Faktorová analýza v kontingenčních tabulkách

Metoda LINDA - A—popis a použití

JAN ŘEHÁK
Ústav pro filozofii a sociologii ČSAV, Praha
IVANA LOUČKOVÁ
katedra aplikované sociologie FFUP, Olomouc

Tradiční analýza kontigenčních tabulek vychází především z *modelu nezávislosti* a testování jeho platnosti proti obecné alternative závislosti. Ve společenských vědách však je prosté konstatování závislosti zcela nedostatečné, kdykoli nás zajímají hlubší analytické závěry o směrech a mechanismech kauzálního působení ve zkoumaných procesech. Prvním krokem k jemnější statistické informaci je měření stupně závislosti pomocí *koeficientů asociace a korelace* (o tom viz například J. Řehák—B. Řeháková [1973; 1975; 1984], J. Řehák [1976] a tam citovaná literatura). Snaha odhalit strukturu asociací vedla ke vzniku metodiky *znaménkového schématu* (viz J. Linhart—Z. Šafář [1976], J. Řehák—B. Řeháková [1978; 1984]) a analýzy reziduí nezávislostního modelu (J. S. Haberman [1973]). Znaménkové schéma je však užitečné v určení asociace každého pole zvlášť vzhledem ke všem ostatním a jeho aplikabilita je vázaná na přehledné asociační struktury. Proto se o další kroky a podrobnější analýzu pokoušíme v celé řadě přístupů, které jsou vzájemně příbuzné a všechny se opírají o kanonickou dekompozici reziduálních odchylek skutečných a očekávaných četností. Patří sem především *model korespondenční analýzy* (vícerozměrné škálování řádkových a sloupcových hodnot ve společném prostoru) (J. P. Benzecri [1969]), modely *vzájemné optimální kvantifikace kategorií řádkové a sloupcové proměnné* (L. Guttman [1941, 1959]), kanonická analýza (L. Lancaster [1957], M. G. Kendall—A. Stuart [1973], J. Lauber — J. Linhart [1978], I. I. Jelisejeva [1982]). Souběžně s těmito modely lze formulovat i interpretační *model faktorový*, jehož základním cílem je odhalit případnou existenci společných latentních faktorů vysvětlujících strukturu reziduálních odchylek (J. Řehák — I. Loučková [1982]).

V tomto článku chceme shrnout možnosti modelu LINDA-A,[1]) v němž jsme se pokusili o jednoduchou syntézu různých výše citovaných modelů a o jejich určité rozvinutí.

Důležitost metodiky spočívá v několika analytických aspektech:

a) Metoda rozvíjí postup znaménkového schématu tím, že hledá zákonitosti a korelovanosti v jeho struktuře jako celku, zatímco znaménkové schéma hodnotí každé políčko zvlášť proti všem ostatním, a provádí rozklad celkové chí-kvadrátové asociace na dílčí samostatné složky.

b) Metodu lze využít nezávisle na statistickém testování hypotéz, tj. i tam, kde testy

1) LINDA (lineární dekompoziční algoritmy) je soubor programů vyvinutý a programovaný autory pro počítač *TESLA 200* v Laboratoři výpočetní techniky UP v Olomouci. Model LINDA-A se týká zpracování asociačních reziduí v kontigenční tabulce. Model a metodu nazýváme ve stati podle zkonstruovaného programu: akronym vychází z matematického základu postupu. (Program LINDA má však širší využití i mimo oblast kontingenčních tabulek a umožňuje řešit i další analytické úlohy v analýze profilů a komparačních tabulek (LINDA-K), analýze rozptylu (LINDA-R) a v analýze binárních relačních dat (LINDA-B).

nezávislosti nejsou vhodné vzhledem k úplným šetřením, nebo tam, kde jsou málo informativní (u velkých souborů).[2])

c) Postup vychází z odhadnutých četností, a proto je aplikabilní i v případech, kdy není splněn předpoklad nezávislého výběru, ale kontigenční tabulka obsahuje vhodné a nezkreslené odhady relativních četností.[3])

d) Metodu lze chápat jako realizaci vícerozměrného škálovacího postupu se dvěma množinami objektů — odhaluje vztahy mezi prvky těchto množin. Má tedy vlastnosti relačních zobrazení: jak číselné, tak grafické výsledky přinášejí rychlý přehled o vlastnostech dat i v případě, že faktorový model interpretačně nevyhovuje nebo se nepodařilo jej nalézt. Grafické zobrazení navíc odhaluje různé zajímavé nelineární vztahy v datech, které jsou jinak obtížně rozpoznatelné.

1. Interpretační schémata v analýze asociací kontingenční tabulky

Asociační struktura v kontingenční tabulce s R řádky a S sloupci je dána koeficienty korelace mezi jednotlivými kategoriemi.

Označíme řádkovou proměnnou $A = \{a_1, a_2, \ldots, a_R\}$, sloupcovou proměnnou $B = \{b_1, b_2, \ldots, b_S\}$ a četnosti jako n_{rs}, $n_{r+} = \sum n_{rs}$, $n_{+s} = \sum n_{rs}$, $n = \sum\sum n_{rs}$ (obdobně p_{rs}). Korelační koeficienty jsou

$$(1) \qquad r(a_r, b_s) = \frac{n n_{rs} - n_{r+} n_{+s}}{\sqrt{n_{r+}(n - n_{r+})\, n_{+s}(n - n_{+s})}} = \frac{z_{rs}}{n}$$

kde z_{rs} jsou skóry znaménkového schématu. Korelační matice, $\mathbf{R} = \|r(a_r, b_s)\|$, tak vyjadřuje informaci o vztazích mezi oběma znaky A, B specifikovanou pro jednotlivé kategorie.

V tabulkách o rozměrech $R \times S$.[4]) obzvláště jsou-li R a S velká, se přehlednost matice R ztrácí. S rostoucím R a S se vztahy mezi kategoriemi komplikují a možnosti vzájemných propojení násobí.

A) Faktorový model

Jestliže neplatí hypotéza nezávislosti, pak faktorový model kontigenční tabulky můžeme založit na předpokladu, že všechny vlivy projevující se v tabulce, se spojují do několika málo souběžných příčinných komplexů, které nazýváme *faktory* a které jsou natolik silné, že se statisticky konzistentně v datech projeví. Cílem popisované metodiky je odhalit tyto empirické konzistentní projevy a z nich vyvodit působení příslušných faktorů. Při interpretaci může však do značné míry působit obtíže ta skutečnost, že drobné příčinné vlivy se v realitě mohou seskupovat k souběžnému působení různými způsoby, neboť vytvářejí nepřeberné množství kombinací, z nichž mnohé mají samostatný obsahový smysl. Lze tak předpokládat, že tutéž situaci je

2) U velkých souborů má přijetí modelu závislosti malou interpretační hodnotu, neboť test je citlivý i na malé, interpretačně nezajímavé odchylky.

3) To se týká např. sdruženého třídění dvou otázek s násobnými odpovědmi („*Vyberte k položek ze seznamu...*"), tj. tabulky, v níž jako jednotka vystupuje vybraná položka, nikoli respondent, a v níž je tedy porušen princip nezávislosti výběru. Jinými případy jsou dále skupinkový výběr jednotek s nevychýlenými odhady pravděpodobností a případ tabulek, v nichž odhadujeme relativní četnosti pomocí složitých váhových funkcí.

4) Situace pro tabulky 2×2 je jednoduchá, neboť asociace je popsána jediným číslem ($r_{11} = r_{22} = -r_{12} = -r_{21}$). U tabulek $2 \times S$ (resp. $R \times 2$) je možno vyjít z komparace sloupcových (resp. řádkových) relativních četností a tyto seskupovat, řadit podle velikosti, hledat v nich trendy apod. U malých tabulek např. 3×3 nebo 5×5 bude metodika efektivní především tehdy, bude-li LINDA-A použita ve smyslu kanonické analýzy, tj. pro hledání takových hodnot pro kategorie, které maximalizují lineární korelaci, a pro měření ordinální korelace tímto způsobem.

možno popsat pomocí obsahově různých systémů nezávisle na sobě působících faktorů.[5]) (Při popisu modelu uvedeme dále dva principy výběru faktorů, které lze v praxi s úspěchem použít — princip hlavních faktorů a princip prosté struktury.)

B) Interpretační schémata

Interpretace faktorů se budou lišit podle toho, jaký vztah vyjadřuje kontigenční tabulka A × B; jaké kauzální schéma přijmeme pro dané proměnné jako nejvhodnější.

Interpretační schémata vztahů kategorií v kontigenčních tabulkách (grafické znázornění viz obr. 1):

a) Asociace A a B odpovídá symetrickému vztahu $A \leftrightarrow B$; příčinný směr zde nelze určit. Jednotlivé vazby $a_r \leftrightarrow b_s$ jsou způsobeny neznámými, *latentními společnými příčinami* $\mathbf{F} = (F_1, F_2, \ldots)$. Modelem je $\mathbf{F} \rightarrow A \times B$, **F** působí asociace v polích tabulky.

b) Asociace A a B odpovídá asymetrickému vztahu $A \rightarrow B$, který však není přímý, ale je zprostředkován nezahrnutými neznámými faktory, $A \rightarrow (F_1, F_2, \ldots) \rightarrow B$. Vazba v poli (a_r, b_s) může být *zprostředkována* některými z $(F_1, F_2, \ldots)$.

c) Asymetrický vztah $A \rightarrow B$ je přímý jen zdánlivě, ve skutečnosti je způsoben systémem latentních faktorů $\mathbf{F} = (F_1, F_2, \ldots)$, který působí na A i B *souběžně*.

d) Hodnoty proměnné A vytvářejí *zobecněné proměnné* $\mathbf{F} = (F_1, F_2, \ldots)$, které souběžně působí na kategorie B a které vznikají seskupením hodnot $\{a_r\}$. (Každá z hodnot a_r se může podílet na více faktorech.)

e) U asociačních tabulek (i závislostních modelů), které nemají zjevné kauzální interpretace, ukazují faktory *korespondenční seskupení* hodnot $\{a_r\}$ a $\{b_s\}$, a tak odhalují pravidelnosti souběžných výskytů jevů.

f) Hodnoty $A = \{a_r\}$ působí jako *kontexty* pro působení jiných proměnných $\mathbf{F} = (F_1, F_2, \ldots)$, které jsou přímo nezjištěné (resp. nezjistitelné); skupiny hodnot $\{a_r\}$ mohou mít roli *indikátorů* pro jednotlivé proměnné systému **F**.

Schémata obr. 1 odpovídají těm interpretačním modelům, které používáme v jednotlivých případech a které jsou v praxi nejčastější a nejdůležitější. Všem je společné to, že v nich vystupují neměřené, ale metodou odhalované faktory, které mají vysvětlující, zprostředkující nebo jen deskriptivní charakter a mají vysvětlit strukturu vztahů mezi A a B. Jejich obsah lze odhalit jen nepřímo — je indikován současným působením na určitá významově určená pole (a_r, b_s) a vzájemnou statistickou (tedy předpokládáme i obsahovou) nezávislostí. Přijetí jednotlivých schémat záleží na myšlenkovém modelu, který výzkumník přijímá pro analýzu každého konkrétního případu, na výsledku analýzy. Záleží tu na směru vztahu, na významu faktorů a na další dodatkové informaci.

C) Analyticko-statistická formulace úlohy

Pro asociační tabulku A × B máme nalézt systém faktorů $\mathbf{F} = (F_1, F_2, \ldots F_M)$ podle některého z modelů obr. 1 tak, aby **F** vysvětloval veškeré asociace tabulky, tedy takový systém, aby parciální korelace ve všech polích vymizela: $r(a_r, b_s/\mathbf{F}) = 0$ pro všechna (a_r, b_s). V analytické praxi měníme tuto formulaci na nalezení systému $\mathbf{F}^* = (F_1, F_2, \ldots, F_L)$, $L \leqq M$ tak, aby vysvětlil všechny podstatné asociace.

K základní úloze nalezení (odhalení, interpretaci) jednotlivých faktorů F_m i faktorového celku **F** patří i další úlohy, které ji nutně doprovázejí a doplňují:

---

5) Tato potíž je dobře známa z faktorové analýzy korelačních matic; jako tam i zde musíme při odhalování faktorů počítat s nejednoznačností řešení i s tím, že se optimálně interpretovatelné faktory nepodaří zjistit.

a) Určení podílu každého F_m na působení **F** jako celku, nalezení koeficientu determinace pro F_m, t.j. měření síly vztahu $F_m \rightarrow (A \times B)$ nebo měření jeho podílu na $A \rightarrow \mathbf{F} \rightarrow B$.

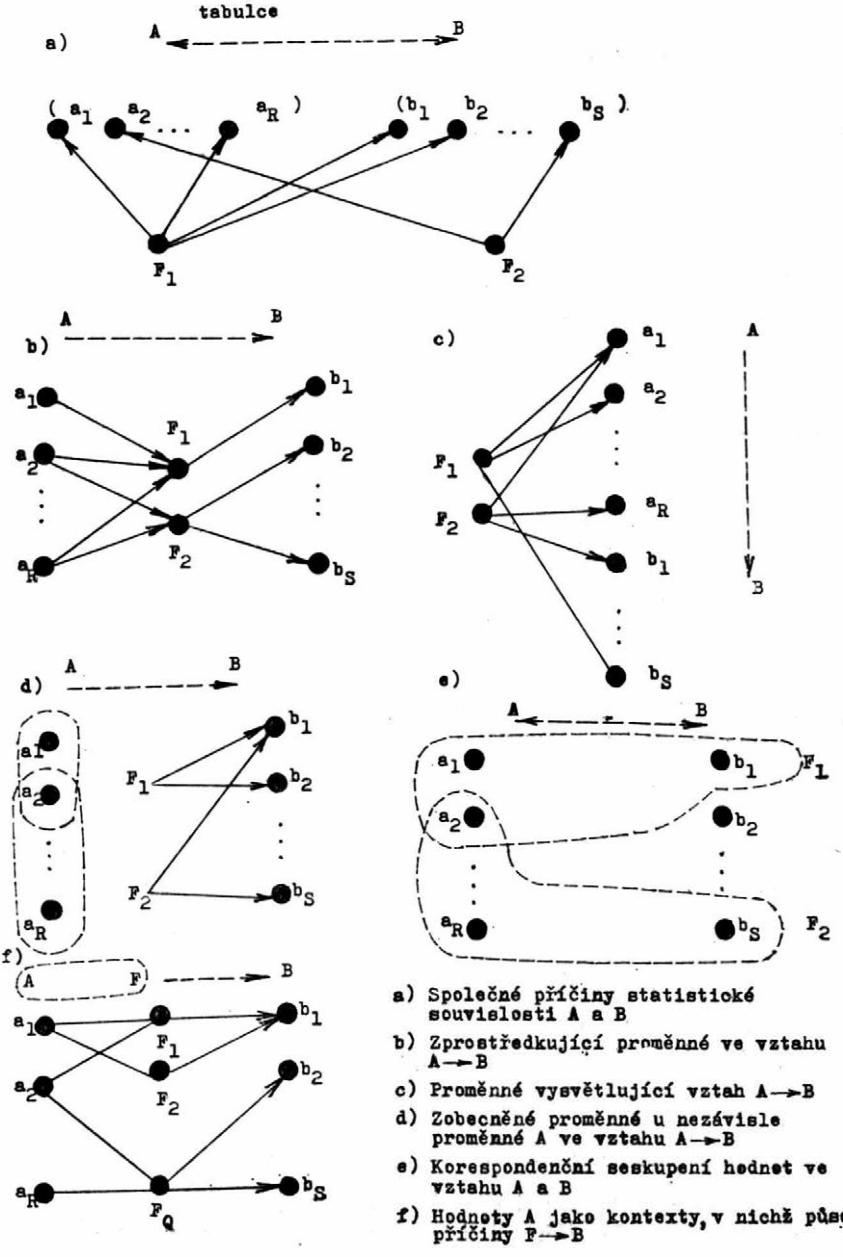
b) Určení vhodného a dostatečného subsystému $\mathbf{F}^* \subset \mathbf{F}$, t.j. vyloučení faktorů, které jsou bez většího systematického vlivu.

c) Určení podílu celkového asociačního projevu **F** na kategorie $\{a_r\}$ a $\{b_s\}$.

d) Určení predikčního vlivu působení jednotlivých faktorů na asociační rezidua.

e) Určení vztahu mezi dvěma různými systémy faktorů $\mathbf{F}^*$ a $\mathbf{G}^*$.

Obrázek 1. Interpretační schémata faktorové analýzy v kontingenční tabulce



a) Společné příčiny statistické souvislosti A a B

b) Zprostředkující proměnné ve vztahu A→B

c) Proměnné vysvětlující vztah A→B

d) Zobecněné proměnné u nezávisle proměnné A ve vztahu A→B

e) Korespondenční seskupení hodnot ve vztahu A a B

f) Hodnoty A jako kontexty, v nichž působí příčiny F→B

Z uvedeného je zřejmé, že jde o úlohu *exploračního charakteru*. O interpretaci faktorů a o její obtížnosti platí vše, co je známo z praxe faktorové analýzy korelačních matic. Metoda faktorové analýzy kontigenčních tabulek není koncipována pro snadné rutinní použití (pokud ji nechceme využít jen dílčím způsobem nebo nejde-li o opakované situace). Předpokladem dobré aplikace je značná *spolehlivost zařazování* jednotek do kategorií (t.j. nízká měřicí a záznamová chyba). V praxi to znamená dobře definované a obsahově oddělené kategorie A i B a přísný způsob zařazování. U větších tabulek je interpretace ulehčena tím, že se faktory mohou projevovat u více kombinací (a_r, b_s), tj. ve více polích, což vede k širšímu obsahovému základu závěrů (na druhé straně však také k možnostem nalezení širších a roztříštěnějších faktorových systémů).

2. Příklady — formulace úloh

Metoda LINDA-A byla vyvíjena a shrnuta jako pracovní prostředek, který má napomoci k řešení praktických úloh sociologie. Proto i příklady[6]) bereme z běžné analytické praxe. Záměrně volíme příklady s jednoduchou a užitečnou, i když ne efektní interpretací, abychom ukázali nejen přednosti, ale i potíže, a aby bylo vidět, že kromě výrazných a vcelku snadno rozpoznatelných pravidelností obsahují výsledky i náznaky různých vztahů, které mají námětový a hypotetický význam.

Příklad 1.[7]) *Statistická souvislost věkové a profesní struktury v průmyslovém podniku*

V tabulce 1 jsou uvedeny absolutní četnosti vyjadřující statistický vztah dvou základních strukturních ukazatelů: profese a věku.

Tabulka 1. Věková struktura podle profesí — počty osob

Profese	Celkem	Věkové kategorie do 20	do 30	do 40	do 50	do 60	nad 60	Věkový průměr
				let				
Strojní zámečník	6951	386	2774	1730	1284	699	78	35
Soustružník	2780	230	1575	518	346	107	4	29
Frézař	1381	153	693	322	154	57	2	29
Brusič	1157	55	475	287	222	102	16	33
Kovář	248	3	69	65	61	48	2	38
Slévač	397	12	125	111	97	49	3	36
Seřizovač	83	0	36	17	20	9	1	35
Zámečník-údržbář	1899	74	520	480	461	310	54	38
Celkem	14896	913	6267	3530	2645	1381	160	

Marginální údaje informují o rozložení věku a rozložení profesí zvlášť. Struktura asociací však může podat informaci o výsledcích procesů technického rozvoje (profese) i organizačně kádrových opatřeních, které s doplňováním profesí souvisí. Testování hypotézy nezávislosti zde nemá význam jednak vzhledem k úplnému šetření a jednak vzhledem k nereálnosti takového předpokladu. Znaménkové schéma ztrácí význam vzhledem k vysokému n (téměř 15 800 jednotek). Strukturu závislosti můžeme zachytit korelační maticí (viz (1)), jejíž koeficienty mají hodnotu na n nezávislou a pomocí metody LINDA-A pokusit se odhalit latentní příčiny, faktory, které nerovnoměrnost korelační struktury způsobily.

Příklad 2. Trvale obydlené byty podle velikosti v obvodech Prahy

Rozložení bytů podle velikosti v pražských obvodech je uvedeno v tabulce 2. Rychlou interpretaci je možno provést z relativních četností či z korelačních koeficientů v jednotlivých polích. Zhod-

6) Další využití viz např. H. Kasalová — J. Řehák [1984], J. Řehák — I. Loučková [1981].

7) Příklad je převzat z knihy J. Řehák — B. Řeháková [1984].
Data jsou ukázkou běžně dostupné tabulky ASŘ v průmyslovém podniku.

nocení společných kontextových vazeb a vlivů a jejich rozklad na nezávislé složky faktorového systému však lze odhalit pouze podrobnější analýzou korelační struktury. Otázky, které se nabízí výzkumníkovi, jsou: Je možno odhalit kontexty, v jejichž rámci se diferencuje skladba velikostí bytů? Jaké vztahy mezi pražskými obvody a typy bytů (podle velikosti) lze ve strukturách asociací nalézt?

Tabulka 2. Počty bytů podle velikosti v Praze

Obvod	Velikost bytů				
	1	2	3	4	5
Praha 1	7 741	6 985	3 641	1 212	405
Praha 2	12 995	9 418	5 523	1 255	270
Praha 3	22 825	11 806	4 435	523	111
Praha 4	29 983	28 091	28 506	6 151	1 442
Praha 5	16 195	11 967	8 341	2 649	1 165
Praha 6	15 545	19 438	10 203	2 812	1 170
Praha 7	12 710	6 696	3 081	607	111
Praha 8	16 100	11 917	15 995	2 220	661
Praha 9	11 261	14 887	13 101	2 601	1 016
Praha 10	21 771	22 991	14 580	2 089	836

Zdroj: Čísla o Praze 1980. Praha. MSÚ 1981.

Struktura velikosti bytů je výsledkem působení dlouhodobých historických procesů budování města i výsledkem soudobých strategií výstavby, mísí se zde vlivy různých stylů a dob, životního způsobu i prostorových možností. Oddělit tyto vlivy detailně nebude možné, neboť znak *„velikost bytu"* je pouze jednou z mnoha vlastností, u níž se navíc pod stejnou kvantitou rozlišují zcela různé kvality i další kvantitativní aspekty. Je tedy zřejmé, že faktorová analýza může dát jen velmi povrchní pohled na problematiku, nicméně může asociační strukturu zpřehlednit a zjednodušit a tak navodit další hypotézy a další explorační analytické kroky.

Příklad 3. Vybavování mladých manželství

Postup nákupu různých předmětů osobní potřeby a předmětů vybavení domácnosti ve vzájemných souvislostech byl zkoumán pomocí součtových proměnných (PPOP — počet předmětů osobní potřeby, PPVD — počet předmětů vybavení domácnosti). Vzájemná korelace je středně vysoká,[8]) $r = .401$, $\tau_b = .283$. Vyvstávají otázky, v jakém vztahu je postup vybavování osobními předměty a předměty pro domácnost: Jdou oba procesy souběžně? Je jeden rychlejší? Má jednoznačnou přednost jeden z nich? Na řadu otázek odpoví podrobná přímá analýza relativních četností či korelačních koeficientů nebo regresní analýza vztahu obou proměnných (u té však musíme zkoušet různé modely). Pokusíme se vztahy mezi kategoriemi odhalit pomocí faktorové analýzy a vícerozměrné škály, která by měla ukázat i nelineární vazby a vstupy. Kladná korelace naznačuje souběžnost obou proměnných, monotónnost podél celé škály však zdaleka není indikována.

Příklady se liší několika statistickými a analytickými aspekty:

1. V příkladech 1 a 2 mají kategorie pevně stanovené hodnoty, jsou dobře rozlišitelné s jednoznačnými pravidly klasifikace jednotek (osoby, byty). Data, až na možné záznamové chyby, jsou vysoce spolehlivá. V příkladu 3 jsou obě proměnné faktické, ale jsou zatíženy značnou výpovědní chybou.[9]) Data se liší *reliabilitou*. Skutečný obsah proměnných *„velikost bytu"* a *„počet přemětů"* je však velmi nejenoznačný (různé předměty skládající stejný počet, různě velké místnosti, typy domu, nezahrnuté příslušenství apod.). Data se tedy liší i *stupněm určenosti obsahu*.
2. V příkladu 2 je dáno úplné šetření stabilních údajů s malou vývojovou variabilitou (kterou může narušit jen nově působící vývojový faktor), v příkladu 1 jde též o úplné šetření, ale na rozdíl od 2 nemůže být plně řízen (obzvláště z hlediska sousedních věkových kategorií), jednotky vstu-

8) Oba koeficienty mají pro $n = 968$ dosaženou významnost menší než .00005; tabulku četností neuvádíme pro úsporu místa; PPOP má 8 hodnot (od *0* do *7 a více*), PPVD má 6 hodnot (*0* až *5 a více*). Populaci tvoří mladá manželství do 30 let.

9) Povaha chyby je v subjektivní výpovědi i metodologické formě otázky (indikace vlastních typů předmětů): působí zde paměť respondenta, jeho subjektivní klasifikace předmětů, zahrnutí či nezahrnutí předmětů vadných či téměř nefungujících, není reflektován násobný počet předmětů apod.

pují do tabulky na základě neidentifikovaných procesů.[10] Data se tedy liší *genezí, stupněm výběrové chyby i smysluplnosti statistické inference.*

3. V případu 1 jde o vztah „kardinální—nominální", v druhém „nominální—ordinální" (kardinalita je pouze zdánlivá vzhledem k různosti kvalit) a ve třetím „kardinální—kardinální". V posledním případu má smysl uvažovat ordinální (monotónní) asociaci, ověřit ji a hledat asociační složky, které jsou na ní nezávislé a doplňují ji. Data se liší *typem znaků.*

4. Příklady se liší *interpretačními schématy*[11])

— pro příklad 1 se jeví jako možné schéma c), v němž profese působí na věkové požadavky zdánlivě, skutečnými příčinami jsou (hypoteticky) nerovnoměrný vývoj profesní struktury, fluktuační proudy a jejich příčiny, organizační a propagační úroveň řídícího aparátu atd.;

— pro příklad 2 se nabízí schéma f) v němž pražské obvody svými situačními charakteristikami ovlivňují působení neurčených faktorů, které ovlivňuje výstavba i její aspekty, jako je velikost bytu;

— u příkladu 3 lze použít schéma a), v němž latentními faktory mohou být hodnotové preference, finanční možnosti, doba trvání domácnosti, tlak rodiny; lze též použít schéma e), jehož výsledkem je prostý zjednodušený popis empirických vztahů, který může naznačit cesty další analýzy.

Všechny tyto okolnosti se projeví v samotné analýze a při formulaci závěrů.

3. Model faktorové dekompozice — popis metody LINDA-A

Interpretace všech výstupů modelu vycházejí z jeho matematické formulace. Metodika řešení je příbuzná faktorové analýze korelačních matic, jmenovitě metodě hlavních komponent, která je tu aplikována (viz. K. Úberla [1980] H. H. Harman [1967]).

A) Matematická formulace modelu

V metodě LINDA-A vycházíme z *poměrového porovnávání* četností (empirických i očekávaných), převod na aditivní tvar je proveden pomocí logaritmování ($f_{rs} = n_{rs}/n$, $\|n_{rs}\|_{R\times S}$ je matice četností kontigenční tabulky):

$$x_{rs} = \ln n_{rs} \quad \text{nebo} \quad x_{rs} = \ln f_{rs} \tag{2}$$

$$\bar{x}_{r+} = \frac{1}{S}\sum_{s} x_{rs}\,, \quad \bar{x}_{+s} = \frac{1}{R}\sum_{r} x_{rs}\,, \tag{3}$$

$$\bar{x}_{++} = \frac{1}{RS}\sum_{r}\sum_{s} x_{rs}\,.$$

Logaritmická asociační rezidua definujeme jako:

$$y_{rs} = x_{rs} - \bar{x}_{r+} - \bar{x}_{+s} + \bar{x}_{++}\,. \tag{4}$$

Při platnosti nulové hypotézy ($n_{rs} = (n_{r+}n_{+s})/n$) jsou všechna y_{rs} nulová ($y_{rs} = 0$ pro všechna r, s); kladná asociace v poli nastává, když $y_{rs} > 0$, záporná, když $y_{rs} < 0$.

Předpokládáme, že reziduum y_{rs} je složeno z příspěvků M faktorů $\mathbf{F} = (F_1, F_2, \ldots, F_M)$:

$$y_{rs} = Q_{1(rs)} + Q_{2(rs)} + \ldots + Q_{M(rs)}\,. \tag{5}$$

Faktory jsou ve svém působení *statisticky nezávislé*. Příspěvek $Q_{m(rs)}$ m-tého faktoru v poli (r, s) vzniká jako *interakční součinový člen*

$$Q_{m(rs)} = a_{rm}b_{sm} = \begin{Bmatrix}\text{vliv } F_m \text{ na kategorii } a_r\end{Bmatrix} \times \begin{Bmatrix}\text{vliv } F_m \text{ na kategorii } b_s\end{Bmatrix}. \tag{6}$$

10) Proto má význam testovat zde hypotézy a interpretační možnosti se zvýší znaménkovým schématem.

11) Jen zřídka je interpretační model jednoznačný, většinou tehdy, je-li vyvíjen postupně a tak navazuje na předchozí podobné analýzy dat. Jeho volba závisí jak na znalostech výzkumníka, tak i na cílech analytického postupu. Interpretační proces nás může podnítit k tomu, že odkryté vztahy nás přivedou ke změně modelu.

Výraz (6) znamená, že příspěvek faktoru F_m v poli (r, s) je úměrný jeho současnému vlivu na kategorii a_r i na kategorii b_s. Koeficienty a_{rm} nazýváme *faktorové efekty řádkových kategorií*, koeficienty b_{sm} jsou *faktorové efekty sloupcových kategorií*. Matice $\mathrm{A} = \|a_{rm}\|$ typu $R \times M$ a matice $\mathrm{B} = \|b_{sm}\|$ typu $S \times M$ obsahují neznámé parametry modelu, které má metoda odhalit a pomocí nichž identifikujeme neznámé faktory **F**.

Hlavní rovnice modelu jsou tedy:

$$y_{rs} = a_{r1}b_{s1} + a_{r2}b_{s2} + \ldots + a_{rM}b_{sM}\,, \quad \begin{matrix} r = 1, 2, \ldots, R\,, \\ s = 1, 2, \ldots, S\,, \end{matrix} \tag{7}$$

přičemž platí $\sum_{\mathrm{r}} a_{rm} = \sum_{\mathrm{s}} b_{sm} = 0$ pro všechna $m = 1, 2, \ldots, M$. Maticově pro $\mathrm{Y} = \|y_{rs}\|$ typu $R \times S$ bude

$$\mathrm{Y} = \mathrm{AB}^T\,. \tag{8}$$

Vliv faktoru F_m na úhrn asociačních reziduí měříme podílem faktoru na vysvětlení *celkové reziduální variability* $\sum\sum y_{rs}^2$, který vyjadřuje *koeficient diferenciační síly faktoru*, resp. *koeficient determinace pro* F_m:

$$r_m^2 = \frac{\sum_{\mathrm{r}} a_{rm}^2}{\sum_{\mathrm{r}}\sum_{\mathrm{m}} a_{rm}^2} = \frac{\sum_{\mathrm{r}} b_{sm}^2}{\sum_{\mathrm{s}}\sum_{\mathrm{m}} b_{sm}^2} = \frac{\lambda_m}{\sum_{\mathrm{m}} \lambda_m}\,, \tag{9}$$

kde λ_m jsou podle velikosti seřazená charakteristická čísla matice YY^T, resp. $\mathrm{Y}^T\mathrm{Y}$; $M = h(\mathrm{Y}) = h(\mathrm{YY}^T) = h(\mathrm{Y}^T\mathrm{Y})$.

Podíl prvních m faktorů [12]) určujeme číslem R_m^2:

$$R_m^2 = \sum_{\mathrm{j}=1}^{\mathrm{m}} r_j^2\,. \tag{10}$$

Pro interpretaci používáme model $(L \leqq M)$:

$$y_{rs} = a_{r1}b_{s1} + \ldots + a_{rL}b_{sL} + z_{rs}\,, \tag{11}$$

v němž zbytek z_{rs} zahrnuje faktory s malým r_m^2 a s nejasnou interpretací.

B) Jednoznačnost řešení

Řešení rovnice (5) existuje vždy, pro $h(\mathrm{Y}) \geqq 2$ *není* jednoznačné. Mezi transformacemi spojujícími jednotlivá řešení jsou důležité především ty, které se nazývají *rotace*, a ty, které se nazývají *překlopení*. Rotace využíváme v postupném hledání prosté struktury. Je-li T libovolná ortogonální matice typu $M \times M$, potom matice $\mathrm{A}^* = \mathrm{AT}$ a $\mathrm{B}^* = \mathrm{BT}$ jsou též řešením modelu (7). Výsledkem je tedy celá třída řešení, z nichž musíme vybrat jedno nebo více podle určitého zvoleného dodatečného kritéria. Obvykle se používají kritéria dvě (viz K. Überla [1980], H. H. Harman [1967], kde jsou i další možné přístupy):

1. *Princip hlavních faktorů*: vybíráme první faktor tak, aby vysvětlil co nejvíce Variability $\sum\sum y_{rs}^2$, tj. aby měl maximální možné r_1^2. Po jeho extrakci (zjištění) odečteme jeho vliv a hledáme faktor, který je na prvním nezávislý a který maximalizuje vysvětlení zbytkové variability $\sum\sum(y_{rs} - a_{r1}b_{s1})^2$). Třetí faktor bude nezávislý na prvních dvou a bude maximalizovat r_3^2 ze zbytkové variance prvních dvou faktorů atd., až vyčerpáme asociační variabilitu úplně.

---

12) Obdobně lze určit celkový podíl vlivu libovolného subsystému faktorů $(F_i, F_j, F_k, \ldots)$ jako $r^2 + r^2 + r_k^2 + \ldots$,)

2. *Princip jednoduché interpretace*: vybíráme takové faktory, které jsou vzájemně nezávislé, ale mají tzv. *prostou strukturu* [13]), tj. každý z faktorů co nejvíce rozlišuje působení na položky (kategorie) a každá položka (kategorie) je ovlivněna co nejmenším počtem faktorů.

Postup při řešení modelu je:

a) *extrakce M* hlavních faktorů,

b) určení $L \leqq M$ faktorů, které dohromady tvoří *dostačující systém* pro vysvětlení variability asociačních reziduí (dostatečně vysoké R_L^2),

c) *rotace* faktorů $\mathbf{F^*} = (F_1, F_2, \ldots, F_L)$ k systému $\mathbf{G^*} = (G_1, G_2, \ldots, G_L)$, který je nejblíže prosté struktuře.

Na rozdíl od běžné faktorové analýzy se v našem případě musíme rozhodnout pro jednu ze tří rotačních strategií.

Strategie I: rotace u řádkových kategorií

Hledáme prostou strukturu matice A, rotujeme A k A^* ($A^* = AT_A$) a matici B transformujeme odvozeně: $B^* = BT_A$. Výsledkem je jednoduchá interpretace vztahu mezi faktory a řádky. Používáme ji především v případě asymetrické asociace $\{řádky\} \rightarrow \{sloupce\}$, v němž jde o odhalení souběžného vlivu řádků a určení efektů tohoto vlivu na sloupce. Při opačném vztahu $\{sloupce\} \rightarrow \{řádky\}$ dává tato rotace naopak shrnutí nezávislých efektů v co nejjednodušší formě a odhaluje jejich příčiny.

Strategie II: rotace u sloupcových kategorií

Hledáme prostou strukturu matice B, rotujeme B k B^* ($B^* = BT_B$) a matici A transformujeme odvozeně; $A^* = AT_B$. Použití je obdobné (ale převrácené) k I.

Strategie III: rotace u řádkových i sloupcových kategorií

Hledáme prostou strukturu matice $C = \left\| \begin{matrix} A \\ B \end{matrix} \right\|$ typu $(R + S) \times L$, která vznikla spojením matic A, B. Rotujeme C k C^*, $C^* = CT_C$. Její řešení je mezi I a II a používá se v případě symetrických vztahů $\{řádky\} \leftrightarrow \{sloupce\}$, u nichž hledáme co nejjednodušší interpretaci faktorů vzhledem k souběžnému vlivu na obě proměnné.

U rotovaných faktorů G* počítáme obdobně koeficienty r_m^{*2} z matic A^*, B^*. Tato čísla však (vzhledem k součtu do L) dávají podíl vlivu faktorů v rámci systému $\mathbf{G^*}$, tj. podíl G_m na působení $\mathbf{G^*}$, jež v celku $\sum\sum y_{rs}^2$ má determinaci R_L^2. (R_L se rotací nezmění.) Podíl G_m na celkové variabilitě zjistíme jako

$$r_{Gm}^2 = r_m^{*2}/R_L^2 . \tag{12}$$

Transformační matice T, která převádí faktorový systém $\mathbf{F^*}$ na $\mathbf{G^*}$ obsahuje korelační koeficienty mezi $\{F_m\}$ a $\{G_m\}$a má proto důležitou podpůrnou interpretační roli (kontroluje vztah mezi interpretacemi F* a G*).

Poznamenejme, že speciálním případem transformace je násobení číslem -1 u kterýchkoli sloupců A a B (současně) a že tudíž znaménka u faktorových efektů

---

13) Matice $K \times L$ ($K > L$) má *prostou strukturu*, když:

1. každý řádek matice má alespoň jednu nulu;
2. každý sloupec má alespoň L nul;
3. pro každou dvojici sloupců nastává případ, že u několika řádků hodnoty vymizí v jednom a nevymizí v druhém sloupci;
4. pro každou dvojici sloupců vymizí v obou velká část hodnot (pro $L > 4$);
5. u každé dvojice sloupců se vyskytuje jen nepatrný počet řádků, u nichž jsou vysoké hodnoty u obou sloupců.

(Viz K. Überla [1980], H. H. Harman [1967].)

téhož faktoru mohou být změněna (a proto také znaménka u a_{rm}, b_{rm} nemají absolutní význam směru — na rozdíl od součinu $a_{rm}b_{rm}$, který takovou faktorovou změnou není dotčen).

C) Postup řešení

Metoda kanonického rozkladu reziduí vede k hlavnímu řešení

krok 1: určíme $\mathrm{Y} = \|y_{rs}\|$ typu $R \times S$;

krok 2: určíme YY^T typu $R \times R$, její charakteristická čísla $\lambda_1 \geqq \ldots \geqq \lambda_M > 0$ a k nim charakteristické vektory $P_1, P_2, \ldots, P_M$, každý o R složkách, které tvoří matici P typu $R \times M$; $M = h(\mathrm{YY}^T)$, označíme $\Lambda^{\frac{1}{2}} = \mathrm{diag}\,(\sqrt{\lambda_m})$ (diagonální matice $M \times M$ s prvky $\sqrt{\lambda_m}$);

krok 3: určíme $\mathrm{A} = \mathrm{P}\Lambda^{\frac{1}{2}}$;

krok 4: obdobně ke kroku 2 určíme $Q_1, Q_2, \ldots, Q_M$ pro matici $\mathrm{Y}^T\mathrm{Y}$, jež má též $h(\mathrm{Y}^T\mathrm{Y}) = M$ a $\Lambda^{\frac{1}{2}}$ stejnou;

krok 5: určíme $\mathrm{B} = \mathrm{Q}\Lambda^{\frac{1}{2}}$;

krok 6: určíme $r_m^2 = \lambda_m/(\lambda_r + \lambda_2 + \ldots + \lambda_M)$ a $R_m^2 = \sum^m r_j^2$.

Kroky 4 a 5 lze nahradit[14]):

krok 4′: $\mathrm{B} = \mathrm{Y}^T\mathrm{A}\Lambda^{-1}$ $(\Lambda = \Lambda^{\frac{1}{2}}\Lambda^{\frac{1}{2}})$.

Algoritmy rotace neuvádíme, jsou dobře popsané ve výše citovaných monografiích.

D) *Model mnohorozměrného škálování — grafická korespondenční analýza*

Považujeme-li asociační rezidua za hodnoty skalárních součinů mezi body A_r a B_s, které reprezentují kategorie a_r a b_s, jsou řádky matic A a B souřadnicemi v M-rozměrném prostoru, které tento skalární součin vytvářejí (viz (7)). Vezmeme-li v úvahu pouze prvních L souřadnic hlavního řešení, pak tyto souřadnice (body) nejlépe aproximují skalární součiny pomocí L dimenzí. To umožňuje zobrazit vztahy mezi kategoriemi do L-rozměrného prostoru graficky, přičemž R_L^2 určuje přesnost aproximace vstupních relací — každý bod je určen L souřadnicemi. Nejjednodušší škála je jednorozměrná ($L = 1$), řadí kategorie na přímce podle jednotlivých efektů.

Dvojrozměrná škála ($L = 2$) zobrazuje body do roviny, přičemž osa x odpovídá faktoru F_1 a osa y odpovídá faktoru F_2. Kromě interpretace jednotlivých škál (průměty bodů na osy) se zde vyjevují korespondenční relace kategorií přizpůsobené současnému působení obou faktorů F_1, F_2. Pro vícerozměrné škály[15]) je nutno buď použít některé ze speciálních metod grafického zobrazení, nebo (jako v případě programu LINDA) tisknout obrázky pro kombinace dvojic faktorů — například tři obrázky pro $L = 3$: (F_1, F_2), (F_1, F_3), (F_2, F_3). Pro větší počet faktorů (škál) je však tento způsob nepřehledný a grafické zobrazení ztrácí svůj význam.

Při rotaci se vztahy mezi body nemění, pouze se celý vzniklý útvar natáčí podle počátku, relace se nezmění ani při transformacích překlopení, které označují zrcadlově převrácený útvar bodů podle některé osy. Každá dvě řešení modelu (7), resp. (11), lze na sebe převést pomocí rotací a překlopení.

---

14) Při použití tohoto kroku je výhodné aplikovat postup pro $R < S$, v opačném případě je vhodnější vyjít v kroku 2 z matice $\mathrm{Y}^T\mathrm{Y}$, zjistit B a potom $\mathrm{A} = \mathrm{YB}\Lambda^{-1}$

15) V praxi autorů se ukázalo, že pro grafické určení relací je ve velké většině případů postačující použít dvě dimenze (nebo dokonce jen jednu dimenzi); v některých případech je výhodné rozdělit škálu na dílčí vícerozměrné škály o menší dimenzionalitě vzhledem k obsahu faktorů.

Osy obrázku jsou jednorozměrné škály, celý obrázek je vícerozměrná škála. Každá škála odpovídá jednomu faktoru: Čím více se faktor F_m projevuje u dané kategorie, tím extrémnější pozici dostává příslušný bod ve směru osy odpovídající m-té složce vektoru souřadnic. Kolmost os odpovídá nezávislosti faktorů (body, které určují pravý úhel s počátkem mají skalární součin, tj. kovarianci, roven nule).

E) Poznámky k interpretaci faktorů

Jednotlivé faktory interpretujeme (tj. určujeme jejich vlastnosti, obsah, význam) jednak podle jejich efektů na jednotlivé kategorie (a tím i v jednotlivých polích) a jednak z modelového předpokladu, že jsou navzájem nezávislé. Postup obsahuje obvykle kroky:

1) vyloučení faktorů, které mají zanedbatelný vliv na asociační strukturu (malé r_m^2) a které nemají vhodnou a rozumnou interpretaci;
2) vyčlenění faktorů, které mají velmi specifické působení a samostatnou interpretaci;
3) přisouzení významů jednotlivým faktorům podle jejich efektů (projevů) a interakčních vlivů;
4) kontrola významů vzhledem k interpretaci nezávislosti a případná úprava či změna předchozích závěrů;
5) rotace zvolených faktorů a provedení kroků 3 a 4 pro nové řešení;
6) kontrola interpretace pomocí transformační matice mezi hlavním a rotovaným řešením.

Jednou ze základních otázek je, kolik hlavních faktorů ponecháme v interpretačním procesu. Je zřejmé, že interakční struktura $\|y_{rs}\|$ je nejen výsledkem systematických působení, ale že se zde projevují též náhodné chyby sběru a záznamu dat, náhodné vlivy v sociálních procesech atp. Tyto vlivy zahrnujeme do slabých a nesystematicky působících posledních faktorů. Náhodné vlivy se však promítnou i do koeficientů (efektů), jejichž numerické hodnoty budou jen odhady skutečných efektů. Proto i relace v obrázku i číselné relace v maticích A a B je nutno chápat z tohoto pohledu a interpretačně či matematicky je vyhladit. Chceme znovu upozornit na okolnost, že výsledkem metody jsou hypotézy a konstrukce závislé na nepřesných údajích a že interpretační proces musí s možnými výkyvy hodnot počítat. I když budou jistě vyvinuty metody pro statistické testování významnosti faktorů, hlavním kritériem přijetí každého jednoho faktoru bude vždy jeho interpretabilita.

F) Shrnutí výstupních parametrů

Metoda poskytuje celou řadu výstupních parametrů:

1. vstupní data $\|n_{rs}\|$ resp. $\|f_{rs}\|$;
2. normalizace řádků, sloupců nebo celé tabulky;
3. logaritmické reziduální odchylky $\|y_{rs}\|$;
4. matice korelačních koeficientů $\|r(a_r, b_s)\|$, Z-skórů znaménkového schématu $\|Z_{rs}\|$, resp. jejich dosažených významností $\|\alpha_{rs}\|$, znaménkové schéma;
5. normalizované charakteristické vektory P, Q a charakteristická čísla λ_m;
6. hlavní efekty x_{r+}, x_{+s}, x_{++};
7. faktory hlavního řešení A, B, jejich koeficienty determinace r_m^2 a R_m^2;
8. matice obsahující postupně příspěvky faktorů v polích $\|Q_{m(rs)}\|$ a dosud nevysvětlená rezidua $\|y_{rs} - Q_{1(rs)} - \ldots - Q_{m(rs)}\|$, popřípadě i zpětný přepočet k relativním četnostem;
9. rotované řešení podle metody VARIMAX, a to podle zvolené strategie I, II nebo III — matice faktorových efektů A*, B*, koeficienty r_m^2, r_{Gm}^2, R_{Gm}, transformační matice T od hlavního k rotovanému řešení;

10. reziduální výstupy jako v 8 pro rotované řešení;
11. grafické zobrazení pro hlavní i rotované řešení: přímky jednorozměrných škál i rovinné zobrazení všech kombinací dvojic volených faktorů.

Výstupy 2 a 4 nepatří k metodě a pouze ji (na přání a volbu uživatele) doplňují; u každého z výstupů je možno volit či potlačit tisk, k základním výsledkům však patří 7, 9, 11; výstupy 5, 8, 10 mají speciální použití, neboť obsahují normalizované faktory pro komparace predikční role faktorů. Výstup 9 je závislý na volbě typů strategie rotace, na volbě počtu použitých faktorů a na volbě kritéria, které počet faktorů pro rotaci určí samo. Popsaná metoda je založena na matici $\|y_{rs}\|$, avšak u 3 může být voleno i jiné hledisko na indikaci reziduí.

4. Příklady — řešení

Řešení příkladů z části 2 je založeno na programu LINDA. Nejsou ilustrovány všechny možné výstupy, jen ty běžné. Nejsou ani popisovány možnosti[16]) praktic-

Tabulka 3. Hlavní faktory odhalující zkorelované vztahy profesí a věkových kategorií

Profese	Efekty pro profese					Marginální řádkový efekt
	F1	*F2*	*F3*	*F4*	*F5*	
Strojní zámečník	−.016	.104	.493	−.070	−.181	2.23
Soustružník	.914	−.179	−.175	−.362	.093	.82
Frézař	1.147	−.176	−.024	.312	−.049	.11
Brusič	.110	−.111	.307	−.018	−.001	.21
Kovář	−.567	.421	−.352	−.097	−.146	−1.44
Slévač	−.047	.358	−.554	.130	.069	−.74
Seřizovač	−1.045	−.782	−.159	.042	.014	−2.32
Údržbář	−.496	.365	.464	.062	.202	1.12
Koeficient determinace	85.6 %	7.6 %	6.3 %	.4 %	.1 %	—
Věk	Efekty pro věk					Marginální sloupcový efekt
	F1	*F2*	*F3*	*F4*	*F5*	
do 20 let	1.354	.407	.432	−.087	.006	−.69
do 30 let	.465	−.826	−.216	−.172	−.049	1.55
do 40 let	.178	.008	−.211	.424	−.108	1.02
do 50 let	−.196	.090	−.320	.032	.281	.75
do 60 let	−.697	.499	−.394	−.218	−.133	−.05
nad 60 let	−1.104	−.178	.710	.031	.003	−2.58
Koeficient determinace	85.6 %	7.6 %	6.3 %	.4 %	.1 %	—

kých kroků, ať už se týkaly použití výsledků v rozhodování, v širších rozborech či návazných analytických postupů a úloh.

Příklad 1. Statistická souvislost věkové a profesní struktury v průmyslovém podniku

Hlavní řešení faktorového modelu pro tabulku 1 je uvedeno v tabulce 3.

Čtvrtý a patý faktor *F4* a *F5* jsou bezvýznamné (0.42 % a 0.07 %). Podstatnou část asociace vysvětluje první faktor *F1* (85.6 %), jehož závislost na věku je vidět z efektů, které se projevují u věkových kategorií. Postupné rozdíly mezi faktorovými koeficienty (.889, .287, .374, .501, .407)

16) V příkladech (vzhledem k účelu článku) nejsou uváděny širší formulace sociologických úloh, z nichž byla data vyňata a v nichž metoda sloužila jako dílčí analytický krok.

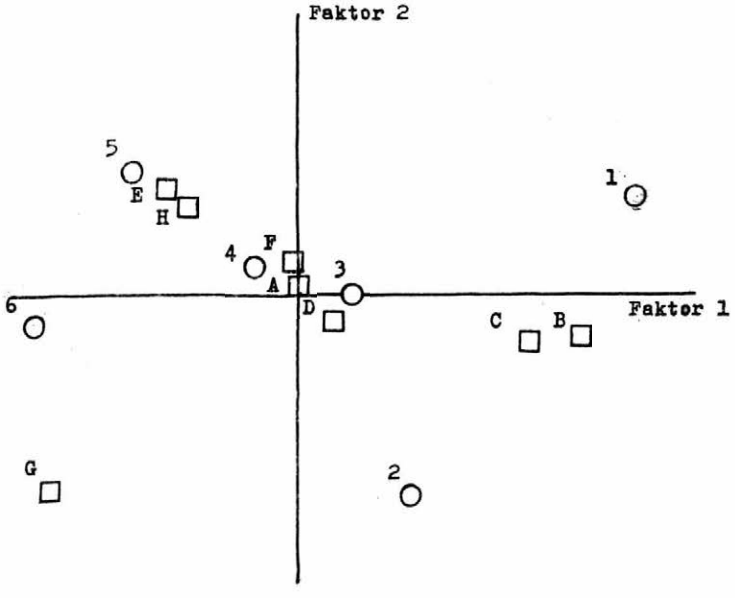
ukazují, že monotónnost (linearita) vlivu faktoru je porušena především u prvních dvou věkových skupin. Faktor působí v plné intenzitě u soustružníků a frézařů, kde indikuje silnější obsazení u mladých kategorií a slabší u starších, zatímco opačně působí se stejnou intenzitou u seřizovačů a s poloviční u kovářů a údržbářů. Tento faktor nepřináší zřejmě žádnou zvláštní novou informaci oproti věkovému průměru — ukazuje se však jeho dominance a nerovnoměrnost vlivu u věkových kategorií.

Druhý faktor *F2* se týká především kategorie *„do 30 let"*, která má tak zvláštní postavení v asociační struktuře. Její odchylka je kompenzována u kategorií *„do 20 let"* a *„do 60 let"*. Působí u seřizovačů (zvýšené procento 21—30letých) a u kovářů, slévačů a údržbářů (opačný vliv).

Třetí faktor se dotýká nejvíce poslední věkové kategorie a ukazuje závislost profesí na důchodcích (a to dodatkově, nezávisle a mimo vliv společného prvního faktoru).

Souvislostní vztahy věkových a profesních kategorií odvislé od *F1* a *F2* ukáže obrázek korespondence (třetí faktor vzhledem k jeho specifické interpretaci nebudeme uvažovat).

Obrázek 2. Souvislostní vztahy věkových a profesních kategorií



Profese:		Věkové kategorie:	
A	strojní zámečník	1	do 20 let
B	soustružník	2	do 30 let
C	frézař	3	do 40 let
D	brusič	4	do 50 let
E	kovář	5	do 60 let
F	slévač	6	nad 60 let
G	seřizovač		
H	údržbář		

V uspořádání profesí (*C*, *B*, *D*, *A*, *F*, *H*, *E*) hraje rozhodující roli první faktor. Profese *G* je mimo tento směr vychýlena druhým faktorem. Věkové kategorie leží podél osy faktoru 1. který je zřejmě uspořádal.

Tabulka 4. Faktory korespondence profesních a věkových kategorií (rotace efektů pro věk metodou VARIMAX)

Věk	Efekty pro věk		Součet čtverců
	G1	*G2*	
do 20 let	1.398	—.212	1.998
do 30 let	.067	—.945	.898
do 40 let	.165	—.069	.032
do 50 let	—.139	.165	.047
do 60 let	—.416	.749	.734
nad 60 let	—1.074	.311	1.251
Vliv faktoru	67.2 %	32.8 %	—
Profese	Efekty pro profese		Součet čtverců
	G1	*G2*	
Strojní zámečník	.031	.101	.011
Soustružník	.749	—.553	.867
Frézař	.961	—.650	1.346
Brusič	.052	—.148	.025
Kovář	—.332	.623	.499
Slévač	.111	.344	.130
Seřizovač	—1.279	—.260	1.702
Údržbář	—.292	.542	.380
Vliv faktoru	67.2 %	32.8 %	—

Celkový podíl obou faktorů na asociační struktuře = 93.2 %.

Tyto údaje lze interpretovat tak, že u každé profese existuje určité věkové posunutí, které se projevuje pravděpodobně jako důsledek dlouhodobých, popřípadě i ustálených procesů stárnutí struktury i její obnovy, i jako důsledek zavádění, rozšiřování, nebo naopak redukce některých profesí. Vedle těchto procesů se uplatňuje i nedávná jednorázová náborová akce pro doplnění profesí (seřizovač ve skupině 21—30letých). Nezávisle na předchozích vlivech se tu ještě projevuje doplňování struktury profesí pomocí pracujících v důchodovém věku.

Rotace řadkových koeficientů pro *F1* a *F2* znamená natočení obrázku 2 tak, aby body *A*, *B*, ..., *H* co nejvíce přiléhaly k osám. Řešení však nepřináší žádnou novou informaci.

Rotace sloupcových koeficientů přináší nový pohled (viz tabulka 4). Znamená natočení obrázku tak, aby body věkových kategorií co nejvíce přiléhaly k osám.

První faktor *G1* zůstává faktorem vlivu podél celé věkové škály, s výjimkou kategorie 21—30letých, jež je z něho eliminována — její vazby jsou zcela přeneseny do faktoru *G2*, který zdůrazňuje její asociaci nejsilněji s kategorií „*51—60 let*" a částečně též „*nad 60 let*". Jestliže vezmeme v úvahu podpůrný koeficient —.212 u kategorie „*do 20 let*". naznačuje *G2* tuto interpretaci: odlišení těch profesních kategorií, u nichž se nedostatek pracovních sil řešil kampaní a intenzívním náborem mladých lidí, a těch kategorií, kde problém ještě řešen nebyl a kde se spoléhá na pracující důchodce. Podíl obou rotovaných faktorů (67.2 % : 32.8 %) ukazuje na to, že vliv kategorie „*21—30 let*" na strukturu je velmi podstatný a že při úvahách o struktuře není druhý faktor zanedbatelný.

Příklad 2. Trvale obydlené byty podle velikosti v obvodech Prahy

Hlavní řešení odhalilo 4 faktory (83.7 %, 10.2 %, 4.4 %, 1.7 %). Poslední faktor budeme považovat za nevýznamný. Předposlední vysvětluje sice jen necelých 5 % reziduální rozmanitosti, mohl by však, vzhledem k povaze dat, mít značný význam — jeho interpretace však byla nejasná, a proto i *F3* byl z dalších úvah vyloučen. První nejsilnější faktor *F1* ukázal zcela pravi-

Tabulka 5. Faktory působící ve struktuře rozložení bytů podle velikosti v obvodech Prahy (řešení VARIMAX rotací řádkových efektů)

Obvod	Efekty pro obvody Prahy		Marginální řádkové efekty
	G1	*G2*	
Praha 1	—.117	.424	—.56
Praha 2	.198	—.052	—.39
Praha 3	1.004	—.291	—.63
Praha 4	—.347	—.243	.98
Praha 5	—.292	.481	.23
Praha 6	—.261	.368	.37
Praha 7	.651	—.107	—.90
Praha 8	—.308	—.445	.21
Praha 9	—.515	—.037	.26
Praha 10	—.014	—.099	.42
Podíl faktoru	69.90 %	30.10 %	
Velikost bytů	Efekty pro velikost bytů		Marginální sloupcové efekty
	G1	*G2*	
1	.998	—.070	1.001
2	.484	.028	.235
3	—.262	—.724	.593
4	—.416	.172	.203
5 a více	—.804	.594	.999
Podíl faktoru	69.90 %	30.10 %	

Celkový podíl obou faktorů na asociační struktuře = 93.8 %.

delné efekty velikosti bytu (.974, .454, —.035, —.449, —.944), a tím seřadil pražské obvody obdobně jako podle průměrů. Druhý faktor *F2* ukázal na výraznou samostatnou roli bytů o velikosti 3 ve struktuře bytového rozmístění v pražských obvodech.

Rotované řešení (prostá struktura řádkových efektů) ukázalo obdobnou strukturu (viz tabulka 5), v níž se o něco ostřeji vyčlenila role kategorie 3.

Podíl rotovaných faktorů 70 % : 30 % ukazuje, že druhý faktor (*G2*) má značně silný vliv. Polarizuje výstavbu bytů o 3 místnostech (2 + 1) oproti pěti a více místnostem (4 + 1 a více), a to nejsilněji v *P8* a dále v *P3* a *P4* proti *P1*, *P5* a *P6*. Oba faktory odrážejí odkaz historické minulosti i strategii bytové výstavby v nedávných letech spolu s procesem udržování a obnovy staršího i historického fondu. Vzhledem k obsahové heterogenitě uvnitř velikostních kategorií je podrobnější a přesnější závěr obtížný, naznačuje však užitečnost dalších analýz obdobných dílčích struktur určených stářím bytového fondu, druhem domu atp. Korespondenční mapa je znázorněna na obrázku 3.

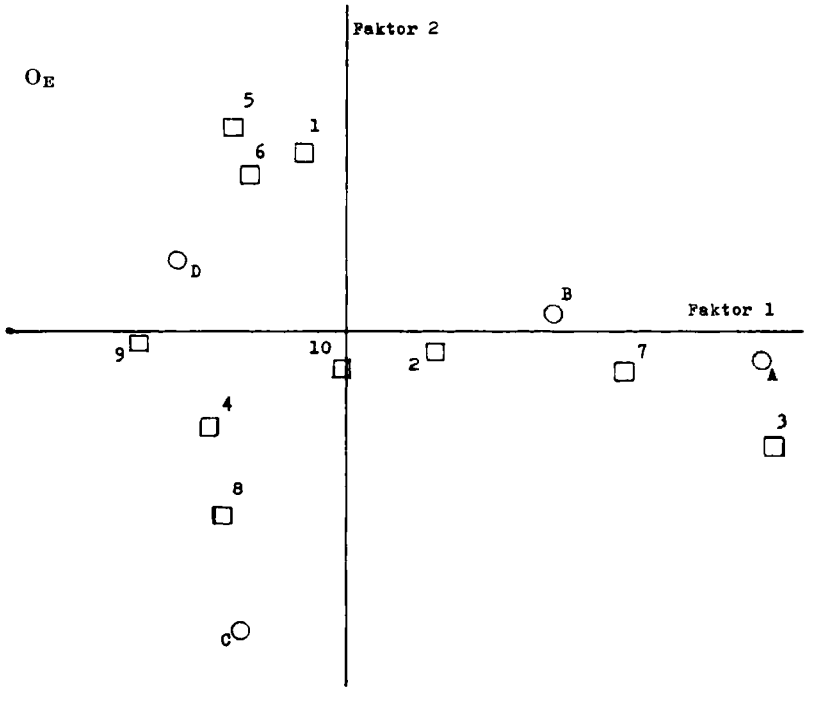
Příklad 3. Vybavování mladých manželství

Kladná korelace obou proměnných není nikterák vysoká a tak se pomocí metodiky LINDA-A pokusíme o hlubší pohled do souvislostní struktury. Hlavní řešení (viz tabulka 6) musíme hodnotit vzhledem ke genezi dat, která jsou zatížena mnoha náhodnými vlivy. Zamítneme faktory *F4* a *F5*, dokonce i *F3* (3% vliv bychom uvažovali jen v případě logické a dobře argumentované interpretace). I u *F1* a *F2* však musíme předpokládat nepravidelnosti[17]) způsobené náhodnými šumy — faktorové koeficienty budou mít indikační roli, jejich vztahy pouze relace v datech naznačují.

První faktor ukazuje silnou přímou korelaci (r_1^2 = .80; odpovídá koeficientu korelace r_1 = = .89), která však nezabírá celé škály obou znaků, nýbrž spojuje hodnoty ***4***, ***5 a více*** u PPVD a ***5***, ***6***, ***7 a více*** u PPOP. U ostatních kategorií se projevuje určitá nerovnoměrnost, která naznačuje, že rychlost vybavování osobními předměty je na začátku pomalejší, později se však (zřejmě po vybavení domácnosti nejnutnějšími předměty) zrychluje, zatímco vybavování domácnosti se zde jeví jako proces rovnoměrný. Spojení kategorií ukazuje, že u horních hodnot je již doplňování předmětů nesystematické a „náhodné".

Druhý faktor ($r_2^2 = .16$, $r_2 = .40$) kontrastuje u PPVD kategorie *0* a *5* a více proti ostatním, u PPOP vyděluje *0*, zatímco u zbytku škály se projevuje doplňkový (zpožděný) vybavovací trend, který se projeví i na konci škály, jenž u *F1* nebyl rozlišen. Interpretace je poněkud nejasná: *F2* může znamenat jiný proces vybavování u části populace nebo jinou posloupnost v kvalitě vybavování, která svými ekonomickými souvislostmi ovlivňuje další vývoj; může také znamenat absenci vnější pomoci (rodiče, půjčky). Je tu však naznačeno štěpení populace na dvě části, u nichž postupuje proces vybavování různými směry.

Obrázek 3. Korespondenční mapa trvale obydlených bytů podle velikosti v obvodech Prahy



Obvody:				Velikost bytů:	
1	Praha 1	6	Praha 6	A	1 místnost
2	Praha 2	7	Praha 7	B	2 místnosti
3	Praha 3	8	Praha 8	C	3 místnosti
4	Praha 4	9	Praha 9	D	4 místnosti
5	Praha 5	10	Praha 10	E	5 a více místností

Body *P1—P8* i body *D1—D6* vytvářejí souběžné vývojové křivky, z nichž druhá je poněkud posunuta. Zvláštní poloha bodů *P1, D1* může být ovlivněna tím, co lidé do manželství přinášejí (svatební dary, vlastnictví předmětů před vstupem do manželství). Upřesnění hypotézy by znamenalo specifické třídění mezi jednotlivými typy předmětů a aplikací jiných analytických modelů (a pochopitelně nejlépe vývojová data, která jsou však k dispozici zřídka).

Po rotaci k prosté struktuře řádkových efektů dostaneme poměr vlivů 57 % : 43 % (v rámci celkových 96 %).

Tyto faktory naznačují jinou hypotézu: *G1* odpovídá procesu postupného vybavování s důrazem na domácnost, procesu, v němž se začíná z ničeho, ale v němž má rodina zásadní přednost, přičemž osobní předměty jsou doplňovány podle možností. Při tomto typu procesu vybavování je zhruba do 30 let věku manželů domácnost vybavena 5 předměty (z použitého seznamu). Druhý typ procesu (reprezentovaný faktorem *G2*) je charakterizován souběžným vybavováním domácnosti i osob; je pravděpodobně umožněn buď lepší finanční situací, nebo pomocí zvnějšku

Tabulka 6. Faktorová dekompozice asociačního vztahu mezi vybavováním předměty pro domácnost (PPVD) *a předměty osobní potřeby* (PPOP) *u souboru mladých manželství (do 30 let) (hlavní řešení)*

Počet předmětů pro domácnost	Efekty pro PPVD				
	F1	*F2*	*F3*	*F4*	*F5*
0	−1.379	−.449	.597	.031	−.077
1	−.919	.307	−.692	−.001	−.204
2	−.272	.344	−.066	−.271	.410
3	.601	.672	.176	.516	.037
4	1.130	.289	.255	−.385	−.228
5 a více	.838	−1.162	−.268	.110	.062
Koeficient determinace	80.1 %	15.9 %	3.1 %	.8 %	.2 %
Počet předmětů osobní potřeby	Efekty pro PPOP				
	F1	*F2*	*F3*	*F4*	*F5*
0	−1.744	−.716	−.163	−.021	.030
1	−.769	.788	.078	.150	−.165
2	−.047	.559	.369	−.212	.255
3	.128	.299	.283	.109	−.061
4	.501	.196	−.692	.030	.263
5	.640	.040	−.362	.072	−.289
6	.559	−.438	.121	−.515	−.123
7 a více	.732	−.728	.366	.387	.092
Koeficient determinace	80.1 %	15.9 %	3.1 %	.8 %	.2 %
Kumulativní podíl	80.1 %	96.0 %	99.1 %	99.8 %	100. %

Tabulka 7. Faktorová dekompozice asociačního vztahu mezi vybavováním předměty pro domácnost (PPVD) *a předměty osobní potřeby* (PPOP) *u souboru mladých manželství (do 30 let) (rotace efektů pro* PPVD *metodou VARIMAX)*

Počet předmětů pro domácnost	Efekty pro PPVD	
	G1	*G2*
0	−1.3 3	.404
1	.587	.770
2	.032	.437
3	.874	.217
4	1.097	−.396
5 a více	.040	−1.432
Podíl faktorů	57.1 %	42.9 %
Počet předmětů osobní potřeby	Efekty pro PPOP	
	G1	*G2*
0	−1.845	.388
1	−.193	1.084
2	.276	.489
3	.274	.175
4	.525	−.120
5	.551	−.327
6	.216	.677
7 a více	.196	−1.014
Podíl faktorů	57.1 %	42.9 %

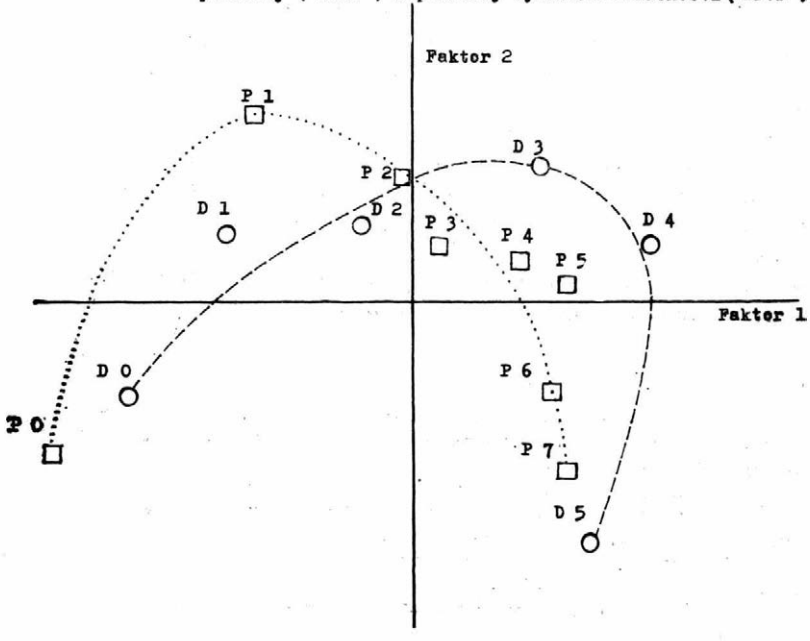
Celkový podíl obou faktorů na asociační struktuře = 96 %.

(jak ukazuje posunutí od kategorie *0* předmětů, která má samostatný význam a naznačuje splynutí dvou různých kvalit do jednoho popisu).

Transformační matice přechodu od hlavního k rotovanému řešení je:

.8270	—.5622
.5622	.8270

Obrázek 4. Korespondenční škála procesu vybavování osobními předměty (PPOP) a předměty vybavení domácnosti (PPVD)



Počet předmětů pro domácnost:	
D 0	0
D 1	1
D 2	2
D 3	3
D 4	4
D 5	5 a více

Počet předmětů osobní potřeby:	
P 0	0
P 1	1
P 2	2
P 3	3
P 4	4
P 5	5
P 6	6
P 7	7 a více

Výsledkem tohoto příkladu jsou tedy jen pouhé (dokonce v určitém smyslu i konkurující si) hypotézy, které jistě mohou být různými výzkumníky formulovány různě, vyžadují proto další analytické zpracování a ověření. V každém případě jsme dostali velmi důležitou informaci o charakteru ordinální a lineární korelace i o doplňkové složce základního hlavního faktoru.

5. Závěry

Uvedený přístup k faktorové dekompozici kontigenčních tabulek je jedním z mož-ných. Pro výrazné působení faktorů se různé přístupy nebudou příliš lišit. Volba logaritmicko-lineárních členů vyplynula z praktických zkušeností i z odtud odvozeného názoru, že poměrové porovnávání relativních četností je vhodnější než rozdílové. Postup LINDA-A *odpovídá metodě hlavních komponent.* Není v něm zahrnuto testování hypotéz ani o počtu faktorů, ani o významnosti faktorových efektů. Nejsou známy ani intervaly spolehlivosti pro keoficienty, ani komparační testy pro paralelní

výsledky u různých souborů. K překonání těchto problémů může přispět další rozpracování metodiky, a to ve dvou směrech: a) *postupy statistické inference* o parametrech modelu, b) *zahrnutím chybového členu* a specificity do modelu[17].

Metodika má především explorační charakter, její výsledky *slouží k formulaci hypotéz obsahového* (sociologického), nikoli tedy statistického typu, a proto hlediskem přijetí modelu i jednotlivých faktorů je především interpretovatelnost a užitečnost pro závěry. Vzhledem k tomu, že výsledná interpretace je (jako u všech typů faktorové analýzy) velmi odvislá od subjektu i od dalších poznatků, je obtížné překročit hypotetický stupeň závěrů; zajímavé výsledky jsou zpravidla konfrontovány s dalšími statistickými informacemi, které data skýtají. Užitečnost metodiky v praxi už se prosadila a může se stát, přes všechny své limity, značným přínosem pro analytický arsenál sociologického výzkumu.

I když se model a některé jeho momenty mohou jevit praktickým sociologům jako obtížné, náročnost na uživatele metodiky není vyšší než u faktorové analýzy. Analytická úloha sama je obtížnější v tom smyslu, že nás zajímají vztahy dvou množin objektů či působení faktorů na dvě množiny objektů současně. Po aplikaci několika případů se uživatel-sociolog seznámí s významem výstupních parametrů a tisků a matematické formulace pak už k ničemu nepotřebujeme.

Literatura

Benzécri, J. P.: *Statistical Analysis as a Tool to Make Patterns Emerge from Data.* In: *Methodologies of Pattern Recognition* (S. Waternabe, Ed.). New York, Academic Press 1969, pp. 35—60.

Guttman, L.: *The Quantification of a Class of Attributes: a Theory and Methods of Scale Construction.* In: *The Prediction Adjustment* (P. Horst, Ed.). New York, Social Science Research Council 1941, pp. 251—364.

Guttman, L.: *Metricizing Rank-Ordered or Unordered Data for Linear Factor Analysis.* Sankhya 21, 1959, pp. 257—268.

Haberman, S. J.: *The Analysis of Residuals in Cross- Classified Tables.* Biometrics 29, 1973, pp. 205—220.

Harman, H. H.: *Modern Factor Analysis.* Chicago, The University of Chicago Press, 1967.

Hill, M. O.: *Correspondence Analysis: A Neglected Multivariate Method.* Applied Statistics 23, 1974, pp. 340—354.

Jelisejeva, I. I.: *Statističeskije metody izměrenija svjazej.* Leningrad, Izdatěl'stvo leningradskogo universitěta 1982.

Kasalová H. — Řehák J.: *Diferenciační tendence v recepci umění.* Sociologický časopis XX, 1984, s. 35—57.

Kendall, M. G. — Stuart A.: *Statističeskije vyvody i svjazi.* Moskva, Nauka 1973.

Lancaster, H. O.: *Some Properties of the Bivariate Normal Distribution Considered in the Form of Contingency Tables.* London, Ch. Griffin 1957.

Lauber J. — Linhart J.: *Kanonická analýza kontingenčních tabulek.* Sociologický časopis XIV, 1978, s. 610—618.

Linhart J. — Šafář Z.: *Programování a třídění statistických výpočtů pomocí samočinných počítačů.* Sociologický časopis III, 1967, s. 435—443.

Řehák J.: *Základní deskriptivní míry pro rozložení ordinálních dat.* Sociologický časopis XII, 1976, s. 416—431.

Řehák J. — Loučková I.: *Korespondenční analýza.* In: *Škálování a kvantifikace* (red. J. Řehák). ČSSS, sekce metod a technik, Mikulov 1981, s. 102—111.

Řehák J. — Loučková I.: *Faktorová analýza kontingenčních tabulek. Robust.* Praha, JČMF 1982, s. 71—75.

Řehák J. — Řeháková B.: *Měření statistické závislosti nominálních znaků.* Sociologický časopis IX, 1973, s. 404—418.

Řehák J. — Řeháková B.: *Míry statistické závislosti pro ordinální znaky.* Sociologický časopis XI, 1975, s. 75—92.

Řehák J. — Řeháková B.: *Analýza kontingenčních tabulek: rozlišení dvou základních typů úloh a znaménkové schéma.* Sociologický časopis XIV, 1978, s. 619—631.

17) To platí u všech výběrových dat a dat se sníženou reliabilitou, proto interpretace faktorové analýzy musí být vždy velmi opatrná.

Řehák J. — Řeháková B.: *Analýza kategorizovaných dat v sociologii.* Praha, Academia (v tisku) 1984.
Uberla K. *Faktornyj analiz.* Moskva, Statistika 1980; česky: *Faktorová analýza.* Bratislava, Alfa, 1976.

Резюме

Я. Ржегак - И. Лоучкова: Факторный анализ в таблицах сопряженности. Метод ЛИНДА-А — описание и применение

Факторный анализ в таблицах сопряженности с помощью метода ЛИНДА-А (линейный декомпозиционный алгоритм для ассоциативных таблиц) основан на разбиении остатков или логарифмических остатков от модели независимости.

В первой части статьи анализируются интерпретационные схемы в анализе ассоциаций таблицы сопряженности. В случае принятия альтернативной гипотези о статистической зависимости строчной и столбцовой переменной одним из методов является анализ факторов, которые совместно действуют на обе переменные и влияние которых проявится в отдельных клетках в том, что остатки неслучайно повысятся. В этой части описаны факторная модель и шесть схем интерпретации, которые наиболее распространены и наиболее важны в задачах анализа данных и аналитико-статистической формулировки задачи.

Во второй части формулируются примеры из аналитической практики: взаимоотношение возрастной и профессиональной структуры в промышленном предприятии, взаимоотношение размера квартиры и районов Праги и наличие предметов личного потребления у молодоженов и оснащение домашнего хозяйства. Эти примеры подвергаются методологическому анализу.

Третья часть содержит модель факторной декомпозиции и описание метода ЛИНДА-А: математическую формулировку, однозначность решения и ротации, порядок решения, модель многомерного шкалирования, замечания по интерпретации факторов и обобщение выходных параметров.

В четвертой части решены раннее сформулированные примеры.

В заключение намечаются пути дальнейшего развития модели и методики.

Summary

J. Řehák — I. Loučková: Factor Analysis in Contingency Tables. LINDA-A Method — Description and Application

The factor analysis in contingency tables with the aid of the LINDA-A method (linear decomposition algorithm for association tables) is based on the decomposition of residuals, or logarithmic residuals, from the model of independence.

The first part of the paper describes interpretation schemes in the analysis of associations of the contingency table. If we accept the alternative hypothesis concerning the statistical dependence of the row and column symbol, one of the methods is the analysis of factors simultaneously affecting both variables; their effect manifests itself in the individual cells by a nonrandom increase of the residuals. The factor model and its six interpretation schemes that are most common and most important in data-analysis are described.

The second part is concerned with the formulation of examples from analytical practice: the relation between age and professional structure in an industrial enterprise, the relation between the size of appartments and the districts of Prague, between the equipment of young married couples with personal and household equipment. These examples are subjected to methodological analysis.

The third part presents a model of factor decomposition as well as a description of the LINDA-A method: mathematical formulation, characterization of the class of solutions, rotation, algorithm, model of multidimensional scaling, comments on factor interpretation, and summary of output parameters.

The fourth part is devoted to solving the previously formulated examples.

In conclusion, ways of the further development of the model and of the respective methodology are outlined.

Obdobným problémem a příbuzným způsobem řešení jako v této stati se zabývá monografie Shizuhika Nishisata: *Analysis of Categorical Data: Dual Scaling and Its Applications.* Toronto, University of Toronto Press 1980.