SHLUKOVÁ ANALÝZA DOMÁCNOSTÍ CHARAKTERIZOVANÝCH KATEGORIÁLNÍMI UKAZATELI

Hana Řezanková, Tomáš Löster

Úvod

Při šetření životních podmínek domácností jsou uplatňovány různé způsoby klasifikace domácností (podle vzdělání, pracovní aktivity, pracovní intenzity, klasifikace podle EU či OECD). V tomto článku naznačíme možnosti klasifikace domácností podle jejich finančních možností, a to na základě shlukové analýzy s využitím kategoriálních ukazatelů.

Český statistický úřad provádí od roku 2005 šetření „Životní podmínky“, což je národní modul šetření EU-SILC (European Union – Statistics on Income and Living Conditions). Jeho cílem je získat přehled o stavu a vývoji sociální situace obyvatelstva. Společný rámec tohoto šetření v evropských zemích upravuje novelizace Nařízení (EC) 1177/2003 a navazující prováděcí nařízení Evropské komise (viz metodické vysvětlivky publikované na webové stránce http://www.czso.cz).

Šetření je zaměřeno jednak na příjmová rozdělení jednotlivých typů domácností, jednak na způsob, kvalitu a finanční náročnost bydlení a vybavení domácností předměty dlouhodobé spotřeby. Poskytuje také údaje o pracovních, hmotných a zdravotních podmínkách dospělých osob v domácnosti. Získaná data slouží jako podklady pro tvorbu a hodnocení sociální politiky státu a pro rozhodování o alokaci finančních prostředků Evropské unie, které mají pomoci při odstraňování sociálních problémů.

Uvedeným šetřením je získáno velké množství údajů, v nichž lze zkoumat různé souvislosti a závislosti, a to buď ověřovat předpokládané, nebo hledat nové. V příspěvku se zaměřujeme na analýzu dat získaných od domácností v rámci šetření „Životní podmínky 2008“. Datový soubor zakoupený od Českého statistického úřadu obsahuje údaje o 11 294 domácnostech. Našim cílem bylo navrhnout jednak postup při stanovení počtu skupin domácností vytvořených na základě vybraných kategoriálních ukazatelů, jednak způsob charakterizování těchto skupin.

K dosažení tohoto cíle byla využita dvoukroková shluková analýza ve statistickém programovém systému IBM SPSS. Ve druhé části uvádíme princip použité metody, způsoby hodnocení výsledných shluků a způsoby stanovení optimálního počtu shluků. Třetí část obsahuje ukázky praktických aplikací s využitím vybraných kategoriálních ukazatelů z výše uvedeného šetření. V závěru naznačujeme další možné přístupy k hodnocení shluků objektů, které jsou charakterizovány kategoriálními proměnnými.

1. Shluková analýza pro kategoriální proměnné

Jedním z prostředků pro zkoumání vztahů ve vícerozměrných datových souborech je shluková analýza, viz [4]. Programové prostředky pro shlukování objektů charakterizovaných nominálními či ordinálními proměnných se praxi vyskytují mnohem méně ve srovnání s prostředky určenými pro kvantitativní data, i když v literatuře již byla pro řešení této úlohy navržena řada přístupů, přehled je uveden v [5]. Pokud jde navíc o rozsáhlý datový soubor, je situace ještě obtížnější. Základním způsobem, který lze aplikovat bez převádění proměnných na skupinu binárních, je vytvoření matice nepodobností pro všechny dvojice objektů, a to na základě koeficientu neshody, a využití této matice v hierarchické shlukové analýze. Tento způsob však v běžných statistických programových systémech není možné realizovat pro rozsáhlejší datové soubory, jakým je například výše uvedených soubor přesahující 10 tisíc objektů.

Příkladem algoritmu, který spojuje možnosti shlukování jak v případě nominálních proměnných (resp. kombinace proměnných různých typů), tak v případě rozsáhlých datových souborů, je dvoukroková shluková analýza implementovaná v systému SPSS od verze 11.5 (nyní systém IBM SPSS Statistics, poslední je verze 21). Algoritmus je určen pro kombinaci proměnných nominálních a kvantitativních spojitých (objekty mohou být samozřejmě charakterizovány též buď pouze nominálními, nebo pouze kvantitativními proměnnými). Vychází z prací [7] a [2] a je realizován ve dvou fázích.

V první se objekty shlukují do malých shluků (podshluků), jejichž počet je podstatně menší než počet objektů původního souboru. Je aplikováno inkrementální shlukování, kdy se hodnotí objekty v pořadí daném datovým souborem. První objekt je základem prvního (na počátku jednoprvkového) shluku. U dalších objektů se posuzuje, zda mohou být zařazeny do již vytvořeného shluku, nebo zda bude vytvořen nový shluk. Ve druhé fázi algoritmu je každý vytvořený podshluk přiřazen do některého z konečných shluků, jejichž počet je předem stanoven. Protože počet podshluků je podstatně menší než počet objektů původního souboru, mohou být již využity tradiční metody shlukování. V systému SPSS se tato fáze realizuje pomocí hierarchické shlukové analýzy, podrobněji viz [6].

V obou fázích se používá stejná míra nepodobnosti. V SPSS jsou implementovány dvě míry, z nichž euklidovská je použitelná pouze pro kvantitativní proměnné. Druhou je věrohodnostní (log-likelihood) míra, která je vhodná pro kategoriální proměnné, může však být použita též pro proměnné kvantitativní spojité, případně pro datový soubor obsahující proměnné obou typů. Vzdálenost mezi dvěma shluky zohledňuje pokles věrohodnostní míry, jenž nastává při spojení dvou shluků do jednoho, tj. vzdálenost mezi *h*-tým a *h´*-tým shlukem je definována jako

$$D_{hh'} = \xi_{\langle h,\, h'\rangle} - (\xi_h + \xi_{h'}), \qquad (1)$$

přičemž symbol $<h, h'>$ označuje shluk vytvořený spojením objektů z *h*-tého a *h´*-tého shluku a

$$\xi_h = n_h \left(\sum_{l=1}^{m^{(1)}} \frac{1}{2} \ln(s_l^2 + s_{hl}^2) + \sum_{l=1}^{m^{(2)}} H_{hl} \right), \qquad (2)$$

kde n_h je počet objektů v *h*-tém shluku, $m^{(1)}$ je počet kvantitativních spojitých proměnných, $m^{(2)}$ je počet kategoriálních proměnných, s_l^2 je výběrový rozptyl *l*-té spojité proměnné, s_{hl}^2 je výběrový rozptyl *l*-té spojité proměnné v *h*-tém shluku a H_{hl} je entropie *l*-té spojité proměnné v *h*-tém shluku, daná vztahem

$$H_{hl} = -\sum_{u=1}^{K_l} \frac{n_{hlu}}{n_h} \ln \frac{n_{hlu}}{n_h}, \qquad (3)$$

kde K_l je počet kategorií *l*-té kategoriální proměnné a n_{hlu} představuje četnost *u*-té kategorie *l*-té kategoriální proměnné v *h*-tém shluku. Je tedy zkoumán rozdíl mezi variabilitou shluku vzniklého spojením *h*-tého a *h´*-tého shluku a součtem variabilit těchto jednotlivých shluků, přičemž variabilita každého shluku je vážena příslušným počtem objektů.

Ačkoliv metody shlukové analýzy jsou používány několik desítek let, nejsou stále zcela vyřešeny některé základní problémy, kterým je například stanovení počtu shluků. K určování optimálního počtu shluků byla v literatuře navržena řada indexů, obsáhlý přehled je součástí knihy [1], ovšem až na výjimky se týkají shlukování objektů charakterizovaných pouze kvantitativními proměnnými.

V systému SPSS mohou být pro datové soubory s kategoriálními proměnnými či proměnnými různých typů využity Schwarzovo bayesovské informační kritérium (BIC) a Akaikeho informační kritérium (AIC) počítané pro všechny počty shluků ze zadaného intervalu. První je dáno vztahem

$$I_{\mathrm{BIC}}(k) = 2\sum_{h=1}^{k} \xi_h + k\left(2m^{(1)} + \sum_{l=1}^{m^{(2)}} (K_l - 1) \right) \ln(n), \qquad (4)$$

kde *k* je počet shluků, druhé vztahem

$$I_{\mathrm{AIC}}(k) = 2\sum_{h=1}^{k} \xi_h + 2k\left(2m^{(1)} + \sum_{l=1}^{m^{(2)}} (K_l - 1) \right). \qquad (5)$$

Z daných vzorců lze odvodit, že informační kritéria jsou založena na vnitroshlukové variabilitě, což je průměrná variabilita uvnitř jednotlivých shluků. Tato vnitroshluková variabilita je násobená hodnotou *2n*, kde *n* je celkový počet objektů. Protože se zvyšujícím se počtem shluků jsou shluky menší a více homogenní, vnitroshluková variabilita se snižuje. Z toho důvodu je ve vzorcích (4) a (5) aplikována penalizace, která jednak znevýhodňuje vyšší počet shluků, jednak zohledňuje počet kvantitativních proměnných a počet kategorií nominálních proměnných. Za optimální počet shluků lze považovat lokální minimum ze vypočtených hodnot

daného kritéria ze zadaného intervalu. V systému SPSS je však tato hodnota pouze pomocná pro sérii dalších výpočtů.

Ve vzorci (4) a (5) je součet počtu kvantitativních proměnných a počtu kategorií nominálních proměnných násobený počtem shluků. Tento součin je ve vzorci (4) násoben přirozeným logaritmem počtu objektů, ve vzorci (5) násoben hodnotou 2. Protože ani přes tuto penalizaci nemusí být v zadaném intervalu nalezeno minimum, používají se v některých jiných programových systémech (například v systému Latent GOLD, který provádí shlukovou analýzu na základě pravděpodobnostních modelů) ještě některá další kritéria, například AIC3, v němž je počet parametrů modelu (pro nominální proměnné odpovídající zde uvedenému součinu) násoben hodnotou 3, a CAIC, v němž je tento součin násoben přirozeným logaritmem počtu objektů zvýšeným o hodnotu 1.

V systému SPSS se pro stanovení optimálního počtu shluků ze zadaného intervalu nejprve vypočtou diference mezi hodnotami kritérií pro po sobě následující počty shluků. Dále bude tento postup naznačen pouze pro kritérium BIC, tj.

$$dI_{\mathrm{BIC}}(k) = I_{\mathrm{BIC}}(k) - I_{\mathrm{BIC}}(k+1). \qquad (6)$$

Za předpokladu, že $dI_{BIC}(1) > 0$, se vypočtou poměry

$$R_1(k) = \frac{dI_{\mathrm{BIC}}(k)}{dI_{\mathrm{BIC}}(1)}. \qquad (7)$$

Dále se stanoví hodnota K jako

$$K = arg_k min R_1(k) \text{ pro } R_1(k) < 0{,}04. \qquad (8)$$

Tato hodnota je základem pro výpočty poměrů

$$R_2(k) = \frac{D_{\min}(P_k)}{D_{\min}(P_{k+1})} \text{ pro } k = 2, \ldots, K, \qquad (9)$$

kde $D_{min}(P_k)$ je vzdálenost dvou nejbližších shluků při rozdělení objektů do k shluků. Ze získaných hodnot $R_2(k)$ se vyberou dvě největší. Pokud je největší hodnota více než 1,15krát větší než druhá největší, pak je jako optimální počet shluků určen ten, pro který byla dosažena největší hodnota $R_2(k)$. V opačném případě se ze dvou počtu shluků, pro které byly vypočteny dvě největší hodnoty, vybere jako optimální větší počet shluků.

Jiným kritériem, které hodnotí získané shluky, je *obrysový koeficient*, jehož detailní popis je uveden v knize [3] a který je již řadu let implementován v programovém systém S-PLUS. Uvedeme zde princip výpočtu tohoto koeficientu, neboť v SPSS je nově nyní také začleněn. Jeho hodnoty se však nezobrazují číselně, pouze je pomocí nich vytvořen graf vyjadřující kvalitu vytvořených shluků.

Označme si vektor charakterizující i-tý objekt symbolem $\boldsymbol{x_i}$ (dále pouze objekt), $i = 1, 2, \ldots, n$, a h-tý shluk symbolem C_h, $h = 1, \ldots, k$. Pro $\boldsymbol{x_i} \in C_h$ vypočítáme hodnoty ψ_i na základě průměrných vzdáleností sledovaného objektu s ostatními objekty v jednotlivých shlucích, a to

$$\psi_i = \frac{\mu_i - \eta_i}{\max\{\eta_i, \mu_i\}}, \qquad (10)$$

kde η_i je průměrná vzdálenost i-tého objektu od ostatních objektů nacházejících se ve stejném shluku a μ_i je minimum z průměrných vzdáleností i-tého objektu od objektů každého dalšího shluku, tj.

$$\eta_i = \frac{\sum_{j \in C_g} D_{ij}}{n_g - 1} \qquad (11)$$

a

$$\mu_i = \min_{h \neq g} \left(\frac{\sum_{j \in C_h} D_{ij}}{n_h} \right). \qquad (12)$$

Obrysový koeficient je pak stanoven jako průměrná hodnota z hodnot ψ_i, tj.

$$\psi = \frac{\sum_{i=1}^{n} \psi_i}{n}. \qquad (13)$$

Může nabýt hodnoty od -1 do 1. Pokud průměrná vzdálenost i-tého objektu od ostatních objektů nacházejících se ve stejném shluku je menší než průměrná vzdálenost s objekty z libovolného jiného shluku, pak obrysový koeficient nabývá kladných hodnot. Čím vyšší je jeho hodnota, tím jsou shluky kompaktnější. V systému SPSS je pro hodnoty nižší než 0,2 rozdělení objektů do shluků označováno jako chabé (poor), pro hodnoty od 0,2 do 0,5 jako uspokojivé (fair) a pro hodnoty vyšší než 0,5 jako dobré (good). Nejvyšší hodnota obrysového

koeficientu pro počty shluků ze zadaného intervalu může sloužit také pro určení optimálního počtu shluků.

Jiným hodnocením výsledků shlukování je určování vlivu proměnných na vytvoření shluků. V systému SPSS je stanovována *důležitost proměnné*, která indikuje, jak dobře mohou být pomocí dané proměnné rozlišeny různé shluky. Stanovení se provádí na základě zjištěné meziskupinové variability. Praktická aplikace bude uvedena v následující části.

2. Shlukování domácností na základě vybraných ukazatelů

Na základě ukazatelů o příjmech, typu, velikosti a vybavení bytu a na základě řady dalších charakteristik týkající se životní úrovně bychom se mohli pokusit vytvořit skupiny domácností, které nemusí být v souladu se stanovenými druhy domácností podle vzdělání, pracovní aktivity, pracovní intenzity apod. U kategoriálních proměnných je vhodné zjistit, jak jsou zastoupeny jednotlivé kategorie. Pokud u některého ukazatele výrazně převažuje pouze jedna kategorie (např. u vybavení pračkou, barevným televizorem či telefonem z více než 96 % převažuje odpověď, že domácnost daný předmět dlouhodobé spotřeby vlastní), nebude užitečné tento ukazatel do analýzy zařadit.

Pro shlukování domácností podle vybavenosti předměty dlouhodobé spotřeby je vhodné zařadit pouze vybavením autem a počítačem. U těchto ukazatelů je dostatečné zastoupení ve všech třech kategoriích, kterými jsou 1 (má vlastní), 2 (nemá – nemůže si dovolit) a 3 (nemá z jiných důvodů/nechce). Protože existuje teoreticky celkem devět možných kombinací těchto kategorií, podle nichž lze vytvořit devět shluků, jde o velmi jednoduchou úlohu. Výsledné shluky jsou homogenní, nelze je dále z hlediska použitých ukazatelů rozdělovat. Četnosti jednotlivých kombinací jsou uvedeny v tabulce 1.

Tab. 1: Četnosti kombinací kategorií proměnných počítač a auto

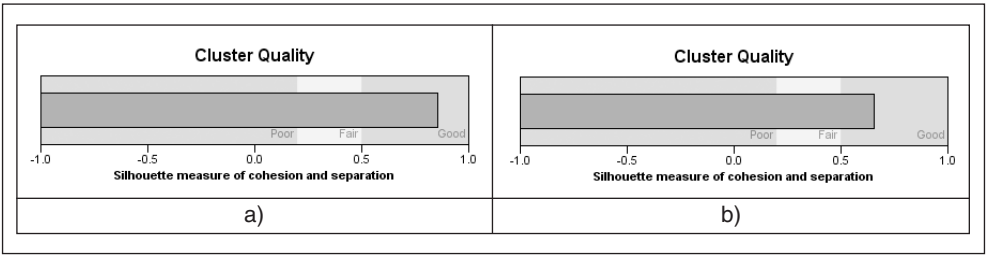
		Auto			Celkem
		má vlastní	nemá – nemůže si dovolit	nemá z jiných důvodů/nechce	
Počítač	má vlastní	4718	385	560	5663
	nemá – nemůže si dovolit	275	511	86	872
	nemá z jiných důvodů/nechce	1940	421	2398	4759
Celkem		6933	1317	3044	11294

Zdroj: ČSÚ (šetření SILC, ČR, 2008), vlastní výpočty v SPSS

V tomto případě bychom mohli bez použití shlukové analýzy vytvořit všechny kombinace daných kategorií a charakterizovat domácnosti, které náleží příslušným skupinám. Pro více kategoriálních proměnných, případně pro kombinace kategoriálních a kvantitativních spojitých, by však takový postup byl poměrně složitý. Úlohou shlukové analýzy je v takovém případě především určit vhodný počet shluků.

Provedli jsme dvoukrokovou shlukovou analýzu v systému SPSS s nastavením vyhledání optimálního počtu shluků pro maximální hodnotu 9 (tj. z intervalu od 1 do 9) s využitím Schwarzova bayesovského informačního kritérium (BIC). Systém vyhodnotil pomocí výše popsaného postupu využívajícího vzorce (6) až (9) jako optimální počet osm shluků, což již je poměrně hodně pro vhodnou interpretaci získaných skupin. Máme možnost snižovat horní hranici intervalu, čímž získáváme jako optimální počty shluků hodnoty 6 (pro 7 a 6 shluků stanovených jako horní interval), 4 (pro maxima 5 a 4 shluků) a 2 (pro maxima 3 a 2 shluky). K největší změně v hodnotách obrysového koeficientu došlo při porovnání čtyř a tří shluků (v porovnání se změnami hodnot pro sousední počty shluků), viz obrázek 1 (pro čtyři shluky je hodnota přibližně 0,8, zatímco pro tři shluky 0,65; pro pět shluků by to bylo 0,88 a pro dva shluky hodnota 0,63), proto bychom jako vhodný počet mohli uvažovat čtyři shluky.

Obr. 1: Grafické znázornění obrysového koeficientu (auto, počítač) pro a) 4 shluky a b) 3 shluky



Zdroj: vlastní výpočty v SPSS

Charakterizujme tedy čtyři shluky získané pomocí dvoukrokové shlukové analýzy. Na obrázku 2 jsou pro shluky seřazené podle velikosti (neodpovídá číselnému označení shluků) uvedeny jednak absolutní četnosti a odpovídající procentní zastoupení objektů vzhledem k celkovému počtu objektů (řádek *Size*), jednak převažující zastoupení kategorií jednotlivých proměnných (řádek *Inputs*). Tři shluky jsou tedy tvořeny pouze jednou kombinací dvou kategorií sledovaných dvou proměnných a jeden shluk zahrnuje více kombinací.

Systém SPSS umožňuje v datovém editoru každý řádek s hodnotami charakterizujícími určitý objekt doplnit o hodnotu vyjadřující příslušnost k jednomu ze stanoveného počtu shluků. Na základě tohoto přiřazení pak můžeme shluky charakterizovat i pomocí jiných proměnných, než na základě kterých byla provedena shluková analýza.

Obr. 2: Charakteristika shluků z hlediska proměnných auto, počítač (4 shluky)

Size	41.8% (4718)	21.2% (2398)	19.8% (2238)	17.2% (1940)
Inputs	auto má vlastní (100.0%)	auto nemá z jiných důvodů/nechce (100.0%)	auto nemá - nemůže si dovolit (58.8%)	auto má vlastní (100.0%)
	počítač má vlastní (100.0%)	počítač nemá z jiných důvodů/nechce (100.0%)	počítač má vlastní (42.2%)	počítač nemá z jiných důvodů/nechce (100.0%)

Zdroj: vlastní výpočty v SPSS

U analyzovaného souboru byly v jednotlivých shlucích sledovány relativní četnosti kategorií následujících ukazatelů: typ obce, oblast (stupeň urbanizace), sociální skupina osoby v čele, počet ekonomicky aktivních (pracujících) členů, počet nezaměstnaných, druh domácnosti podle pracovní aktivity, druh domácnosti podle vzdělání. Tyto četnosti byly získány tak, že pro každý ukazatel byla sestavena kontingenční tabulka vyjadřující vztah ukazatele k rozdělení objektů do shluků. Pokud se procentní zastoupení u některé kategorie zajímavě lišilo od procentního zastoupení v celém souboru, je uvedeno v tabulce 2, a to vyšší hodnoty normálním typem písma a nižší hodnoty kurzívou.

Tab. 2: Charakteristika shluků pomocí vybraných kategorií vybraných ukazatelů

	Významné kategorie	Relativní četnost v rámci shluku (v %)	Relativní četnost pro celý soubor (v %)
Shluk 1 (vlastní auto i počítač)	alespoň 1 ekonomicky aktivní člen	91,1	61,4
	(z toho 2 ekonomicky aktivní členové)	46,5	26,6
	vyšší zaměstnanec v čele	39,3	23,3
	alespoň 1 z partnerů VŠ vzdělání	23,4	14,0
	nepracující důchodci	*7,4*	*35,7*
Shluk 2 (vlastní auto, nechce počítač)	nepracující důchodci	52,5	35,7
	alespoň 1 z partnerů SŠ vzdělání	86,2	75,4
	venkovské obce	46,5	36,3
	řídce obydlená oblast	55,3	43,0
	nízká úroveň vzdělání	*6,8*	*10,6*
	alespoň 1 z partnerů VŠ vzdělání	*7,0*	*14,0*
Shluk 3 (nechce auto ani počítač)	nepracující důchodci	81,7	35,7
	nízká úroveň vzdělání	28,6	10,6
	alespoň 1 z partnerů SŠ vzdělání	*66,7*	*75,4*
	alespoň 1 z partnerů VŠ vzdělání	*4,8*	*14*
Shluk 4 (nemá na auto)	1 nezaměstnaný člen	9,2	4,6
	2 nezaměstnaní členové	1,3	0,5
	hustě obydlená oblast	38,8	31,3

Zdroj: ČSÚ (šetření SILC, ČR, 2008), vlastní výpočty v SPSS

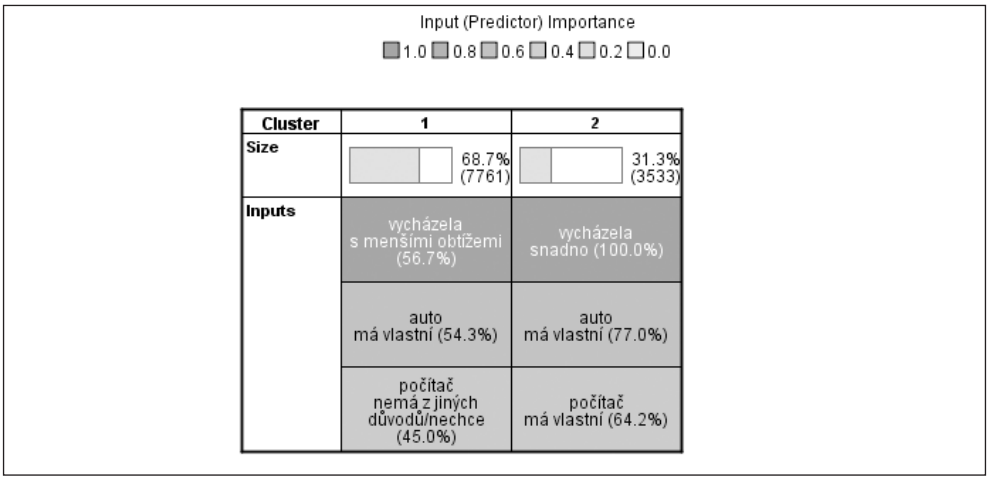
Obdobným způsobem bychom mohli vytvářet skupiny domácností podle množiny jiných ukazatelů. V některých případech je vhodné některé proměnné s málo zastoupenými kategoriemi překódovat. Takovou proměnnou je například ukazatel, jak domácnost vycházela s příjmy na šestibodové škále, od kategorie „s velkými obtížemi" po kategorii „velmi snadno". Tuto proměnnou jsme překódovali do tří kategorií, a to 1 (s obtížemi), 2 (s menšími obtížemi) a 3 (snadno).

Aplikací dvoukrokové shlukové analýzy na proměnné *počítač, auto* a *vycházela* získáme jako optimální výsledek 12 shluků. Pokud chceme popsat jen malý počet shluků, můžeme začít od dvou shluků a například sledovat důležitost proměnných pro vytvořené shluky. Z obrázku 3 je zřejmé, že na vytvoření shluků měla největší vliv [*Input (Predictor) Importance*] proměnná vycházela a nejmenší vliv proměnná *počítač*. Tento poznatek by mohl vést k mylnému závěru, že proměnné nebyly vhodně vybrány a že nemá smysl zkoumat rozdělení objektů do jiných počtů shluků. Na obrázku 4, který charakterizuje tři shluky, jsou již všechny tři proměnné ohodnoceny stejnou důležitostí (hodnotou jedna, legenda není z důvodu úspory místa součástí obrázku). Rozdíl v hodnotách obrysového koeficientu je přitom přibližně stejný, jako rozdíl pro jiné po sobě následující počty shluků.

Získané tři shluky můžeme tedy charakterizovat následujícím způsobem. V největším shluku (5 530 domácností) převažují domácnosti, které vlastní auto i počítač a vycházely s příjmy s menšími obtížemi. Druhým shlukem z hlediska velikosti (3 533 domácností) je skupina domácností, které (všechny) vycházely s příjmy snadno. V rámci nich převažují domácnosti, které vlastní auto i počítač. Ve třetím shluku (2 231 domácností) jsou zastoupeny domácnosti, které (všechny) auto nemají z jiných důvodů než finančních. V rámci nich převažují domácnosti, které také nemají počítač z jiných důvodů než finančních a vycházely s příjmy s menšími obtížemi.

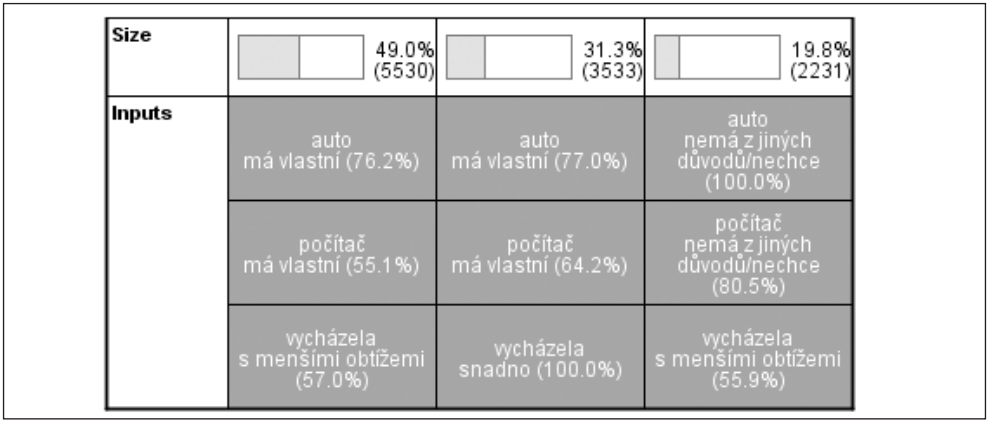
Při větším počtu proměnných zařazených do analýzy by bylo třeba zohlednit intenzitu jejich závislosti, aby skupina velmi závislých proměnných neměla na shlukování větší vliv než proměnné ostatní. Ve výše uvedených případech jsme se však zaměřili pouze na dvě a tři proměnné, u nichž šlo především najít významné kombinace kategorií, tudíž jsme výše zmíněný aspekt nesledovali.

Obr. 3: **Charakteristika shluků z hlediska proměnných auto, počítač, vycházela (2 shluky)**



Zdroj: vlastní výpočty v SPSS

Obr. 4: **Charakteristika shluků z hlediska proměnných auto, počítač, vycházela (3 shluky)**



Zdroj: vlastní výpočty v SPSS

Závěr

Na příkladu shlukování domácností charakterizovaných kategoriálními ukazateli jsme se pokusili ilustrovat problematiku stanovení vhodného počtu shluků. Prostředky pro určení počtu shluků jsou v komerčních statistických programových systémech implementovány zřídka. Pokud systém zahrnuje nějakou podporu, počítají se pouze hodnoty koeficientů, které jsou případně graficky znázorňovány pro různé počty shluků. Obvykle je tato možnost určena pro shlukování objektů charakterizovaných kvantitativními proměnnými a optima stanovená pomocí různých koeficientů se často liší.

Ucelený přístup pro stanovení optimálního počtu shluků objektů, které jsou charakterizovány nominálními proměnnými, případně

proměnnými různých typů, poskytuje systém SPSS v rámci dvoukrokové shlukové analýzy. Uživatel by ovšem měl zadat horní hranici počtu shluků, pro které má být optimum nalezeno (standardně je nastavena hodnota 15). Způsob výpočtu zajišťuje nalezení optimálního počtu z jakéhokoli zadaného intervalu od 1 do zadané maximální hodnoty počtu shluků. Je tedy zřejmé, že pro různé intervaly mohou existovat různé „optimální" počty.

Největší zodpovědnost leží ovšem stále na uživateli, aby určil, kolik přibližně chce získat shluků, aby mohly být vhodně popsány a interpretovány. Programem stanovené hodnoty mohou být pouze podpůrným prostředkem. Pro určení vhodného počtu shluků je užitečné aplikovat více nástrojů. V tomto článku byly kromě postupu založeném na informačním kritériu BIC použity hodnoty obrysového koeficientu a zkoumání důležitosti proměnných z hlediska, jak dobře pomocí nich mohou být rozlišeny různé shluky.

Na základě našich zkušeností při analýzách různých datových souborů jsme vyhodnotili BIC kritérium jako nástroj, který může být v řadě případů využit pro stanovení vhodného počtu shluků přímo, tj. podle minimální hodnoty ze zadaného intervalu, aniž by musely být prováděny dopočty podle vzorců (6) až (9). V případě AIC kritéria jsme se s takovou možností nesetkali. V praxi by bylo vhodnější použít spíše výše zmíněná kritéria AIC3 či CAIC, které však v systému SPSS nejsou implementována.

V dřívějších verzích SPSS (od 11.5) se zobrazovaly jednak hodnoty zvoleného informačního kritéria, jednak diference (6) a poměr (7). V současných verzích uživatel žádné z těchto hodnot k dispozici nemá, naopak přibyl ve výstupu graf zobrazující hodnotu obrysového koeficientu s indikací kvality rozdělení objektů do shluků. Jedním z možných způsobů stanovení vhodného počtu shluků je tak sledovat diference v hodnotách obrysového koeficientu a uvažovat takový počet shluků (nejlépe z oblasti Good), kdy pro nižší počet dojde k většímu poklesu hodnotu ve srovnání s poklesem z většího počtu shluků.

Ve svém dalším výzkumu se zaměříme na návrh a analýzu vlastností některých jiných koeficientů, založených na principu vícenásobné analýzy rozptylu s využitím speciálních měr variability pro nominální a ordinální proměnné. Plánujeme též analyzovat ostatní modifikace stávajících koeficientů určených pro kvantitativní data s využitím speciálních měr nepodobností pro data kvalitativní a data smíšeného typu.

Tento článek byl zpracován za podpory prostředků institucionální podpory na dlouhodobý koncepční rozvoj vědy a výzkumu na Fakultě informatiky a statistiky VŠE v Praze v roce 2010.

Literatura

[1] GAN, G., MA, C. a WU, J. *Data Clustering: Theory, Algorithms, and Applications*. Philadelphia: ASA-SIAM, 2007. ISBN 978-0-898716-23-8.

[2] CHIU, T., FANG, D., CHEN, J., WANG, Y., JERIS, C. A Robust and Scalable Clustering Algorithm for Mixed Type Attributes in Large Database Environment. In *Proceedings of the seventh ACM SIGKDD international conference on knowledge discovery and data mining*. San Francisco: ACM, 2001.

[3] KAUFMAN, L., ROUSSEEUW, P. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, Hoboken, 2005. ISBN 0-471-73578-7.

[4] KLÍMEK, P. Shlukovací metody v data miningu. *E+M Ekonomie a Management*. 2008, roč. 11, č. 2, s. 120–125. ISSN 1212-3609.

[5] ŘEZANKOVÁ, H. Cluster analysis and categorical data. *Statistika*. 2009, roč. 89, č. 3, s. 216–232. ISSN 0322-788X.

[6] ŘEZANKOVÁ, H., HÚSEK, D., SNÁŠEL, V. *Shluková analýza dat*. 2. vyd. Praha: Professional Publishing, 2009. ISBN 978-80-86946-81-8.

[7] ZHANG, T., RAMAKRISHNON, R., LIVNY, M. BIRCH: An efficient data clustering method for very large databases. In *Proceedings of the ACM SIGMOD Conference on Management of Data*. Montreal: ACM, 1996.

prof. Ing. Hana Řezanková, CSc.
Vysoká škola ekonomická v Praze
Fakulta informatiky a statistiky
Katedra statistiky a pravděpodobnosti
hana.rezankova@vse.cz

Ing. Tomáš Löster, Ph.D.
Vysoká škola ekonomická v Praze
Fakulta informatiky a statistiky
Katedra statistiky a pravděpodobnosti
tomas.loster@vse.cz

Doručeno redakci: 24. 11. 2010
Recenzováno: 4. 1. 2011, 28. 1. 2011
Schváleno k publikování: 4. 7. 2013

Abstract

CLUSTER ANALYSIS OF HOUSEHOLDS CHARACTERIZED BY CATEGORICAL INDICATORS

Hana Řezanková, Tomáš Löster

In the paper we deal with evaluation of the results of cluster analysis which is applied to data files in which objects are characterized qualitative variables. We describe methods of clustering, determination of optimal cluster numbers, and evaluation of obtained clusters implemented in the procedure for two-step cluster analysis in the SPSS statistical software package. These techniques are applied to the selected household indicators gathered in the SILC (Statistics on Income and Living Conditions) survey in the Czech Republic in 2008.

We clustered households characterized by the indicators expressing if a household owns a computer and a car as an example. We discuss the problem of determination of optimal cluster numbers by the approach based on information criteria (we use the Bayesian information criterion) and determine number of clusters by means of the silhouette coefficient. Then we describe four obtained clusters on the basis of indicators of working activity, degree of education and degree of urbanization. Moreover, we extended characterizing variables to the recoded indicators expressing how the household goes well with its income. On the basis of this example we illustrate investigation of variable importance. In this case we describe obtained three clusters by three variables used in the analysis.

In conclusion we mention some other approaches to evaluation of clustering objects characterized by categorical variables. They consist in both coefficients based on multivariate analysis of variance with using specialized variability measure for nominal and ordinal data, and modification of some other coefficients for qualitative data. The problem of mixed type variables is also mentioned.

Key Words: *cluster analysis, number of clusters, qualitative variables.*

JEL Classification: *C19, C49, D19.*