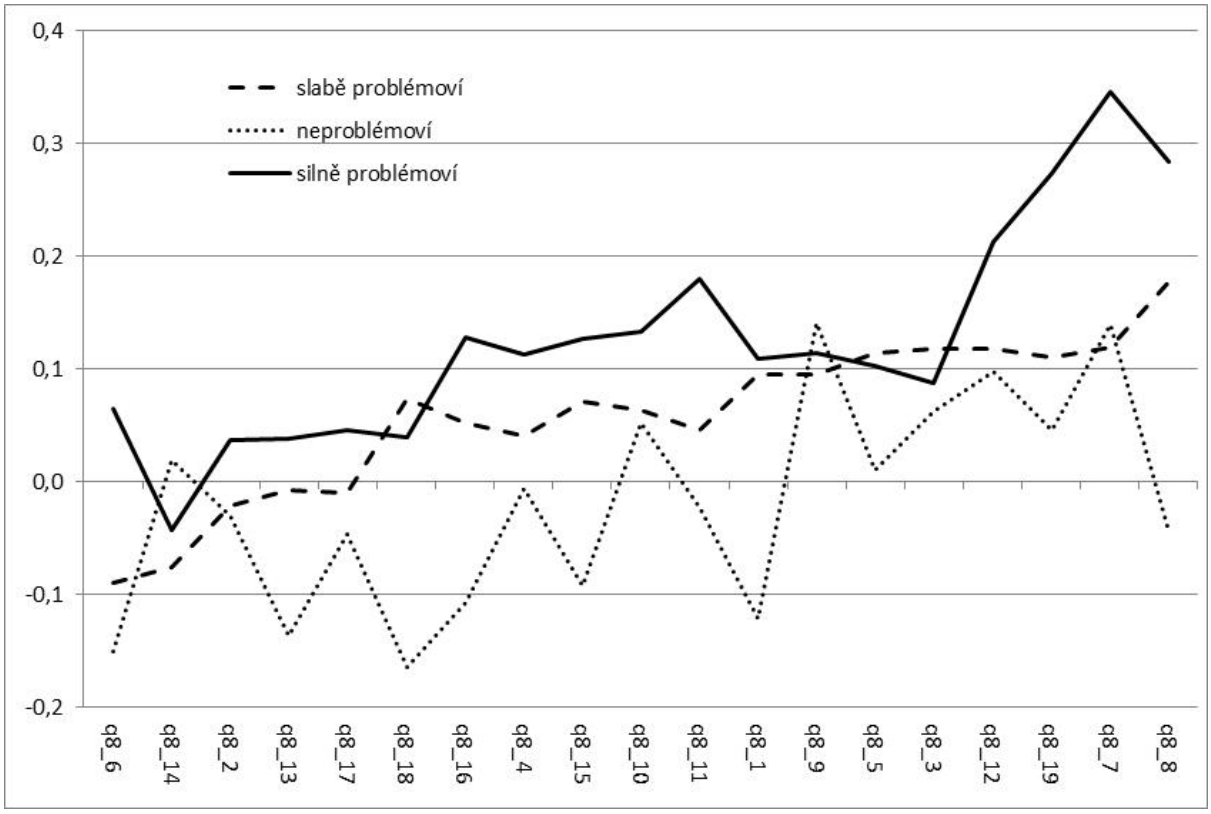
„Klasifikace je uspořádávání objektů podle jejich podobností... a objekty chápeme v nejširším slova smyslu včetně procesů a aktivit – cokoliv, co lze připojit k vektoru deskriptorů, takže klasifikace… je základní vlastností žijících organismů. Pokud organismy nejsou schopny seskupovat podněty do podobných druhů a určit podle nich třídy, na něž lze reagovat příznivě nebo vyhýbavě, jsou špatně přizpůsobeny k přežití “. (Sokal, 1977:1)

Shluková analýza se v kriminologickém výzkumu běžně používá už mnoho desítek let. Namátkou lze uvést několik příkladů: k určení psychologického profilu delikventů (Hindelang & Weis, 1972; Spaans et al, 2009) a alkoholiků (Bühler a Bardeleben, 2008), k identifikaci potenciálních problematických adolescentů/delikventů (Mun et al., 2008), ke zkoumání vlivu sousedství na riziko viktimizace (Product management group, 2004), ke geografické lokalizaci míst se zvýšenou kriminalitou (Neema, I. & D.Böhning, 2010; Griffith, 2007) a k předvídání budoucí kriminální kariéry na základě školního hodnocení (Juon et al., 2006). Aplikace vícerozměrných statistických postupů v kriminalistice včetně

shlukové analýzy se dokonce učí na některých amerických universitách nebo ve Spojeném království (např. Old Dominion University v Norfolku).

Začínajícím čtenářům lze doporučit jako vstup do problematiky shlukové analýzy osvěžující kurzy Andy Fielda (2000 a 2009) nebo webovou stránku Karla Wuensche (2012), kde si lze stáhnout názorná pojednání i datové soubory jako příklady k procvičení.

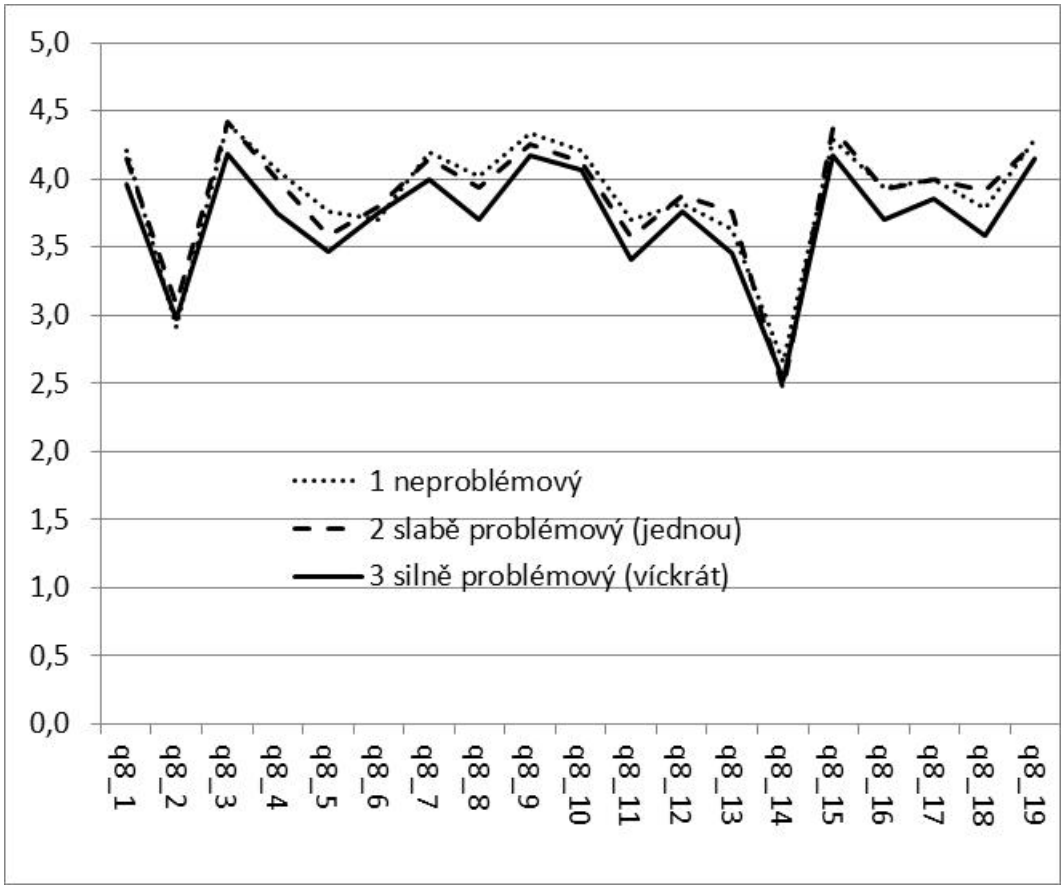
Jedinečnost přístupu shlukové analýzy lze demonstrovat na příkladu z výzkumu potenciální kriminality mládeže. **Příklad č. 8. Názory mladých lidí v ČR na stupeň problematičnosti některých negativních jevů** (Večerka et al., 2011). Respondenti měli vyjádřit, do jaké míry považují různé společenské jevy za problém: počínaje týráním členů rodiny, kyberšikanou, přes užívání psychotropních látek, korupci, terorizmus, nezaměstnanost atd. až k obecné kriminalitě. Korelace uvedených jevů k potenciální deviaci skupin mládeže, měly rozdílný profil. Za neproblémové jsme považovali ty mladistvé, kteří bez oprávnění nikdy neřídili motorové vozidlo. K velmi problémovým byli zařazeni ti, kteří bez řidičského průkazu jezdili v autě víckrát a ti, kteří řídili bez oprávnění jenom jednou, byli označeni za slabě problémové. Zřetelně podle tvaru křivek obr. 4 vidíme, že bližší jsou si slabě a velmi problémoví, kdežto neproblémoví (vytečkovaní) se vyčleňují a budou asi patřit do jiného shluku.



Obr. 5: Korelace potenciální deviace k dalším negativním jevům podle vnímaného stupně jejich problémovosti (data pro názornost upravena)

Legenda k Obr. 5, 6 a 7: q8_1 Týrání žen v rodinách; q8_2 Týrání mužů v rodinách; q8_3 Týrání dětí v rodinách; q8_4 Týrání starých lidí; q8_5 Kyberšikana; q8_6 Nefunkční policie; q8_7 Fetování; q8_8 Alkoholismus; q8_9 Nezaměstnanost; q8_10 Korupce a podplácení; q8_11 Rozpad rodin; q8_12 Poškozování přírody a životního prostředí; q8_13 Terorismus; q8_14 Počítačová negramotnost; q8_15 Zneužívání dětí k výrobě pornografie; q8_16 Agresivita vztazích; q8_17 Nebezpečí nakažení se AIDS/HIV; q8_18 Úmyslné, zlovolné pronásledování a obtěžování jiné osoby; q8_19 Kriminalita

Shluková analýza dokáže podobné rozdíly zachytit, přestože je založena na analýze vzdáleností a nesleduje korelace jako v příkladu uvedeném na obr. 5. Kdybychom se naopak řídili pouze dosaženými průměrnými skory, křivky by nám svým tvarem a blízkostí skoro splývaly (obr. 6).



Obr. 6: Průměrné hodnocení závažnosti negativních jevů podle problémovosti respondentů

Typy shlukové analýzy a příklad (věcné) obsahové identifikace shluků

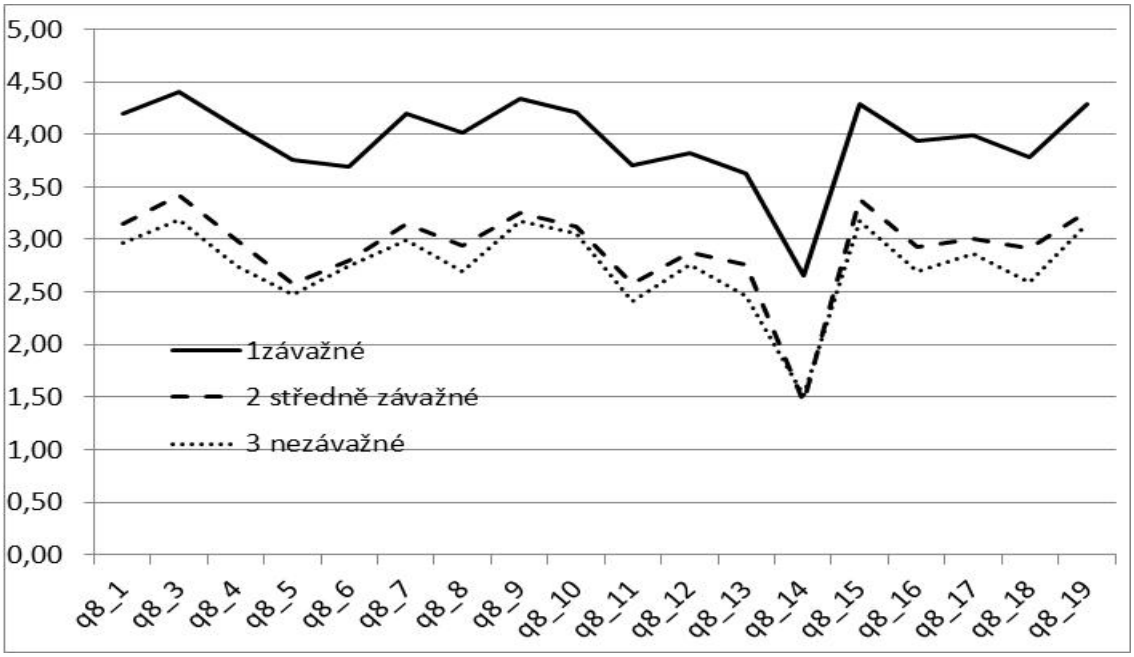
Metody shlukové analýzy se různě třídí. Jak už bylo shora zmíněno, v zásadě jde o

a) **hierarchickou metodu**, kdy se využívá jednou již vzniklých shluků k vytváření (dekompozici- v SPSS Hierarchical Cluster Analysis nebo agregaci- v jiných SW) dalších na zbylém souboru a

b) **nehierarchické metody** (v SPSS jako Quick cluster, tzv. k-means a pak Analýza nejbližšího souseda, Nearest Neighbour Analysis). Shluky jsou vytvářeny najednou, nikoliv postupně jako v hierarchické metodě. Vzájemně se vylučují buď úplně (=každý respondent patří jen do jednoho shluku) nebo aspoň zčásti (=respondent může patřit více, méně nebo stejnou měrou do různých shluků).

c) Existuje ještě případně jedna metoda, spočívající v **kombinaci** předešlého (v SPSS Two-step cluster analysis), kdy v prvním kroku se vytvoří nehierarchické shluky a rozsáhlý výběrový soubor se tak zredukuje na menší sadu k-means shluků. A teprve tehdy, na menším souboru do 100-200 případů, se s nimi v druhém kroku pracuje metodou hierarchické analýzy, která není jinak použitelná při zpracovávání (nepřehledných) tisícovkových souborů dat.

Společné všem přístupům je to, že vždy jde o metody současného hledání podobností a rozdílů, sdružování dat do smysluplných celků (shluků), kdy případy patřící do stejného shluku jsou si maximálně podobné, přičemž jsou minimálně podobné případům patřících do jiného shluku. Dalším charakteristickým rysem shlukové analýzy je fakt, že hledá podobnosti a rozdíly, strukturu v datech, aniž vysvětluje, proč tyto vztahy existují. To musí výzkumník zjistit podrobným dalšími analýzami.



Obr. 7: Průměry hodnocení závažnosti společenských jevů podle nalezených skupin (data pro názornost upravena)

Příklad č. 9: **Hodnocení stupně závažnosti negativních společenských jevů českou veřejností** (Zeman et al., 2011b). Pomocí shlukové analýzy jsme v rámci tohoto výzkumu identifikovali skupiny lidí, které různě hodnotily závažnost negativních společenských jevů. Nejdříve nám procedura k-means nabídla prosté zjištění, že shluků může být víc. Vybrali jsme z nich 3 nejstabilnější a zjistili (frekvencí v SPSS), že do prvního shluku patří 38,3 %, do druhého 40,3 % a do třetího 21,4 % respondentů.

K tomu, co jsou to za shluky, jak je nazvat, jsme dospěli až další analýzou, při níž jsme použili různé procedury včetně analýzy rozptylu (ANOVA). Tyto analýzy ukázaly, že prvé dvě skupiny se příliš neliší pořadím položek, na které kladou důraz: obě skupiny shodně hodnotily např. jako nejzávažnější týrání dětí a zneužívání dětí k pornografii.

Třetí skupina kladla na první místo nezaměstnanost, dále alkoholismus, fetování a týrání dětí. Hlavní rozdíl jsme spatřovali v intenzitě, s jakou skupiny považují dané jevy za problém (Obr. 7): jestli jsou pro ně závažné, středně závažné anebo nezávažné. Dalšími korelacemi ke složení a k názorům respondentů jsme zjistili, že skupina se sklony podceňovat negativní společenské jevy je potencionálně delikventní (obsahuje významně více problémových osob). Naopak skupina hodnotící negativní jevy jako závažné je tvořena výlučně neproblémovými respondenty.

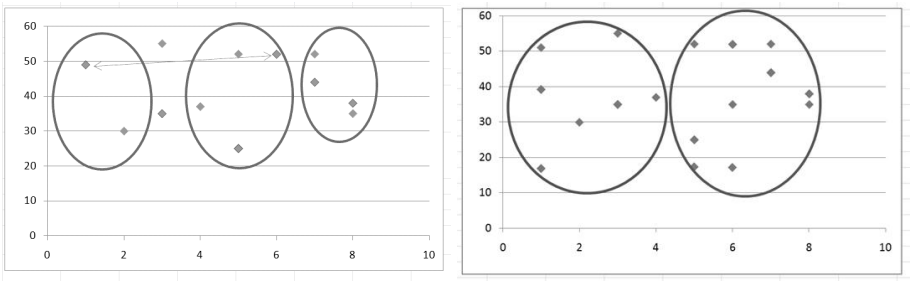
Koncept vzdálenosti a podobnosti

Podobnost vyjadřujeme koeficientem korelace nebo asociace. Blízkost či nepodobnost je na našem příkladu (Obr. 8) dobře vidět, protože je vyjádřena průměry. Ty však už jsou počítány na základě výsledků shlukové analýzy. Blízkost a nepodobnost sledujeme na stovkách a tisících konkrétních pozorování, skorů a jejich konstelací u jednotlivých respondentů. Potřebujeme pro ně nezkreslenou, nevyhlazenou míru blízkosti a nepodobnosti. Tu představují jednak různé distanční koncepty a také koeficienty podobnosti nebo nepodobnosti. V odborné literatuře se hovoří o měření a o koeficientech zachycujících **vzdálenost** (distance) nebo **nepodobnost** (dissimilarity) v n-rozměrném prostoru. Na osách tohoto prostoru leží různé odstupňované vlastnosti, např. agresivita v mezilidských vztazích, alkoholismus nebo týrání dětí. Nejde ve skutečnosti o vzdálenosti prostorové, ale o prostorovou projekci, o prostorové znázornění přisuzovaných kvalit (atributů).

Základní typy vzdáleností (eukleidovská a statistická) jsme zmínili již dříve. Tyto základní typy se různě obměňují, takže se mj. používá také tzv. dvounormová vzdálenost (umocněné z-skory), manhattanská (součet všech vzdáleností-jako když se jde okolo bloku domů), Pearsonovská (d=1-r) aj. Výběr vhodné vzdálenosti se řídí typem použitých škál (kategoriálních nebo metrických).

Problém nastane v okamžiku, kdy máme vyjádřit vzdálenost mezi alespoň dvěma shluky, které mají více než jeden člen. Aniž bychom chodili do podrobností, které zmiňuje každý běžný manuál, objasníme si pouze princip na jedné z metod: **nejbližšího souseda (**jinak také nazývaná metoda prosté vazby). Provádí se ve skutečnosti maticovým výpočtem shluků.

Můžeme si zjednodušeně představit 1. krok vytvoření shluků podle nejbližšího středu jako d (C_i C_j)= $d_{ij.}$ a 2. další krok jako d (C_{ij} C_m)=d(Min (C_i C_m) Min (C_j C_m)). Body sobě nejbližší se přeskupí a vytvoří nový shluk. Vzdálenost různých shluků je určována nejdříve vzdáleností dvou nejbližších bodů, objektů z různých shluků A, B a C (viz obr. 6 vlevo). Při použití metody nejbližšího souseda jsou pak objekty taženy k sobě, výsledkem jsou dlouhé řetězy (znázorněno přiblížením shluků a jejich redukcí na A a B: obr. 8 vpravo).



Obr. 8: Schematické znázornění shluků podle bodů, které jsou si nejblíž

Na podobném principu je založeny metody nejvzdálenějšího souseda, dále metoda opírající se o průměr všech vzdáleností, o jejich součet, centroidů nebo mediánová. Každá metoda má své výhody i nevýhody, které zde nebudeme rozebírat, protože jsou zmiňovány v každé učebnici vícerozměrné statistiky (Kelbel a Šilhán, 2002; Hebák et al., 2005; Kahounová, 1994; aj.).

Hierarchická shluková analýza

Hierarchické modely jsou založeny na propojení vzdáleností. Hierarchickou metodou se hledají postupně další a další shluky na základě předtím již určených shluků, které se samy už dál nemění. V tom je síla i slabost této metody. Mj. proto se musí stabilita a spolehlivost výsledků testovat, analýza se musí různým způsobem opakovat, abychom nalezli a potvrdili si podobné výsledky, jejich shodný vzorec. Výhodou hierarchické shlukové analýzy je možnost pracovat také s nominálními a ordinálními daty, pokud zvolíme vhodnou metriku.

Jednou možností je postup odzdola nahoru, aglomerativně, jako když sledujeme rampouchy od konečků směrem ke střeše, jak postupně splývají do jednoho celku. Tento tzv. aglomerativní postup (použitý v SPSS) začíná tak, že každý prvek, každá proměnná nebo respondent tvoří oddělený shluk. Postupně se spojují do větších celků-shluků.

Druhá možnost je opačná, postupuje se odshora dolů, rozděluje, jako když se kmen stromu postupně větví do různých kořenů a kořínků. Algoritmus rozdělování nejdříve pracuje s celkem, se všemi proměnnými nebo respondenty. Ten pak postupně dělí na menší shluky.

V SPSS se **hierarchická shluková analýza** provede takto: zvolíte statistiku, která kvantifikuje vzdálenost nebo podobnost mezi (dvěma) případy. Například eukleidovskou nebo chí-kvadrátovou vzdálenost. Nastavíte grafické výstupy, zpravidla tzv. dendrogram a rampouchový graf (icicle plot). Dále se rozhodnete, kolik shlukových řešení uložit do datového souboru. Můžete zvolit určité rozmezí, řekněme řešení se 2 až 10 shluky. Pokud máte přesnou představu, uložíte jen jedno vámi preferované řešení. Naposledy posoudíte, které shlukové řešení (kolik skupin) nejlépe reprezentuje vaše data. Někdy například postačí k takovému posouzení porovnat frekvence shluků. Posuzuje se přitom, jak jsou si po přidání nebo rozdělení dalšího shluku řešení podobná, jak jsou jejich skupiny početně vyvážené, pokud jde o zastoupení členů s tím, že výrazný nepoměr, např. 50 respondentů ve shluku A a 8 ve shluku B, zpravidla signalizuje, že máte před sebou špatné řešení. Dále se samozřejmě posuzuje smysluplnost nalezených shluků, jejich vztah k výchozí teorii výzkumníka. Oporou může být také dendrogram ukazující blízkost a vzdálenost jednotlivých proměnných.

Hierarchická shluková analýza vyžaduje srovnávání každého případu s každým, což sice přináší obvykle přesnější výsledky než jiné metody (má méně „šumů“ než např. k-means analýza), ale u velkých datových souborů je náročná na čas a přináší nepřehledné výsledky, je tedy prakticky bez užitku.

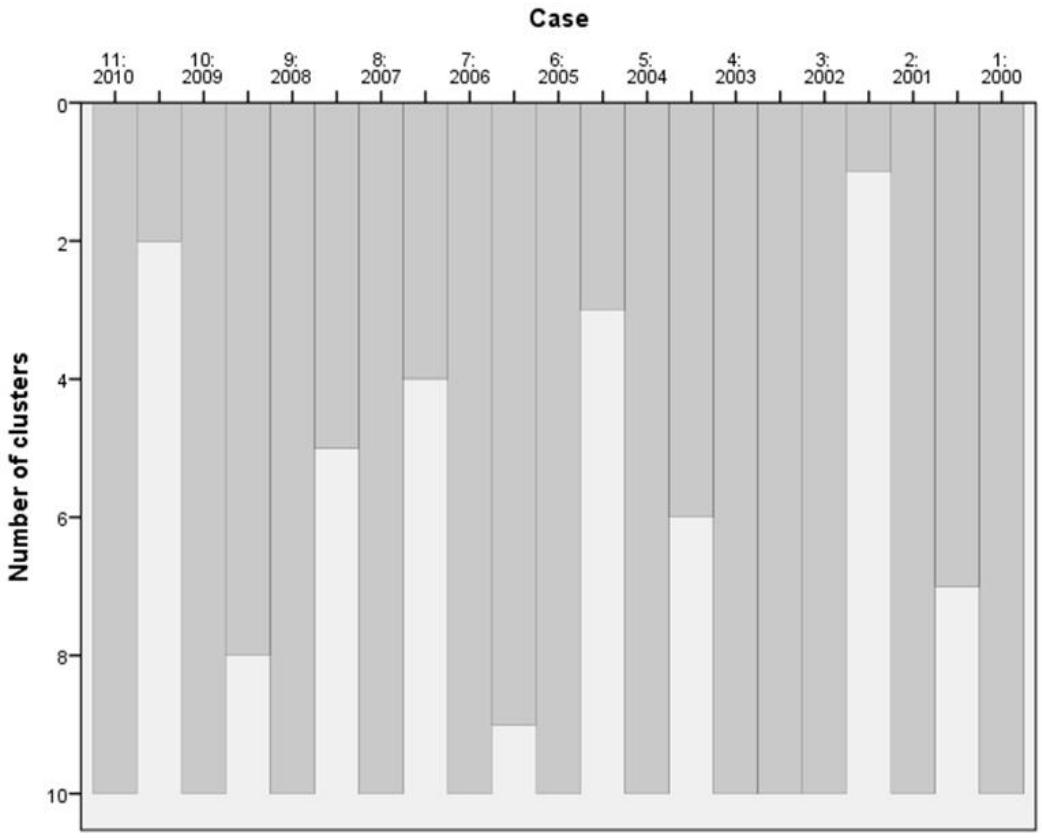
Příklad č. 10: Analýza trendů kriminality za poslední desetiletí (Marešová et al., 2011).
Z uvedené publikace byly čerpány údaje o těchto statistikách (počtech) z období let 2000-2010: tr.č. evidované, tr.č. objasněné, evidovaná majetková tr.č., evidovaná násilná tr.č., osoby stíhané/vyšetřované policií, z toho recidivisté, z toho ženy, z toho muži, z toho dospělí, z toho do 18 let, soudně stíháno, obžalováno a odsouzeno. Tato data jsme museli převést na podobnou stupnici (centrovat s průměrem 0 a směrodatnou odchylkou 1), aby byla srovnatelná. Bez této transformace by převážily nejvyšší statistiky nad nejnižšími, tj. např. počty evidovaných trestných činů nad počty odsouzených a shluky by vznikaly podle počtu evidovaných trestných činů.

Položili jsme si otázku, zda se uvedené jedenáctileté období nečlení do nějakých kratších několikaletých celků, které mají svérázné složení a časovou souvislost, návaznost v letech. Podobnost, pokud se nějaká v datech nachází, může vzniknout z organizačních i legislativních důvodů. Může se odvozovat z rostoucí nebo cyklicky upadající zkušenosti

příslušného policejního sboru (včetně fluktuace); z ekonomických příčin, jako následek migrace osob s kriminálními sklony, ministerských opatření nebo ze změny metodiky evidence. Vysvětlení vyžaduje znalost dalších fakt nebo statistik, kterými zde nedisponujeme, ale v každém případě se zdají být výsledky hierarchické shlukové analýzy smysluplné.

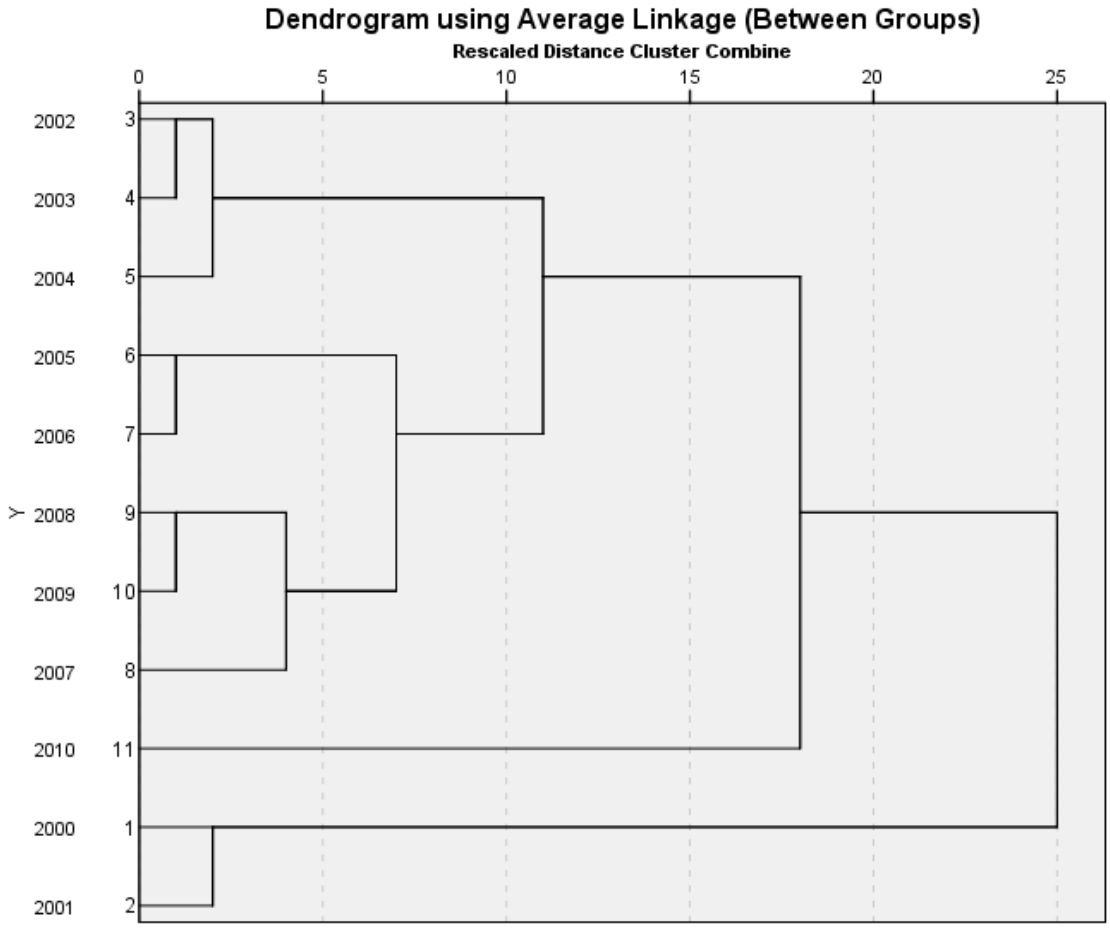
Rampouchový graf (icicle plot) je jeden z možných výstupů této analýzy (Obr. 9). Čte se odspoda nahoru, tj. od největšího počtu shluků a je zároveň podle let seřazen vzestupně odprava doleva. První údaj odspoda (v grafu se nezobrazuje) představuje stav, kdy každý rok (případ) je shlukem sám o sobě. Na takovémto grafu je mj. patrné, proč se hierarchická analýza praktikuje jenom na malých souborech. Kdybychom disponovali výsledky za podstatně delší období, působily by velmi nepřehledně (představte si analýzu za 100 nebo více let). V druhé řádce obr. 8, již na grafu viditelné, je celkem 10 shluků, přičemž léta 2002-2003 jsou sloučena do jednoho shluku. Pokud se podíváme na řádku 8 grafu 9 (z levé strany je označena číslem 8), dospějeme k 8 shlukovému řešení. Zde jsou sloučeny do jednoho shluku roky 2008-2009 a 2005-2006. Samozřejmě také roky 2002-2003, protože jak víme z předešlého výkladu, jednou již založený shluk se dál nerozkládá. Zbylých 5 let stojí samostatně. Pohlcování jednotlivých let směrem vzhůru pokračuje, až se všechna poslučují do jednoho shluku (na úrovni řádky 0).

Z výsledků jsme jako smysluplné vybrali dvou shlukové řešení s jedním odlehlým pozorováním (na grafu 9 je na řádce 3, tj. mezi 2 a 4), protože má oproti jiným řešením nejvyváženější charakter. V pořadí odleva doprava: je to shluk roku 2010-2005, 2004-2001 a 2000 (odlehlé pozorování). Stačí si jen přehodit pořadí letopočtů na obvyklejší: 2000, 2001-2004 a 2005-2010.



Obr. 9: Rampouchový graf z hierarchické shlukové analýzy statistik kriminality 2000-2010

Obr. 10 je zpravidla názornější než rampouchový graf. Je na něm tzv. dendrogram, tj. graficky zobrazená blízkost a vzdálenost jednotlivých proměnných (zde: roků vykazované kriminality) s projekcí do dvourozměrného prostoru a přeškálovanou stupnicí. Nejblíže k sobě mají statistiky kriminality za léta 2002, 2003, popř. 2004, dále 2005 - 2006, 2008 - 2009, popř. ještě 2007 a 2000-2001. Nejdále rok 2000 a ostatní léta s výjimkou 2001. Podíváme-li se na toto rozčlenění podle rampouchového grafu (obr. 8) tak vidíme, že „jádro" shluku 2001-2004 tvoří sobě navzájem nejbližší období 2003-2004 a dále, ve shluku 2005-2010 jsou si nejbližší dvě spojení (2005 - 2006 a 2008 - 2009). Rok 2000 tvoří odlehlé pozorování v obou grafech, ale přechody mezi jednotlivými shluky dendrogram neukazuje tak ostře a jednoznačně jako rampouchový graf- především se jedná o jisté vyčlenění větší distancí od „svých" shluků let 2001 a 2010. V tabulce 13 níže je patrný důvod.



Obr. 10: Dendrogram z hierarchické shlukové analýzy statistik kriminality 2000-2010

Tabulka 13: Průměrné statistiky kriminality podle období (2 shluků let a odlehlého roku) s vyčleněním let 2001 a 2010

	2000	2001-2004 průměr	2005-2010 průměr
tr.č. evidované	391469	**360072**	337985
tr.č. objasněné	172245	**147086**	130171
evidovaná majetková tr.č.	284296	**252346**	219081
evidovaná násilná tr.č.	21996	**22800**	18874
osoby stíhané/vyšetřované policií	130234	**123687**	121625
z toho recidivisté	38664	49055	**54718**
z toho ženy	15483	15189	**16059**
z toho muži	114751	**108498**	105566
z toho dospělí	112430	110230	**113534**
z toho do 18 let	17804	**13456**	8096
soudně stíháno	110808	110080	**111393**
obžalováno	86074	92272	**98135**
odsouzeno	63211	64964	**72039**

Pozn.: zvýrazněná okénka označují roční průměry období 2001 - 2004 a 2005 - 2010. Jsou vždy viditelně vyšší než průměry druhého z obou shluků.

Tabulka 13 nám alespoň přibližně pomůže vysvětlit rozdíly mezi shluky. Rok 2000 je charakteristický vysokou aktivitou policie (nejvíc evidovaných a objasněných trestných činů za sledované 11 leté období), avšak nižším stíháním recidivistů než v následných obdobích, vyšším stíháním mladistvých a vcelku nižší (možná jako důsledek předešlých let) efektivitou soudního výkonu, protože je vykázáno méně obžalovaných a odsouzených.

V období 2001 - 2004 je zaznamenána v průměru o něco slabší aktivita policie než v r. 2000 (s výjimkou evidované násilné trestné činnosti). Jsou ale více postihováni recidivisté a soudy dosahují vyšší efektivity. Období 2001 - 2004 je ve srovnání s obdobím 2005 - 2010 v průměru „efektivnější“ ve sledovaných položkách práce policie, ale liší se strukturou stíhaných osob: stíhá se méně recidivistů a méně žen a více mladistvých než dospělých. Rok 2001 má podle dendrogramu (Obr. 10) velmi blízko k roku 2000 zřejmě z těch důvodů, že se blíží jeho stavu v položkách vykazované recidivy podle pohlaví a věku i soudního stíhání a odsouzení.

Období 2005 - 2010 je charakteristické nejvyšší průměrnou stíhaností recidivistů, dospělých a žen za poslední 11 leté období a také relativně nejvyšší efektivitou práce soudů.

Jistá problematičnost postavení roku 2010 ve zmíněném shluku je dána poklesem evidované a objasněné kriminality.

Tato zjištění je samozřejmě zapotřebí vysvětlit a komentovat z odborného hlediska, doložit případně dokumenty (vývoj zákonů, reorganizace policie) a dalšími statistikami (např. ukazateli stavu, mobility a kvalifikace policejních sil a jejich speciálních útvarů). Náš příklad (9) zde jen posloužil k ilustraci podnětů, jaké může přinést taková analýza jdoucí za hranice běžného studia párových asociací.

Shluková analýza založená na k-means

K-means poprvé použil v r. 1967 James B. MacQueen. (MacQueen je emeritním profesorem na Kalifornské universitě v Berkeley). Princip jeho metody spočívá v tom, že daná (n) pozorování se rozdělují na (k) shluky. Příslušná pozorování (respondenti a jejich skory) patří vždy ke shluku s nejbližším průměrem. Algoritmus rozdělí pozorování do (k-) množin a vyhledává pak takové řetězce skorů (vektory), které mají vždy nejmenší rozdíl od průměru každé k-množiny (nejnižší sumu čtverců). Na začátku se zadají středy co nedál od sebe navzájem. Pak se k nim přiřadí jednotlivá pozorování podle toho, kterému průměru se nejvíc blíží. V druhém kroku se z těchto shluků vypočtou nové středy a znovu se k nim přiřazují ty hodnoty, které jsou nejblíže. Tak se pokračuje, až už žádné pozorování, žádný skor necestuje a najde se finální řešení. Počet (k-) shluků si zadává sám výzkumník. Vzdálenosti jsou v SPSS vyjádřeny eukleidovskou mírou. Ale v jiných softwarech, např. v CCEA (Orme, 2008:11), se pracuje se třemi typy vzdáleností při tvorbě výchozích bodů: eukleidovskou (Distance-based starting points), hierarchickou (Hierarchical-based starting points) a založenou na hustotě bodů (Density-based starting points). Dále se užívá také smíšeného přístupu (Mixed strategy: "osciluje kolem všech metod zadání počátečního bodu"). Navíc lze tímto softwarem také vyloučit odlehlá pozorování nebo centrovat a normalizovat data, což může být důležité při shlukování nesourodých škál (různě oskórovaných, čili různě dlouhých škál nebo proměnných založených na smíšených typech, tj. jak kategoriálních, tak i metrických). SPSS uvedené úkony provádět podobně jako CCEA sice neumí, ale zato dokáže pracovat se soubory s chybějícími údaji po párech. CCEA a mnohé další shlukovací SW používají výlučně jen data bez chybějících údajů, což může

způsobovat značný problém, protože respondenti obvykle ve více otázkách alespoň něco vynechají a tím jsou z analýz vyloučeni a výsledky analýzy zpravidla ztrácí reprezentativitu.

Shlukovací analýza se spolu se statistickou teorií rozhodování, lineární algebrou, korespondenční analýzou a mnoha jinými technikami účastní na úlohách nového vědního odvětví: rozpoznávání vzorců (pattern recognition), tj. vypátrání anebo rozpoznání struktury v datech. Často jde o data založená na sluchových nebo zrakových signálech, např. při vyhledávání podezřelých osob pomocí kamer v davu v metru nebo na letišti. Nebo se jedná o proměnné měřené ve dvou či troj-rozměrném prostoru.

Ze zběžného přehledu odborné literatury vyplývá, že v kriminologii se často shluková analýza používá:

1) při klasifikaci dětí nebo adolescentů do skupin podle problémovosti chování s cílem předvídat jejich budoucí kriminální kariéru;
2) k typologii kriminální činnosti, k nalezení případné kriminální „specializace“ některých pachatelů;
3) k vyhodnocení psychologického profilu souvisejícího s různými typy kriminální činnosti (s použitím klasických psychologických testů a zvláštních testů pro pachatele tr. činů)- často jde o hierarchické metody;
4) ke klasifikaci pachatelů podle jejich životní filosofie a rodinné historie;
5) k mezinárodnímu srovnávání (shlukování) zemí podle násilných a majetkových tr. činů a užívání psychotropních látek- často jde o hierarchické metody.

Příklad č. 11: Evropská studie hodnot a ospravedlnitelnost chování (EVS 2008-2010). Jedna pasáž European Value Survey (EVS) se týká názorů na ospravedlnitelnost různých přestupků nebo kontroversních projevů, počínaje vylákáním neoprávněné státní podpory, šizením na daních až po rozvod, genetickou manipulaci potravin a trest smrti.

Standardizovat není třeba: všechny položky jsou hodnoceny na stejné metrizující stupnici 1-10 (1= nikdy až 10=-vždy ospravedlnitelné). Zajímá nás, zda existují odlišné skupiny lidí se sklonem schvalovat nebo neschvalovat přestupky (jejich určitý typ) a kromě toho jak velké jsou takové skupiny v populaci ČR.

Jak upozorňují odborníci (např. Norušis, 2002), výsledky k-means shlukové analýzy mohou být ovlivněny pořadím pozorování v souboru (pořadím zařazení respondentů). Je to dáno tím, že výběr počátečních shluků určuje další průběh: přepočte se ihned těžiště (průměr) shluku a tím vstupuje do hry původní pořadí případů. Proto je třeba analýzu několikrát opakovat (s náhodnou rotací a s rozdělením souboru na polovinu) než se dojde ke stabilnímu řešení. Při výběru počtu shluků je třeba vycházet z nějaké teorie nebo experimentovat a vylučovat postupně nestabilní řešení, popř. ta řešení, v nichž jsou obsažena vzdálená pozorování („úlety“ s malým počtem členů ve shluku). V příkladech dále prezentujeme už prověřené stabilní výsledky.

V následující tabulce 14, vytvořené k příkladu 11, jsou vlevo znázorněny zvolené středy na počátku analýzy. Např. pro položku „Ospravedlňujete (neoprávněné) požadování státních příspěvků“ algoritmus zvolil jako střed u prvního shluku na škále 1 až 10 skor 1, u druhého shluku skor 10 a u třetího shluku skor 7.

Druhá tabulka 14 (Iteration History) ukazuje, že k finálnímu řešení (konvergenci) bylo zapotřebí 15 kroků (iterací): od prvního „nastřelení“ středů až po poslední 15. řešení, jímž se zjistilo, že se od předchozího významně neliší. Konvergence tedy zároveň znamená, že už mezi shluky k žádným přesunům případů (zde respondentů) nedochází.

Jeden z výstupů dále zobrazuje velikost shluků: byly zařazeny všechny případy s výjimkou 12 respondentů, kteří baterii zcela vynechali. Procenta zastoupení jednotlivých shluků uvádíme dále v závorce za každým shlukem, získáme je v SPSS jednoduše frekvencí shlukové proměnné.

Tabulka 14: Ospravedlnitelnost chování podle EVS: počáteční středy shluků a průběh iterací

Počáteční středy shluků			
Je ospravedlnitelné:	Shluk		
	1	2	3
Q68A|nedůvodné nárokování státních příspěvků	1	10	7
Q68B|šizení na daních	1	10	5
Q68C|někomu z legrace odjet jeho autem	1	10	8
Q68D|užívání měkkých drog	1	10	6
Q68E|lhaní ve vlastním zájmu	1	10	9
Q68F|nevěra	1	1	10
Q68G|přijímání úplatku	1	1	10
Q68H|homosexualita	10	1	9

Iterační historie			
Iterace	Změna ve středech shluků		
	1	2	3
1	12,744	15,336	11,528
2	1,045	3,703	2,399
3	1,266	1,717	0,892
4	0,487	0,535	0,168
5	0,294	0,235	0,169
6	0,152	0,067	0,173
7	0,126	0,067	0,127
8	0,079	0,045	0,075
9	0,057	0,035	0,056
10	0,032	0,035	0,000
11	0,041	0,037	0,021
12	0,021	0,000	0,031
13	0,026	0,000	0,050
14	0,016	0,000	0,029
15	0,000	0,000	0,000

Tabulka 15 nazvaná v SPSS „Konečné středy shluků“ ukazuje rozložení středů podle skorů respondentů na shlukovaných proměnných. Zvolit stručné a výstižné názvy jednotlivých shluků je někdy velmi obtížné (doporučuje se překonat týmovou prací).

Např. shluk 1 jsme nazvali **moderní tolerantní** (40 %) podle toho, že je charakteristický vysokým ospravedlňováním (tolerancí) homosexuality, potratů, rozvodů a euthanasie na jedné straně (v průměru vysoké známky znamenající toleranci: 6-7) a větší či

slabší intolerancí vůči finančním, daňovým přestupkům, drogám a dalším anti společenským jevům na straně druhé (v průměru nízké známky mezi 2-3 znamenající spíše netoleranci).

Tabulka 15: Ospravedlnitelnost chování podle EVS: konečné středy shluků

Konečné středy shluků			
Je ospravedlnitelné:	Shluk		
	1	2	3
Q68A|nedůvodné nárokování státních příspěvků	2	2	5
Q68B|šizení na daních	2	2	5
Q68C|někomu z legrace odjet jeho autem	2	2	4
Q68D|užívání měkkých drog	2	1	5
Q68E|lhaní ve vlastním zájmu	3	2	6
Q68F|nevěra	3	2	6
Q68G|přijímání úplatku	2	2	5
Q68H|homosexualita	6	2	7
Q68I|potrat	7	3	7
Q68J|rozvod	7	3	7
Q68K|euthanasie	7	3	6
Q68L|sebevražda	4	2	6
Q68M|hotovostní platby bez dokladu a daní	3	2	6
Q68N|provozování příležitostného sexu	3	2	7
Q68O|černé pasažerství	3	2	6
Q68P|prostituce	3	2	6
Q68Q|experimenty s lidskými embryji	3	2	5
Q68R| genetická manipulace potravin	3	2	5

Shluk 2, **nesmiřitelní** (38 %), je charakteristický malou tolerancí vůči všem i tzv. menším přestupkům (středy se známkami 2, občas 3). Můžeme ho také chápat jako moralistický nebo bigotní.

Shluk 3, **amorální** (22 %), je poměrně názorově nejvíc tolerantní, protože vysoce či značně toleruje nejenom homosexualitu, potraty, rozvody a euthanasii (průměrné známky 6-7), ale také finanční a daňové přestupky a trestné činy (známka 5), lež, nevěru a další drobné přestupky (známky 6-7) apod. Tato skupina je potenciálně deviantní.

K těmto výsledkům bychom bez vícerozměrné analýzy nedospěli. Poznatky o rozložení shluků nám pak za splnění jistých dalších předpokladů (ověření, že stupnice, jejich metrika i konstelace ve všech zemích fungují stejně) umožňují porovnávat různé země.

Můžeme tak např. dospět k různé velikosti vůči kriminalitě tolerantních a netolerantních skupin mentalit v mezinárodních podmínkách a k vyhodnocení států s nejvyšším procentem kriminalitu netolerujících obyvatel. Citovaný EVS totiž pokrývá reprezentativně 46 evropských zemí. Můžeme uvažovat o dalších faktorech, které působí na vznik a dynamiku těchto skupin uvnitř jedné země. Mezinárodní charakter zjištění také umožňuje přejít na různé objektivizované statistiky za jednotlivé země a vztáhnout je k výskytu shlukovaných mentalit v jednotlivých státech. Z ověřených zdrojů Eurostatu, World Bank nebo Amnesty International načerpáme např. údaje o hrubém národním produktu, svobodě vyjadřování, religiozitě, očekávání délky života po narození atd. a ty korelujeme s výsledky výběrového šetření při využití vícerozměrných statistik jako je SEM nebo hierarchické (vícerozměrné) regresní analýzy.

Poslední Obr. 11 našeho příkladu č. 11 ukazuje 1. a 2.shluk (moderní tolerantní a nesmiřitelní) jsou si navzájem blíž (8,6), kdežto 3. shluk („amorální“) je jim vzdálenější, zvláště pak od shluku „nesmiřitelných“ (10,2 a 15,9).

Distances between Final Cluster Centers

Cluster	1	2	3
1		8,651	10,252
2	8,651		15,964
3	10,252	15,964	

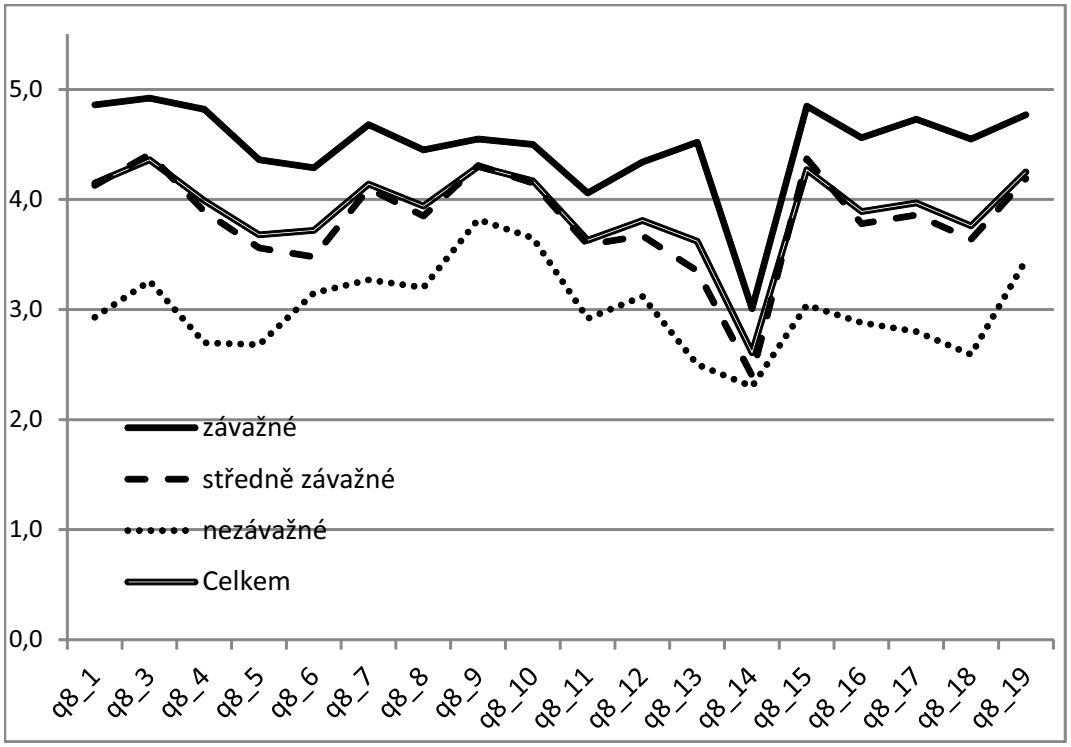
Obr. 11: Vzdálenosti shluků

Vrátíme se nyní k **příkladu č. 8. Názory mladých lidí v ČR na stupeň problematičnosti některých negativních jevů** (Večerka et al., 2011). V tomto výzkumu respondenti (ve věku 15-24 let) hodnotili na stupnici 1=malý, zanedbatelný problém až 5=velký, zásadní problém 19 různých negativních společenských jevů jako je týrání členů rodiny, kyberšikana, nefunkční policie, fetování a alkoholismus, nezaměstnanost, korupce, terorismus, kriminalita ap. Také zde nebylo třeba data při pokusu o shlukovou analýzu normalizovat, protože všechny skory byly dosaženy na stejné stupnici.

Z předešlých analýz jsme předpokládali, že vzniknou skupiny kolem problémů, které mladí lidé rozporně hodnotí. K těmto problémům patřily agresivita navenek, domácí násilí, civilizačně-ekonomické dysfunkce a užívání psychotropních látek. To se ovšem nepotvrdilo a

výsledky shlukové analýzy byly mnohem jednodušší. Jedna skupina, tzv. seriozní, hodnotila všechny negativní jevy jako velký, závažný problém, druhá, kompromisní, jako střední problém – kolem celkového průměru a třetí, zlehčující, jako malý, nezávažný problém vcelku bez ohledu na posuzovaný typ chování. Skupiny se sice poněkud lišily pořadím priorit: zvláště pak zlehčující skupina měla na prvním místě problematičnosti nezaměstnanost, dále alkoholismus a fetování a týrání dětí, kdežto obě zbývající skupiny dětskou pornografii, týrání dětí a jiných členů rodiny. Ale všeobecně rozlišujícím kritériem byla intenzita, s jakou byla danému problému přisouzena závažnost.

Potencionálně nebezpečná může být skupina, která negativní společenské jevy zlehčuje, tj. podceňuje nebo snižuje jejich problematičnost, tedy asi pětina cílové skupiny 15-24 let. Z dalších údajů vyplývá, že 4 z 10 respondentů v této skupině už neoprávněně řídilo motorové vozidlo. Relativně nejvíc právě tato skupina dostávala na základní škole z chování známky 2-3. U této skupiny se dostavuje potenciálně tzv. „testosteronový“ efekt (jde většinou o muže považující manželství za přežitek). Navíc značná část se cítí sociálně-ekonomicky deprivovaná, tj. pochází podle jejich vlastního odhadu z domácností v podprůměrné majetkové situaci a nemá sociální cítění (považuje staré lidi za ekonomické břemeno).



Obr. 12: Průměry hodnocení závažnosti společenských problémů podle shluků

Legenda k Obr. 12: q8_1 Týrání žen v rodinách; q8_3 Týrání dětí v rodinách; q8_4 Týrání starých lidí; q8_5 Kyberšikana; q8_6 Nefunkční policie; q8_7 Fetování; q8_8 Alkoholismus; q8_9 Nezaměstnanost; q8_10 Korupce a podplácení; q8_11 Rozpad rodin; q8_12 Poškozování přírody; q8_13 Terorismus; q8_14 Počítačová negramotnost; q8_15 Dětské pornografie; q8_16 Agresivita vztazích; q8_17 Nebezpečí nakažení se AIDS; q8_18 Zlovolné pronásledování jiné osoby; q8_19 Kriminalita

Ve stejném výzkumu autoři (Večerka et al., 2011) zkoumali také kriminální citlivost. Analyzovali jsme vztahy ke kriminální citlivosti, zahrnující 43 položek od hrubého a výtržnického chování, přes pomluvu, surovost, vandalství, korupci, lichvu a podvádění, sexuální uvolněnost, prostituci až po nelegální stahování autorsky chráněných výtvorů. Skupina podceňující společenské problémy měla ve všech položkách nejnižší citlivost.

Další nehierarchické metody

Metoda k-means může přinést podobně jako její předchůdci (např. Q-metoda faktorové analýzy) velmi přesvědčivé výsledky, pokud je použita s obezřetností, tj. „rozbije se“ její závislost na pořadí pozorování v datovém souboru variantním přístupem, různými zkouškami stability a pokud možno také reliability (nezávislým přezkoumáním). Je vhodná k rychlému vytřídění velkého množství případů, pokud zároveň předpokládáme, že shluky se nebudou navzájem překrývat.

PAM (Partitioning around medoids) je další metoda nehierarchické shlukové analýzy. Vychází z minimalizace nepodobnosti (dissimilarities). Nejdříve se vypočtou tzv. medoidy (Kaufman & Rousseeuw, 1990). Jsou to typičtí představitelé hledané struktury. Dále se určí střed shluku a ten je vždy vyjádřen nějakým konkrétním případem. Všechny případy v daném shluku jsou si minimálně nepodobné, mají ke svému středu minimální vzdálenost.

Právě to, že se při PAM minimalizují nepodobnosti (místo sumy čtverců eukleidovských vzdáleností jako u k-středového shlukování) propůjčuje této metodě větší přesnost, protože není zkreslena vzdálenými pozorováními nebo šumem. Je ovšem vhodná jenom pro malé výběrové soubory. Algoritmus PAM je nadprogramový pro R-jazyk a je tudíž zdarma přístupný komukoliv, kdo jazyk R ovládá. K dispozici je také v řadě komerčních softwarů (jako NCSS).

MDP (Multivariate divisive partitionning)- tzv. klasifikační a regresní strom. Analytik zvolí závislou proměnnou, např. počet násilných trestných činů a přidává postupně potenciálně ovlivňující proměnné, např. rodinné zázemí, skory iritability až zjistí, které proměnné nejlépe rozdělují výchozí skupinu a na kolik dalších skupin. Tato technika je zvlášť vhodná k analýze velkých datových souborů. Vyvinula ji forma Claritas pro USA a pod názvem PRIZM ji provozuje Nielsen. V SPSS je podobná procedura nazvaná strom (TREE), která nabízí tři shlukovací a klasifikační metody včetně CRT (Classification and Regression Tree). Ze zadaných hypoteticky ovlivňujících proměnných tato procedura sama vybere jen ty skutečně významné.

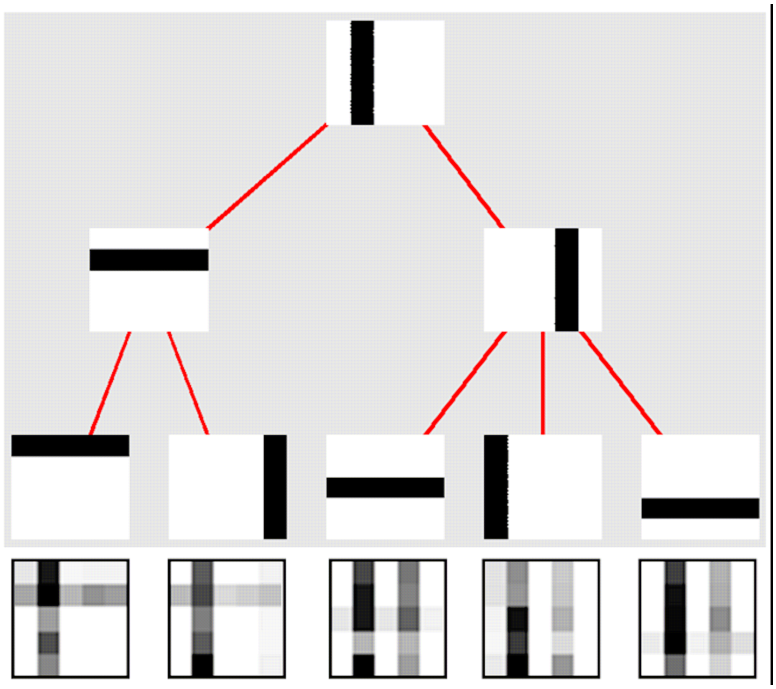
Jedním z omezení shlukové analýzy k-means je charakter vstupních dat: musí být alespoň ordinální nebo dichotomická, ideálně intervalová nebo spojitá. Toto i další omezení se pokouší překonat metoda **Maximalizace-očekávání (EM:** expectation-maximalization). Je založena na výpočtu pravděpodobnosti členství ve finálních (tedy latentních) shlucích podle logaritmické pravděpodobnosti rozložení dat.

EM připouští také použití kategoriálních dat. Nejdříve se kategoriálním proměnným (každé kategorii) náhodně přiřadí různá pravděpodobnost (váha) pro každý shluk. Pak se v postupných krocích maximalizuje pravděpodobnost členství ve shlucích. Algoritmus EM vysvětlili a popsali A.Dempster, N.Laird a D. Rubin (1977). Tato metoda není v SPSS k dispozici, ale například se nabízí v programu Statistica, v řadě komerčních softwarů nebo zdarma v balíčcích pro jazyk R.

Dvou kroková shluková analýza

Dvou kroková shluková analýza (two step cluster analysis) je zvláště vhodná pro velké výběry, kdy předem nepředpokládáte určitý počet shluků a máte k dispozici metrická nebo kategoriální data nebo obojí. Data mají být na sobě nezávislá (otestuje se např. korelacemi, kontingenčními tabulkami a příslušnými testy, analýzou variance). Metrické proměnné by měly být normální (ukáže procedura explore). Kategoriální data by měla mít kolísavý trend, tj. představíme-li si křivku spojující vrcholy (četnosti) jednotlivých kategorií, neměla by mít tvar přímky, ale křivky s jedním nebo více vrcholy. Nezávislost kategoriálních párů se otestuje CHI^2.

Nejdříve se v prvním kroku vytvoří předběžné shluky, tj. případ po případu se zařazuje tzv. strom shlukových rysů (CF tree, tzv. Clustering Feature, Obr. 13). Ten má dva parametry: větvící faktor (na obrázku černý proužek) a práh (jeho umístění v okénku).



Obr. 13: Příklad CF tree (Sivic et al., 2008)

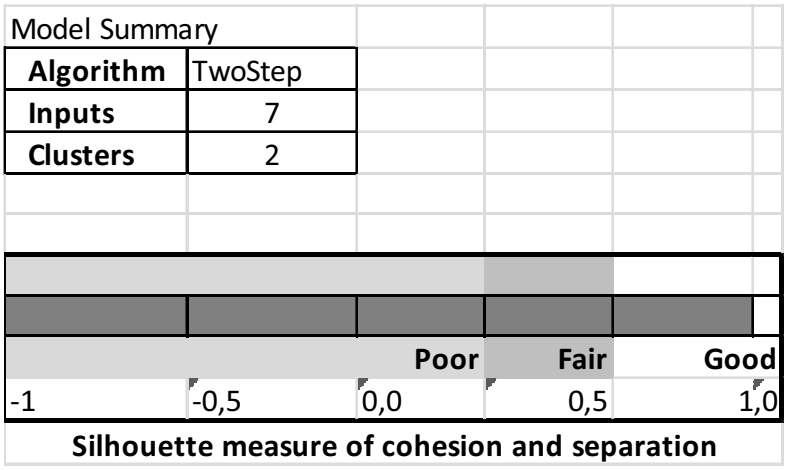
Algoritmus pak tyto předběžné shluky v druhém kroku znovu shlukuje hierarchickou metodou (na obrázku 5 oddělených okének dole). Podmínka, že data by měla být rozložena normálně nebo multinormálně (u metrických nebo jejich kombinace s kategoriálními) nebo mít tzv. polynomický trend a být nezávislá (u kategoriálních dat) není striktně vyžadována. Není tedy bezpodmínečně nutné testovat předpoklady, ale je to žádoucí, protože procedura dosahuje lepší výsledky, pokud jsou zmíněné předpoklady splněny.

U většiny typů shlukové analýzy testujeme stabilitu (někdy výsledky závisí na seřazení případů ve výběrovém souboru). Doporučuje se soubor rozdělit na polovinu a v ní analýzu zopakovat. Zkorelované výsledky pak vybrat jako spolehlivé.

Zpět k **příkladu 4 – Zkušenost české veřejnosti s vyjmenovanými psychotropními látkami** (Zeman et al., 2011a). V tomto výzkumu reprezentujícím českou veřejnost od 15 let výše se mj. zjišťovala zkušenost veřejnosti s měkkými a tvrdými drogami. Použita byla stupnice užívání: 1=ano, v posledních 30 dnech, 2=ano v posledních 12 měsících, 3 naposledy před více než 12 měsíci, 4=nikdy.

Kvalita shluků podle dvoustupňové shlukové analýzy, kterou jsme prováděli v SPSS, byla naměřena jako 100 % (tzv. Silhouette measure of cohesion and separation byl plně v poli „Good“- Obr. 14).

Dále, dvoustupňová shluková analýza v SPSS nabízí výstupy – tabulky i grafy – interaktivně a tak musí výzkumník dobře s výstupem pracovat, přepínat různé volby zobrazení. Navíc může také interaktivně měnit shlukující proměnné podle pruhového grafu, který ukazuje, jestli mají dostatečně silný vliv na diferenciaci shluků.



Obr. 14: Výstup dvoustupňové shlukové analýzy: potvrzení kvality shlukování

V následující tabulce 16 jsou popsány 3 shluky, vzniklé ve dvoustupňové shlukové analýze z citovaného příkladu 8. První shluk tvoří lidé drogy vůbec (nikdy) neužívající. U každého typu drogy mají proto v průměru čistých 4 (=nikdy) a to bez odchylek (Std. Odchylka = 0). Další, druhý shluk tvoří dotázaní, kteří už někdy déle než před rokem užili marihuanu (čistý skor 3) a žádné další drogy (vždy 4). Poslední shluk tvoří uživatelé drog, kteří převážně už letos užili marihuanu a k tomu ještě nějakou další drogu (nejčastěji extázi nebo halucinogenní houby: 3,37 a 3,57).

Tabulka 16: Shluky podle užívání drog: výsledky dvou krokové analýzy

q17c Vyzkoušel/a jste někdy v životě ... kdy to bylo naposledy?	Statistika	Shluk			Celkem
		1 neužívá, nikdy nevyzkoušel(a)	2 kdysi zkusil(a) marihuanu	3 užívá marihuanu a občas i jné drogy	
marihuanu či hašiš	Průměr	4,00	3,00	2,30	3,67
	Směr.odchylka	0,00	0,00	1,06	0,73
extázi	Průměr	4,00	4,00	3,38	3,91
	Směr.odchylka	0,00	0,00	0,83	0,38
pervitin, amfetaminy	Průměr	4,00	4,00	3,73	3,96
	Směr.odchylka	0,00	0,00	0,60	0,24
kokain	Průměr	4,00	4,00	3,86	3,98
	Směr.odchylka	0,00	0,00	0,47	0,18
heroin	Průměr	4,00	4,00	3,89	3,98
	Směr.odchylka	0,00	0,00	0,45	0,17
LSD, "krystal", "trip", "papír"	Průměr	4,00	4,00	3,80	3,97
	Směr.odchylka	0,00	0,00	0,51	0,20
halucinogenní houby	Průměr	4,00	4,00	3,57	3,94
	Směr.odchylka	0,00	0,00	0,67	0,29

Podle výsledků asi ¾ veřejnosti není a nikdy nebylo uživatelem drog, asi 1 z deseti užil někdy měkké drogy (marihuanu) a asi 14 % je nedávným uživatelem marihuany a také zároveň uživatelem (nedávným nebo s dávno minulou zkušeností) dalších drog.

Tabák a alkohol jsou sice mnohem rozšířenějšími psychotropními látkami, ovšem nezahrnuli jsme je do shlukové analýzy, protože ta by pak ztratila rozlišovací schopnost. Nicméně konzumace tabákových výrobků a požívání alkoholu jsou rovněž odstupňovány od nejnižší úrovně ve skupině drogami nezatížené až po nejvyšší u širších uživatelů drog. Vysoce zkorelované (ve stejném směru) s těmito typy jsou zkušenosti respondentů s uživateli a prodejci drog v nejbližším okolí bydliště.

Analýza pomocí rozhodovacích stromů (procedura Tree)

Rozhodovací stromy patří ke statistickým postupům tzv. dolování dat, při nichž se využívají grafické metody zobrazování vztahů (stromový diagram, Tree diagram). Jejich podstatou je redukce nesourodé množiny skorů/záznamů „na menší počet sourodých skupin s využitím zaměřeného objevování poznatků“ (Al Ghoson, 2010: 57). Toto poznávání je cílově zaměřené: hledá pro danou konstelaci, např. pro uživatele drog, statisticky významné konstelace dalších proměnných, např. věku, pohlaví a informovanosti o boji s drogami. Pokud takové významné konstelace za pomoci Rozhodovacích stromů (Tree) najdeme, můžeme na jejich základě předvídat jevy, např. šikanu (Brewer, 2010), násilnou trestnou činnost a bankovní loupeže (Gerristsen et al., 2012), pojišťovací podvody (Gepp et al., 2012), selhání při výchově dětí (Barlow et al., 2012), znásilnění (Goodwill, 2007) a mnohé jiné. Rozhodovací stromy jsou součástí nově se formujícího odvětví tzv. výpočetní kriminologie (EISIC, 2011).

Pokud jsme rozhodovací stromy vytvořili, můžeme na základě jejich pravidel předvídat také z dat, ve kterých zkoumaná cílová konstelace chybí, např. z věku a pohlaví (a dalších proměnných) lze usuzovat na drogovou závislost, i když tento údaj samotný v datech chybí. Při užití Rozhodovacího stromu se hledá nejlepší strategie, jak vysvětlit nebo dospět k cílové konstelaci. Může se tedy jednat o klasifikaci, založenou na kategoriálních proměnných, např. zjišťujeme, co přispívá k zařazení do cílové kategorie drogově závislých. Nebo jde o tzv. regresní strom, kdy se na základě metrických proměnných, popř. jejich kombinace s kategoriálními proměnnými, předpovídá nějaké konkrétní číslo, např. počty recidivistů nebo pachatelů ekonomické kriminality.

V 80. letech minulého století uvedli Breiman et al. (1984) poprvé pro tyto techniky souhrnný název „klasifikační a regresní strom“. Dále se zaměříme v příkladu jen na jednu metodu růstu stromu, tzv. CHAID (CHI-square Automated Interactive Detecting), který jako jednu ze tří metod SPSDS nabízí (kromě toho ještě CRT=klasifikační a regresní strom a QUEST=rychlý účinný nezkreslený statistický strom pro nominální data).

Vrátíme se nyní k dobře popsanému **příkladu 4 – Zkušenost české veřejnosti s vyjmenovanými psychotropními látkami** (Zeman et al., 2011a): můžeme si mj. položit

otázku, zda existuje mezi někdejšími a nynějšími uživateli drog nějaký podstatný významný demografický rozdíl nebo rozdíl ve stupni informovanosti o protiopatřeních v oblasti drog a pokud takové rozdíly existují, tak jak se projevují.

Závislou proměnnou v našem příkladu jsou typy uživatelů drog (N=457) a prediktory mohou být podle našeho výchozího očekávání pohlaví, věk, vzdělání, sociálně-ekonomické postavení, výše příjmu, velikost místa bydliště a informovanost o protiopatřeních v oblasti drogové problematiky (v horní části Obr. 15 uvedeny jako závislé a nezávislé proměnné). Umíme si na základě dřívějších výzkumů a logiky představit, že spíše muži a mladí lidé, dále méně vzdělaní a informovaní o účincích drog, pocházející spíše z míst s menší sociální kontrolou (větší města s větší anonymitou než jakou poskytuje venkovské a maloměstské prostředí), hůře ekonomicky situovaní, vykonavatelé méně kvalifikované práce mají častěji sklony k užívání drog. Naproti tomu spíše ještě mladší muži i ženy, vzdělanější, lépe situovaní atd. mají tendenci k dočasnému experimentování s marihuanou.

SPSS pomocí procedury Tree vybere jen ty nezávislé proměnné, které podstatně ovlivňují diferenciaci respondentů, a tedy podle našeho příkladu rozliší uživatele drog na dřívější s dávnou zkušeností s marihuanou a nynější závislejší, užívající více drog než jen marihuanu. Tyto vybrané „strategické“ proměnné jsou uvedeny na Obr. 15 dole v řádcích výsledky (results): pohlaví a věk. Tedy už v prvním výstupu vidíme, že informovanost o protiopatřeních v oblasti drogové problematiky a řada dalších proměnných podstatný „štěpící“ vliv prokazatelně nemá.

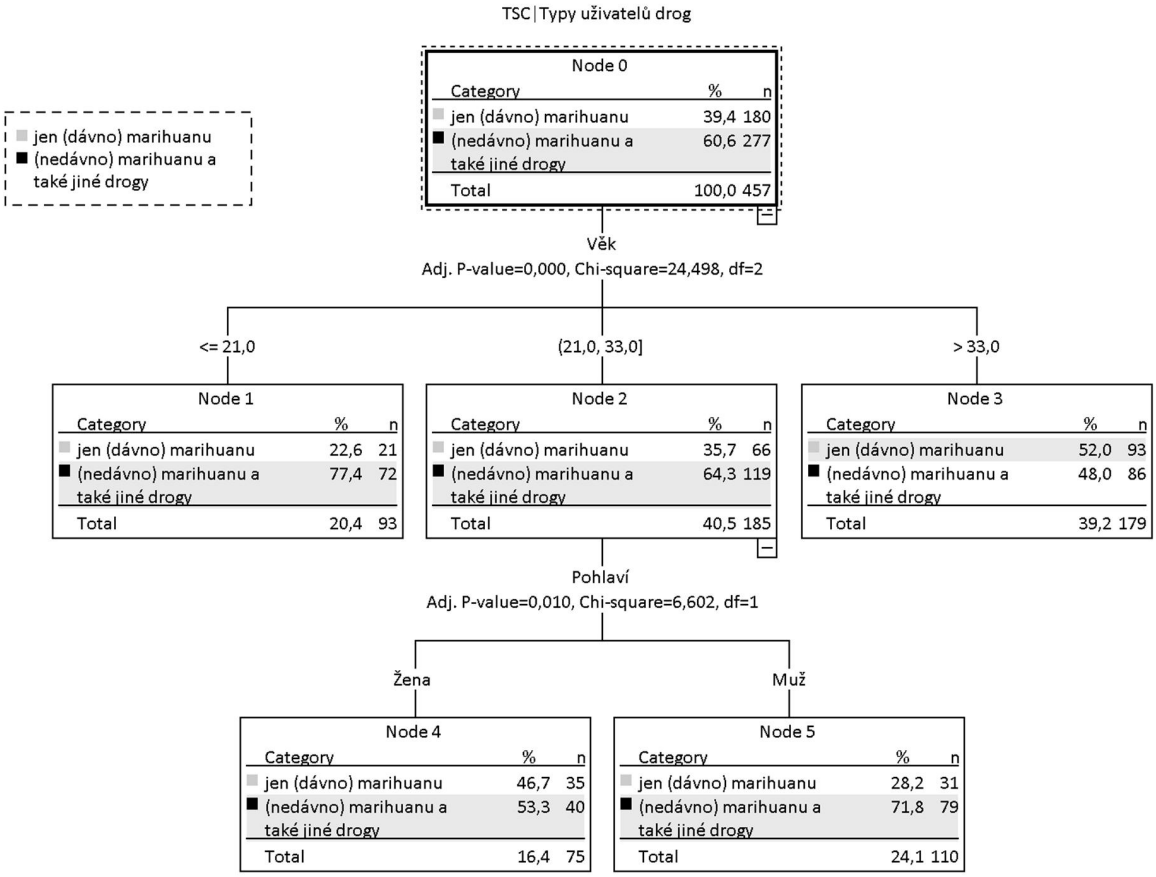
Tabulka na Obr. 15 také obsahuje popisné informace o použité metodě růstu stromu (CHAID) a celkové hloubce stromu (3 stupně), počtu kořenových uzlů celkem (6) a konečných (4) a jejich hloubce (2).

Model Summary		
Specifications	Growing Method	CHAID
	Dependent Variable	TSC_2 TSC|Typy uživatelů drog
	Independent Variables	d1 Pohlaví, d2 Věk, d3x Vzdělání, d5 Socio-ekonomické postavení, d8 Hrubé příjmy Vaší domácnosti, d4 Velikost místa bydliště respondenta, q1_si Q1_3I|Testovaná informovanost o protiopatřeních v oblasti drog
	Validation	None
	Maximum Tree Depth	3
	Minimum Cases in Parent Node	100
	Minimum Cases in Child Node	50
Results	Independent Variables Included	d2 Věk, d1 Pohlaví
	Number of Nodes	6
	Number of Terminal Nodes	4
	Depth	2

Obr. 15: Modelové shrnutí k uživatelům drog

Stromový diagram (Obr. 16) ukazuje, že v našem výběru připadají na 2 někdejší uživatele marihuany 3 aktuální uživatelé také dalších drog. Následuje níže první rozlišující úroveň, která je dána věkem.

Uživatelé se rozkládají na 3 věkové skupiny: do 21 let, 21 - 33 let a více než 33 let. Ve skupině do 21 let, která je už koncovým „kořenovým uzlem“ (node), převažují aktuální uživatelé drog (77 %) a to bez rozdílu pohlaví. Ve skupině nad 33 let jsou oba typy zhruba v rovnováze: ti co přestali užívat marihuanu i aktuální uživatelé dalších drog. Ve skupině 21-33 let mírně převažují (64 %) aktuální uživatelé. Tato skupina se ovšem dále štěpí pod vlivem pohlaví: na mužskou část, kde aktuální uživatelé drog dominují (72 %) a ženskou část, kde jsou zhruba v rovnováze s „experimentujícími“ někdejšími uživatelkami marihuany. V souhrnu tedy z hlediska užívání drog a drogové závislosti nejrizikovější je věk do 21 let, mezi 21 - 33 lety je výrazně rizikovější pro muže než pro ženy a po 33 letech jsou šance přestat s drogami nebo v užívání pokračovat zhruba vyrovnané a jak pro muže, tak i pro ženy se neliší.



Obr. 16: Stromový diagram popisující uživatele drog

Tree Table

Node	2 jen (dávno) N	2 jen (dávno) Percent	3 (nedávno) N	3 (nedávno) Percent	Total N	Total Percent	Predicted Category	Parent Node	Primary Independent Variable: Variable	Sig.[a]	Chi-Square	df	Split Values
0	180	39,4%	277	60,6%	457	100,0%	3 (nedávno) marihuanu a také jiné drogy						
1	21	22,6%	72	77,4%	93	20,4%	3 (nedávno) marihuanu a také jiné drogy	0	d2 Věk	,000	24,498	2	<= 21,0
2	66	35,7%	119	64,3%	185	40,5%	3 (nedávno) marihuanu a také jiné drogy	0	d2 Věk	,000	24,498	2	(21,0, 33,0]
3	93	52,0%	86	48,0%	179	39,2%	2 jen (dávno) marihuanu	0	d2 Věk	,000	24,498	2	> 33,0
4	35	46,7%	40	53,3%	75	16,4%	3 (nedávno) marihuanu a také jiné drogy	2	d1 Pohlaví	,010	6,602	1	Žena
5	31	28,2%	79	71,8%	110	24,1%	3 (nedávno) marihuanu a také jiné drogy	2	d1 Pohlaví	,010	6,602	1	Muž

Growing Method: CHAID

a. Bonferroni adjusted

Obr. 17: Stromová tabulka (včetně 6 koncových kořenových uzlů)

Výchozí hypotézy o vlivu příjmu, informovanosti o drogách, sociálně-ekonomického postavení, velikosti bydliště apod. na užívání drog jsme neměli možnost v plném rozsahu testovat, protože jsme do analýzy nemohli zahrnout neuživatele drog. Tvoří ohromnou většinu výběrového souboru a strhávají na sebe veškerou varianci, což znemožňuje identifikovat rozdíly vzhledem k uživatelům. Zkusili jsme proto vybrat náhodně z neuživatelů přiměřeně velkou skupinu (N=431) a porovnat ji s aktuálními uživateli drog (N=277). Celkem jsme získali pro kořenový diagram více skupin než v předchozím případě (10 oproti 6) a také koncových „kořenových uzlů" bylo víc (6 oproti 4).

Nicméně ze všech vkládaných nezávislých proměnných se opět jako v předchozí analýze prosadily pouze pohlaví a věk. Tabulka na Obr. 17 podrobně charakterizuje každou skupinu stromového diagramu (ten samotný zde již neuvádíme): počínaje od nulté s nasazením zhruba 3 neuživatelů (61 %) na 2 aktuální uživatele (39 %) drog a konče 9. uzlem, který je koncový, a v němž převládají neuživatelé (97 %), ženy ve věku od více než 52 let.

Ke koncovým skupinám patří také uzly 4 a 5 pro respondenty do 33 let, 6 a 7 pro respondenty ve věku 33-52 let a 8-9 pro respondenty nad 52 let. Část tabulky na Obr. 17 se záhlavím Primární nezávislá proměnná informuje o tom, která proměnná bezprostředně skupinu štěpí. U všech koncových uzlů je to pohlaví, u těch v pořadí blíže kořeni (1 až 3) je to věk. Všechna věková štěpení jsou velmi významná (alfa<.001), kdežto štěpení podle pohlaví jsou nepatrně za okrajem významnosti (které je alfa=.05). Přesto se je pokusíme interpretovat a to jako možnou hypotézu pro další výzkum. Ve věkové skupině do 33 let (uzel 4-5) je drogové riziko obecně nejvyšší a to častěji pro muže (75 %) než pro ženy (60 %). Ve skupině 33-52 let (uzel 6-7) převažují neuživatelé, opět častěji mezi ženami (76 %) než muži (59 %). Ve skupině nad 52 let (uzel 8-9) neužívání naprosto převažuje, přestože o něco víc mezi ženami (97 %) než muži (87 %). Opět se tedy potvrdilo, že nejrizikovější období představuje věk do 33 let. Navíc odtud víme, že po 52 letech věku se riziko výrazně snižuje, přestože rozdíl mezi muži a ženami se projevuje i nadále.

Procedura Tree poskytuje na výstupu také možnost posoudit, jak moc se model v porovnání s daty osvědčil. Riziko chyby je v našem případě (Tabulka na Obr. 18) celkově skoro čtvrtinové, tj. ze 4 odhadů podle věku a pohlaví se model osvědčuje ve 3 případech a v jednom ne (75,3 %).

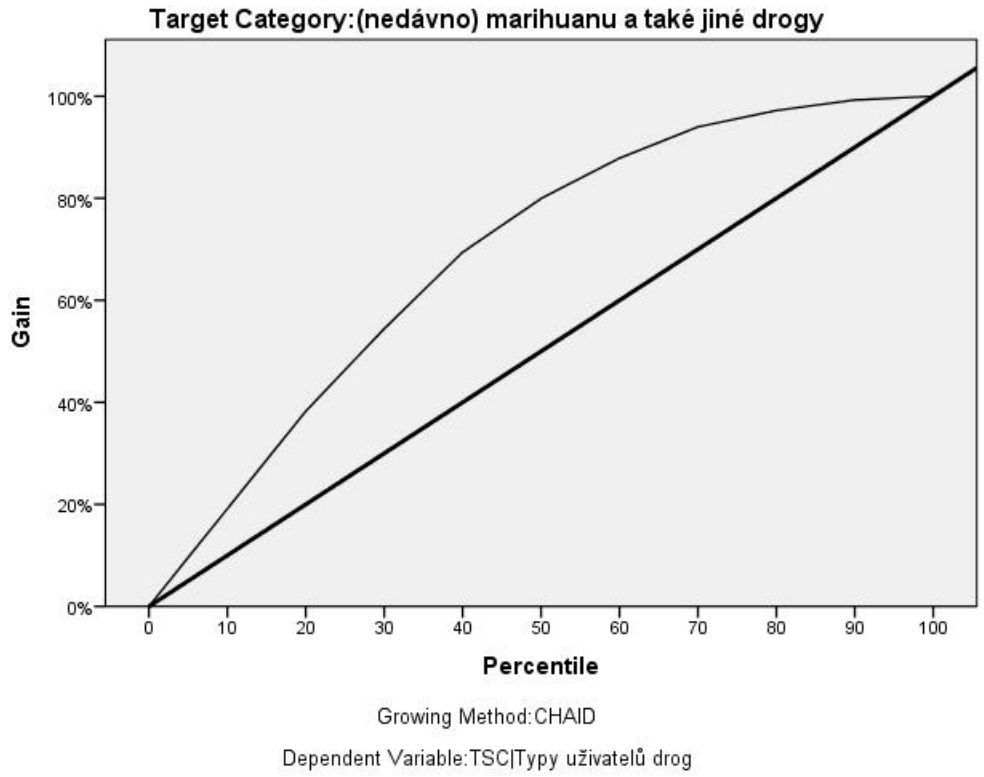
Klasifikace

Observed	Očekávaná		
	1 není uživatel	3 (nedávno) marihuanu a také jiné drogy	Procento správně
1 není uživatel	1191	350	77,3%
3 (nedávno) marihuanu a také jiné drogy	86	191	69,0%
Celkové procento	70,2%	29,8%	76,0%

Metoda růstu CHAID
Závislá proměnná: TSC_2 TSC|Typy uživatelů drog

Obr. 18: Klasifikace

Pro úplnost lze uvést, že dalším kritériem úspěšnosti analýzy je ziskový graf (Obr. 19), týkající se shody modelu s daty u cílové kategorie závislé proměnné (zde šlo o uživatele drog, které jsme chtěli blíže specifikovat). Zisk je počet případů, které připadají v dané skupině (uzlu) na cílovou kategorii. Křivka tedy znázorňuje kumulativní růst procenta uživatelů drog v závislosti na rostoucím počtu respondentů, začíná vždy na nule a končí 100 procenty. Jako v našem příkladu, má shoda modelu s daty po decilech prudce stoupat a pak se poblíž 100 % odchýlit směrem k diagonále. Nesmí se s diagonálou prolínat nebo ztotožňovat, což se nestalo, jak je na Obr. 19 dobře vidět.



Obr. 19: Ziskový graf

Náš příklad aplikace procedury TREE není ani tak dokonalý (možnost selhání modelu je dost vysoká, ne všechna očekáváná kritéria se osvědčila), jako názorný. Ale protože se výsledky v zásadě shodují s dříve uvedenými analýzami a přesvědčivě je dokreslují, čímž dobře navazují na předchozí uváděné příklady, uvádíme je zde spíše než jiné možné výstupy (v jiné analýze pro IKSP – viz Blatníková & Zeman, 2012 bylo TREE již velmi úspěšně uplatněno). Nabízíme tento postup jako perspektivní, moderní a zajímavou možnost práce s daty.

Výsledky analýzy lze validizovat porovnáním výsledků z rozdělených částí datového souboru, přičemž lze použít dva postupy: rozpůlení a rozdělení na více částí. Procedura také počítá s chybějícími hodnotami a umí je různým způsobem zpracovávat nebo nahrazovat. O jiných metodách růstu stromu, které SPSS rovněž nabízí, jsme se shora zmiňovali.

4. Korespondenční analýza

Pozadí

V 70. letech 20. století korespondenční analýzu vyvinul Jean-Paul Bénzecri (1973), francouzský statistik libanonského původu.

Koncepčně je velmi blízká faktorové analýze (analýze hlavních komponent), je to jedna z vícerozměrných technik analýzy dat. Na rozdíl od FA, která může používat jenom metrické nebo quasi metrické proměnné (např. věk, známková škála) není KA omezena: pracuje i s čistě nominálními proměnnými (jako je muž/žena, 14 regionů ČR), s ordinálními, diskrétními daty nebo s kombinací nominálních a diskrétních dat.

K variantám korespondenční analýzy patří:

a) detrendovaná korespondenční analýza (detrended correspondence analysis)
b) kánonická korespondenční analýza
c) mnohonásobná korespondenční analýza (multiple CA) uzpůsobená pro více nominálních proměnných zároveň – (Pierre Bourdieu, 1984; Greenacre a Jöerg, 2006)
d) diskriminační korespondenční analýza (discriminant CA).

Bénzecri si kladl za cíl dokázat, že z cizího, úplně neznámého jazyka lze odvodit jeho gramatická pravidla a význam slov podle četnosti, s jakou jsou užívány jednotlivé výrazy (slova). Nejdříve odlišil slovesa od podstatných jmen z lingvistického „korpusu“ cizího jazyka (např. z literatury-knih, záznamů řeči). Pak z něho sestavil tabulku četnosti společného výskytu podstatných jmen a sloves. Ve sloupcích tabulky jsou různá slovesa, v řádcích četnost jejich spojení s různými podstatnými jmény, např. „běží“ v řádce s „muž“ 200x a „žena“ 130x; kdežto „spí“ v řádce „muž“ 400x a „žena“ 150x.

Relativní důležitost kontextu podstatného jména zjistíme výpočtem profilu: „muž“ celkem 200+400=600, „běží“=200/600=0,33 (tj. 33 %) a „spí“ =400/600=0,666 (=67 %)

Pokud mají slova stejný „profil“, jde o souznačná, synonymní slova. Např. spí a „sní“ v řádce s „muž“ 399x , tj. skoro 66 % a „utíká“ 199x, tj. skoro 33 %). Otázkou ovšem je, jak si představit či zobrazit celou sadu různých profilů.

Bénzecri řešil tak, že sčítal souznačné řádky („muž“ sní+spí=399+400=799 a běží+ utíká= 399+199=598) s tím, že se jejich poměr vzhledem k celku nezmění ((400+399)/ (600+598)= 799/1198=0,666 (=67 %). Nazval tento jev „princip distribuční ekvivalence“. Podle tohoto principu pak lze odvodit také „distribuční vzdálenost“ (místo sčítání odečítáme a vyjadřujeme jako čtverec, tj. eukleidovskou vzdálenost): vzdálenost spí od sní = $(400-399)^2/1198=0,0008$.

Korespondenční analýza se uplatnila rovněž v kriminologii a to v tak rozmanitých oblastech jako je zkoumání životní dráhy usvědčených vrahů (Dobash et al., 2007), vizualizace souvislostí mezi různými druhy majetkových trestných činů a jejich postihem (Richardson, 2009), zobrazení zahraničních aktivit organizovaného zločinu (Varese, 2012) nebo typologie kriminálních kariér (Smith et al., 1986). Je zmiňována také v učebnicích statistiky pro kriminology (např. Walker a Maddan, 2005). Reess-Jones (2007:141) dokonce spojuje pionýrské využívání korespondenční analýzy, resp. předchůdcovských podob s Pearsonem a s Fisherovou prací o kontingenčních tabulkách z roku 1940.

Objasňující příklad

Zpět k **Příkladu 2: Hodnocení orgánů činných v trestním řízení** (Zeman et al., 2011b), o kterém jsme již shora pojednali. Ve výzkumu byly mj. využity souhrnné indexy (0 -100), které vyjadřují hodnocení:

- spravedlnosti soudního procesu,
- smysluplnosti práce policie,
- efektivity detence a
- převýchovy stíhaných osob.

Ve všech případech tyto ukazatele (včetně těch původních nesumarizovaných) korelovaly s krajem, což pochopitelně vzbudilo zájem výzkumníků.

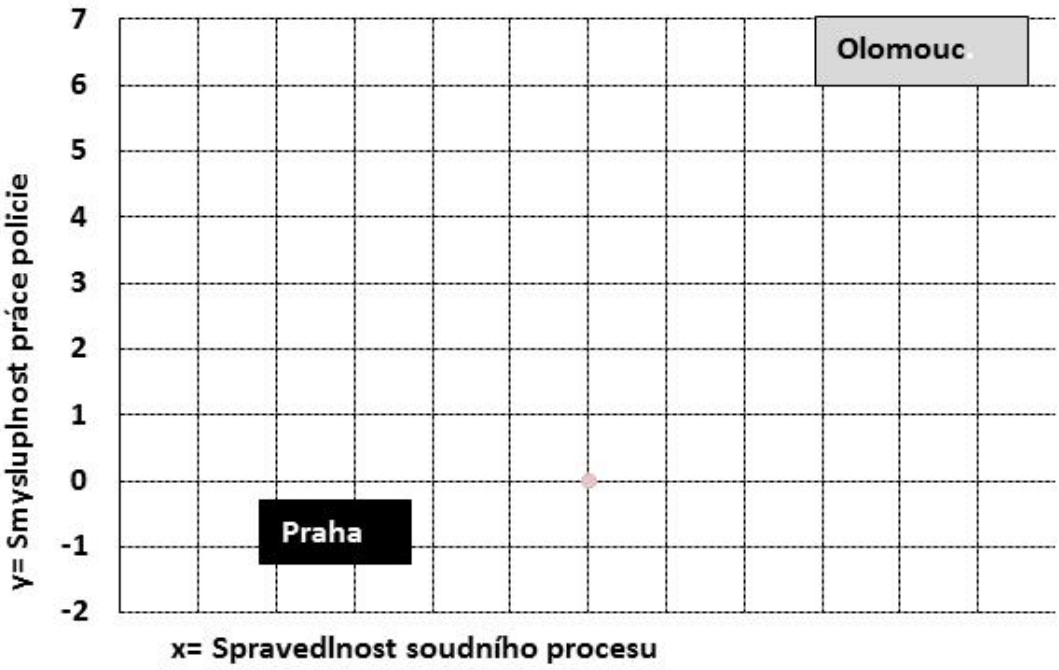
Ke studiu a interpretaci máme celkem 4 x 14=56 průměrů (viz Tabulka 17 níže), popř. rozdělíme-li skory na 5 intervalů od nejnižšího (0-20) po nejvyšší (80-100) hodnocení, máme 56x5 (=280) okének.

Tabulka 17: Souhrnné standardizované průměrné indexy hodnocení orgánů činných v trestním řízení podle krajů

Kraj	f26_1 Spravedl-nost soudního procesu	f26_2 smyslupl-nost práce policie	f26_3 Detence stíhaných osob	f26_4 Rescocializace pachatelů
1 Praha	44,03	47,46	70,24	52,08
2 Středočeský kraj	46,85	48,20	65,90	51,91
3 Jihočeský kraj	42,14	48,52	64,41	43,90
4 Plzeňský kraj	46,06	49,74	64,86	50,17
5 Karlovarský kraj	41,60	47,60	62,31	39,95
6 Ústecký kraj	44,85	47,14	61,96	46,19
7 Liberecký kraj	55,82	60,69	73,26	55,95
8 Královéhradecký kraj	42,35	47,78	58,46	43,58
9 Pardubický kraj	47,79	50,63	65,42	54,66
10 Kraj Vysočina	38,84	42,16	59,43	43,93
11 Jihomoravský kraj	43,51	49,94	64,74	48,50
12 Olomoucký kraj	51,71	55,61	67,08	54,46
13 Zlínský kraj	47,85	52,53	63,44	44,18
14 Moravskoslezský kraj	48,54	48,31	63,54	48,55
Total	45,77	49,28	64,94	48,92
ANOVA (sig.F)	<.001	<.001	<.001	<.001
Eta	.199	.192	.190	.184

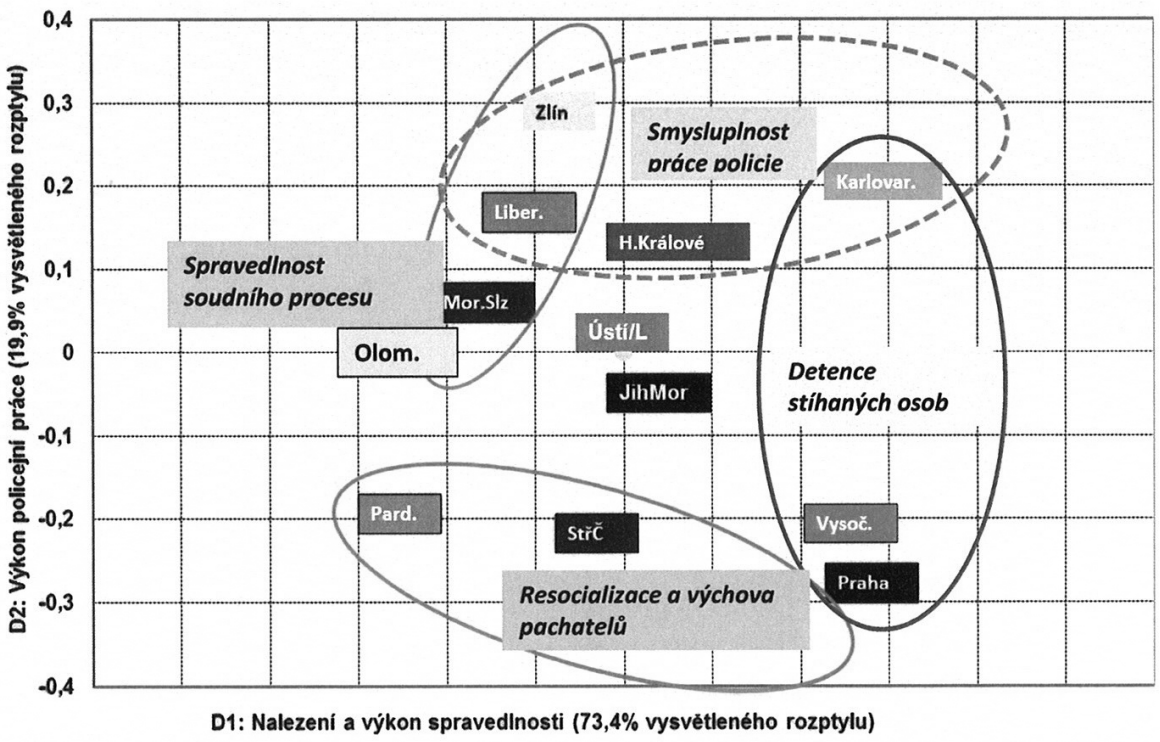
Tabulka 17 ukazuje, že nejlépe hodnotí Liberecký a Olomoucký kraj (u detence také Praha) a nejhůře Vysočina. Testováním průměrů, např. post-hoc testy získáme málo přehlednou spleť pořadí v hodnocení.

Další možnosti nabízí grafické zobrazení. Podle dosažených skorů můžeme spočítat vzdálenosti a zobrazit jednotlivé kraje jako body ve 4-rozměrném prostoru. Např. souřadnice pro dvě dimenze (spravedlnost soudního procesu a smysluplnost práce policie) získáme odečtením od celkové hodnoty (=Total v tab. 17). Praha bude mít polohu x=44,03-45,77=-1,75 a y=47,46-49,28=-1,82; a Olomouc x=51,71-45,77=5,94 a y=55,61-49,28=6,33



Obr. 20: Poloha dvou krajů v prostoru spravedlnosti soudů vs. smysluplná práce policie

Ovšem takové grafy jako na Obr. 20 musíme udělat pro všechny dvojice, tj. celkem ((n*n-1)/2)= 6 grafů (F26_1 s F26_2, F26_3, F26_4, dále F26_2 s F26_3 , F26_4, a F26_3 s F26_4). V každém z 6 grafů pak sledovat pozice 14 bodů - 14 krajů. Mnohem přehlednější je korespondenční analýza s 18 průměty zobrazenými mezi dvěma osami (faktory).



Obr. 21: Graf korespondenční analýzy hodnocení práce orgánů činných v trestním řízení podle krajů

Na Obr. 21 např. pro Prahu máme bod, který leží v různé vzdálenosti od spravedlnosti, smysluplnosti práce policie, resocializace pachatelů a má nejblíže ke kladně hodnocené detenci pachatelů. Navíc leží také v různé vzdálenosti od ostatních krajů (např. Vysočina a Jihočeský kraj mají k detenci stíhaných osob ještě blíže než Praha). Na stejné dvourozměrné mapě jsou zobrazeny i další kraje a jejich vzájemné vzdálenosti a vzdálenosti ke čtyřem průměrným hodnocením.

Další charakteristika korespondenční analýzy

Při korespondenční analýze lze použít nominální, kategoriální data: odstraní se tak nevýhoda, že nemetrické údaje (např. profese respondenta, kraj) nemůžeme analyzovat příliš náročnými, ale přínosnými vícerozměrnými technikami, jako je faktorová analýza.

Korespondenční analýza úsporně zobrazuje řádkové a sloupcové profily (tj. relativní sloupcové a řádkové četnosti). Hledá (podobně jako ve faktorové analýze) menší počet rozměrů než jsou počty sloupců, popř. řádek původní tabulky. V příkladu jsme měli tabulku 4 x 14 a hledáme tudíž řešení se 2 rozměry (zobrazení ve 4 rozměrech by nepřineslo nic nového). Matematicky se pracuje tedy s relativními řádkovými, sloupcovými a celkovými (tabulkovými) četnostmi. Korespondenční tabulka P je definována jako matice celkových relativních četností.

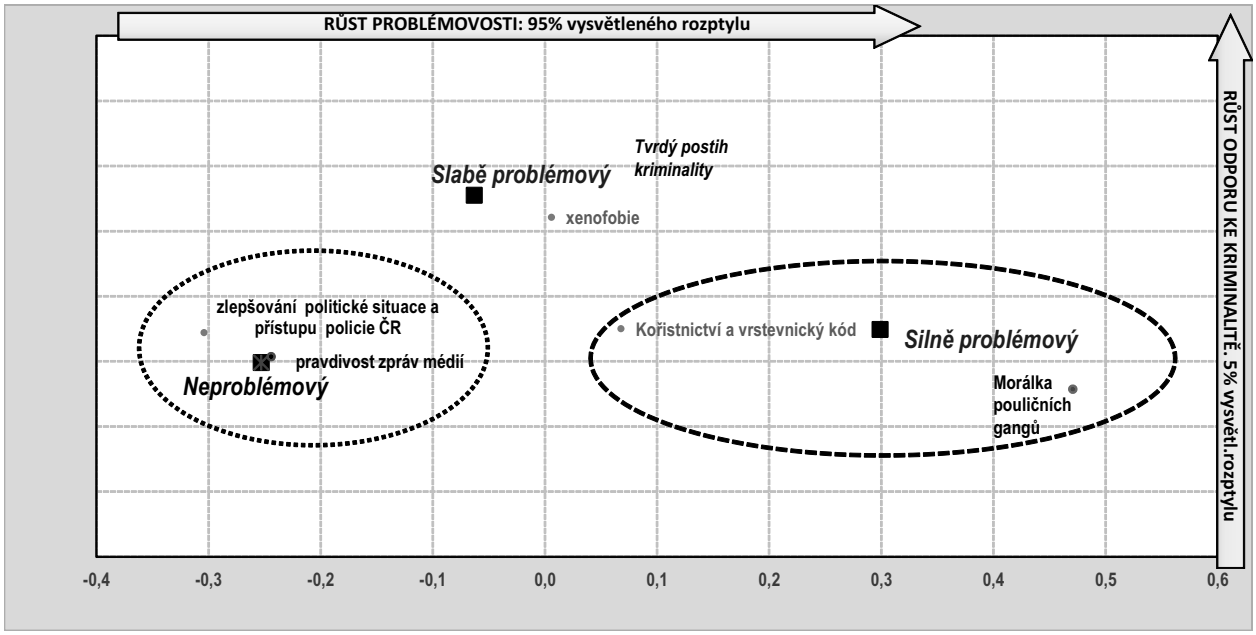
Protože chceme zobrazit graficky vzdálenosti mezi řádkovými (nebo sloupcovými) profily, hledáme nejdříve společné těžiště . Středem sloupců je průměr jejich „mas“ (=marginálních relativních ř.četností) , obdobně střed řádek (průměr marginálních sl.četností). Společný střed, těžiště sloupců i řádek najdeme „vycentrováním“ P podle středů řádků a sloupců (P-ř*sl).

Dále pak od s těmito vycentrovanými rezidui (p_{ij}) provádíme matematické operace související s příslušnými relativními řádkovými a sloupcovými četnostmi. Následuje obtížnější část, jádro celé metody KA: rozklad standardizovaných reziduí (tabulky či matice A), tzv. SVD (singular value decomposition). Reziduím v matici A se přiřadí pořadová čísla podle velikosti (tzv. jedinečné hodnoty – singular values) a ty se seřadí sestupně. Hledá se pak vyjádření řádkového profilu v jednotkách sloupcového profilu a naopak (převod jako u rovnic z jedné strany na druhou) a to podle jedinečných hodnot a dále se vypočtou souřadnice jednotlivých proměnných. Získáváme tak grafické zobrazení řádků a sloupců tabulky v jednom, zpravidla dvourozměrném prostoru.

Korespondenční analýza je velmi pružná**:** můžeme ji provádět s absolutními četnostmi, procenty, hodnotícími skory i s nestejnorodými daty.

Příklad č. 12: Životní zásady, s nimiž se ztotožňují potenciálně problematičtí a neproblematičtí lidé v ČR (Večerka et al., 2011). V tomto výzkumu byla nalezena pomocí vícerozměrných analýz proměnná rozdělující respondenty podle jejich potencionální problémovosti (viz shora, příklad č. 8. Názory mladých lidí v ČR na stupeň problematičnosti některých jevů). Provedli jsme také faktorovou analýzu 24 životních zásad a názorů, s nimiž se lidé v různé míře ztotožňují. Dospěli jsme k 7 faktorům a vyhodnotili je v kombinaci s potencionální problémovostí respondentů pomocí korespondenční analýzy.

Na obr. 18 vidíme, že neproblémoví respondenti měli nejblíže k pozitivním názorům na politickou situaci ČR, práci policie a také dobře hodnotili pravdivost zpráv uveřejňovaných v médiích. Slabě problémoví respondenti vnímali situaci v ČR jako chaotickou, málo kontrolovanou, vyznávali zásady odvety („oko za oko"), byli pro udělování trestu smrti a měli velmi kritický postoj k práci policie. Byly jim také blízké výhrady vůči cizincům, bezdomovcům a Romům. Silně problémoví respondenti tíhli k násilným řešením, nemorálním postojům a osobní mstě. Vyznávali kořistnické zásady: cenili si peněz, nevěřili v možnost, že lze poctivě zbohatnout a byly jim blízký kód v oblékání mladých lidí.



Obr. 22: Názory neproblémových a problémových mladých lidí na životní zásady

Korespondenční analýza nemusí sloužit jenom k tomu, abychom si potvrdili naše domněnky, jako v příkladu 12 (Obr. 22) ve více nebo méně přehledné spleti různých proměnných. Pomocí KA můžeme někdy dospět k neočekávaným výsledkům, k odhalení

nových souvislostí: korespondenční analýza není jenom popisnou grafickou obdobou kontingenční tabulky **- může mít explorativní charakter: „**Tato metoda je explorační i deskriptivní“ (Rees-Jones, 2007, 142)**.**

KA není vázána na nějaký matematický model, ale její výsledky mohou posloužit jako **úvod** ke složitějšímu **vícerozměrnému modelování kategoriálních dat**, např. k diskriminační korespondenční analýze, k porovnávání faktorových skorů různých analýz nebo jako v našem případě k možným návazným regresním analýzám. Navíc, pokud je korespondenční analýza používána spolu s dalšími metodami, pak „skýtá velmi užitečný, možná málo využívaný prostředek analýzy sociálních vztahů“ (Rees-Jones, 2007: 148).

Ale pozor: KA je jen jedna z možností - je vhodné ji doplnit dalšími metodami analýzy kategoriálních dat a nezapomínat, že jde pouze o popisně-explorační (zobrazovací) techniku! Významnost korespondenční analýzy sama o sobě není statisticky testována, KA není založena na parametrech, rozdělení hodnot.

5. Mnohorozměrné škálování (multidimensional scaling)

Pozadí a příklad

Účelem mnohorozměrného škálování (MDS) je vyjádřit odlehlost objektů, popř. názory na podobnost objektů nebo preference objektů jako vzdálenosti. (Za objekty jsou považovány názory, podněty, produkty, uchazeči o něco, volby ap.) Tyto vzdálenosti jsou pak zobrazeny v mnohorozměrném prostoru. Např. názor „nejlepší je neplést se do ničeho, co se mne netýká“ může ležet blíže skupině A, vzdáleněji od skupiny B a C podle toho, do jaké míry s tímto názorem každá skupina souhlasí. Jako klasický výukový příklad se v literatuře uvádí vzájemná poloha velkých měst, myšlená vzdušnou čarou v kilometrech (Giguere, 2007).

Odlehlost anebo stupeň souhlasu lze vyjádřit jako různou vzdálenost a zobrazit graficky na jednom obrázku, zachytit tzv. perceptuálním mapováním. Zachytit a zobrazit tak lze velké množství informací, celé profily charakteristik vztahů mezi objekty a to vše ve velmi přehledné a názorné formě percepční mapy. Na těchto obrázcích jsou zachyceny vztahy mezi proměnnými také v případech, kdy tyto vztahy vyjadřujeme různými nelineárními koeficienty (nikoliv pomocí korelací). Kromě grafů jsou při mnohorozměrném škálování v SPSS k dispozici také tabulky, které jsou oporou pro interpretaci výsledků, popř. poslouží jako podklad k dalším výpočtům. SPSS nabízí v současnosti tři techniky mnohorozměrného škálování (Alscal, Prefscal a Proxscal), z nichž se zde zaměříme na relativně nejjednodušší z nich, tj. ALSCAL.

MDS je založeno na srovnávání objektů (věcí, představ, událostí, lidí). Každý objekt má objektivizované a přisuzované vlastnosti. Počty zaznamenaných trestných činů mezi rokem 2000 a 2009 klesly o téměř 60 tisíc – to je příklad objektivizované vlastnosti vývoje kriminality v ČR. Veřejnost ale tuto skutečnost nemusí znát. Příkladem zcela odlišně vnímané, přisuzované vlastnosti může být fakt, že bezmála 8 respondentů z 10 vývoj kriminality za toto období vnímá jako výrazně nebo mírně rostoucí.

Takže přisuzované vlastnosti objektů nemusí souhlasit s objektivními vlastnostmi a někdy dokonce naopak, přisouzené vlastnosti mohou odporovat objektivní realitě. Veřejnost vnímá např. růst kriminality pod vlivem skoro denně zveřejňovaných, silně medializovaných případů násilné nebo majetkové trestné činnosti. Jak objektivní tak i přisuzované vlastnosti mohou být předmětem analýzy metodou mnoharozměrného škálování.

MDS má kořeny v psychometrice 20. - 60. let 20. století. Zpočátku byla metrická data používaná k vyjádření podobnosti nebo nepodobnosti. Z daných quasi vzdáleností mezi body se usuzovalo na jejich polohu a na uspořádání (konstelaci) bodů, které bylo těmito vzdálenostmi způsobeno.

Někteří autoři hovoří o „nemetrické revoluci“ (Coxon, 2004), kterou přisuzují Clydu H. Coombsovi (1912-1988) a J. B. Kruskalovi (1928-2010). Kruskalův algoritmus (Kruskal & Wish, 1978; Kruskal, 1964b) a tzv. minimal spanning tree umožnily pracovat nejenom s metrickými, ale také s ordinálními daty. Mnohorozměrné škálování pak bylo silně rozvinuto rozvojem výpočetní techniky v 60. letech s uplatňováním tzv. iterativních postupů.

Multidimenzionální škálování se v kriminologii používá jako jedna z metod profilování zločinců (Kocsis, 2010). Vzhledem k možnosti pracovat i s velmi malým počtem případů se tato technika využívá např. k určení typů znásilňovatelů žen (Hendrix a Scimone, 2007) a také k určení vražedných sklonů delikventů (Fisher a Salfati, 2007). Někteří autoři pomocí multidimenzionálního škálování např. profilují sexuální násilníky (Goodwill et al., 2009) nebo identifikují potřeby podmíněně propuštěných vězňů (Brown, 2004). Šíře možné aplikace těchto technik je velká a zdaleka jsme ji zde uvedenými příklady nevyčerpali.

Podle některých autorů (Kruskal a Wish, 1978) je mírou či testem „špatné shody“ modelu s daty tzv. s-stress. Špatná shoda >0,20 až výborná shoda ≤0,05 a dokonalá shoda <0,01. V našem příkladu jsme dosáhli dokonalé shody modelu s daty. Index čtvercové korelace (RSQ) je rovněž mírou shody mezi daty a modelem. Za přijatelnou se považuje výše RSQ≥0,60 (Garson, 2012: 5).

ALSCAL nevyžaduje na rozdíl od faktorové analýzy normalitu rozdělení. Dále vzdálenosti mezi body se měří celými koeficienty. (Ve faktorové analýze parciálními

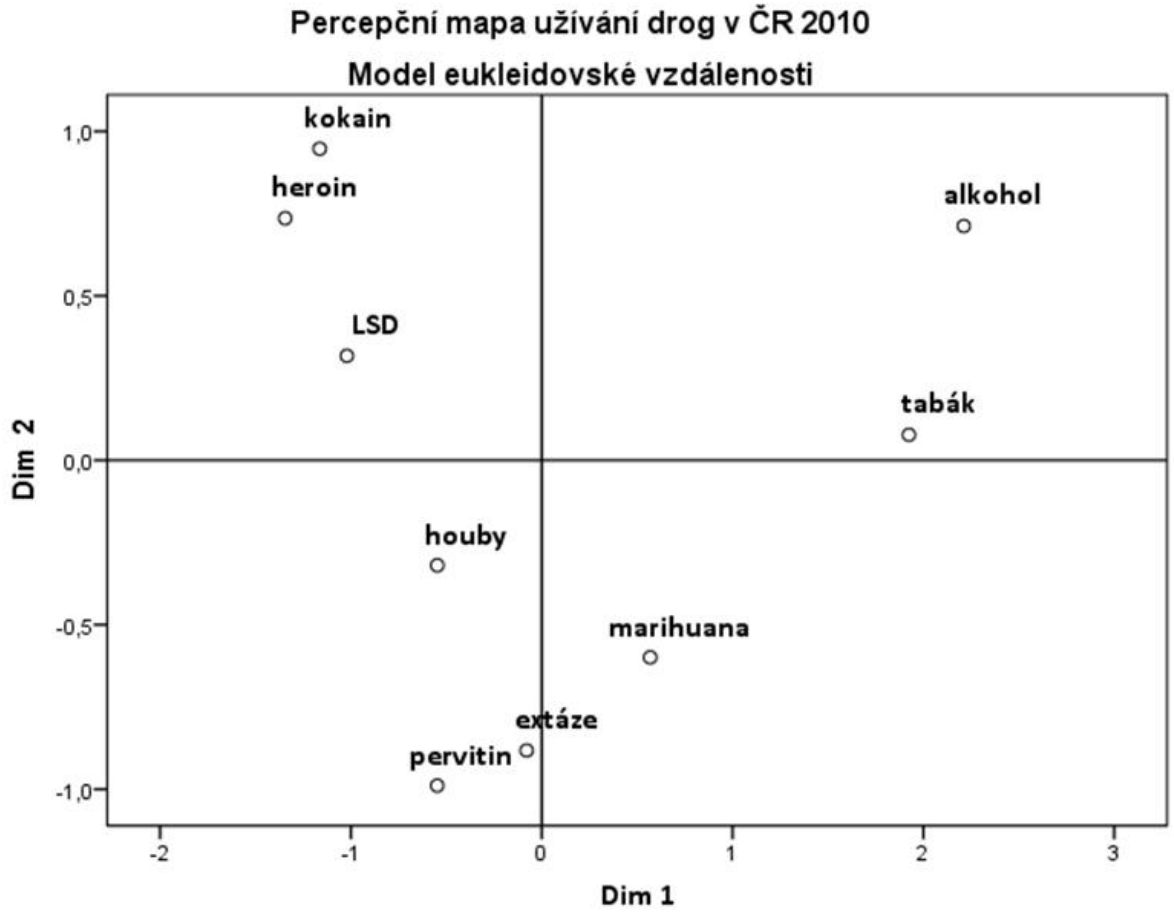
korelacemi, takže FA kontroluje příbuznost analyzované vzdálenosti se vzdálenostmi dalších proměnných.)

Vstupní údaje vhodné pro mnohorozměrné škálování

Jaké údaje jsou pro multidimenzionální škálování vhodné? Garson (2012b) shrnuje dřívější literaturu a podle typu dotazování poukazuje na 4 varianty: 1) preferenční – je A podobnější B nebo C? ; 2) získané metodou párového srovnání – porovnejte A a B na stupnici od 0=žádná podobnost až po 10=naprostá podobnost; nebo mám raději A než B; 3) s využitím balíčků karet – respondenti dávají vždy sobě podobné nebo stejně preferované karty (s příslušným názvem produktu nebo osoby) na stejnou hromádku; 4) přímé seřazení – respondenti očíslují objekty podle preferencí od prvního až do posledního pořadovými čísly.

Uvedené typy dotazování se v kriminologických výzkumech moc často neobjevují, pokud vůbec ano. Naproti tomu pro mnohorozměrné škálování jsou vhodná ***data pátého typu***, z kterých se často ve výzkumech a analýzách IKSP čerpá. Jsou to data ***získaná objektivními metodami***, ze statistik kriminality, policejních a soudních dokumentů a registrů, popř. obsahovou analýzou dalších zdrojů. Obecněji jako příklad takových dat můžeme uvést metrické vzdálenosti (např. míst se zvýšenou hustotou kriminality), frekvence výskytu nějakého jevu (např. počty policií evidovaných a objasněných trestných činů, počty stíhaných a odsouzených osob celkově a v různých kategoriích, počty soudniček/zpráv a pořadů s tematikou kriminality v médiích za určité období atd.), počty transakcí (např. počty rozsudků, počty odvolání k instancím vyššího stupně). Výhodou multidimenzionálního škálování je již jednou zmíněný fakt, že lze pracovat s malými soubory i s individuálními respondenty do 100 případů, což je častý případ v analýzách IKSP (faktorová analýza naproti tomu vyžaduje velké datové soubory- vhodné při výzkumech dotazujících veřejnost).

Také z korelační matice (Pearsonových korelací r) lze získat matici nepodobností nebo vzdáleností (dissimiliarity matrix), pokud např. odečteme r od 1. Následující percepční mapa užívání drog z **příkladu č. 4** (Zeman et al., 2011a), týkajícím se osobní zkušenosti české veřejnosti s užíváním drog, (Obr. 23) je založena na nepodobnostech.



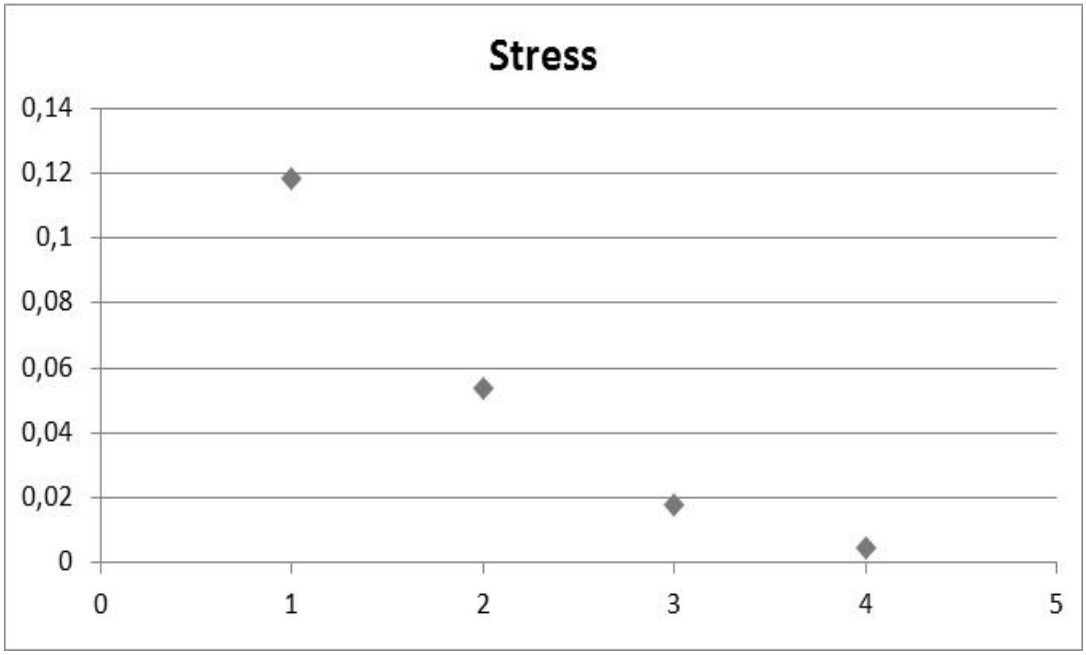
Obr. 23: Percepční mapa užívání drog v ČR roku 2010

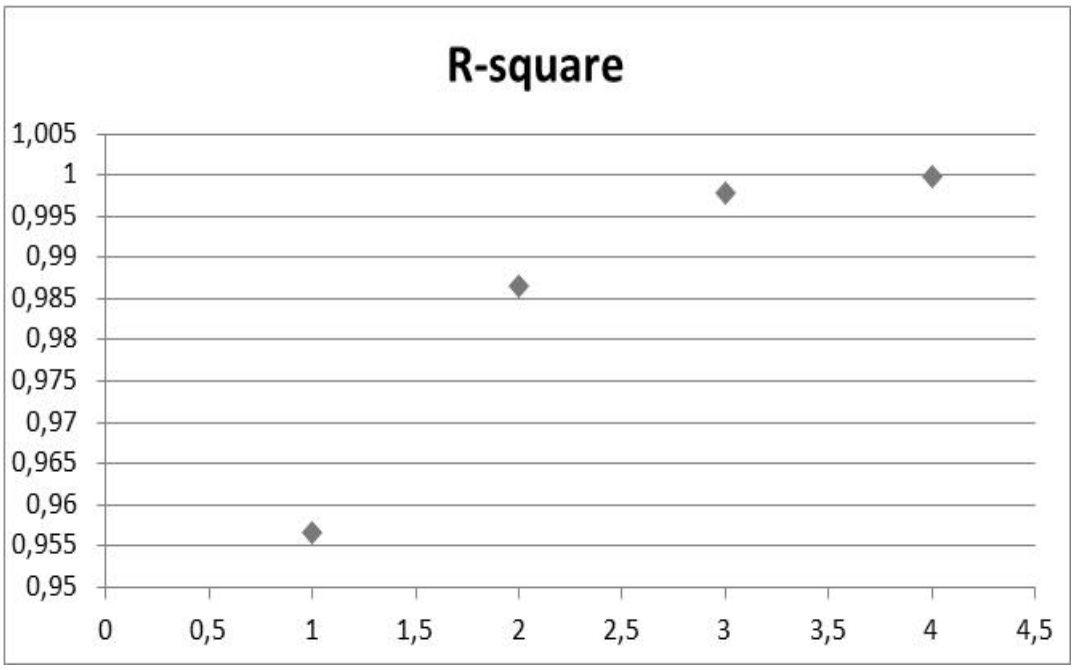
Celkem zřetelně se zde rýsuje v záporné polovině Obr. 23 shluk pokrývající kokain, heroin, halucinogenní houby a LSD, popř. pervitin a extázi, tedy tvrdé drogy. Další skupina, měkkých drog, je tvořena marihuanou, tabákem a alkoholem v kladné polovině grafu. Z druhého pohledu jak podle obrázku 19 tak i podle souřadnic v tabulce 17 patří tabák a marihuana více k podskupině měkkých „tanečních" a „rekreačních" drog jako jsou pervitin a extáze, kdežto alkohol je vnímán spíš v kontextu tvrdých drog. Stress (.05395) ukazuje celkem dobrou, ale nikoliv naprostou shodu modelu se vstupními daty a RSQ (=.98656) je přiměřeně vysoké.

Tabulka 17: Souřadnice užívání psychotropních látek pro dvourozměrné zobrazení v ALSCALu

Název	Dimenze	
	1	2
tabák	1,8641	-0,358
alkohol	2,4623	0,3447
marihuana	0,4474	-0,3596
extáze	-0,1416	-0,7774
pervitin	-0,7217	-0,861
kokain	-1,2355	1,0035
heroin	-1,4327	0,5526
LSD	-0,9325	0,0879
houby	-0,3096	0,3674

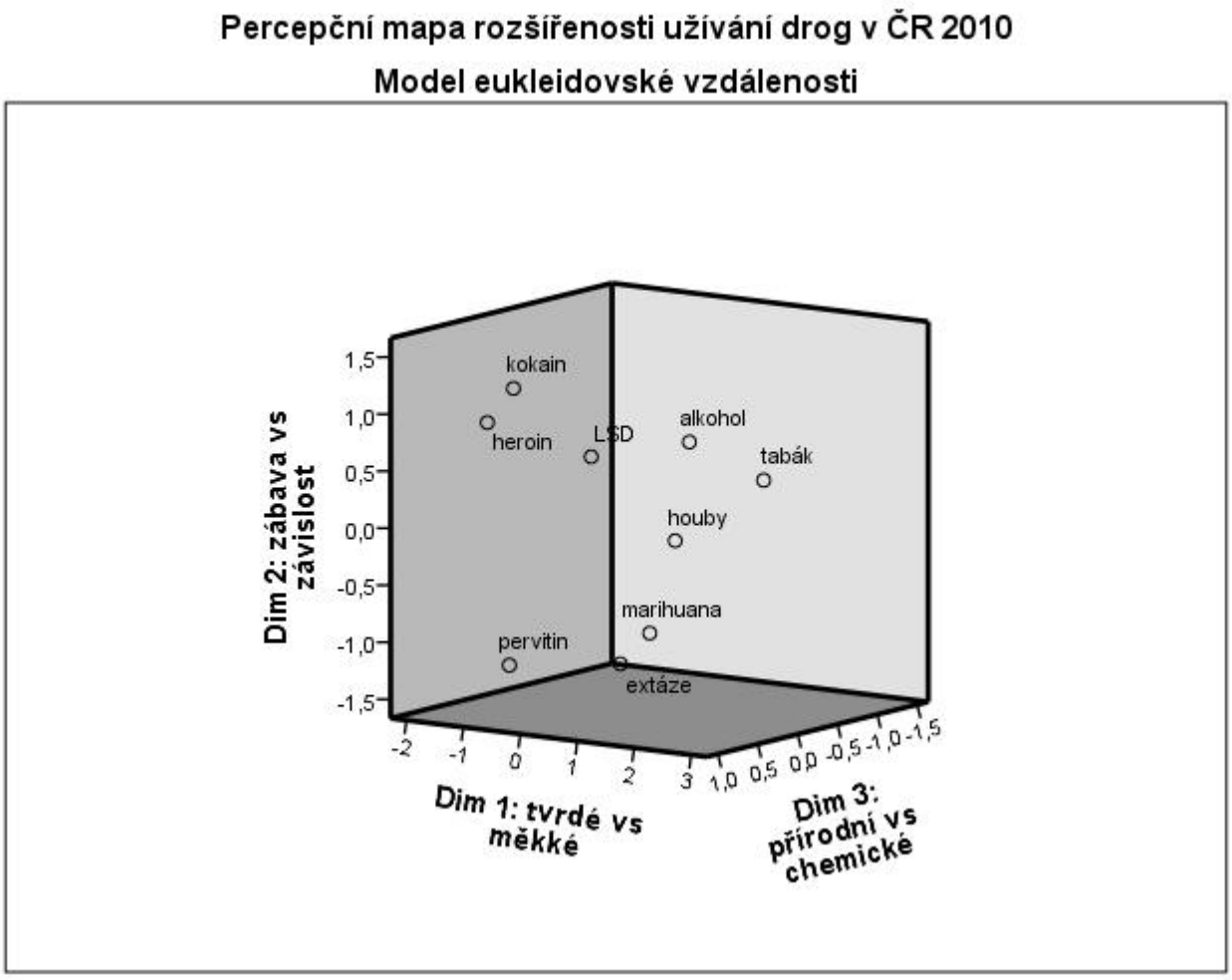
Jako východisko pro výpočet matic nepodobnosti byla použita stupnice užívanosti 1=nikdy-2=před více než 12 měsíci-3= v posledních 12 měsících a 4=v posledních 30 dnech. Abychom přiblížili model více datům, můžeme rozhodnout o počtu dimenzí na základě sutinového grafu. Postupně provedeme větší počet řešení a hodnoty stressu a r^2 vyneseme do grafu. Podle něho se jeví jako optimální trojrozměrné řešení, i když samozřejmě nelze rozhodovat mechanicky, ale podle toho, které řešení dává lepší smysl.





Obr. 24: Sutinové grafy shody modelu s daty a procenta vysvětleného rozptylu

Sutinový graf na Obr. 24 v obou případech ukazuje, že optimální je trojrozměrné řešení: rozdíl mezi dvou a trojrozměrným řešením je ještě značný, mezi tří a čtyř už malý; další řešení neexistují. Obecně platí, že při zvyšování dimenzionality stress musí klesat a naopak % vysvětleného rozptylu (r^2) stoupat.



Obr. 25: Percepční mapa rozšířenosti užívání drog v ČR (Stress =.01768 ; RSQ = .99775)

Při pokračování analýzy příkladu 4 technikou ALSCAL vidíme na Obr. 25 v prvním vodorovném rozměru, že se stejně jako u dvourozměrného řešení na kladné straně soustředily měkké „drogy“ (včetně alkoholu, tabáku a marihuany) a na záporné straně tvrdé drogy heroin, kokain, pervitin, LSD, extáze a halucinogenní houby. Tato dimenze diferencuje užívanost od nejrozšířenějšího alkoholu a tabáku až po nejméně rozšířený heroin a kokain.

Druhý svislý rozměr od nejzápornější extáze po nejkladnější kokain má v záporu přístupnější měkčí a „rekreační“ drogy: pervitin, marihuana, extáze a halucinogenní houby. V kladné části jsou tvrdé a také nejběžnější drogy: kokain heroin, LSD, tabák a alkohol. Liší se celkem nepatrně (pozicí alkoholu) od dvourozměrného řešení. V tomto pohledu jsou uživatelské závislosti na tabáku a alkoholu východiskem k užívání tvrdších drog, kdežto tzv. taneční drogy (extáze) jen zřídkavě.

V třetím opět vodorovném rozměru se látky diferencují od nejzápornějších hub po nejkladnější pervitin. Houby, LSD, tabák, marihuana a extáze tvoří zápornou, převážně přírodní část a kokain, heroin, alkohol a pervitin kladnou chemickou část (znaménko dimenze nemá žádný kvalitativní smysl, nic se jím nehodnotí). Tato dimenze také definuje přístupnost s tím, že „přírodní" látky si snáze opatří nebo zpracuje uživatel sám, kdežto ty chemické zpravidla vyvolávají závislost na dodavateli. Záleží na výzkumníkovi, jestli se spokojí s dvourozměrným řešením, nebo jestli chce získat ještě detailnější pohled na uživatelské zvyklosti a rozšířenost drog podle jejich typu.

Tabulka 19: Souřadnice trojrozměrného řešení užívanosti drog v ČR 2010

Název	Souřadnice podnětů		
	Rozměry		
	1	2	3
tabák	2,173	0,430	-0,363
alkohol	2,343	0,958	0,690
marihuana	0,468	-0,977	-0,150
extáze	-0,040	-1,276	-0,145
pervitin	-0,477	-1,127	0,935
kokain	-1,438	1,113	0,197
heroin	-1,451	0,869	0,516
LSD	-1,187	0,392	-0,598
houby	-0,390	-0,380	-1,082

Tento příklad ukázal mj. také to, že ALSCAL je přinejmenším stejně vhodný jako faktorová analýza a to zejména pro data ordinálního charakteru, která nemají normální rozložení. V obou případech, jak při faktorové analýze tak i pomocí ALSCAL jsme dospěli ke třem faktorům, které mají hodně obsahových přesahů, protože jsme vycházeli z podobné matice (korelační). ALSCAL v našem příkladu ovšem přispěl něčím navíc. Při FA se LSD, marihuana a halucinogenní houby odchylovaly od jednoduché struktury, měly dost vysoké podvojné zátěže. Navíc, vedle tvrdých a měkkých drog, vznikl faktor „běžně/ji dostupných" psychotropních látek (tabák, alkohol, marihuana). V ALSCALu (Obr. 25) se látky odlišily nejenom do „tvrdé" a „měkké" kategorie (nebo kategorie běžněji dostupných látek: alkohol, tabák a marihuana) jako při faktorové analýze, ale navíc také podle geneze závislosti (rekreační vs. silná závislost přinášející drogy/látky) a přístupnosti (či způsobu výroby: „udělej si sám" vs. dealer).

Samozřejmě záleží hodně na fantazii a zkušenostech výzkumníka, jak jednotlivé konfigurace „podnětů“ interpretují. Odborná literatura (např. Garson, 2012:5) doporučuje různé pomůcky. Jednou z nich je týmová práce nad interpretací výsledků, kdy si výzkumník vyslechne, shrne a vyhodnotí různé varianty vysvětlení. Další nezřídka používanou metodou jsou regresní analýzy. Např. souřadnice dimenze měkké-tvrdé drogy (viz shora v Tabulce 19 sumární hodnoty mezi 2,34 a -1,19) za náhodně vybraných 100 respondentů jsou vloženy do regresní rovnice jako závislá proměnná.

Růst spotřeby „měkkých drog vs. tvrdých drog“ může záviset na věku, pohlaví, informovanosti o zdravotních rizicích, legálnosti pěstování marihuany pro vlastní potřebu, postojích k prodeji a užívání drog na veřejnosti ap. Toto mohou být ovlivňující, nezávislé proměnné podle toho, jaká data jsou dostupná a jakou imaginací výzkumník vládne. Lze se mj. domnívat, že mezi uživateli tvrdých drog jsou častěji mladší respondenti a chlapci než dívky a starší respondenti; nebo že respondenti ve středním a starším věku budou tíhnout víc k „měkčí“ části této dimenze, tj. k alkoholu, tabáku a marihuaně. Takovéto a podobné hypotézy lze dále testovat regresemi nebo MANOVA a ANCOVA modely.

Závěr

Doufám, že toto skromné pojednání povzbudí jak aktivní výzkumníky, tak i uživatele kriminologických výzkumů k širšímu zájmu o vícerozměrné statistiky. Také bych si přál, abych touto cestou přispěl ke snížení nadměrného respektu vůči těmto technikám.

Popsané metody podle mého názoru dnes tvoří ve světě všeobecně využívanou moderní výbavu, bez které bude stále těžší se obejít. Ve zprostředkování aspoň části tohoto metodologického bohatství by měl spočívat přínos tohoto pojednání.

Souhrn

V našem kriminologickém výzkumu, zvláště v kvantitativních výběrových šetřeních, se začíná teprve odnedávna při zpracování dat prosazovat vícerozměrná statistika. Přitom v amerických, francouzských, holandských nebo britských kriminologických pramenech se setkáváme s 20-40 letou zkušeností s aplikací různých vícerozměrných technik statistické analýzy. Nejedná se o samoúčelný prestižní problém, ale o to, že jednorozměrné a dvourozměrné statistiky, na něž se ještě i dnes u nás příliš spoléhá, nepřinášejí vždy optimální řešení. Bývají např. zdrojem nesprávně potvrzovaných nebo naopak vyvracených souvislostí mezi proměnnými a odtud i možným podnětem k nedoloženým závěrům a prognózám týkajících se kriminálních jevů.

Tato práce sestává z pěti částí. V první obecně vícerozměrnou statistiku charakterizujeme, ve druhé se zabýváme faktorovou analýzou, ve třetí shlukovou a ve čtvrté korespondenční analýzou a v poslední části mnohorozměrným škálováním.

Pokud statisticky zkoumáme vztahy tří a více proměnných zároveň a hledáme mezi nimi hlubší souvislosti, jde o vícerozměrnou statistiku. Komplexní charakter této statistiky otvírá nové a mnohem spolehlivější perspektivy. Zpravidla můžeme s její pomocí mnohem přesněji a všestranněji zjistit a popsat zkoumané jevy než při aplikaci dvourozměrných statistik. Také s jejich pomocí dokážeme mnohem lépe zachytit a popsat proměny studovaných subjektů a jejich chování v čase, ať už jde o kriminální kariéru recidivistů nebo např. o nácvik ovládání trestaných agresivních osob.

Vícerozměrná statistika se opírá o a) testovací techniky, b) o rozdělení náhodných veličin, proti nimž se testuje významnost průměrů, procent a jiných námi zjištěných parametrů a o c) koncepci vzdáleností mezi studovanými jevy v pomyslném prostoru. V tomto pojednání se zabýváme jen některými testovacími technikami: faktorovou, shlukovou a korespondenční analýzou a zčásti také mnohorozměrným škálováním. Okrajově zmiňujeme rozhodovací stromy a modelování pomocí strukturálních rovnic. Na různých příkladech z výzkumů IKSP dále uvádíme aplikace i některá úskalí těchto technik.

Faktorová analýza, pojednávaná ve druhé části, je statistická metoda, kterou se četnost pozorovaných či manifestních proměnných omezí na menší počet latentních (nepozorovaných, konstruovaných) faktorů. Tyto faktory vysvětlují rozptyl měřených (manifestních) proměnných; jejich násobky (lineární kombinace) se rovnají měřené proměnné plus chybě (tj. chyba měření a nevysvětlená část rozptylu - chyba modelu).

FA ve formě analýzy hlavních komponentů před 100 lety vynalezl Spearman. Rozvinula se v 60. letech 20. století a rozšířila se do mnoha oborů přírodních i společenských věd. Východiskem explorativní FA jsou korelace nebo kovariance množiny měřených proměnných. Na jejich základě se statistickým způsobem určuje, zda a které z těchto manifestních proměnných si jsou navzájem blízké a patří tedy k sobě nebo zda patří k jinému společnému faktoru.

Příklad 2 popisuje výzkum mínění veřejnosti o aktivitě orgánů činných v trestním řízení. Faktorová analýza z 19 položek předložených veřejnosti k hodnocení vytvořila podle očekávání 4 faktory: hodnocení práce policie, soudů, státních zástupců a pracovníků vězeňské služby. Na výchozích korelačních maticích a na maticích parciálních korelací se vysvětlují mezietapy faktorové analýzy. Protože FA má určité požadavky na vstupní data, uvádí se také příklad (č. 3) týkající se názorů veřejnosti na to, zda vyjmenované skutky by měly být trestné. Zde výchozí data tyto požadavky nesplňují, ale přesto lze FA provést, pokud se vstupní korelační matice přetvoří na tzv. tetrachorické koeficienty.

Po provedení uvedené transformace jsme získali k příkladu 3 dvoufaktorové řešení s komponenty činů poškozujících společnost a poškozujících osobní integritu. Popisuje se, jakými metodami lze odhadnout optimální počet faktorů (hlavních komponentů) a ukazuje na tomto případě extrémně málo využívaná optimální metoda (Hornovy) paralelní analýzy. Ukáže se však odlišný výsledek, pokud použijeme „nejzavedenější“ metodu odhadu počtu faktorů podle Kaiserova pravidla a výchozí matici Pearsonových korelací.

Vysvětluje se rozdíl metody extrakce hlavních komponentů od metod faktorové analýzy. Příklad č. 5 zaměřený na souhlas mladých lidí v ČR s vybranými životními zásadami není vhodný pro faktorovou analýzu s extrakcí metodou maximální věrohodnosti, přestože jeho vstupní matice byla přizpůsobena a vyhovují požadavkům spojitých metrických dat. Proto je provedena analýza alternativní přípustnou metodou analýzy podle hlavních os.

Škály získaných jednotlivých (tří) faktorů jsou také podrobeny analýze spolehlivosti podle Cronbacha.

Na příkladu č. 6 (Názory mladých lidí na chování svědků ve vybraných situacích souvisejících s trestným činem), který se opírá o k tomuto účelu vhodnější stupnice, se demonstruje faktorová analýza provedená metodou extrakce maximální věrohodnosti a její výhody. Podle některých expertů je faktorová analýza na rozdíl od analýzy hlavních komponent vhodnější. Zvláště pak doporučují šikmé rotace a metodu extrakce maximální věrohodnosti, jejíž výsledky nejsou vázány na charakteristiku konkrétního datového souboru. Přidržujeme se tohoto doporučení s tím, že vstupní data by mohla pravděpodobně být lépe uzpůsobena pro použití této metody, kdyby se už při přípravě metodiky (dotazníku, expertní škály) počítalo se statistickým zpracováním.

Poslední část kapitoly o faktorové analýze vyzvedává jednu z nejvšestrannějších a nejpružnějších metod rozboru vzájemných souvislostí proměnných, modelování pomocí strukturálních rovnic (SEM). Nejčastější aplikací SEM bývá tzv. konfirmační faktorová analýza, jíž testujeme výchozí teorii nebo strukturu nalezenou pomocí explorační faktorové analýzy. Příklad č. 6 (Názory mladých lidí na chování svědků ve vybraných situacích souvisejících s trestným činem), resp. výsledky explorační faktorové analýzy, jsou podrobeny KFA. Model odvozený z předchozích FA, s nímž jsme vstoupili do analýzy, se osvědčil a plně vyhovuje datům a potvrdilo se, že respondenti mají různý pohled na chování svědka v situacích, kdy se dostává do popředí buď pachatel anebo spíše oběť trestného činu. Ověřili jsme si ovšem, že tento rozdíl není příliš velký.

Shluková analýza (SA) představuje třetí část pojednání po úvodní charakteristice vícerozměrné statistiky a po faktorové analýze. Na příkladu č. 8 (Názory mladých lidí v ČR na stupeň problematičnosti některých negativních jevů) objasňujeme „korelační profil“ a podstatu SA. Na příkladu č. 9 (Hodnocení stupně závažnosti negativních společenských jevů českou veřejností) ukazujeme, jak je nutné nalezené shluky identifikovat obsahově a dále je charakterizovat jako reálně existující skupiny lidí s distinktními sklony a chováním. Hierarchická shluková analýza je provedena na příkladu č. 10 (Analýza trendů kriminality v posledním desetiletí). Je vhodná pro menší, přehledné datové soubory s metrickými daty. Ukazuje se, že jednotlivá období vykazované kriminální statistiky se mohou postupně (hierarchicky) slučovat do sobě podobných celků. Za pomoci rampouchového grafu

a dendrogramu se krok po kroku tyto statistiky slučují, přičemž se nacházejí a vysvětlují jejich společné rysy. Dále je navrhován směr, ve kterém by se měla hledat vysvětlení existence těchto shluků, jaké otázky mohou shluky vyvolat a na jaké problémy mohou poukázat.

K-means (někteří čeští uživatelé hovoří o k-středové shlukové analýze) je metoda shlukové analýzy používaná na velkých souborech s alespoň ordinálními, v ideálním případě metrickými daty. Příklad č. 11 (Evropská studie hodnot a ospravedlnitelnosti chování) ukazuje, jak shluky v postupné analýze (iteracích) vznikají a jakým způsobem lze zvolit jejich názvy. Přitom počet shluků si zadává výzkumník sám a musí pečlivě zvažovat a vybírat nakonec ty z nich, které jsou stabilní, početně vyvážené a rozumně vysvětlují zkoumanou realitu. Nalezené shluky lze použít k mezinárodnímu srovnávání, pokud stupnice tvořící jejich základ mají stejnou konfiguraci a metriku v různých zemích.

Dobře vybrané a identifikované shluky by měly mít smysluplné korelace k dalším proměnným výzkumu. V našem příkladu č. 8 (Názory mladých lidí v ČR na stupeň problematičnosti některých negativních jevů) byli mladí lidé rozděleni na skupinu seriózních (problémy vnímány jako závažné), kompromisních (středně závažné) a zlehčujících závažnost (nezávažné). Tyto skupiny měly mj. významně odlišnou kriminální citlivost, sledovanou na 43 ukazatelích a už ve škole byli odlišně známkováni za chování (relativně nejhůř skupina zlehčujících) apod.

Dvoustupňová shluková analýza (SPSS) je kompromisem mezi hierarchickou analýzou a k-means. Vhodně se uplatní při práci s velkými výběrovými soubory a může využívat kategoriální i metrická data. Na příkladu č. 4 (Zkušenost české veřejnosti s psychotropními látkami) jsme ukázali trojshlukové řešení, kde byli s využitím ordinálních dat odlišeni neuživatelé od lidí s náhodnou ojedinělou zkušeností a s častějším užíváním drog.

Okrajově zmiňujeme problematiku rozhodovacích stromů, která je v SPSS zastoupena procedurou TREE. Touto procedurou můžeme zpracovávat jakýkoliv druh dat. Podrobili jsme příklad č. 4 (Zkušenost české veřejnosti s vyjmenovanými psychotropními látkami) této analýze s cílem zjistit, zda různé typy uživatelů drog vznikají na základě sociálních a demografických rozdílů (jako je věk, pohlaví, vzdělání, příjem, velikost bydliště atd.) a zda

na rozdíly mezi nimi spolupůsobí informovanost o účincích a protiopatřeních v oblasti užívání drog.

Korespondenční analýze je věnována čtvrtá část. Její vznik je spjat s lingvistikou. Na příkladu č. 2 (Hodnocení orgánů činných v trestním řízení) sledujeme, jak se názorně uspořádají jinak zcela nepřehledné údaje o hodnocení jednotlivých orgánů v závislosti na jednotlivých krajích ČR. Korespondenční analýza má výhodu v tom, že nevyžaduje od dat splnění takových předběžných podmínek jako je normalita a pracuje s jakýmkoliv typem dat. V příkladu 12 (Životní zásady, s nimiž se ztotožňují potenciálně problematičtí a neproblematičtí lidé v ČR) jsme různě problémovým skupinám lidí od potenciálně deviantních až po zcela neproblémové přisoudili pomocí korespondenční analýzy smysluplně rozdílné životní zásady a názory, shrnuté předchozími analýzami do 7 kodexů a „filosofií" chování od morálky pouličních gangů až po pozitivní neproblematický vztah k médiím, práci policie a k politické situaci.

Pátá část se zabývá mnohorozměrným škálováním, resp. věnuje se jedné z jeho technik, ALSCAL. Data vhodná k těmto analýzám bývají metrického rázu, např. kriminální statistiky nebo data získaná z analýz obsahu dokumentů nebo komunikace. Tato data, převedená do matic podobností nebo nepodobností, se pak zpracovávají mj. formou tzv. percepční mapy jako příklad 4 (Osobní zkušenosti české veřejnosti s užíváním drog). V ní jsou patrné vzdálenosti a dimenzionalita jednotlivých zkoumaných jevů, jako je užívání měkkých a tvrdých drog. V příkladu ukazujeme důležitý krok analýzy, jak zjistit optimální počet dimenzí, protože výzkumník si tento počet zadává v SPSS sám. V trojdimenzionálním řešení příkladu 4 se nám na základě stejné výchozí matice jako při faktorové analýze (Pearsonovy korelace) podařilo zjistit dodatečné poznatky. Kromě rozdílu charakterizujícího sklony k užívání tvrdých nebo měkkých drog je odkryta navíc dynamika vzniku náchylnosti k tvrdým drogám (podnětem se zdají být více konzumace tabáku a alkoholu než měkkých drog) a stupeň závislosti na přírodních nebo chemických (dodavatelských) látkách. Analýza tak může vygenerovat další hypotézy, které lze testovat za pomoci jiných vícerozměrných technik.

A to je také hlavní důvod, proč vznikl tento text. Nejen, aby ukázal možná řešení, která se nabízejí v kriminologickém výzkumu, ale také aby podnítil tvořivost a vybídl k novým otázkám a hypotézám.

Summary

Košťál, Jaroslav: SELECTED MULTIVARIATE STATISTICS METHODS (with a special focus on criminological research)

Multivariate statistics have only recently started to be used in Czech criminology while Western criminology has used them for two or even four decades. This is not just an issue of prestige as "multivariate" is in; this is also a matter of being right since one- or two-dimensional statistics (still too much relied upon) may lead to incorrect conclusions. They can, for instance, be the source of incorrectly accepted or rejected hypotheses about relationships between variables and therefore can lead to such conclusions and predictions of criminal phenomena that are not supported by real evidence.

This treatise consists of 5 parts. The first is devoted to a general characterization of multivariate statistics, the second part focuses on factor analysis, the third on cluster and correspondence analyses and the last two describe multidimensional scaling methods.

Multivariate approach is applied when we concurrently search for deeper relationships among three or more variables. Using multivariate rather than bivariate analyses generally allows to identity and describe phenomena in a more precise and at the same time more complex and reliable manner. This complexity also increases confidence on which we can base our judgment, e.g., findings with regard to criminal behaviour across time, or evaluations of anger management training in aggressive offenders.

Multivariate statistics is based on a) test techniques, then not always b) on comparative random distributions against which the significance of estimated parameters (means, counts, variances etc.) are tested and c) the concept of distances between the observed phenomena in (extrapolated) space. This text is limited to description of factor, cluster and correspondence analysis and to a certain degree also to multidimensional scaling, marginal mention of decision trees and the use of structural equation modeling. In addition,

we demonstrate application of these techniques to various data examples of Institute of Criminology (ICSP) projects and discuss possible pitfalls.

Factor analysis, addressed in the second part, is a statistical method which reduces the number of observed or manifest variables to a smaller number of latent (unobserved, construed) factors. These factors explain the variance of measured (manifest) variables. If multiplied and added, i.e., by the sum of their linear combinations, these factors equal the measured variables plus error terms (measurement error and the unexplained part of variance –the error of the model).

FA was invented by Spearman in the form of principal component analysis 100 years ago. Significant development of the method took place in the 1960s and it has since been applied across many fields of natural and social sciences. The point of departure of explorative FA is the correlation or covariance matrix of a set of the measured variables. This is the basis for statistical decision whether and which of these manifest variables are closely related to each other and therefore belong together or whether they belong to a different common factor.

Example 2 describes a survey focused on the opinion of Czech adults about the authorities involved in criminal proceedings. As expected, the factor analysis of responses to 19 proceeding items produced 4 factors: evaluations of work performed by the police, the courts, the state prosecution and by prison officers. The preliminary steps of factor analysis depend on correlation and partial correlation matrices. Examples are used to demonstrate FA requirements on input data, and desirable transformation of input correlation matrices into tetrachoric coefficients.

Following data transformation, factor analysis was executed for Example 3 with a two-factor solution and principal components (factors) dividing activities to socially or individual harmful categories. Strategies for estimating of optimal number of factors (principal components) are discussed with a stress on Horn´s parallel analysis which is an optimal yet rarely applied option. Shortcomings of other popular choices, such as Kaiser criterion and Pearson correlation matrix are discussed.

The difference between the principal component extraction method and the factor analysis method is explained. Example 5 focuses on the degree of consensus among young people in the Czech Republic with regard tomoral principles (codes of behaviour). Data of this type are not suitable for factor analysis treatment by the maximum likelihood extraction, despite the fact that the input matrix has been adapted and meets the requirements of continuous metric data. For that reason, data are analysed using an alternative acceptable method of principal axis factoring. The scales corresponding to three individual factors are also submitted to Cronbach's Alpha reliability analysis.

Example 6 deals with young people's opinion on the conduct of crime witnesses. It demonstrates factor analysis performed by maximum likelihood estimate and its advantages. for this particular case and notes on controversy which relates to the use of factor analysis versus principle component analysis. Some scholars strongly recommend oblique rotation and the maximum likelihood extraction method as the results are not dependent on the peculiarities of the concrete data set. We stand behind this recommendation and also suggest to plan for the statistical analyses at the time of the construction of research methods (in the questionnaire, expert scale).

The last segment of the factor analysis chapter highlights structural equation modeling (SEM), which is one of the most multifaceted and flexible methods of analysis of the interrelated variables. SEM is most frequently used in so-called confirmatory factor analysis, typically applied to test a current theory or structure revealed by exploratory factor analysis. Example 6 (Young people's opinion on the conduct of crime witnesses) a result of exploratory factor analysis, is submitted to CFA. The model derived from preceding FAs with was entered to the analysis proved itself fully compatible with the data and confirmed that respondents' views on witness behaviour depend on the context, i.e., whether it is the perperrator or the victim who is in the focus.

Cluster analysis (CA) is a subject of the third part. In example 8 (Opinion of young people in the Czech Republic on how much problematic certain negative phenomena are), we explain the "correlation profile" and the essentials of CA. In example 9 (Rating the degree of severity of negative social phenomena by the Czech adults), we illustrate the necessity to identify the clusters'content and subsequently to characterise them as actually existing groups of people with distinct inclinations and behaviour. Hierarchical cluster analysis is described by example 10 (Analysis of crime rates over the past decades). CA is suitable for metric data

sets that are small and easy to keep in mind. Some periods of criminal statistics may gradually (hierarchically) merge into units with similar pattern. Icicle charts and a dendrograms may identify and explain their change and common features.

K-means is a method of cluster analysis used for large samples with at least ordinal, and ideally metric data. Example 11 (European Values Study and justifiability of behaviour) shows how clusters emerge during gradual analysis (iteration steps) and how properly name them. Importantly, the number of clusters is decided by the researcher himself/herself and he/she must carefully consider and select those which are stable, balanced in number and which provide the best explanation of the subject matter. The suitability of clusters for international comparisons is discussed (particularly the same understanding of test batteries). It is highlighted, that the same configuration and metrics (invariance) of the battery across various countries may be tested by SEM.

Well identified clusters tend to meaningfully correlate with other research variables. Example 8 (Opinion of young people in the Czech Republic on how much problematic certain negative phenomena are) divided respondents into groups who voiced problems are serious, semi-serious and those who minimized their importance. These groups significantly differed by their criminal sensitivity indicated by 43 items. These groups also differed by their anamnestic data (e.g., grades of conduct at school).

The two-step cluster analysis (SPSS) provides a compromise between hierarchical analysis and k-means. Its particularly suitable for work with large samples and can process both categorical and metric data. Example 4 (Czech adults' experience with psychotropic substances) demonstrates a three-cluster solution where non-users were differentiated from people with casual, isolated usage and from habitual drug users.

A shorter segment is devoted to decision trees, the TREE procedure in SPSS, suitable for processing and clear presentation of any type of data.Tree procedure was applied in case of example 4 (Czech adults' experience with psychotropic substances) in order to assess the significance of social, cognitive (drug awareness) and socio-demographic variables (such as age, gender, education, income, size of village or town etc.) for drug usage differences. And whether such differences in drug user career are affected by awareness of the effects and countermeasures taken in this area.

Part four is dedicated to correspondence analysis. Its origins are linked with linguistics. Example 2 (The opinion of Czech adults about the authorities involved in criminal proceedings) was used to organize data which otherwise appeared to be chaotic. Graphic seizure provided by correspondence analysis allowed insight with respect to rating of individual authorities in various regions of the Czech Republic. Conveniently correspondence analysis works with any type of data, it does not require the data to comply with such assumptions as normality of distribution. Example 12 (Life principles of potentially problematic and unproblematic inhabitants of the Czech Republic) categorised respondents from the point of view of their social mal/adaptation. Using correspondence analysis we assigned those groups significantly different principles of conduct and opinions from negative extreme of a street gang moral code to the other pole in well socially adjusted attitudes to media, police and politics.

The concluding fifth part deals with multidimensional scaling, above all with the ALSCAL technique. ALSCAL is particularly suitable for metric data (e.g., criminal statistics) or data gained from content analysis of documents or correspondence. These data, transformed to proximity or dissimilarity matrices, are subsequently processed into so-called perception maps such as in example 4 (Czech adults' experience with psychotropic substances). The maps clarify the distances and dimensionality of individual phenomena, such as soft and hard drug usage. We demonstrate how to identify the optimum number of dimensions. This is important because it is the researcher who must determine their number and enter it into SPSS. The three-dimensional solution to example 4 revealed more than factor analysis in which a comparable initial matrix had been used (the Pearson correlation). In addition to the assessment of different characteristics irelated to usage of hard or soft drugs, ALSCAL revealed also the dynamics of proneness to hard drugs (e.g., that the facilitaing agents tend to be tobacco and alcohol rather than soft drugs), and the intensity and type of dependency. Analysis thus may generate further hypotheses which can be tested by other multivariate techniques. And this is also the main raison d'être for this text—not just to illustrate the solutions in criminological research but to encourage creativity, new questions and hypotheses.

Translated by: Presto

Seznam literatury

Al Ghoson, A. (2010). Decision Tree Induction & Clustering Techniques In SAS Enterprise Miner, SPSS Clementine, and IBM Intelligent Miner: A Comparative Analysis. *International Journal of Management & Information Systems 14*, 3. 57-70

Arbuckle, J.A. (2011). *Amos 20 User's Guide.* Amos Development Corporation. Chicago, Illinois

Barlow, J., Fisher, J. D., & Jones, D. (2012). *Systematic review of models of analysing significant harm. Research Report DFE-RR199*. Oxford University. https://www.education.gov.uk/publications/.../DFE-RR199.pdf

Benzécri, J.-P. (1973). *L'Analyse des Données. Volume II. L'Analyse des Correspondances.* Paris, France: Dunod.

Blatníková, Š., & Zeman, P. (2012). *Násilná sexuální kriminalita v ČR. Předběžné výstupy ke grantovému projektu.* Praha: IKSP.

Bourdieu, P. (1984). *Distinction: A Social Critique of the Judgement of Taste*. London: Routledge.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees.* Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software.

Brewer, S., Iannacchione, B. M., & Pantaleo, K. (2010). Predicting bullying: Logistic regression or decision trees? Paper presented at the *Annual Meeting of the American Society of Criminology*, San Francisco, CA.

Brown, J. (2004). Managing the Transition from Institution to Community: A Canadian Parole Officer Perspective on the Needs of Newly Released Federal Offenders. *Western Criminology Review 5* (2), 97-107

Bühler, K.H., & Bardeleben, H. (2008). Heuristic cluster analysis of alcoholics according to biographic and personality features. *Addiction Research and Theory 16* (5), 453–473.

Cattell, R. B. (1943). The description of personality: Basic traits resolved into clusters. *Journal of Abnormal and Social Psychology, 38*, 476-506.

Coombs, C.H. (1964). *A Theory of Data.* New York: Wiley.

Costello A.B., & Osborne, J.W. (2005). Best Practices in Exploratory Factor Analysis: Four Recommendations for Getting the Most From Your Analysis, *Practical Assessment, Research & Evaluation,* 10, 7. 1-9. Available online: http://pareonline.net/getvn.asp?v=10&n=7

Coxon, A. P. M. (2004). Multidimensional Scaling In M.S. Lewis-Beck, A. Bryman, T. F. Liao, *The Sage Encyclopedia of Social Science Research Methods*, Thousand Oaks, CA, Sage.

Dawes, R. N., & Tversky, A. (1989). Coombs obituary. *American Psychologist, 44* (11), 1415-1416.

Dempster, A. P. N., Laird, M., & Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological), 39*(1), 1-38.

Dobash, R.P., Cavanagh, K., Smith, D., & Medina-Ariza, J. (2007). Onset of Offending and Life Course Among Men Convicted of Murder. *Homicide Studies 11*, 243-271.

EISIC (2011). European Intelligence and Security Informatics Conference. Athens. http://www.eisic.eu/eisic2011/pbrantingham.aspx

Everitt, B.S., & Dunn, G. (1991) *Applied Multivariate Data Analysis*. London: Arnold.
EVS, European Value Survey, http://www.europeanvaluesstudy.eu/
Field, A. (2000). Postgraduate Statistics: Cluster Analysis.
< *www.statisticshell.com/docs/**cluster**.pdf* >
Fisher, C., & Salfati, G. C. (2007). Classifying Bias: Utilizing Multidimensional Scaling Analytic Techniques to Examine Bias-motivated Homicides. Paper presented at the *Annual Meeting of the ASC, Atlanta Marriott Marquis*, Atlanta, Georgia.
Fiske, D. W. (1949). Consistency of the factorial structures of personality ratings from different sources. *Journal of Abnormal Social Psychology, 44*, 329-344.
Garson, G. D. (2012a). *Multidimensional Scaling (Statistical Associates Blue Book Series)* Kindle Edition. North Carolina State University.
Garson, G. D. (2012b), *Factor Analysis, Statistical Associates Publishing, Blue Book Series.* Latest update 4.1.2012.
Garson, G. D. (2011). Multivariate GLM, MANOVA, and MANCOVA. (Statistical Associates Blue Book Series) Kindle Edition. North Carolina State University.
Gepp, A. Wilson, J. H., Kumar, K., & Bhattacharya, S. (2012). A Comparative Analysis of Decision Trees Vis-a-vis Other Computational Data Mining Techniques in Automotive Insurance Fraud Detection. *Journal of Data Science 10*, 537-561.
Gerritsen, C., & Hoogendoorn, M. (2012). Avoidance of norm violation in multi-agent oranizations. www.few.vu.nl/.../paper-ECMS-norm-violation.
Giguere, G. (2007). Collecting and analyzig the data in multidimensional scaling experiments: A guide for psychologists using SPSS. *Tutorials for Quantitative Methods in Psychology*, *2*(1), 26-37.
Goodwill, A. M., & Alison, L. J. (2007). When is profiling possible? Offense planning and aggression as moderators in predicting offender age from victim age in stranger rape. *Behav Sci Law 25*(6), 823-40.
Goodwill, A. M., Alison, L. J., & Humann, M. (2009). Multidimensional scaling and the analysis of sexual offence behaviour: A Reply to Sturidsson et al. *Psychology, Crime & Law*, *15*, 517-524.
Greenacre, M., & Jörg, B. (Eds.) (2006). *Multiple Correspondence Analysis and Related Methods.* London: Chapman & Hall/CRC.
Griffith, E. (2007). Geographic information systems (GIS) and spatial analysis. In M. Williams & W. P. Vogt (Eds.), *Innovation in Social Research Methods* (pp. 442-464) London: Sage.
Hagenaars, J. A., & McCutcheon, A. L. (2002). *Applied Latent Class Analysis.* Cambridge: Cambridge University Press.
Hebák, P., & Hustopecký, J. (1987). *Vícerozměrné statistické metody s aplikacemi.* Praha: SNTL Alfa.
Hebák, P., Hustopecký, J., Pecáková, I., Průša, M., Řezanková, H., Svobodová, A., & Vlach, P. (2005). *Vícerozměrné statistické metody (3).* (pp. 120-144). Praha: Informatorium.
Hendl, J. (2004). *Přehled statistických metod zpracování dat*. Praha: Portál
Hendrix, C., & Scimone, B. (2007). *Structure and Offender Behavior in Swedish Rape Cases: A Multidimensional Scaling Approach*. University essay from Lunds universitet/ Institutionen för psykologi.
Hindelang, M. J., & Weis, J. G. (1972) Personality and self-reported deliquency: An application of cluster analysis. *Criminology, 10* (3), 268-294.
Hlavsa, T. (2006). Metody shlukové analýzy. Aplikované kvantitativní metody pro zemědělskou praxi. Acta Universitatis Bohemiae Meridionales. Skripta.
Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30*, 179-185.

IBM SPSS Statistics Base 20 (2001). IBM Corporation, Chicago, Illinois 122-145
John, O. P., Naumann, L. P., & Soto, C. J. (2008). Paradigm Shift to the Integrative Big-Five Trait Taxonomy: History, Measurement, and Conceptual Issues. In O. P. John, R. W. Robins, & L. A. Pervin (Eds.), *Handbook of personality: Theory and research* (pp. 114-158). New York, NY: Guilford Press.
John, O. P. (1999). The Big Five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research*. 2nd ed., (pp. 102-138). New York: Guildford.
Juon, H.-S., Doherty, E. E., & Ensminger, M. E. (2006). Childhood Behavior and Adult Criminality: Cluster Analysis in a Prospective Study of African Americans. *Journal of Quantitative Criminology 22*, 193–214.
Choi, K.-S. (2008). *Structural Equation modeling Assesment of Key Causal Factors in Computer Crime Victimization*. Ann Arbor: ProQuest Information and Learning Company.
Kaiser, H. (1960). The Application of Electronic Computers to Factor Analysis. *Educational and Psychological Measurement, 20,* 141-151.
Kahounová, J. (1994). *Měření podobnosti struktur.* Skripta. Praha: VŠE.
Kaufman, L., & Rousseeuw, P. I. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, New York.
Kelbel, J. & Šilhán, D. (2002). Shluková analýza. Praha: s.n., Dostupné z WWW: < http://gerstner.felk.cvut.cz/biolab/X33BMI/slides/KMeans.pdf>
Kline, R.B. (2004). Beyond Significance Testing. American Psychological Association, Washington
Kocsis, R. N. (2010). *Criminal Profiling: Principles and Practice*. New Jersey: Humana Press Inc.
Kruskal, J. B. (1964a). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika, 29*, 1-28.
Kruskal, J. B. (1964b). Nonmetric multidimensional scaling: A numerical method. *Psychometrika, 29*, 115-129.
Kruskal, J. B., & Wish, M. (1978). *Multidimensional Scaling, Sage University Papers on Quantitative Applications in the Social Sciences*, 07011. Beverly Hills: Sage Publications.
Ledesma, R.D., & Valero Mora, P. (2007). Determining the Number of Factors to Retain in EFA: an easy-to-use computer program for carrying out Parallel Analysis. Practical Assessment, *Reasearch and Evaluation, 12*, 2, Available online: http://pareonline.net/getvn.asp?v=12&n=2
Ledesma, R.D.&Molina, J.G. (2009). Classical item and test analysis with graphics: The ViSta-CITA program. Behavioral Research methods, 41 (4), 1161-1168; http://www.mdp.edu.ar/psicologia/vista
Lukasová, A., & Šarmanová, J. (1985). *Metody shlukové analýzy*. Praha: SNTL.
MacQueen, J. B. (1967). Some Methods for classification and Analysis of Multivariate Observations, *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, University of California Press, 1, 281-297
Mardia, K. V., Kent, J. T., & Bibby, J. M. (1979). Multivariate Analysis. London: Academic Press
Marešová, A., Cejp,M., Martinková, M., Scheinost, M., & Vlach, J. (2011). Analýza trendů kriminality v roce 2010, Praha: IKSP.
Meloun, M., & Militký, J. (1994). *Statistické zpracování experimentálních dat v chemometrii, biometrii, ekonometrii a v dalších oborech přírodních, technických a společenských věd*, Praha: PLUS.

Mislevy, R. J., & Bock, R. D. (1990). *BILOG 3: Item analysis and test scoring with binary logistic models*. Mooresville: Scientific Software.
Mulaik, S. A. (1990). Blurring the Distinctions between Component Analysis and Common Factor-Analysis. *Multivariate Behavioral Research, 25* (1), 53-59.
Mun, E. Y., Windle, M., & Schainker, L. M. (2008). Model-based cluster analysis approach to adolescent problem behaviors and young adult outcomes. *Development and Psychopathology 20,* 291-318.
Neema, I., & Böhning, D. (2010). Improved methods for surveying and monitoring crimes through likelihood based cluster analysis. *International Journal of Criminology and Sociological Theory 3* (2), 477-495.
Norušis, M. J. (2002). *Cluster Analysis*, Chapter 16. IBM SPSS Statistics Guides: Straight Talk about Data Analysis and IBM SPSS Statistics. www.norusis.com/pdf/SPC_v13.pdf
Orme, B. (2008). *CCEA v3. Software for Convergent ClusterEnsemble Analysis*. Sawtooth Software, Wahington.
Osborne, J. W., & Costello A. B. (2009). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Pan-Pacific Management Review 12* (2), 131-146.
Product management group (2004). *Statistical Methods in Criminological Sciences using Systat.* www.intesoft.com/produits/.../StatCrimino.pdf
Rees-Jones, I. (2007). Correspondence analysis: A case for methodological pluralism? In M. Williams & W. P. Vogt (Eds.), *Innovation in Social Research methods* (pp. 139-149). London: Sage.
Richardson, A. (2009). *Visualising sentencing space: correspondence analysis of a criminology data set.* Canberra: Faculty of ISE. Quant. Researchers Club.
Sivic, J., Russell, B. C., Zisserman, A., Freeman, W. T., & Efros, A. A. (2008). Unsupervised Discovery of Visual Object Class Hierarchies. www.di.ens.fr/~josef/publications/sivic08.pdf
Sneath, P. H. A., & Sokal, R. R. (1973). *Numerical taxonomy*. San Francisco: W. H. Freeman & Company.
Sokal, R.R. (1977). Clustering and Classification: Background and current directions. In J.van Ryzin (Ed.), *Classification and Clustering* (1-15). New York: Academic Press,
Spaans, M., Barendregt, M., Muller, E., de Beursa, E., Nijmanc, H., & Rinn, T. (2009). MMPI profiles of males accused of severe crimes: a cluster analysis. *Psychology, Crime & Law 15* (5), 441-450.
Spearman, C. (1904). General intelligence, objectively determined·and measured. *American Journal of Psychology 15*, 201-293.
Škaloudová, A. (2012), *Faktorová analýza*, web CUNI. http://userweb.pedf.cuni.cz/kpsp/skalouda/fa/
Smith, W. R., Smith, D. R., & Norma, E. (1986). The multidimensionality of crime: A comparison of techniques for scaling delinquent careers. *Journal of Quantitative Criminology 2* (4), 329-353.
Tabachnick, B. G., & Fidell, L. S. (2001). *Using Multivariate Statistics*. Boston: Allyn and Bacon.
Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications*. Washington, DC.: American Psychological Association.
Thurstone, L.L. (1934). The Vectors of the Mind. Address of the president before the American Psychological Association, Chicago meeting, September, 1933. *Psychological Review 41*, 1-32.

Trávníčková I., & Zeman P. (2010). *Kriminální kariéra pachatelů drogové kriminality*, Praha: IKSP.
Tucker, W. H. (2009). *The Cattell Controversy: Race, Science, and Ideology*. Urbana, IL: University of Illinois Press.
Tupes, E. C., & Christal, R. E. (1961). *Recurrent Personality Factors Based on Trait Ratings (ASD-TR-61-97).* Lackland Air Force Base, TX: Aeronautical Systems Division, Personnel Laboratory.
Tryon, R. C. (1939). *Cluster analysis.* Ann Arbor: Edwards Brothers.
Überla, K. (1974). *Faktorová analýza*. Bratislava: ALFA.
Varese, F. (2012). How Mafias Take Advantage of Globalization. *British Journal of Criminology 52* (2), 235-253.
Večerka, K. a kol. (2011). Výzkum potencionální kriminality mládeže. Praha: IKSP.
Velicer, W. F., & Fava, J. L. (1998). Effects of variable and subject sampling on factor pattern recovery. *Psychological Methods 3* (2), 231-251.
Velicer, W. F., & Jackson, D. N. (1990). Component Analysis Versus Common Factor-Analysis – Some Further Observations. *Multivariate Behavioral Research 25* (1), 97-114.
Walker, J. T., & Maddan, S. L. (2005). *Statistics in Criminology and Criminal Justice: Analysis and Interpretation* 2nd Edition. 301-323. Boston: Jones and Bartlett.
Widaman, K. F. (1990). Bias in Pattern Loadings Represented by Common Factor-Analysis and Component Analysis. *Multivariate Behavioral Research 25* (1), 89-95.
Wuensch, K. (2012). *Dr. Karl L. Wuensch's SPSS-Data Page*. http://core.ecu.edu/psyc/wuenschk/spss/spss-Data.htm
Zeman, P., Trávníčková, I., & Štefunková, M. (2011a). *Vybrané aspekty drogové problematiky z pohledu občanů*. Praha: IKSP.
Zeman, P. a kol. (2011b). *Názory a postoje občanů v oblasti trestní politiky*, Praha: IKSP

V ediční řadě Vybrané metody kriminologického výzkumu dosud vyšly monografie:

Cejp, M.: *Aplikace výzkumných metod a technik v kriminologii. Obecná část.* (Vybrané metody kriminologického výzkumu - svazek 1). Praha: IKSP 2011. ISBN 978-80-7338-108-0

Marešová, A.: *Resortní statistiky - základní zdroj informací o kriminalitě v České republice.* (Vybrané metody kriminologického výzkumu - svazek 2). Praha: IKSP 2011. ISBN 978-80-7338-110-3

Blatníková, Š.: *Aplikace klinických a testových metod v kriminologickém výzkumu.* (Vybrané metody kriminologického výzkumu - svazek 3). Praha: IKSP 2011. ISBN 978-80-7338-109-7

VYBRANÉ METODY VÍCEROZMĚRNÉ STATISTIKY

(se zvláštním zaměřením na kriminologický výzkum)

Autor:	Jaroslav Košťál
Vydavatel:	Institut pro kriminologii a sociální prevenci Nám. 14. října 12, 150 21 Praha 5
Určeno:	Pro odbornou veřejnost
Tiskárna:	Vydavatelství KUFR s.r.o. Naskové 3, Praha 5
Do tisku:	únor 2013
Edice:	Studie
Řada:	Vybrané metody kriminologického výzkumu
Svazek:	čtvrtý
Vydání:	první
Náklad:	100 ks

www.kriminologie.cz

ISBN 978-80-7338-128-8