SZIGMA

Matematikai közgazdasági folyóirat

A Magyar Közgazdasági Társaság Matematikai-Közgazdasági Szakosztályának lapja

E szám szerzői:

Csicsman József, a KSH munkatársa, Dr. Chikán Attila, a Marx Károly Közgazdaságtudományi Egyetem adjunktusa, Dr. Futó Péter kandidátus, az Építéstudományi Intézet tudományos csoportvezetője, Dr. Füstös László, az MTA Szociológiai Kutató Intézet tudományos munkatársa, Hamza Lászlóné, a SZÁMKI munkatársa, Losonczy Istvánné, a SZÁMKI munkatársa, Meszéna György, a Marx Károly Közgazdaságtudományi Egyetem docense, a Matematikai és Számítástudományi Intézet osztályvezetője, S. Benedikt Vera, a SZÁMKI tudományos munkatársa, Simonné Dr. Mosolygó Nóra, az OT Tervgazdasági Intézet főelőadója, Subicz Péter, a SZÁMKI munkatársa, Vári Anna, a SZÁMKI tudományos munkatársa, Zsellér Gyula, a SZÁMKI tudományos munkatársa.

Szerkesztőség: Budapest XI., Budaörsi út 43–45.

Levélcím: 1361 Budapest, Pf. 11.

Terjeszti a Magyar Posta. Előfizethető bármely postahivatalnál, a kézbesítőknél, a Posta hírlapüzleteiben és a Posta Központi Hírlap Irodánál (PKHI 1900 Budapest V., József nádor tér 1.) közvetlenül vagy postautalványon, valamint átutalással a PHI 215–96162 pénzforgalmi jelzőszámára. Egyes példányok beszerezhetők az 1055 Budapest V., Bajcsy-Zsilinszky út 76. sz. alatti hírlapboltban

Előfizethető és példányonként megvásárolható: az AKADÉMIAI KIADÓ-nál, 1363 Budapest V., Alkotmány u. 21. Telefon: 111–010. Pénzforgalmi jelzőszámunk: 215–11488., és az AKADÉMIAI KÖNYVESBOLT-ban, 1368 Budapest V., Váci u. 22. Telefon: 185–612. Előfizetési díj egy évre: 40,– Ft

Külföldön terjeszti a KULTÚRA Külkereskedelmi Vállalat, H-1389 Budapest Pf. 149

A Szigmának ezt a számát egyetlen témának szenteltük. Tárgyunk a cluster analízis (tanító nélküli osztályozás, automatikus osztályozás). Némi vonakodással tartottuk meg a diszciplína angol nyelvű megnevezését, de elfogadott magyar elnevezése nincs. A kérdéssel most ismerkedő olvasónak a Füstös—Mészéna—Simonné szerzőhármas első cikke nyújt bevezetést, de más cikkek is tartalmaznak bevezető részeket. Közöttük az átfedéseket nem tudtuk kiküszöbölni, de még ellentmondás is akadhat. A további cikkek főképp a cluster analízis hazai alkalmazási tapasztalataival foglalkoznak, Futó Péter cikke egy új cluster modellt és technikát mutat be.

FÜSTÖS LÁSZLÓ—MESZÉNA GYÖRGY—SIMONNÉ MOSOLYGÓ NÓRA

Cluster analízis: fogalmak és módszerek.

1. Bevezetés

A számítógépes adatfeldolgozás elterjedésével gyakorlati lehetőség nyílt a különböző sokváltozós matematikai statisztikai módszerek széles körű alkalmazására az empirikus vizsgálatok eredményeinek értékelésénél. A klasszifikációs technikák – igen változatos körülmények között – a megfigyelt objektumok osztályokba sorolását teszik lehetővé. Az objektumokat a lehető legáltalánosabban értelmezhetjük, objektumok összességének tekintünk minden kvantitatív vagy kvalitatív jellemzőkkel definiált egyedekből álló rendszert. Az osztályok meghatározása tanulási folyamat eredménye, melynek két fő típusát különböztetjük meg:

1. Tanulás tanítóval,
2. Tanulás tanító nélkül (cluster analízis).

Az első esetben a gép kiértékelt tananyagot kap és – a megfelelő algoritmus segítségével – ennek az információnak az alapján végzi az osztályozást. A tanító nélküli tanulásnál az osztályokat kizárólag a tananyag (a minta) felhasználásával alakítják ki, valamilyen előre megadott osztályozási kritérium alapján. A döntési szabálynak a probléma szempontjából legfontosabb információkat kell tartalmaznia, az alakfelismerésnek ezt a lépését *lényegkiemelésnek* nevezzük, s ez magában foglalja:

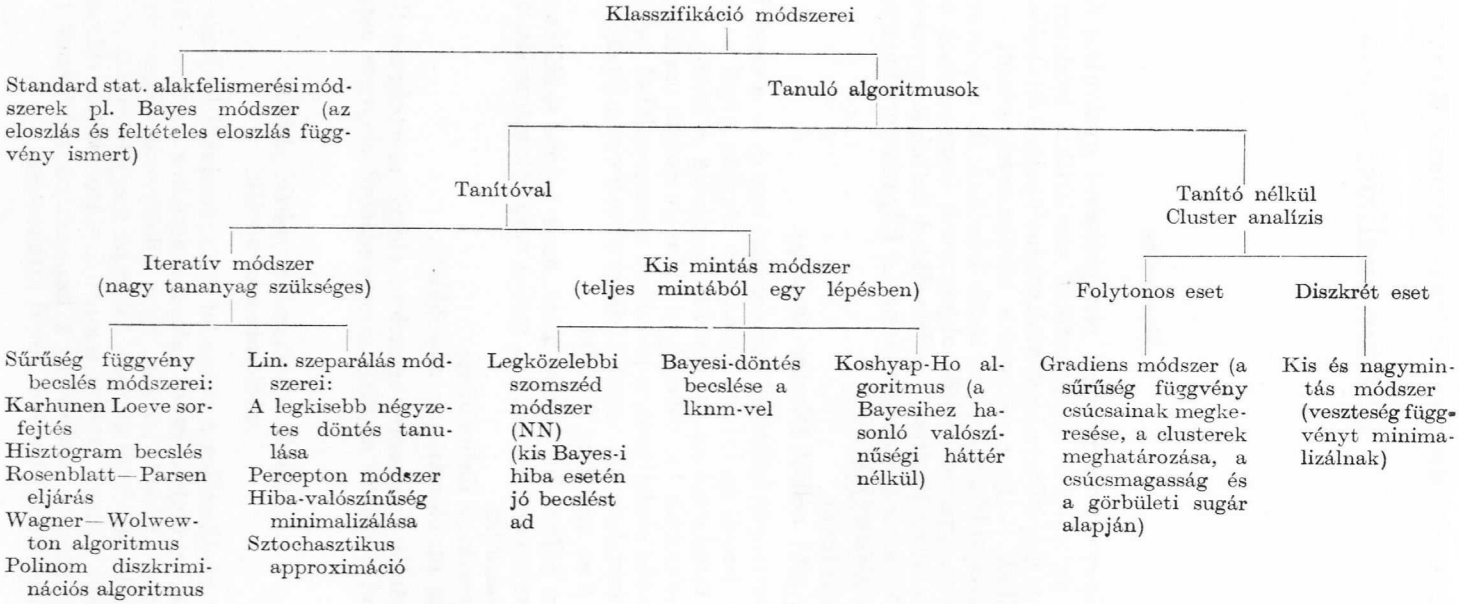
1. A lényeges jellemzők kiválasztását, azok mérési skálájának definiálását.
2. Az adatrendszer egységesítését, azaz a változók mérési skáláinak szükséges transzformációját.
3. A mértékrendszer definiálását.
4. A súlyozás problémájának megoldását.

A klasszifikációs módszerek elmélete eléggé szerteágazó (l. 1. táblázat), a továbbiakban a *cluster analízis* szempontjából lényeges aspektusokat tárgyaljuk.

Jellemzők kiválasztása, mérési skálák, skálatranszformációk

A jellemzők kiválasztása a vizsgálat szempontjából lényegesnek ítélt tulajdonságok számbavételét, ezen tulajdonságokhoz mérési skálával ellátott változók hozzárendelését jelenti. Bár tulajdonság lehet bármely ismérv, mennyiségi érték, minőségi állapot, földrajzi megjelölés stb. mégis a tulajdonságokhoz rendelt változó értékét tekintjük matematikai változónak. A változók típusától függenek elsősorban a kapcsolatok mérésének módszerei. Célszerű ezért a különböző típusok rövid áttekintése.

1. táblázat


Klasszifikáció módszerei
Standard stat. alakfelismerési módszerek pl. Bayes módszer (az eloszlás és feltételes eloszlás függvény ismert)
Tanuló algoritmusok
Tanítóval
Tanító nélkül
Cluster analízis
Iteratív módszer (nagy tananyag szükséges)
Kis mintás módszer (teljes mintából egy lépésben)
Folytonos eset
Diszkrét eset
Sűrűség függvény becslés módszerei:
Karhunen Loeve sorfejtés
Hisztogram becslés
Rosenblatt—Parsen eljárás
Wagner—Wolwewton algoritmus
Polinom diszkriminációs algoritmus
Lin. szeparálás módszerei:
A legkisebb négyzetes döntés tanulása
Percepton módszer
Hiba-valószínűség minimalizálása
Sztochasztikus approximáció
Legközelebbi szomszéd módszer (NN) (kis Bayes-i hiba esetén jó becslést ad
Bayesi-döntés becslése a lknm-vel
Koshyap-Ho algoritmus (a Bayesihez hasonló valószínűségi háttér nélkül)
Gradiens módszer (a sűrűség függvény csúcsainak megkeresése, a clusterek meghatározása, a csúcsmagasság és a görbületi sugár alapján)
Kis és nagymintás módszer (veszteség függvényt minimalizálnak)

Alapvetően két szempont alapján teszünk különbséget: az értékkészlet nagysága és a mérési mód szerint.

2. táblázat

A változók típusai

Mérési mód	Értékkészlet nagysága		
	folytonos	diszkrét	bináris
Nominális	—	születési hely	nő — férfi igaz — hamis
Ordinális	hangintenzítás, fényerősség	tanulmányi eredmény munkahelyi beosztás	kicsi — nagy jó — rossz
Intervallum	hőmérséklet C°-ban	jövedelem	feleség száma (0 vagy 1)
Arány	életkor	családonkénti gyermekek száma	két különböző egységpár

[1] L. [1]. 27. o.

A változók típusainak differenciált megkülönböztetésével és a különböző típusú változók közötti skálatranszformációk segítségével lehetővé vált a kevert változótípusok együttes kezelése. Az egységesítéshez alkalmazandó skálatranszformációt a változórendszer struktúrája határozza meg, általános alapelv az információveszteség minimalizálása.

2. A kapcsolatok mérésének módszerei

Legyen adva egy n elemű statisztikai sokaság, amelyet S-el jelölünk, S az osztályozandó objektumok véges, nem üres halmaza:

$$S = \{s_1 \dots s_n\}.$$

Adottnak tekintjük még a vizsgálat céljára kiválasztott tulajdonságok

$$T = \{X_1 \dots X_m\}$$

m-elemű halmazát.

Az osztályozás kiinduló adatbázisa az objektumokat és azok tulajdonságait tartalmazó T, $n \times m$-es adat-mátrix.

	T				
S	X_1	...	X_j	...	X_m
s_1	x_{11}	...	x_{1j}	...	x_{1m}
.					
.					
.					
s_i	x_{i1}	...	x_{ij}	...	x_{im}
.					
.					
.					
s_n	x_{n1}	...	x_{nj}	...	x_{nm},

ahol x_{ij} az i-edik objektum j-edik tulajdonságának megfigyelt vagy mért értéke. Az osztályozás tényleges input adata a T mátrixból a sorok vagy oszlopok páronkénti összehasonlításával keletkező szimmetrikus DC mátrix (dissimilarity coefficient). Az előbbi esetben a DC mátrix az objektumok közötti, az utóbbi esetben pedig a tulajdonságok közötti ún. taxonomikus távolsági vagy hasonlósági mérőszámokat tartalmazza. A továbbiakban a T → DC transzformáció kérdését vizsgáljuk. A kapcsolatok mérési módszerét egyrészt a T mátrix változóinak típusa határozza meg, másrészt pedig az, hogy az objektumokat vagy a tulajdonságokat kívánjuk összehasonlítani. Ez utóbbi esetre a matematikai statisztika számos jól használható mérőszámot dolgozott ki. Bizonyos esetekben úgy tűnhet, hogy az ilyen fajta feladatok nem is képezik az automatikus osztályozás feladatát. Azonban a botanikában, zoológiában, pszichológiában egyre elterjedtebben használják a cluster analízis módszereit az ismérvek osztályozására. Az ilyen típusú mérőszámok rövid áttekintése azért is indokolt, mert bizonyos esetekben – más értelmezéssel – ezeket is felhasználhatjuk az objektumok közötti kapcsolatok mérésére.

Az ismérvek közötti hasonlósági mutatók

A taxonomikus hasonlósági mérőszámok általános (de nem minden esetben érvényes) tulajdonságai az alábbi formában írhatók fel, ha s_i, s_j két tetszőleges összehasonlítandó objektum és $A(s_i, s_j)$ a hasonlósági mérőszám, akkor

1. $A(s_i, s_j) = A(s_j, s_i)$ (szimmetria),
2. A értéke általában a $0 \leq A \leq 1$ vagy a $-1 \leq A \leq 1$ intervallumba esik,
3. $A(s_i, s_i) = 1$.

A mérési módszereket a változók egyes típusaira külön-külön ismertetjük.

Nominális és ordinális változók

A mérés alapja a statisztikából ismert kontingencia tábla

A \ B	1	2	...	q	
1	f_{11}	f_{12}	...	f_{1q}	$f_{1.}$
2	f_{21}	f_{22}	...	f_{2q}	$f_{2.}$
.	.	.		.	.
.	.	.		.	.
.	.	.		.	.
r	f_{r1}	f_{r2}	...	f_{rq}	$f_{r.}$
	$f_{.1}$	$f_{.2}$	...	$f_{.q}$	n

,

ahol f_{ij} az i és j tulajdonság együttes előfordulásának – az n elemű mintából számított – gyakorisága.

A mérési módszerek jelentős része az ismert χ^2 statisztikára épül. A kontingencia táblázatból számítható és a változók közötti függetlenséget feltételező χ^2 formulából:

$$\chi^2 = n\left(\sum_{i=1}^{r}\sum_{j=1}^{q}\frac{f_{ij}^2}{f_{i.}f_{.j}} - 1\right)$$

látható, hogy χ^2 közvetlenül függ a tábla méretétől és n növekedésével minden határon túl nő. Ezért χ^2-nek különféle normált értékei jöhetnek számításba. Ilyen normalizáló faktor nyilvánvalóan az n, a kapott érték 0 és 1 közé esik. Ezt figyelembe véve javasolta *Pearson* a P, *Csuprov* a T, *Cramer* a C kontingencia együtthatót.

$$P = \left(\frac{\Phi^2}{1+\Phi^2}\right)^{1/2}, \quad \text{ahol} \quad \Phi^2 = \frac{\chi^2}{n};$$

$$T = \left(\frac{\chi^2}{n(r-1)(q-1)}\right)^{1/2};$$

$$C = \left(\frac{\chi^2}{n \cdot \min[(r-1),(q-1)]}\right)^{1/2},$$

Kendall és *Stuart* mutatott rá a χ^2 statisztikán alapuló mértékek torzító hatásainak okaira. Ezek a mértékek arra a hipotézisre épülnek, hogy a kontingencia tábla olyan kétváltozós normális eloszlást reprezentál, amelyre teljesül az alábbi összefüggés:

$$\lim_{n\to\infty} P^2 = r^2, \qquad \text{(ahol } r \text{ a korrelációs együttható).}$$

A gyakorlatban ez a feltevés általában nem jogos, így ezek a mértékek csak korlátozottan alkalmasak az asszociáció mérésére.

Másik hiányosságukra *Goodman* és *Kruskal* mutatott rá: (hiv. *Anderberg*, 740 o.) a változó párok egymás között nem összehasonlíthatók e mértékek alapján.

Ebből kiindulva javasolták a γ statisztika bevezetését; ez az asszociációs mérték az optimális osztály becslésén alapul.

Nominális változók esete

Ha a kontingencia tábla minden elemét n-el elosztjuk a kapott értékek relatív gyakoriságok, amelyek n növelésével jól közelítik a megfelelő valószínűségeket, indokolt tehát a következő jelölések bevezetése

$$p_{ij} = \frac{f_{ij}}{n}; \qquad p_{.j} = \frac{f_{.j}}{n}.$$

A következő valószínűségi modellt használjuk: válasszunk ki az n elemű sokaságból véletlenszerűen egy elemet. Becsüljük meg a lehető legkisebb hibával, hogy melyik A_i, ill. B_j ismérv-osztályba tartozik. A becslést két esetben végezzük el:

1. csak azt tudjuk a kiválasztott elemről, hogy besorolható valamelyik két osztályba;
2. ismerjük a kiválasztott elem A_i osztályát.

Nyilvánvaló, hogy az utóbbi esetben több információnk van; az elkövetett hiba legfeljebb akkora lehet, mint az első esetben.

Legyen P_1 a besorolás hibájának valószínűsége az 1. esetben
P_2 a 2. esetben
Ekkor a Γ_B asszociációs mérőszámot így definiáljuk:

$$\Gamma_B = \frac{P_1 - P_2}{P_1},$$

Γ_B a besorolási hiba valószínűségének azt a relatlv csökkenését mutatja, amely az A_i osztály ismeretéből származó információtöbbletből ered. Ha bevezetjük a $p_{.m} = \max_j p_{.j}$ és a $p_{im} = \max_j p_{ij}$ jelöléseket, akkor

$$P_1 = 1 - p_{.m}, \quad P_2 = 1 - \sum_i p_{im},$$

$$\Gamma_B = \frac{\sum_{i=1}^{r} p_{im} - p_{.m}}{1 - p_{.m}}.$$

Ha nem az A_i, hanem egy B_j osztály azonosítható a véletlenszerűen kiválasztott elemmel, akkor a Γ_A hasonlósági mérőszám a fentivel teljesen analóg módon definiálható:

$$\Gamma_A = \frac{\sum_{j=1}^{q} p_{mj} - p_{m.}}{1 - p_{m.}}.$$

A fenti gondolatmenetet megismételhetjük akkor is, ha az A és B ismérvváltozatok között a kapcsolatoknak nincs kitüntetett iránya. Tehát egy tetszőleges elem kiválasztásakor $\frac{1}{2}$ valószínűséggel az A_i vagy a B_j osztályba tudjuk sorolni. Ekkor ismeretlen prediktor osztály esetén a hiba valószínűsége

$$P_1 = 1 - \frac{1}{2}(p_{.m} + p_{m.}),$$

ismert prediktor osztály esetén pedig

$$P_2 = 1 - \frac{1}{2}\left(\sum_{i=1}^{r} p_{im} + \sum_{j=1}^{q} p_{mj}\right).$$

Az asszociációs mutató értéke:

$$\Gamma = \frac{\frac{1}{2}\left(\sum_i p_{im} + \sum_j p_{mj} - p_{m.} - p_{.m}\right)}{1 - \frac{1}{2}(p_{.m} - p_{m.})}.$$

Γ értékei a $\Gamma_A \leq \Gamma \leq \Gamma_B$ intervallumba esnek.

Az asszociációs mutató tulajdonságai

a) Γ akkor és csak akkor nem határozható meg, ha az egész sokaság egy osztályba tartozik; egyébként $0 \leq \Gamma \leq 1$,

b) $\Gamma = 1$ akkor és csak akkor, ha A_i ismerete egyértelműen definiálja a megfelelő B_j osztályt, azaz függvényszerű kapcsolat van a két változó között.

c) $\Gamma = 0$, ha a vizsgált osztályok statisztikailag függetlenek (nem megfordítható állítás).

d) Γ invariáns a kontingencia tábla sorainak (vagy oszlopainak) permutációjára.

A cluster analízisben a nominális változók közötti kapcsolatok jellemzésére jól használhatók még a

– kanonikus korreláció és az
– entrópia elméleten alapuló mértékek.

A γ statisztika ordinális változók esetén

Most az A és B ismérvváltozatok közül legalább az egyik természetes módon rendezhető. Így a kontingencia táblázat sorainak vagy oszlopainak permutációjára γ nem lehet invariáns. A valószínűségi modell: válasszunk ki a sokaságból véletlenszerűen (visszatevéssel) két elemet. Tegyük fel, hogy az első valamilyen $(A_{i1}; B_{j1})$, a második pedig valamilyen $(A_{i2}; B_{j2})$ kategóriába tartozik, ahol $1 \leq i_k \leq r$ és $1 \leq j_k \leq q$ $(k = 1, 2)$. Függetlenség esetén joggal várhatjuk, hogy az i_k indexek rendezettsége nincs összefüggésben a j_k indexek rendezettségével, míg kapcsolat esetén ez a rendezettség általában megegyezik.

Jelöljük a hasonló rendezettség valószínűségét P_h-val

$$P_h = P\{i_1 < i_2 \quad \text{és} \quad j_1 < j_2, \quad \text{vagy} \quad i_1 > i_2 \quad \text{és} \quad j_1 > j_2\},$$

az eltérő rendezettség valószínűségét P_e-vel

$$P_e = P\{i_1 < i_2 \quad \text{és} \quad j_1 > j_2, \quad \text{vagy} \quad i_1 > i_2 \quad \text{és} \quad j_1 < j_2\},$$

valamint az azonosság valószínűségét P_a-val

$$P_a = P\{i_1 = i_2, \quad \text{vagy} \quad j_1 = j_2\}.$$

Az egyértelműség kedvéért ez utóbbi esetet a statisztika definiálásakor nem engedjük meg, vagyis P_h és P_e helyett az $\{i_1 = i_2$ vagy $j_1 = j_2\}$ esemény negáltjára vonatkozó feltételes valószínűségeket tekintjük. Pl. P_h helyett a $P_h/(1 - P_a)$ valószínűséget.

Az asszociációs mutató:

$$\gamma = \frac{P_h - P_e}{1 - P_a}.$$

A γ mutató tulajdonságai

a) γ nem határozható meg, ha a kontingencia tábla nem nulla elemei egy sorban vagy egy oszlopban vannak.
b) $-1 \leq \gamma \leq 1$.
c) $\gamma = 1$, ha a nem nulla elemek a $p_{11} \rightarrow p_{rq}$ irányú átlóban vannak; ekkor $P_e = 0$;
d) $\gamma = -1$, ha a nem nulla elemek a $p_{r1} \rightarrow p_{rq}$ irányú átlóban vannak;
e) $\gamma = 0$, ha teljesül a függetlenség. Ez az állítás nem megfordítható (kivétel a 2×2-es tábla).

Arány és intervallum változók

A T mátrix elemei mérhető értékek, feladatunk két tetszőleges oszlop hasonlóságának a mérése. Jelöljük a T mátrix két oszlopát X és Y vektorokkal, $X,\ Y \in R^n$.

A hasonlóság mértékéül a vektorok hajlásszöge és a szorzatmomentum korrelációs együttható tekinthető.

$$A(X, Y) = \cos \alpha = \frac{X * Y}{|X| \, |Y|} .$$

Képezzük az $\hat{X} = X - \bar{X}$ és $\hat{Y} = Y - \bar{Y}$ nulla átlagú vektorokat. A szorzatmomentum korrelációs együtthatója

$$r = r(X, Y) = \frac{\operatorname{cov}(X, Y)}{\sqrt{\operatorname{var}(X) \cdot \operatorname{var}(Y)}} ,$$

ahol $\operatorname{cov}(XY) = \dfrac{\hat{X}\,\hat{Y}}{n}$ és $\operatorname{var}(X) = \dfrac{X * X}{n}$.

Könnyen belátható, hogy $r(X, Y) = A(\hat{X}, \hat{Y})$.

$A(X, Y)$ invariáns a nyújtásra, $r(X, Y)$ pedig a nyújtásra és az eltolásra. Ebből adódik, hogy $A(X, Y)$ az arány, az $r(X, Y)$ pedig az intervallum változók esetén alkalmazható eredményesen.

Bináris változók

A bináris változók sajátos tulajdonsága miatt célszerű a külön kiemelés mert:

– az előző formuláknak bináris esetre általában létezik egyszerűbb alakja,
– a tulajdonságok asszociációs mérőszámai bizonyos esetekben, mint említettük alkalmazhatók objektumok összehasonlítására is, ez különösen bináris változókra áll fenn.

A T mátrix most csak a 0 és az 1 számokat tartalmazza. Két tetszőleges oszlop összehasonlítása nyilvánvalóan minden esetben redukálható egy 2×2-es táblára.

A \ B	1	0	
1	a	b	$a+b$
0	c	d	$c+d$
	$a+c$	$b+d$	$a+b+c+d=n$

Példaként a már bevezetett *Csuprov*-együttható bináris alakját mutatjuk be:

$$A_{es} = \frac{ad - bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}} .$$

A bináris mérőszámok konstruálásánál a problémát egyrészt a d érték figyelembevétele jelenti, ez ugyanis a közös tulajdonságok hiányát méri, másrészt az, hogyan súlyozzuk az illeszkedéseket és nem illeszkedéseket.

A mutatókat a felvetett problémák szerint osztályozva a 3. táblázat foglalja össze.

3. táblázat

Súlyozás	0—0 Illeszkedés a nevezőben	0—0 Illeszkedés a számlálóban	
		nem szerepel	szerepel
Egyenlő súlyok	szerepel	1. *Russel* és *Rao* $\frac{a}{a+b+c+d} = \frac{a}{n}$	2. *Sokal* és *Michner* $\frac{a+d}{a+b+c+d} = \frac{a+d}{n}$
	nem szerepel	3. *Jaccard* $\frac{a}{a+b+c}$	4. —
Dupla súlyozás a kapcsolódó pároknál	szerepel	5. *nem ajánlott* $\frac{2a}{2(a+d)+b+c}$	6. $\frac{2a+d}{2(a+d)+b+c}$
	nem szerepel	7. *Dice* $\frac{2a}{2a+b+c}$	8. —
Dupla súlyozás nem kapcsolódó pároknál	szerepel	9. *nem ajánlott*	10. *Rogers—Tanimoto* $\frac{a+d}{a+d+2(b+c)}$
	nem szerepel	11. $\frac{a}{a+2(b+c)}$	12. —
A kapcsolódó párok kizárva a nevezőből	—	13. *Kulczynski* $\frac{a}{b+c}$	14. $\frac{c+d}{b+c}$

[2] L. [1]. 89. o.

3. Az objektumok közötti távolság, hasonlóság mértékei

Ebben a fejezetben a T mátrix sorainak páronkénti összehasonlításával foglalkozunk. Számos alkalmazási területen kizárólag így vetődik fel a kérdés. A tárgyalást a mérhető változókkal kezdjük. Az eddigiekhez képest alapvetően új eljárásokkal ezek esetében találkozunk, mert az objektumok között értelmezhető a taxonomikus távolság fogalma.

A taxonomikus távolság metrikus mértékei

A cluster analízis kvantitatív módszereinek gyakorlati alkalmazásánál az egyik központi probléma a pontok, ill. ponthalmazok közötti távolság definiálása. A távolság megfelelő megválasztása legalább olyan körültekintést igényel, mint az adekvát osztályozási algoritmus kiválasztása. Tételezzük fel, hogy a T mátrix elemei mérhető változók. Minden objektum egy pontnak tekinthető a p dimenziós absztrakt térben [4]. E pontok között értelmezhetők metrikus tulajdonsággal rendelkező távolságmérő függvények.

Jelöljük az (x, y) pontpár távolságát $d(x, y)$-al, amely minden $x, y, z \in M$ esetén, az alábbi tulajdonságokkal rendelkezik:

1. $d(x, y) = d(y, x)$,
2. $d(x, x) = 0$,
3. $d(x, y) > 0$, ha $x \neq y$,
4. $d(x, y) \leq d(x, z) + d(y, z)$.

Ezek a metrikus tér általános tulajdonságai és az ezeket kielégítő $d(x, y)$ függvényt metrikus függvénynek vagy röviden *metrikának* nevezzük. Ha a 3. feltétel nem teljesül, akkor d-t *pszeudo metrikának* nevezzük.

Érdemes megvizsgálni, hogy milyen zavarhoz vezet ha a 4. tulajdonság az ún. háromszögegyenlőtlenség nem teljesül. Ezt az esetet *szemimetrikának* nevezzük. Azon metrikákat, amelyek kielégítik a $d(x, y) \leq \max[d(x, z) + d(y, z)]$ egyenlőtlenséget *ultrametrikának* nevezzük.

Tegyük fel, hogy 5 pontunk van, az i-edik és a j-edik távolságát jelöljük d_{ij}-vel. Távolságaink legyenek a következők:

$$\begin{array}{lll} d_{12} = 2 & d_{23} = 10 & d_{15} = 1 \\ d_{13} = 10 & d_{24} = 10 & d_{25} = 100 \\ d_{14} = 10 & d_{34} = 2 & d_{35} = 1{,}5 \\ & & d_{45} = 100 \end{array}$$

Az első két oszlop alapján két jól elkülöníthető osztályt kapunk, az $S_1 = \{x_1, x_2\}$ és az $S_2 = \{x_3, x_4\}$ osztályokat, de hova soroljuk az x_5 pontot? Ha S_1-hez vesszük hozzá, akkor x_2-től távolabb lesz, mint x_3-tól, pedig az előbbivel egy osztályba tartozik. De ugyanilyen észszerűtlen S_2-be sorolni is, mert akkor x_1-hez lesz aránytalanul közelebb, mint x_4-hez. Marad még egy lehetőség, x_5 külön osztályt alkot. De ez sem kielégítő megoldás, mert x_1-től is és x_3-tól is kisebb távolságra van, mint a velük egy osztályba tartozó x_2, ill. x_4 pontok.

Ha ismerjük adatrendszerünk valószínűség eloszlását, jól alkalmazható a következő mérték.

Tegyük fel, hogy az $x_1 \ldots x_n$ pontokat egy p-dimenziós valószínűségi vektorváltozó értékeinek tekintjük és rögzített k mellett az $x_1^k \ldots x_n^k$ számok a megfelelő egydimenziós valószínűségi változó értékei.

Legyen D a p-dimenziós változó szórásmátrixa (kovariancia mátrixa),

$$D = \begin{pmatrix} D_{11} & D_{12} \ldots D_{1p} \\ D_{p1} \ldots & D_{pp} \end{pmatrix},$$

ahol $D_{ij} = M[(X_i - M(X_i))(X_j - M(X_j))]$;

D^{-1} legyen D inverze. Ekkor a következőképpen definiálhatunk pontjaink között egy metrikát:

$$d_{ij} = \sqrt{(x_i - x_j)^* S' D^{-1} S(x_i - x_j)},$$

ahol S diagonális elemeket tartalmazó súlymátrix.

Ha változóinkat nem akarjuk súlyozni, akkor S elhagyható a formulából. Ezt a metrikát akkor célszerű alkalmazni, ha ismerjük az eloszlást, ill. ha a minta elengedően nagy ahhoz, hogy D értékét kielégítő pontossággal becsüljük. Ebben az esetben a d_{ij} távolság nemcsak az x_i és x_j pontok koordinátáitól függ, hanem az összes többi ponttól is, ellentétben a következőkben ismertetendő metrikákkal. Az sem elhanyagolható, hogy figyelembe vettük a különböző változók kapcsolatát is. Ha a változók páronként korrelálatlanok, akkor $D_{ij} = 0$, ha $i \neq j$, és ennek megfelelően a D diagonális mátrix, D^{-1} pedig olyan diagonális mátrix, amelynek főátlójában az egyes változók szórásának reciproka áll, d_{ij} ekkor a következő egyszerűbb alakba írható

$$d_{ij} = [w_1(x_i^1 - x_j^1)^2 + \ldots + w_p(x_i^p - x_j^p)^2]^{1/2},$$

ahol a w_i-k tetszőleges nem negatív súlyok.

Az egyes változók súlyozása valamennyi metrikánál elvégzendő, de annak megítélése, hogy milyen súlyrendszert alkalmazzunk, elsősorban a kutató feladata. Meg kell jegyeznünk, hogy ha minden változót azonos súllyal akarunk figyelembe venni, akkor is el kell végezni a súlyozást; egyenlő súlyozáshoz akkor jutunk, ha minden változót a szórásával normálunk. A továbbiakban feltesszük, hogy változóink már normáltak.

A Minkowski-metrika és speciális esetei

Az egyik legáltalánosabb metrikaosztály, amely tetszőleges $1 \leq r < \infty$ érték mellett egy-egy metrikát ad, a következőképpen definiálható:

$$d_r(x, y) = \left(\sum_{i=1}^{p} | x_i - y_i |^r \right)^{1/r}.$$

A Minkowski metrika $r = 2$ esetben az ismert euklidesi metrikával azonos.

$$d_2(x, y) = \sqrt{\sum_{i=1}^{p} (x_i - y_i)^2}.$$

$r = 1$ esetén a távolságmérő függvény a koordinátánkénti eltérések összegével egyenlő

$$d_1(x, y) = \sum_{i=1}^{p} | x_i - y_i | .$$

A cluster analízis során ezt a két esetet szokták alkalmazni. Sok esetben a metrika helyett a pontok közötti hasonlóságot definiáljuk. A hasonlóság is egy nemnegatív szám, de a metrikával ellentétben célszerű úgy megválasztani, hogy értékei nulla és egy közé essenek. h-val jelölve a hasonlóságot, megköveteljük, hogy tetszőleges x pontra $h(x, x) = 1$ legyen, azaz minden pont (objektum) saját magához hasonlítson a legjobban. Az ismertetendő módszerek szempontjából közömbös, hogy az adott objektumok között távolságot vagy hasonlóságot értelmezünk, csak arra kell ügyelni, hogy a minimális távolság maximális hasonlóságnak felel meg és fordítva.

Ha egy, a pontpárok távolságán értelmezett, monoton csökkenő függvényt adunk meg, amelynek értékei 0 és 1 közé esnek, akkor a metrikához egy hasonlóságot rendelünk hozzá.

A legegyszerűbb hasonlósági mérték az

$$R_{ij} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{[\Sigma(x_i - \bar{x})^2 \Sigma(y_i - \bar{y})^2]^{1/2}}$$

korrelációs együttható.

Mivel $-1 \leq R_{ij} \leq 1$, a megfelelő hasonlósági mértéket pl. az $R'_{ij} = (1 + R_{ij})/2$ egyenlőséggel definiálhatjuk.

Egy további lehetőség, ha a pontok távolságát a hozzájuk tartozó vektorok hajlásszögével mérjük. A hajlásszög koszinusza,

$$\cos(x, y) = \frac{\Sigma x_i y_i}{(\Sigma x_i^2 \cdot \Sigma y_i^2)^{1/2}} \qquad x, y \neq 0$$

tekinthető a két pont hasonlóságának, távolságuk pedig a

$$d_{ij} = \sqrt{1 - \cos^2(x, y)} \qquad x, y \neq 0$$

képlettel definiálható, d_{ij} ebben az esetben pszeudo-metrika lesz, vagyis a 3. feltétel itt nem teljesül; két pont távolsága akkor lesz nulla, ha a vektorok egy egyenesbe esnek, így $x \neq y$ pontok távolsága is lehet nulla.

Bináris változók esetében – az eddigieken kívül – más távolságot is használhatunk.

A kétdimenziós tábláknál szokásos jelölésekkel

$$d_{ij} = \frac{a + d}{a + b + c + d}$$

egy lehetséges metrika.

De mérhető a pontok távolsága a

$$d_{ij} = \frac{2a}{2a + b + c}$$

értékkel is.

Ez utóbbi metrikánál alaposan mérlegelni kell a clusterezés alkalmazási szempontjait, mert a dichotom változók nulla és egy értékeinek megválasztása gyakran esetleges (pl. a nemeknél férfi = 0, nő = 1 vagy fordítva), ez pedig azt jelenti, hogy a jelöléstől függően más lesz a pontok távolsága. Ha pl.

$$x = (1,0,0,0,0,0), \qquad y = (1,1,1,0,0,0),$$

akkor $a = 1$; $b = 0$, $c = 2$, $d = 3$ alapján: $d_{ij} = \frac{1}{2}$; ha most a változók értékeiben felcseréljük a nulla és egy jelölést, akkor:

$$x = (0,1,1,1,1,1), \qquad y = (0,0,0,1,1,1),$$

vagyis $a = 3$, $b = 2$, $c = 0$, $d = 1$ és ennek megfelelően $d_{ij} = \frac{3}{4}$.

A korábban tárgyalt metrikáknál ilyen probléma nincs, a d_{ij} távolság ott független a változók értékeinek számozásától.

A Minkowski metrika alkalmazásakor figyelemmel kell lenni arra, hogy a metrika a változók különböző tartalmát, dimenzióját nem változtatja meg, ezt a kapott távolsági érték interpretációjánál kell számba venni. A metrika a változók közötti függetlenséget tételezi fel, így előfordulhat, hogy a változók közötti kapcsolatok esetén egyfajta hatást többször veszünk figyelembe.

Nem metrikus mértékek

A nem metrikus mértékek egyik típusa az objektumok között relációkat definiál és ezek relációelméleti alapon történő feldolgozásával csoportosít. A másik típus tulajdonképpen feltételez valamilyen – az előzőektől eltérő – metrikát, amelyre támaszkodva az objektumok egy rendezését végzi, de a rendezésnek már nincsenek meg a metrikától megkövetelt tulajdonságai.

Intervallum változók. Calhoun-távolság

Ez a távolságfogalom a kérdéses két pont és a vonatkoztatási koordinátatengelyek iránya által meghatározott hiperfelületek közé eső többi pontra épül. Pl. az x_1 és x_2 pont közötti távolságot a bevonalazott területbe eső pontok segítségével határozhatjuk meg (1. ábra).

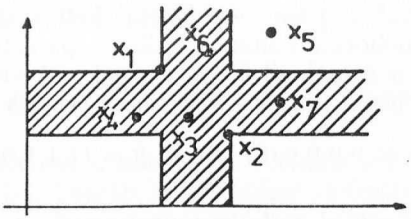
Két pont *Calhoun*-távolsága definíció szerint:

$$D_c = 6N_i + 3N_b + 2N_z,$$

ahol N_i – azon pontok száma, amelyek a két pont által meghatározott hipersíkba vagy meghosszabbításába esnek, legalább egy változójuk szerint.

N_b – azon pontok száma, amelyek egyetlen dimenzióban sem esnek a két pont közé, de egy vagy több változó szerint határra esnek.

[3] L. [1]. 111. o.



1. ábra

N_z – azon pontok száma, amelyek egy vagy több változó szerint mindkét ponttal azonos értékűek, de nem esnek a hipersík belsejébe vagy a határra.

Ha N az alappontok száma, akkor D_c maximális értéke: $6(N-2)$.

A Calhoun-távolság, mint mérték, nem felel meg a metrika követelményeinek, mert két pont távolsága akkor is lehet nulla, ha a két pont nem esik egybe. Ez a mérték hasznos lehet olyan esetekben, ha a clusterek egy vagy több változó szerint átfedik egymást.

Lance és Williams-féle mérték

Két objektum közötti távolság definíciója:

$$d_{LW} = \frac{\Sigma \mid x_i - y_i \mid}{\Sigma \, (x_i - y_i)} .$$

A számláló Minkowski metrika, $r = 1$ esetre, a nevező pedig a maximális kiterjedés mértéke. Bináris változó esetén a következő alakú:

$$d_{LW} = \frac{b + c}{(a + b)\,(a + c)} = 1 - \frac{2a}{2a + b + c} .$$

Ezek a mértékek ritkán használatosak, speciális eseteket elégítenek ki.

Nominális változók (l. [1] 123. o.)

Az ismérvek közötti hasonlóság mérésére alkalmazott mutatók az objektumok közötti hasonlóság mérésére is alkalmasak.

Ekkor az összehasonlítás azon alapul, hogy összeszámoljuk a közös és eltérő ismérveket. Ha az ismérvek jelenléte vagy hiánya egyértelműen megállapítható, akkor a bináris változóknál bevezetett 2×2-es táblához jutunk. Előfordulhat, hogy egyes ismérvek nem jellemzőek a kérdéses objektumra, ezt figyelembe véve a két lehetséges alternatíva (0 és 1) helyett hármat bevezetve, a bináris változók kiterjesztéséről beszélhetünk.

Legyen:

n_{a+d} azon ismérvek száma, amelyekben a két objektum megegyezik;

n_d azon ismérvek száma, amelyekkel a két objektum nem jellemezhető;
n_{b+c} azon ismérvek száma, amelyekben a két objektum eltér egymástól.

A 3. sz. táblázat képleteit felhasználva megkapjuk a 4. táblázatban összefoglalt kapcsolódási együtthatókat.

4. tábla (L. [1]. Table S. S.)

Kapcsolódási együtthatók nominális változók esetén

1.) $\dfrac{n_{a+d} - n_d}{n_{a+a} + n_{b+c}}$ 2.) $\dfrac{n_{a+d}}{n_{a+d} + n_{b+c}}$

3.) $\dfrac{n_{a+d} - n_d}{n_{a+d} - n_d + n_{b+c}}$

6.) $\dfrac{2n_{a+d}}{2n_{a+d} + n_{b+d}}$

7.) $\dfrac{2(n_{a+d} - n_d)}{2(n_{a+d} - n_d) + n_{b+c}}$

10.) $\dfrac{n_{a+d}}{n_{a+d} + 2n_{b+c}}$

11.) $\dfrac{n_{a+d} - n_d}{n_{a+d} - n_d + 2n_{b+c}}$

13.) $\dfrac{n_{a+d} - n_d}{n_{b+c}}$ 14.) $\dfrac{n_{a+d}}{n_{b+c}}$

(A formulák sorszáma a 3. sz. táblázat megfelelő számozására utal.)

Az ordinális változóknak kitüntetett szerepe van a cluster elemzésében, mert a legáltalánosabb osztályozási algoritmusok monoton invariánsak. Ebből következik, hogy ezek alkalmazása esetén elegendő a távolsági értékek helyett azok egymáshoz viszonyított rangsorát tekinteni.

A változók súlyozása

A T mátrix két oszlopának (két ismérv) összehasonlításakor két egyenként homogén koordinátájú vektorunk van, míg két sor (két objektum) egybevetésekor az egyes koordináták gyakran különböző tartalmú és típusú változókat reprezentálnak. Ebből adódik, hogy a mértékeket befolyásolják a nagyságrendek, és felmerül a különböző mértékek additivitásának problémája.

E problémák megoldására szolgál a súlyozás. Például, ha a vizsgálat körébe bevont változók nem egyformán fontosak az adott kérdés szempontjából: szubjektív súlyozást alkalmazunk.

Más jellegű a probléma, amikor a változók különböző dimenziójából adódó esetlegességet kívánjuk kiszűrni. A statisztikában leggyakrabban használt standardizálást itt is alkalmazhatjuk, ekkor minden változó azonos súlyú. Euklideszi távolság esetén ez a normalizálás $w = 1/s^2$ súlyozást jelent. Ez azzal a veszéllyel jár, hogy éppen azon ismérvek szerepét csökkentjük, amelyek a

legalkalmasabbak a csoportok megkülönböztetésére. Lényegesen elfogadhatóbb eredményre jutnánk, ha a teljes szórás helyett a csoporton belüli szórással normálnánk – a vizsgálat előtt ez persze nem ismert.

Az apriori ismeretek hiánya miatt bírálhatók azok a módszerek is, amelyek a súlyozással a korreláció hatását kívánják kikapcsolni, ilyen a *Mahalanobis* távolság-fogalom is.

Megfelelően szűri ki a korrelációnak a kapcsolatok mérését torzító hatását a faktoranalízis főfaktor módszere, ami azzal a további előnnyel is jár, hogy az eredeti adatmátrix mérete jelentősen csökken.

Meg kell jegyezni, hogy a súlyozás a különböző típusú változók együttes megjelenéséből adódó problémákat nem oldja meg, ehhez a megfelelő skálatranszformációt kell elvégezni.

Clusterek távolsága

A ponthalmazok között is többféle távolságot értelmezhetünk, itt azonban a metrika 1–4. követelményei közül általában csak a szimmetria teljesül. Legkézenfekvőbb megoldás, ha a ponthalmazok távolságán a legközelebbi, ill. a legtávolabbi pontjaik távolságát értjük.

1. $D(S_i, S_j) = \min_{m,k} d(x_{im}, x_{jk})$,

2. $D(S_i, S_j) = \max_{m,k} d(x_{im}, x_{jk})$.

Itt d a pontok között értelmezett tetszőleges távolság lehet.

A továbbiakban az i-edik és j-edik osztály távolságát D_{ij}-vel jelöljük, a most definiált távolságokat pedig $D_{\min}$, ill. $D_{\max}$ távolságoknak nevezzük. Szokásos az osztályok távolságának a centroidok távolságát tekinteni. Az n_i pontból álló S_i osztály centroidja a

$$C_i = \frac{1}{n_i} \sum_{k=1}^{n_i} x_{ik}$$

vektor, ahol x_{ik} az S_i osztály k-adik pontja. Ekkor az S_i és S_j osztályok távolsága

$$D_{ij}^c = d(C_i, C_j).$$

Az osztályok távolsága az egyes osztályokból vett pontpárok átlagos távolságaként is értelmezhető.

$$D_{ij}^r = \frac{1}{n_i n_j} \left[\sum_{m,k} \left(d(x_{im}, x_{jk}) \right)^r \right]^{\frac{1}{r}}.$$

r tetszőleges értékére az osztályok közötti távolság egy-egy lehetséges definícióját kapjuk. $r = 1$ esetben D^1 az előbbi távolsággal azonos. Ha $r \to \infty$, akkor $D^\infty \to D_{\max}$, míg ha $r \to -\infty$, akkor $D^{-\infty} \to D_{\min}$.

Az osztályozás során nemcsak két osztály távolságát kell meghatározni, hanem – különösképpen az összevonásos módszereknél – ismerni kell két egymással összevont osztálynak egy harmadiktól vett távolságát is. Az egyes összevonások után célszerű a távolságokat ezek felhasználásával számítani. A D^r távolságnál pl. ha az S_i és S_k osztályokat vonjuk össze S_{jk}-ba, akkor ennek egy S_i-től vett távolsága:

$$D^r(S_i, S_{jk}) = \left[\frac{n_j(D_{ij})^r + n_k(D_{ik})^r}{n_j + n_k}\right]^{\frac{1}{r}}.$$

Ha $r = \pm \infty$, a $D_{\min}$ és $D_{\max}$ távolságoknál ez a formula nem használható, helyette az alábbi összefüggést adjuk meg:

$$D(S_i, S_{jk}) = a\ D_{ij} + b\ D_{ik} + c\ D_{jk} + d|\ D_{ij} - D_{ik}|\ ,$$

ahol $a = b = \frac{1}{2}$, $c = 0$, $d = -\frac{1}{2}$, ha $D = D_{\min}$;

$a = b = \frac{1}{2}$, $c = 0$, $d = \frac{1}{2}$, ha $D = D_{\max}$;

$a = \frac{n_j}{n_j + n_k}$, $b = \frac{n_k}{n_j + n_k}$, $c = d = 0$, ha $D = D^1$.

4. A cluster analízis módszerei

A tudományos klasszifikációtól megköveteljük, hogy az identifikációt objektív kritériumok alapján lehessen elvégezni; ez a követelmény általában csak igen nagy megszorításokkal teljesül. Ennek egyik oka, hogy egy osztályozási sémáról önmagában nem lehet eldönteni, hogy jó-e vagy rossz, ez azon múlik, hogy a felosztás a vizsgálat szempontjából mennyire megfelelő. Ugyanakkor meghatározhatunk általános követelményeket.

1. Objektivitás
2. Stabilitás
3. Prediktivitás

Az első feltétel alatt azt értjük, hogy az adott kutatási terület szakemberei a vizsgált objektumokat jellegében azonos módon csoportosítsák. A második feltétel azt jelenti, hogy a klasszifikációtól megköveteljük, hogy új adatok kevéssé befolyásolják; ez akkor kerül előtérbe, amikor új adatok és eredmények a régi fogalmi struktúrát megkérdőjelezik és ezáltal a klasszifikációs rendszer instabillá válik. A harmadik feltétel az osztályozás igen magas kvalitását jelzi, s ennek megfelelően csak ritkán teljesül (pl. *Mendelejev* periódusos rendszere).

A klasszikus logikára épülő osztályozási modellek két alaplépése különböztethető meg:

– típus és koncepcióalkotás, kategóriák definiálása,
– az események kijelölése a már definiált kategóriákhoz.

A klasszikus logika azonban csak olyan kategóriák definiálását adja meg, amelyeknek minden egyede minden szempontból ekvivalens. Az ilyen elven alapuló osztályozást monotetikus osztályozásnak nevezzük. Ez a módszer sokváltozós, nagy méretű adatrendszer esetén még számológéppel sem kezelhető, de ha ez mégis sikerülne, az eredmények áttekinthetetlen felaprózottsága miatt a gyakorlatban használhatatlan lenne. Ezért nagy jelentőségű a cluster analízis, amely utat nyit a sokváltozós nagy minták áttekinthető numerikus értékeléséhez.

A cluster analízis modellje három szempontból is eltér a klasszikus osztályozási modellektől:

a) Nem definiál típusokat mielőtt kijelölné a mintaegyedeket; az eljárások során a típusokat definiáló fogalmak hozzárendelődnek az osztályozással kialakított csoportokhoz. Ez a meggondolás az alábbi feltételezéseken alapul:
   - léteznek típusok
   - a típusfogalom ismerete nélkül is létezik olyan kritérium, amelynek felhasználásával a clusterek felismerhetők;
   - a felismert clusterhez az egyedek ismérvei alapján megadhatók a típus-jellemzők.

Mindez szemléletesen azt jelenti, hogy az n-dimenziós térben az egyes típusokat elkülönítő hipersíkok akkor válnak láthatókká, ha az azonos clusterbe tartozó elemeket meghatároztuk, és ez azt jelenti, hogy a csak empirikusan előforduló típusok is felismerhetők.

b) A cluster analízis megengedi a politetikus osztályokat. Politetikusnak tekintünk egy osztályt, ha elemei több, de nem minden jellemző szerint ekvivalensek vagy hasonlóak. Az osztályhatárokat nem előre határozzuk meg. Ez a cluster analízisnek a gyakorlat szempontjából további előnyös tulajdonsága, hogy ugyanis minden jellemzőt figyelembe véve is jelentősen csökkenthető az osztályok száma.

c) A klasszikus modellek csak diszkrét változókkal dolgoznak, a cluster analízis megenged folytonos, sőt vegyes változó típusokat is.

Az osztályozás kritériumai

Az alábbiakban megadjuk az osztályozás gyakorlati követelményeit. A módszerek nem mindegyike teljesíti az összes feltételt egyszerre, de a konkrét értékelés alapján megítélhető a használhatóságuk.

1. *Egyértelműség*

Adott adathalmazból mindig ugyanazt az eredményt kapjuk egy adott M rendszer esetén.

2. *Monoton invariáns*

Ha az osztályozás végeredménye csak a DC-k (DC = az osztályozás input mátrixa) sorrendjétől (rangjaitól) függ, akkor a módszer monoton invariáns:

$$[Mf(d)][f(h)] = (Md)(h), \quad \forall\, h \geq 0\text{-ra},$$

ahol f a rangot előállító leképezés.

3. *Skála függetlenség*

Ha $k > 0$ skalár konstans, akkor ez a feltétel az

$$M(kd) = k\,M(d)$$

egyenlőség teljesülését jelenti.

4. *Stabilitás*

Az adatok kis változtatása az eredményben is kis változást jelentsen, vagyis az

$$M: C(S) \to U(S)$$

leképezés folytonos.

5. *A csoportok megőrzése (monotonitás)*

Legyen $h > 0$ rögzített és $S_h \subset (Md)(h)$ egy kialakult osztály a h szinten, akkor a monotonitás követelménye:

$$S_h \subset S_l \quad \forall\, l > h \text{ esetén.}$$

Ez más alakban is felírható. Jelölje:

$$d' \leq d, \text{ hogy } d'(A, B) \leq d(A, B) \qquad \forall\, A; B \in S.$$

Ekkor a monotonitás így írható:

$$M(d) \leq d.$$

6. *Optimalitás*

Az osztályozandó objektumok kapcsolatairól a legtöbb információt az input DC mátrix hordozza, ezt egy többlépéses transzformációnak vetjük alá, amíg kialakulnak az osztályok. Az eljárás folyamán információt vesztünk, az input és az output DC mátrix között legyen a lehető legkisebb az eltérés; azaz ha $d \in U(S)$ és $M(d) \leq d' \leq d$, akkor teljesüljön a $d' = M(d)$ egyenlőség.

Nehezen osztályozható elemek; reprezentatív elemek

A kvantitatív osztályozási elveknél általában számszerűen mérhető, hogy egy elem milyen mértékben (statisztikai elvek szerinti osztályozásnál milyen valószínűséggel) sorolható egyik vagy másik osztályba. Nehezen osztályozható egy elem, ha több osztályba is közel azonos mértékben sorolható. Ilyenkor két lehetőségünk van:

1. eljárásunkat instabilnak minősítjük és evvel összhangban megváltoztatjuk a klasszifikáció elvét, vagy
2. a nehezen besorolható elemet „zaj" elemnek tekintjük (mérési vagy kódolási pontatlanság).

Reprezentatív elemnek az olyan elemeket nevezzük, amelyeket viszonylag egyértelműen, illetve minimális kockázattal tudunk osztályozni. Ez a meghatározás a reprezentáció tetszőleges értelmezését magába foglalja, ha meg-

felelően definiáljuk az osztályozás kritériumait. Például a centroid módszereknél, ahol az osztályokat a hozzájuk tartozó objektumok súlypontjával jellemezzük, az egyes súlypontokhoz – centroidokhoz – közel eső objektumok lesznek a reprezentatív elemek, másszóval az adott osztály közelítőleg átlagos tulajdonságú objektumai. Természetesen az osztályozás szempontjain változtatva a reprezentativitás fogalma is más lesz.

A döntésfüggvények típusai

Döntésfüggvény, ill. döntési kritérium alatt azt az elvet értjük, amely szerint a vizsgálandó objektumokat rendezve az n-dimenziós térben az osztályok kialakulnak.

A döntésfüggvény mérheti:

– a clusteren belüli elemek hasonlóságát,
– a clusterek közötti különbséget.

Az eddigiekből megállapítható, hogy a hasonlóság és különbözőség fogalma többféleképpen is értelmezhető; a tárgyalt különféle mértékek két objektum közötti hasonlóság, ill. távolság mérésére alkalmasak. A döntésfüggvény feladata ennél összetettebb, egyszerre több egyed több jellemző szerinti hasonlóságát, ill. különbözőségét kell mérnie vagy becsülnie.

Az alkalmazott módszerek szerint megkülönböztethetünk:

– sűrűségfüggvény-becslésen alapuló eljárásokat,
– valószínűségeloszlások keverékének szétválasztásán alapuló eljárásokat
– „kevert modell",
– csoporton belüli variancia becslésén alapuló módszereket,
– csoportok közötti diszkriminancia becslésén alapuló és
– gráfelméleten alapuló eljárásokat.

A különböző döntési kritériumokhoz különböző cluster fogalmak kapcsolódnak.

5. Sűrűségfüggvény-becslésen alapuló eljárások

Tekintsük megfigyeléseinket egy r-dimenziós valószínűségi vektorváltozó realizációinak. A felvetődő kérdésekre a sűrűségfüggvény ismeretében válaszolhatunk. Az elméleti sűrűségfüggvény ismeretlen, de a minta alapján becsülhető (ha az elengendően nagy). Jelentős engedmény, hogy a cluster analízis által felvetett problémák megoldásához nincs szükség a sűrűségfüggvény alakjának teljes ismeretére – a típusalkotás célja olyan clusterek körülhatárolása, ahol a pontok koncentrációja viszonylag sűrű – mert a clusterek a sűrűségfüggvény csúcsait magukba foglaló tartományok.

Tehát elegendő a sűrűségfüggvény csúcsait, lokális maximum-helyeit megkeresni. A csúcsok adják a clusterek magját, a körülhatároláshoz a csúcsmagasság és csúcsponti fő görbület értéke használható. A sűrűségfüggvény ezen jellemzőit a sztochasztikus approximáció alapján a gradiens módszer sztochasztikus változata segítségével becsüljük.

Legyen $\xi_i \in F^r$ a megfigyelt valószínűségi vektorváltozó,
$f(x)$ az együttes sűrűségfüggvény.

A becsléshez, mivel a gradiens vektor is ismeretlen, annak egy becslését használjuk.

$$z_{n+1} = z_{n+1}(Y_n, \xi_1 \ldots \xi_n),$$

ahol $Y_n \in F^r$ valószínűségi változó.

Az algoritmus:

$$Y_{n+1} = Y_n + \gamma_n z_{n+1}.$$

Az Y_n kezdeti érték tetszőleges lehet, de a lépésnagyságnak ki kell elégítenie az alábbi feltételeket.

$$\sum_{n=1}^{\infty} \gamma_n = +\infty; \qquad \sum_{n=1}^{\infty} \gamma_n^2 < \infty.$$

A sztochasztikus approximáción alapuló eljárások jó becslést adnak, a konvergencia sebessége általában lassúbb, mint gradiens algoritmusoknál. Ennek oka, hogy a gradiens vektor ismeretlen és egy becslését használjuk. (Az állítás bizonyítását lásd [15]-ben.)

A sűrűségfüggvény becslésén alapul két fontos döntési kritérium:

a) a súlypontok módszere,
b) a sűrűségfüggvény csúcsainak lokalizálása.

a) *Súlypontok módszere*

Tegyük fel, hogy ismert a kívánt clusterek száma: s; a cluster-rendszert pedig jelölje: $C_1, C_2, \ldots C_s$. A sűrűségfüggvény olyan speciális függvény, hogy a C_i cluster egy M_i pontjával – a súlypontjával – azonosítható.

Ez teljesül, ha $f(x)$ a C_i összefüggő halmazokon konstans, rajtuk kívül eltűnik, és minden cluster átmérője kisebb a súlypontjához legközelebb eső más cluster súlypontjától vett távolságnál. Azaz, ha feltételezzük, hogy a clusterek diszjunkt rendszerén belül a sokaság egyenletes eloszlású. A clustereket elkülönítő kritérium az, hogy az egyenletes eloszlás jellemző paraméterei clusterenként eltérőek. E feltételeket kielégítő döntésfüggvény a következő ún. veszteségfüggvény:

$$J(M) = \sum_{i=1}^{s} \int_{C_i} || x - M^{(i)} ||^2 f(x)\, dx = \int_{E^k} \min_{1 \le i \le k} || x - M^{(i)} ||^2 f(x)\, dx.$$

Optimális a cluster-rendszer, ha $J(M)$ minimális. A veszteségfüggvény az egyes clustereken belül a súlypont és a többi pont közötti eltérés várható értékét becsüli.

A J függvény differenciálható, deriváltja a gradiens vektor, jelölje $U^{(i)}(M)$; $i = 1,2, \ldots s$.

$$U^{(i)}(M) = 2 \int_{E^r} \varepsilon^{(i)}(x, M)\,(M^{(i)} - x)\, f(x)\, dx,$$

ahol $\varepsilon^{(i)}(x, M) = \begin{cases} 1, \text{ ha } || x - M^{(i)} || < || x - M^{(j)} ||, \text{ ha } i \neq j, \\ 0, \text{ egyébként.} \end{cases}$

Ha adott az $\xi_1, \ldots, \xi_n \ldots$ minta, akkor

$$Z_{n+1}^{(i)} = 2\varepsilon^{(i)}(\xi_{n+1}, M)\,(M^{(i)} - \xi_{n+1})$$

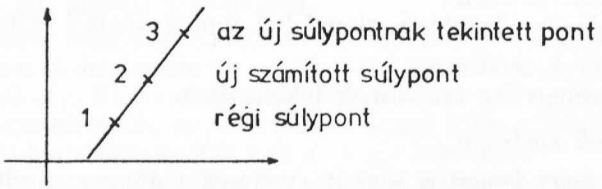
torzítatlan becslése $U^{(i)}(M)$-nek.

Tehát J minimumát $M_0 \in E_k$ kezdőérték mellett az

$$M_{n+1}^{(i)} = M_n^{(i)} - 2\gamma_n \varepsilon^{(i)}(\xi_{n+1}, M_n)(M_n^{(i)} - \xi_{n+1})$$

algoritmus adja, ahol $M_n = (M_n^{(1)}, M_n^{(2)} \ldots M_n^{(s)})$ az n-edik lépésben kialakított súlypontrendszer és $\gamma_n > 0$ eleget tesz a korábbi követelményeknek. Ezen a klasszikus súlypont módszeren alapul *Forgy* [6] konvergens nem hierarchikus módszere. Minden egyedet besorol a legközelebbi súlypontú clusterbe, majd az egész adatrendszer osztályozása után kiszámítja az új súlypontot és újra indítja az osztályozó algoritmust. Ha két egymásutáni ciklus cluster struktúrájában nincs változás, akkor megkaptuk a döntésfüggvény minimumát, az osztályozás optimális.

A konvergencia sebességének gyorsítása céljából a módszer több módosítása ismert. *Jancey* az előző ciklus súlypontjának az új súlypontra vonatkozó tükörképét tekinti az új lépés súlypontjának (2. ábra).



2. ábra

A módosítás alapgondolata, hogy az 1. pontból a 2.-ba húzott egyenes a veszteségfüggvényben az adott súlypont melletti gradiens becslése. Ha ebbe az irányba mozdul el a súlypont, akkor csökkenthető a legjobban a veszteségfüggvény értéke.

Egy másik módszert *MacQueen* [24] javasolt a konvergencia gyorsítására.

$$M_{n+1}^{(i)} = M_n^{(i)} + \gamma_n^{(i)} \varepsilon^{(i)}(\xi_{n+1}, M_n)(M_n^{(i)} - \xi_{n+1}),$$

ahol $\gamma_n^{(i)} = (1 + \omega_n^{(i)})^{-1}$; $M_0 \in E_k$

$$\omega_0^{(i)} > 0; \quad \omega_{n+1}^{(i)} = \omega_n^{(i)} + \varepsilon^{(i)}(\xi_{n+1}, M_n).$$

Teljes indukcióval belátható, hogy

$$M_{N+1}^{(i)} = \frac{\omega_0^{(i)} M_0^{(i)} + \sum_{n=0}^{N} \varepsilon^{(i)}(\xi_{n+1}, M_n)\, \xi_{n+1}}{\omega_0^{(i)} + \sum_{n=0}^{N} \varepsilon^{(i)}(\xi_{n+1}, M_n)}.$$

Ha $\omega_0^{(i)} = 1$, akkor $M_{N+1}^{(i)}$ éppen az $M_0^{(i)}$ kezdőpont és a korábban a C_i osztályba sorolt tanulópontok számtani átlaga súlypontja. Az eljárás minden elem clusteresítése után módosítja a súlypontot. Az eljárás *MacQueen*-féle közép módszerként ismert.

A súlypont-módszerek alkalmazásánál problémát jelent, hogy az induló súlypont-rendszert (magpontokat) előre kell megadni. Ennek feloldására születtek meg a súlypont-módszer olyan változatai, ahol a clusterszám előzetes megadása helyett elegendő bizonyos paraméterek megadása (pl. a clusterek maximális átmérője vagy elemszáma); a clusterek számát maga az algoritmus határozza meg. A súlypont módszer ilyen változata a MacQueen által kidolgozott paraméter becsléssel működő algoritmus, létezik ennek *Wishart* [42] által kidolgozott optimalizáló módosítása, valamint ismertek *Ball* és *Hall* ISODATA néven ismert módszerei [4].

b) *Sűrűségfüggvény csúcsainak lokalizálása*

A súlypont módszernél általánosabb feltételek mellett keresi az optimális rendszert. Míg a súlypont módszer alaphipotézise szerint az $f(x)$ sűrűségfüggvény a cluster egy pontjával jellemezhető és így egy egyszerű szerkezetű veszteségfüggvény definiálható, addig ez az algoritmus a clustereket a sűrűségfüggvény csúcsainak, lokális maximum-helyeinek környezeteként értelmezi.

Legyen $\xi_1, \xi_2 \ldots \xi_n \ldots$ független, azonos eloszlású r-dimenziós vektorváltozók sorozata. Az $f(x)$ sűrűségfüggvény differenciálható és gradiens vektora $U(x)$ egyenletes Lipschitz feltételnek tesz eleget, azaz létezik olyan L konstans, hogy

$$|| U(x) - U(y) || \leq L || x - y || \text{ és}$$

$$|| U(x) || \leq M(1 - || x ||), \qquad \text{ahol } M < \infty.$$

$f(x)$ maximum helyeinek megkeresésére az

$$Y_{n+1} = Y_n + \gamma_n Z_{n+1}, \qquad Y_0 \in E^r$$

gondolatmenet használható, ahol

$$Z_{n+1} = U(Y_n) + \lambda_{n+1}.$$

A számítógépes algoritmust *Wishart* dolgozta ki e módszer alapján [43].

6. „Kevert modell"

Nagy minta esetén jól alkalmazható modelltípus. A hipotézis az egyes csoportokra azonos típusú valószínűségeloszlást tételez fel, az egyes csoportok a paraméterekben különbözhetnek. Az algoritmus az eloszlás típusát ismertnek tételezi fel, és az optimális cluster struktúrát az eloszlások paramétereinek becslése alapján határozza meg. A többváltozós normális eloszlás feltételezése mellett *Wolfe* dolgozott ki számológépes algoritmust (NORMIX; NORMAP).

7. Csoporton belüli variancia becslésén alapuló eljárások

Az eljárás célja olyan osztályok létrehozása, amely rendszer a csoportokon belüli variancia összsokaságbeli összegét minimalizálja.

A többváltozós variancia-elemzés lineáris modellje a következő azonosságból indul ki:

$$x_{ij} = m + (m_j - m) + (x_{ij} - m_j),$$

ahol x_{ij} az i-edik megfigyelés a j-edik csoportban,
m a teljes mintaátlag,
m_j a j-edik csoportátlag.
Átrendezve a fenti formulát

$$x_{ij} - m = (m_j - m) + (x_{ij} - m_j),$$

vagyis az i-edik egyednek a j-edik csoportban a teljes átlagtól való eltérése két részre bontható: a csoportátlagnak a mintaátlagtól való eltérésére és az x_{ij} eltérésére a csoportátlagtól.

Ennek alapján a clusterezés céljára a többváltozós variancia elemzés alapegyenlete a következő:

$$\underbrace{\sum_{j}^{s}\sum_{i}^{n_j}(x_{ij}-m)(x_{ij}-m)'}_{T} = \underbrace{\sum_{j}\sum_{i}(m_j-m)(m_j-m)'}_{K} + \underbrace{\sum_{j}\sum_{i}(x_{ij}-m_j)(x_{ij}-m_j)'}_{B}$$

$$T = K + B$$

$$i = 1 \ldots n_j;\ j = 1, \ldots s,$$

ahol n_j a j-edik csoport elemeinek száma,
s a csoportok száma.

A cluster analízis célja a csoporton belüli homogenitás növelése, azaz a csoporton belüli variancia minimalizálása, ill. a csoportok közötti variancia maximalizálása. Ez a cél többféle típusú döntésfüggvény segítségével valósítható meg.

a) A B mátrix nyomának minimalizálása. Ez a kritérium a csoporton belüli átlagtól mért eltérések négyzetösszegének minimalizálását jelenti. A hierarchikus módszerek között mind agglomeratív, mind divizív algoritmus készült a $tr[B] \rightarrow$ min döntésfüggvény alapján (medián módszer, *Ward*-módszer).

b) A hiba minimalizálása a teljes sokaság varianciájához viszonyítva a *Wilks*-féle Λ döntésfüggvény alapján,

$$\Lambda = \frac{|B|}{|T|}.$$

Az algoritmus olyan cluster struktúrát alakít ki, amelynél Λ értéke minimális.

8. Diszkriminancia analízisen alapuló eljárások

A diszkriminancia analízis lényege, hogy az eredeti ponthalmazt a diszkriminancia függvény segítségével egy olyan térre vetítjük, ahol a pontokból kialakítható csoportok a legjobban elkülönülnek, azaz a pontok közötti diszhomogenitás a lehető legnagyobb. Ez a $\frac{|K|}{|B|}$ hányados maximalizálását

jelenti. A többváltozós variancia elemzés eredményeinek felhasználásával a diszkrimináns függvény:

$$\lambda = \frac{v' Kv}{v' Bv} \to \max!$$

Átalakítva a fenti egyenlőséget

$$(B^{-1} K - \lambda E) v = 0$$

alakúra a λ és v egyszerűen meghatározható. λ a $B^{-1}K$ mátrix sajátértéke, v pedig a sajátvektora. A különböző sajátértékek száma határozza meg a diszkriminációs függvények számát.

A diszkriminációs hatás próbájára is a *Wilks*-féle Λ-t használják, ami ekvivalens az alábbival:

$$\Lambda = \prod_{j=1}^{m} \frac{1}{1 + \lambda_j}.$$

A hierarchikus módszerek közül *Casetti*, *Hung* és *Dubes* módszere épül a diszkriminációs döntésfüggvényre.

9. Gráfelméleti alapokon álló eljárások

A hierarchikus módszerek két típusa gráfelméleti probléma megoldásán alapul.

a) *Egyszerű lánc-módszerek* (single linkage)

Legyen adott az $x_1 \ldots x_n$ pontrendszer az X absztrakt térben, és egy d metrika.

Állítsuk elő az adott pontok által kifeszített minimális fát. A fa konstruálásához a következő algoritmust alkalmazzuk:

- sorbarendezzük a $d(x_i, x_j)$ távolságokat és minden lépésben két pontot összekötünk, amely az alábbi két feltételnek tesz eleget:
- eddig még nem köti össze él a két pontot,
- az összekötött pontokon keresztül nem juthatunk el x_i-ből x_j-be.

Az előbbi két feltételt kielégítő pontpárok közül x_i és x_j távolsága a legkisebb. Az eljárás eredményeként kapott gráf összefüggő részei az egyes osztályok. Az algoritmus során a clusterek száma fokozatosan csökken, végeredményként megkapjuk az adatrendszer által kifeszített minimális gráfot.

A gráf sok értékes információt ad az adatrendszerről, de nagy elemszám esetén nehezen áttekinthető, ezért születtek meg a minimális fát újra felosztó eljárások. A clusterek számát ekkor előre meg kell adni s döntésfüggvény lehet a következő:

$$u = \frac{1}{s} \sum_{j=1}^{s} \bar{d}(j) \to \min,$$

ahol s a clusterek (összefüggő részgráfok) száma,
$\bar{d}(j)$ a j-edik részgráf éleinek átlagos hossza.

b) *Teljes láncmódszerek* (complete linkage)

Két cluster között a hasonlóságot a clusterek legtávolabbi elemei közötti távolsággal mérik, a clusterek közötti összes lehetséges párosításhoz kiszámítják ezt az értéket, és ahol ez minimális, azt a két osztályt vonja össze az algoritmus. Az eljárás lényegében a kialakuló clusterek „átmérőjét" minimalizálja.

10. Osztályozási módszerek. Cluster technikák

A döntési kritériumok sokfélesége, alaphipotéziseikben is lényegesen különböző típusai jelzik a módszertan változatosságát; tovább bővítik ezt a kört a különböző speciális szempontokat kielégítő cluster technikák. Az eljárások különbözősége azonos döntésfüggvény mellett is tovább differenciál (l. 5. táblázat).

5. táblázat

Cluster technikák	Egyszintű módszerek	Egyszintű felosztó módszerek	Objektív módszerek	6. tábla
			Szubjektív módszerek	7. tábla
		Egyszintű optimalizáló módszerek	Iteratív módszerek	8. tábla
			A hierarchia optimális felosztását végző módszerek	Minimálgráfot osztó algoritmus (PAGE)
	Hierarchikus módszerek	Hierarchikus divizív módszerek	Monotetikus módszerek	9. tábla
			Politetikus módszerek	
		Hierarchikus agglomeratív módszerek	Feltételes optimumot kereső eljárások	10. tábla
			A hierarchia adta keretek közt sem optimalizáló eljárások	Centroid módszer nagy adatrendszerekre (WOLFE)

6. táblázat

Objektív módszerek	*Outlierek = egyedülálló súlypontok*	Átlagos lánc Súlypont-módszer *k*-közép módszer (MCQUEEN)
	Egyszerű lánc	Elemi cl. analízis (MCQUITTY)
	Teljes lánc	Legtávolabbi szomszéd módszer "Rank order typal" elemzés (MCQUITTY)

7. táblázat

Szubjektív módszerek	*Egyszerű lánc*	Elemi cl. Paraméter: min. hasonlósági szint Adat átfedő clusterek Elemi cl. analízis (SOKAL és SNEATH)			
		Cluster magot kialakító eljárások	Paraméter: min. hasonlósági szint. Lehetnek átfedő cl.-ek A cl.-mag kialakítása variancia analízisen alapul	Objektív	Max: csoportok közötti diszhomogenitás (ZUBIN, FLEISS, BURDOOK)
					Max: csoporton belüli homogenitás (SNEATH)
				Szubjektív	Max: csoportok közötti diszhomogenitás (BAILEY)
					Max: csoporton belüli homogenitás (SAWREY, KELLER, CONGER)
			Paraméterek: min. hasonlóság és sűrűség Outlierek száma k-tól függ (WISHART)		
	Teljes lánc	A legtávolabbi szomszéd módszere Paraméter: min. hasonlósági szint (SRENSEN)			
	Átlagos lánc	Szakaszonként egy egyed clusteresítését engedélyezi Súlyozza a változókat (SOKAL és MICHENER)			
		Szakaszonként több egyed összevonását is engedélyezi Csoport-pár módszer (SOKAL és SNEATH) Egyenlőtlen súlyozással (LANCE és WILLIAMS)			

8. tálbázat

Iteratív módszerek — Egyszerű cl.-mag. — Politetikus cl.-ek — Természetes cl.-ek — Outlierek megengedettek — Tárolt hasonlósági mátrix — Súlypont-módszerek	*Adott a clusterek száma*	Fix magpontok (RORGY, JANCEY)
		Mozgó magpontok, konvergens k-közép módszer (MCQUEEN)
	A clusterek száma az eljárás során alakul ki	k-közép módszer (MCQUENN)
		A k-közép módszer egy változata (WISHART)
		„Isodata" (BALL és HALL)

A cluster technikák többféle szempont alapján osztályozhatók, leggyakoribb elv a következő:

a) átfedéses osztályozás,
b) diszjunkt osztályozás.

9. táblázat

Monotetikus módszerek (Outlierek nagyszámban lehetségesek)	Felosztás egyetlen, — a leginkább elkülönülő —, változó szerint (LAMBERT és WILLIAMS)
	A jellemzőket dichotomizálja, úgy hogy a max. információ-veszteség feltétele teljesüljön (LANCE és WILLIAMS)
	A csoporton belüli négyzetes hibaösszeget minimalizálja (SONQUIST és MORGAN)
Politetikus módszerek (A outliereket nem választja külön)	Mesterséges clusterek, diszkriminancia analízis (MAYER)
	Természetes clusterek, diszkriminancia analízis (CASETTI, HUNG, DUBES)
	Természetes clusterek, variancia analízis, nyomkritérium (EDWARD és CAVALLI—SFORZA)

10. táblázat

Feltételes optimumot kereső eljárások Tárolt hasonlósági mátrixszal	*Egyszerű lánc*		Legközelebbi szomszéd módszer Hierarchikus cl. analízis (MCQUITTY)
	Teljes lánc		Legtávolabbi szomszéd módszer (SAUNDERS, SCHUEMAN)
	Átlagos lánc	Egyedek páronkénti hasonlósága	Távolsági mérték min. a csoporton belüli (ORLÓCI) Korrelációs mérték max. a csop. belül (ORLÓCI) HOLZINGER-féle B-koeff. (TYRON)
		Centroid módszerek	WARD LANCE és WILLIAMS SOKAL és SNEATH Medián módszer (GOWER)
Tárolt adatrendszerrel (*átlagos lánc*)	WARD-módszer		
	Csoporton belüli variancia min.		
	Csoporton belüli eltérés négyzetösszegének min.		
	Centroid módszerek		
	Minden adat külső tárolón		Rendszerezett hasonlósági mátrix, egyszerű lánc, legközelebbi szomszéd módszer
			Módosított centroid módszer (PARK)

Az átfedéses osztályozás gyakorlati jelentősége kisebb, elmélete is kevésbé kidolgozott, relációelméleti alapon közelíthető meg.

Részletesen a diszjunkt átfedéseket nem tartalmazó osztályozási módszerekkel foglalkozunk, ezt gyakorlati jelentőségük és elméleti kidolgozottságuk indokolja. Az ilyen osztályozásnak két nagy csoportja van:

1. Nem hierarchikus osztályozás: az alapsokaságot k számú osztályra bontja.
2. Hierarchikus osztályozás: kezdetben minden elemet külön osztálynak tekint, majd az osztályok összevonásával lépésről-lépésre újabb osztályozási szinteket alakít ki, mindaddig, amíg az összes elem egyetlen osztályba nem kerül. (Lehetséges természetesen e gondolatmenet fordított irányú alkalmazása is.)

Általános tulajdonságok

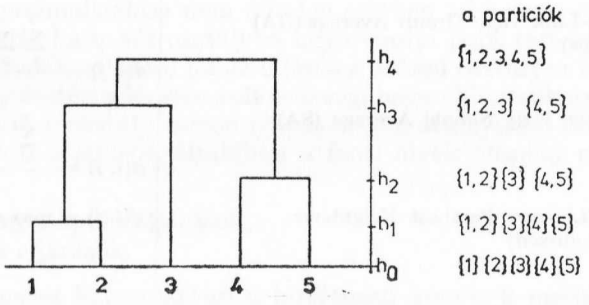
Az osztályozási algoritmus input adatának, a DC mátrixnak az előállítási módjaival a korábbi fejezetekben foglalkoztunk.

Az osztályozási algoritmus outputja az S halmaz diszjunkt partícióinak egy véges sorozata. Olyan fa struktúrával ábrázolható, ahol a fa csomópontjaihoz

$$h \in [0, \max d(A, B)]$$

értékek tartoznak.

Az elmondottakat az alábbi egyszerű példával illusztráljuk, ahol az egész számok objektumokat reprezentálnak (3. ábra).



3. ábra

S-nek egy adott partíciója a fa struktúra egy h szintjéhez tartozik, a sorozat első eleme az izolált pontok halmaza, az utolsó pedig az összes objektumokból álló halmaz. Az ilyen típusú fa-struktúrát *dendogramnak* nevezzük, és a következőképpen definiáljuk.

Jelölje $E(S)$ az S-halmazon értelmezett ekvivalencia relációkat, amelyek egyértelműen meghatározzák a halmaz diszjunkt partícióit, az ekvivalencia osztályokat.

A dendogram olyan

$$C: [0, \infty \to E(S)]$$

leképezés, amely az alábbi tulajdonságokkal rendelkezik:

1. Monotonitás: $C(h) \subseteq C(h')$, ha $0 \leq h \leq h'$.
2. Létezik a két triviális partíció.

3. A partíciók sorozata jól definiált, azaz adott $h > 0$-hoz létezik olyan $\delta > 0$, hogy

$$C(k + \delta) = x(h).$$

Minden objektumpárhoz azt a szintet rendeljük, ahol először egyesültek a dendogramban. Egy adott h szinten azok az objektumok vannak relációban, amelyek között a távolság kisebb vagy egyenlő mint h.

A kiinduló DC mátrix az iterációs lépések során mindig megváltozik, és az S elemei közötti távolságot minden iterációban újra kell számolni.

Ha az i-edik és j-edik csoport elemeinek páronkénti távolságát $d(i_l, j_m)$ jelöli, és ezen távolságok halmazát D_{ij}, akkor az i-edik és j-edik csoport közötti taxonomikus távolságot a módszerre jellemző

$$d(i, j) = f(D_{ij})$$

függvénnyel számítjuk ki.

Az eddigiekből következik, hogy az egyes módszerek a választott $f(D_{ij})$ függvényben különböznek, vagyis abban, hogyan értelmezzük a csoportok közötti távolságot. A leggyakrabban használatos definíciókat táblázatban foglaljuk össze.

Megnevezés	$d(i, j) = f(D_{ij})$
Single-Link (SL) vagy Nearest Neighbour (legközelebbi szomszéd)	$d(i, j) = \min d(i_l, j_m)$
Weighted-Average-Link vagy Group Average (GA) (súlyozott átlagos)	$d(i, j) = \dfrac{\sum_{l}^{r} \sum_{m}^{s} \mid i_l \mid \mid j_m \mid d(i_l, j_m)}{\mid i \mid \mid j \mid}$
Unweighted-Average vagy Simple Average (SA) (átlagos)	$d(i, j) = \dfrac{\sum_{l}^{r} \sum_{m}^{s} d(i_l, j_m)}{r\, s}$
Complete-Link (CL) vagy Farthest Neighbour (legtávolabbi szomszéd)	$d(i, j) = \max d(i_l, j_m)$

11. Nem hierarchikus osztályozás

A módszer lényegét a következő meggondolás alapján érthetjük meg. Induljunk ki abból, hogy objektumaink S halmazát két osztállyá kell bontani S_1 és S_2-re, a felbontást $2^{n-1} - 1$ féleképpen végezhetjük el. Az osztályozás hatékonyságát a

$$Q = \sum_{j=1}^{n_1} \sum_{k=j+1}^{n_1} d^2(x_{1j}, x_{1k}) + \sum_{j=1}^{n_2} \sum_{k=j+1}^{n_2} d^2(x_{2j}, x_{2k})$$

számmal mérjük, vagyis az azonos clusterbe tartozó objektumpárok eltérésének négyzetösszegével. Optimális a felbontás, ha Q értéke minimális. Minthogy az összes objektumpár eltérésének négyzetösszege

$$\sum_{i=1}^{n} \sum_{k=i+1}^{n} d^2(x_i, x_k) = c$$

konstans, ezért az előbbi feltétel ekvivalens a következővel, a:

$$Q' = \sum_{j=1}^{n_1} \sum_{k=1}^{n_2} d^2(x_{1j}, x_{2k})$$

szám legyen maximális. A $Q + Q' = c$ összefüggés miatt a két követelmény ugyanannál az S_1, S_2 felbontásnál teljesül.

A problémát csak az jelenti, hogy milyen módszerrel találjuk meg a felbontást. Az összes lehetséges esetek száma $2^{n-1} - 1$ és csak kis n értékekre végezhető el az enumeráció még számítógép segítségével is. Éppen ezért közelítő módszereket alkalmazunk. Q értékét jól közelíti a

$$Q_1 = n_1 n_2 d^2(C_1, C_2)$$

szám, elegendő ezt maximalizálni. További egyszerűsítés, ha Q_1 helyett a centroidok távolságának négyzetét maximalizáljuk:

$$Q_2 = d^2(C_1, C_2) \to \max.$$

A megoldás során egy tetszőleges felbontásból kiindulva sorra áthelyezünk egy-egy pontot a másik clusterba; az algoritmus véget ér, ha az áthelyezés nem változtat a centroidok távolságán.

A Q' és Q_2 maximalizálása nem minden esetben ad azonos eredményt, általában csak akkor, ha a két osztályba ugyanannyi pont tartozik.

A módszer általánosítható, ha az eljárás k számú osztályra bontja a sokaságot. Két osztály esetén n lépésre volt szükség, hogy eldöntsük az osztályozásról, hogy optimális-e; k osztály esetén $(k-1) \cdot n$-féle áthelyezést kell megvizsgálni.

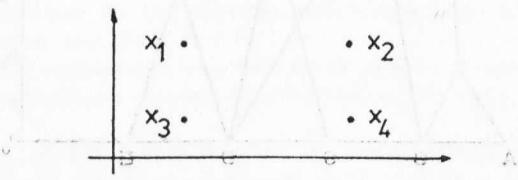
Az alkalmazott eljárások általában a fenti elvek alapján működnek, ilyenek az:

- összevonáson alapuló eljárások,
- reallokációs eljárások.

E két módszerrel kapcsolatban a következő kérdések merülnek fel:

1. Független-e az eredmény az induló osztályozástól?
2. Van-e olyan szempont, amely szerint az osztályozás optimálisnak tekinthető?

Az említett módszerek nem teljesítik a feltételeket, ha pl. a 4. ábrán levő pontokat tekintjük.



4. ábra

Az induló osztálybasorolás legyen $S_1 = \{x_1, x_2\}$, $S_2 = \{x_3, x_4\}$. Az algoritmus az első lépésben végetér, mert minden pont közelebb van a saját osztálya centroidjához, mint a másik osztályéhoz. Ugyanakkor nyilvánvaló, hogy az osztályozás nem minősíthető jónak. A centroid módszerek nem elégítik ki az egyértelműség és monotonitás követelményét.

12. Hierarchikus osztályozás

A módszer kialakítása *Jardin* és *Sibson* nevéhez fűződik. A hierarchikus módszereknek két fő típusa van

a) *Összevonó (agglomeratív) eljárások:*
Indulásnál minden pontot külön clusternek tekintünk és az egyes lépések során mindig két osztályt egyesítünk.

b) *Felosztó (divizív) eljárások*
Az előbbi módszerrel ellentétben itt az indulásnál egyetlen osztálynak tekintjük az objektumok halmazát, és az egyes iterációk során valamelyik osztályt mindig két osztállyá bontjuk.

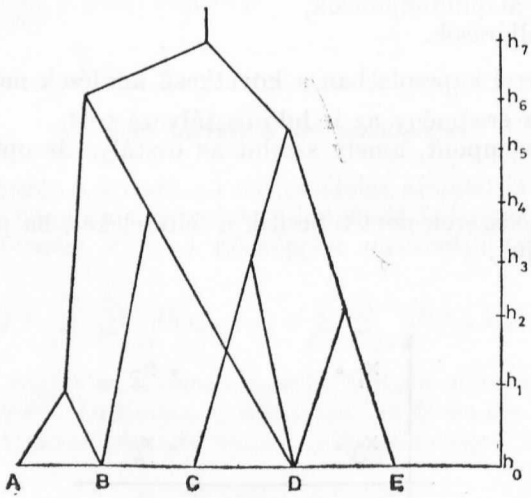
Az említett módszerek minden egyes h szinten ekvivalencia osztályokat határoznak meg az S halmazon. Általánosabb az ún. B_k-módszer, amely az egyes csoportok között átfedéseket is megenged, az átfedés mértékét a k paraméter határozza meg.

Legyenek $S_i, S_j \subset S$ a h szinten a B_k által kialakított csoportok, akkor

$$| S_i \cap S_j | \leq k - 1, \quad \forall h > 0\text{-ra.}$$

A B_k módszer $k = 1$ esetén az SL módszert adja.

A polihierarchikus elnevezés az osztályok kialakulását reprezentáló k-dendogram alapján indokolt. Ez olyan fa-struktúra, ahol minden csomópontból k számú út vezethet a magasabb szinteken levő pontokhoz. Az 5. ábrán egy 3-dendogramot mutatunk be.



5. ábra

A k-dendogram $k > 1$-re mint

$$c_n: [0, \infty) \to \Sigma(S)$$

leképezés adható meg, ahol (S) az S-en értelmezett szimmetrikus reflexív relációk halmaza. A k-dendogramnak olyan $M(d)$ feleltethető meg, amely az ún. gyenge k-ultrametrikus egyenlőtlenséget teljesíti az ultrametrikus helyett. Azaz, ha

$$P \subseteq S \text{ és } |P| = k, \text{ akkor } \forall (A, B) \in S\text{-re},$$

$$d(A, B) \leq \max \{d(x, y) \mid x \in PU \{A, B\}; y \in P\}.$$

Ezt az összefüggést a gyakorlatban előforduló DC mátrixok közül lényegesen több elégíti ki, mint az ultrametrikus egyenlőtlenséget.

Rohlf olyan minimális élhosszúságú fák előállításán alapuló számológépes algoritmust dolgozott ki, ahol fennáll, hogy az input DC mátrix azon elemei, amelyek kielégítik a gyenge k-ultrametrikus egyenlőtlenséget, nem változnak az osztályozás folyamán. Számos alkalmazásnál azonban nem engedhető meg az osztályok átfedése, ugyanakkor igény a valóban homogén osztályok megtalálása. Ebben az irányban jelent továbbfejlesztést a k-ad fokú hierarchikus osztályozás.

13. Értékelési szempontok

A különböző klasszifikáló módszerek minden esetben valamilyen osztályozást létesítenek az objektumok összességében. A kapott osztályozást többféle szempont szerint minősíthetjük.

1. Az eljárás eredményéhez valamilyen mérőszámot rendelünk. Pl. az egyes clusterek belső szórásának négyzetösszegét viszonyítjuk a teljes szórásnégyzethez. Ennek közelítő értéke

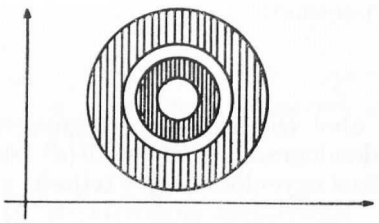
$$h = \frac{\sum_{i}^{k} \sum_{j}^{n_i} d^2(C_i, x_{ij})}{\sum_{i=1}^{n} d^2(C, x_i)},$$

$0 \leq h \leq 1$; h akkor lesz nulla, ha minden objektum egy-egy osztályt alkot, és akkor egy, ha minden osztály centroidja azonos.

Nyilvánvalóan nem állíthatjuk, hogy az osztályozás annál jobb, minél kisebb h értéke. Érdemi összehasonlításra csak akkor van lehetőség, ha az osztályok számát (k) rögzítjük, de még ezzel a megszorítással sem fogadható el, hogy h minden esetben az osztályozás hatékonyságát méri. Pl. a 6. ábra adekvát osztályozása esetében $h = 1$.

Ha két különböző osztályozásunk van és az osztályok száma azonos, akkor a h értékek összehasonlítása alapján dönthetünk egyik vagy másik osztályozás mellett.

Az értékelés egy másik szempontjánál az osztályozástól azt várjuk, hogy maximális információt adjon az objektumokról, vagyis a pontok eloszlása az



6. ábra

osztályok között legyen egyenletes. Ebben az esetben alkalmazható mérték az osztályozásnak, mint valószínűségi változónak az entrópiája:

$$h = -\sum_{i=1}^{k} \frac{n_i}{n} \log \frac{n_i}{n}, \qquad 0 \leq h \leq \log k.$$

$h = 0$ esetén egyetlen osztályunk van és semmi információt nem kapunk. Ha $h = \log k$, akkor minden osztályban azonos számú $\frac{n}{k}$ objektum található; ekkor maximális az átlagos információ.

A legtöbb esetben több szempontot kell egyidejűleg figyelembe vennünk, amikor egy osztályozást minősítünk. Így előfordulhat, hogy egy osztályozás valamely szempontból jobb, egy más szempontból rosszabb a másiknál. Kereshetünk tehát olyan kritériumot is, amely szerint nem lehet tetszőleges osztályozásokat összehasonlítani, de bizonyos osztályozások között mégis egyértelműen dönthetünk. Ez a meggondolás az alábbiak szerint általánosítható:

Legyen $S_1 \ldots S_n$ és $Z_1 \ldots Z_k$ az X halmaz két felbontása, és tegyük fel, hogy az $x_1 \ldots x_n$ elemek átrendezhetők egy $x_{i1}, x_{i2} \ldots x_{in}$ sorozattá oly módon, hogy ha x_j és x_k azonos osztályba tartoznak az S felbontásban, akkor a nekik megfelelő x_{ij} és x_{ik} pontok is azonos osztályba tartoznak a Z felbontásban és fordítva. Ilyen feltételek mellett azt mondhatjuk, hogy az S felbontás legalább olyan jó, mint a Z, ha

a) azonos osztályba tartozó tetszőleges pontpárra

$$d(x_j, x_k) \leq d(x_{ij}, x_{ik}) \text{ és}$$

b) különböző osztályokba tartozó tetszőleges pontpárra

$$d(x_j, x_k) \geq d(x_{ij}, x_{ik}).$$

Ha legalább egy helyen határozottan egyenlőtlenség áll fenn, akkor az S felbontás jobb, mint Z.

Az osztályozások között tehát egy részben rendezési relációt értelmezhetünk. Tehát ahhoz hogy két osztályozás összehasonlítható legyen szükséges (de nem elegendő) feltétel, hogy az osztályok megfeleltethetők legyenek egymásnak, abban az értelemben, hogy a megfelelő osztályokba ugyanannyi objektum tartozzék.

2. A klasszifikációs módszerektől megköveteljük, hogy az eredmény független legyen a kiinduló osztályozástól. Egy további gyakori követelmény, hogy a módszer a lineáris transzformációkkal szemben invariáns legyen: ha az $x_1 \ldots x_n$ pontok helyett az $ax_1 + b, \ldots ax_n + b$ pontokra alkalmazzuk az eljárást, az eredmény ne változzék. E követelmény teljesülése azon is múlik, hogyan definiáljuk a távolságot pontjaink között. Ha a vektorok hajlásszögének koszinuszát tekintjük távolságnak, akkor az $ax + b$ transzformáció hatására a pontok közötti távolság nem arányosan fog változni. A módszerek stabilitásának értékelésére egy lehetőség: Tegyük fel, hogy egy eljárással az $S_1 \ldots S_i \ldots S_k$ felbontást kaptuk, természetes követelménynek látszik, hogy ha az S_i osztály objektumait elhagyva megismételjük az eljárást, akkor az $S_1 \ldots S_{i-1}, S_{i+1} \ldots S_k$ osztályozást kell kapnunk.
3. Magának az adathalmaznak az értékelésére általánosan használható módszer nem adható. Adhatók ugyan absztrakt definíciók, hogy mikor „jó" egy ponthalmaz struktúrája, de ezek a feltételek a gyakorlatban szinte sohasem teljesülnek. Amit tehetünk: többféle eljárással megismételjük az elemzést, s ha eredményeink összhangja megfelelő, elfogadjuk azokat. Ellenkező esetben munkánkat tovább folytatjuk. Feladatunk ilyenkor a módszerek s az adathalmaz kölcsönhatásának felderítése, a reálisan működő algoritmus kiválasztása. Mindehhez célszerűen segítségül hívhatjuk a sokváltozós adatelemzés további módszereit is.
4. Minősíthetjük változóinkat is, olyan szempontból, hogy az egyes változóknak mekkora szerepük van az osztályok kialakításában. Ha az $x_1, x_2 \ldots x_r$ változók mellett a pontok osztályozását is egy további x változónak tekintjük, akkor vizsgálhatjuk x és a többi változó kapcsolatát úgy, hogy az $R(x, x_i)$ korrelációs együtthatók értékei szerint rendezzük a változókat. Ennél hatékonyabb ha az $I(x, x_i)$ kölcsönös információval mérjük, hogy az x_i változó milyen mértékben határozza meg az osztályozást. Ha $I(x, x_i) = 0$, az x_i változónak semmilyen szerepe nem volt az osztályok kialakításában; az ilyen változókat elhagyhatjuk.

Ha a változóinkat úgy számozzuk meg, hogy $i < j$ esetén $I(x, x_i) \leq I(x, x_j)$ legyen, akkor magukat a változókat is osztályozhatjuk. A metrika ebben az esetben:

$$d(x_i\, x_j) = |\, I(x, x_i) - I(x, x_j)\, |.$$

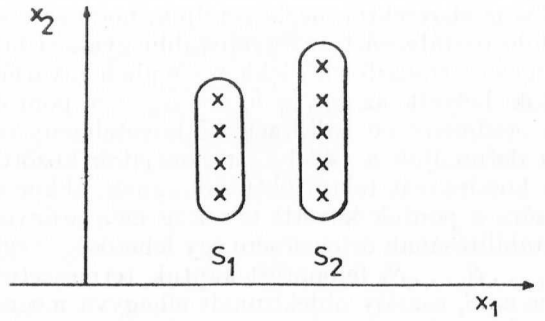
A legegyszerűbb eset, ha két csoportra osztjuk a változókat:

$$S_1 = x_1 \ldots x_j\text{: irreleváns változók,}$$

$$S_2 = x_{j+1} \ldots x_r\text{: releváns változók.}$$

A változók osztályozása után célszerű megismételni az objektumok osztályozását csak a releváns változók alapján. (Hasonló célt érhetünk el, ha az osztályozás előtt elvégezzük a változók faktoranalitikus vizsgálatát.) A 7. ábrán jól látható, hogy az x_1 változó teljes mértékben meghatározza az osztályozást, x_2 független az osztályozástól. Ha az objektumokat az x_1 tengelyre vetítjük, akkor a távolságok megváltoznak ugyan, de ugyanazok a pontok fognak egy osztályba tartozni.

Az ábrán jól látható, hogy x_2 elhagyása nemcsak azért indokolt, mert nincs szerepe az osztályozásban, hanem mert az osztályozás minősége is javítható: a külső és belső szórások hányadosa jelentősen csökken.



7. ábra

Az $I(x, x_i)$ kölcsönös információkat felhasználhatjuk a változók súlyozására is: ha az x_i változó w_i súlyát éppen $I(x, x_i)$-nek választjuk, akkor megismételhető az osztályozás ezekkel a súlyokkal. Az új felosztásnak, mint x' változónak, újra kiszámíthatjuk a kölcsönös információját az egyes változókkal; a következő lépésben a súlyokat $w_i' = I(x', x_i)$-re módosíthatjuk. Az iteráció végén kapott súlyok fejezik ki, hogy az egyes változóknak milyen szerepük van a végső osztályozásban.

(Beérkezett: 1977. aug. 15-én.)

IRODALOMJEGYZÉK

1. Anderberg, M. R.: Cluster Analysis for Applications. Academic Press, New York, 1973.
2. Everitt, B.: Cluster Analysis. London, 1974.
3. Ball, G. H.: Classification Analysis. California, 1970.
4. Ball – Hall – Isodata, A.: A Novel Method of Data Analysis and Pattern Classification. California, 1968.
5. Beale, E. M.: Euclidean cluster analysis. North Holland C.; Amsterdam, 1970.
6. Forgy, E. W.: Cluster Analysis of Multivariate Data. Biometrics. Vol. 21. No. 3. p. 768.
7. Birnbaum and Maxwell: Classification procedures based on Bayes formula. University of Illionois Press, 1965.
8. Bryan, J. K.: Classification and clustering using density estimation. Columbia, 1971.
9. Csibi-Gulyás: A számítógépek tanítása. Természet Világa, 1973. VIII.
10. Csibi, S.: Optimális döntésfüggvények iteratív tanulásáról. Preprint TKI, 1971.
11. Gulyás, O.: Tanuló algoritmusok reprodukáló magú Hilbert terekben. Szemináriumi Közlemények TKI, 1971.
12. Győrfi, L.: A potenciálfüggvényes algoritmusok konvergenciája. TKI, 1971.
13. Cole, A. J.: Numerical Taxonomy. Academic Press. New York, 1969.
14. Goodman and Kruskal: Measures of association for cross classifications. I. Amer. Statist. Assoc. Vol. 49.
15. Fritz, J.: Az alakfelismerés statisztikus módszerei. MTA Budapest, 1974.
16. Harman, H. H.: Modern Factor Analysis. University of Chicago Press, Chicago, 1960.
17. Harrison, I.: Cluster analysis. Metra, 1968.
18. Jardine and Sibson: The structure and construction of taxonomic hierarchies. Math. Biosciences. Vol. 1. No. 2. 1967.
19. Jardine, N.: Algorithm, methods and models in the simplification of complex data. Comput J. 13.
20. Jardine and Sibson: Mathematical Taxonomy. New York, 1971. p. 289.

21. KRUSKAL, J. B.: Multidimensional scaling by optimizing goodness-of-fit to a nonmetric hypothesis. Psychometrika 29, 1 – 27.
22. KRUSHAL, W. H.: Ordinal measures of association. J. Amer. Statist. Assoc. Vol. 33. 814 – 861.
23. MACNAUGHTON – SMITH – WILLEAMS: Numerical Classification of Sequences. The Australian Computer J. Vol. 2. No. 1.
24. MACQUEEN, J. B.: Some methods for Classification and Analysis of Multivariate Observations. Math. Statist. Vol. 1. 281 – 297.
25. ORLOCI: Összevonáson alapuló módszer növényi ökológiai rendszerek osztályozására. Journal of Ecology. 1967. 55. 193 – 206.
26. ROMNEY – SHEPARD: Multidimensional Scaling. Seminar Press. New York, 1972.
27. MAHALANOBIS: On the generalized distance in statistics. Proc. Natl. Inst. Sci. Vol. 12.
28. MARRIOT, F. H.: Practical problems in a method of cluster analysis. Biometrics Vol. 27. 501 – 514.
29. MORRISON, D. G.: Measurement problems in cluster analysis. Managment Sci. 13. 775 – 780.
30. PARKS, J. M.: Classification of mixed mode data by r-mode factor analysis and q-mode cluster analysis on distence functions. Academic Press. New York, 1969.
31. PARZEN, E.: On estimation of a probability density function and mode. Ann. Math. Statist. Vol. 33. 1065 – 1072.
32. RAO, M. R.: Cluster Analysis and Mathematical Programming. Amer. Statist. Assoc. Vol. 66. 1971.
33. ROHLF – SOKAL: Coefficient of correlation and distance in numerical taxonomy. Kausas University Sci. Vol. 45. 8 – 27.
34. SEBESTYÉN, G.: An algorithm for nonparametric pattern recognition. Electronic Computers Vol. 15. No. 6.
35. SOKAL – SNEATH: Principles of Numerical Taxonomy. San Francisco, 1963.
36. SPEARMAN, C.: Correlations of sums and differences. Brit. J. Psychol. 5. 417 – 426.
37. WALLACE – BOULTON: An information measure for classification. Comput J. Vol. 11. p. 185.
38. TOU – GONZALEZ: Pattern Recognition Principles. Addison – Wesley Publishing C. London – Amsterdam – Ontario – Sydney, 1974.
39. WARD, J. H.: Hierarchical grouping to optimize an objective function. J. Amer. Statist. Assoc. Vol. 58. No. 301.
40. LANCE – EILLIAMS: A general theory of classificatory sorting, strategies. Computer J. Vol. 3.
41. LANCE – WILLIAMS: Mixed-Data Classificatory Programs. The Australian Computer J. Vol. 1. No. 2.
42. WISHART, D.: Mode analysis: A generalization of nearest neighbor, which reduces chaining effects. Academic Press. New York, 1969.
43. WISHART, D.: An algorithm for hierarchical classifications. Biometrics. Vol. 22. No. 1.
44. ZADEH, H. A.: Fuzzy sets. Information and Control Vol. 8. 338 – 353.
45. JARDINE – SIBSON: Mathematical Taxonomy. London, 1971.
46. KEREKES ÁGNES – KISS PÉTER: A cluster analízis és egy lehetséges közgazdasági alkalmazása. (Szakdolgozat, Bp. 1977.)

CLUSTER ANALYSIS: CONCEPTS AND METHODS

In the introductory part a survey is given on classification methods, measuring scales of various type, their transformations as well as on the concepts of similarity both among criteria and objects and on measurement possibilities. All this is based on *Anderberg*'s book: Cluster Analysis for Applications.

In the subsequent parts the methodological aspects of cluster analysis are discussed. This includes also a summary of classification criteria and the types of decision functions. The authors briefly review various procedures based on durity functions estimation, on the concept of the "mixed model", on the estimation of variance within groups and on discriminancy analysis, as well as relying on graph theory, simple and complete chain methods, respectively.

In certain respects – e.g. dendogramme – technical details are also dealt with. Hierarchic, nonhierarchic, agglomerative, divisive procedures are discussed and the most relevant points of view of essessment reviewed.

The article is closed by a fairly comprehensive block diagramme reflecting systems approach and a rich bibliography.

КЛАСТЕРНЫЙ АНАЛИЗ

Во вводной части на основании книги Андерберга: „Cluster Analysis for Applications" дается обзор методов классификации, различных типов измерительных шкал, их трансформации, возможностях измерения а также понятия подобия критериев и объектов.

В дальнейшем рассматриваются методологические аспекты кластекрного анализа. Здесь происходит также и обобщение критериев классификации, типов функций принятия решений. Кратко излагаются методы, основанные на оценке функции частоты, опирающиеся на представления «смешанной модели», использующие вариационную оценку в рамках группы, в основе которых находится дискриминационный анализ, теория графов, простые методы цепей, а также полные методы цепей.

В некоторых случаях затрагиваются и, например, в отношении дендограммы, технические, неиерархические, агломеративные и дивизивные методы, излагаются наиболее важные точки зрения оценки.

Данная статья завершается довольно объемной блок — диаграммой, отражающей системный подход, а также и обширным списком литературы.

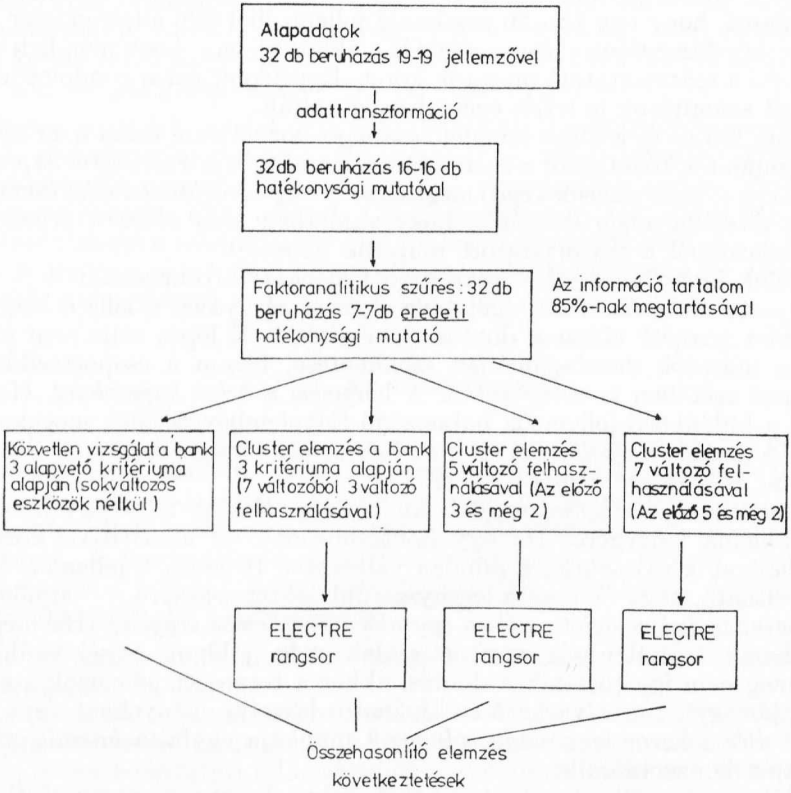
Füstös László—Meszéna György—Simonné Mosolygó Nóra

Beruházási javaslatok csoportosítása, rangsorolása

A tanulmány keretében sokváltozós segédeszközöket használunk: a faktoranalízist, clusteranalízist s az ELECTRE skálázó eljárást. A faktoranalízist ismertnek tételezzük fel, a clusteranalízisről előző cikkünkben részletesen írtunk, az ELECTRE eljárást röviden itt ismertetjük.

Vizsgálatainkat az alábbiakban először vázlatosan tekintjük át, majd kissé részletesebben írjuk le.

A gyakorlatban a döntéshozók ismételten találkoznak az alábbi problémával. *Nagyobb számú beruházási alternatíva közül kell egyet vagy többet kiválasztani*



1. ábra

s ezután megvalósítani. A *konkrét szituációk* tág határok között változhatnak. Lehetséges, hogy *egy meghatározott beruházás* valójában sok szóbajöhető változatban valósítható meg. Más esetben *különböző célú beruházásokról* kell a rendelkezésre álló pénzeszközök erejéig dönteni, a megvalósítandó javaslatokat kiválasztani.

Alapvető *induló feltevésünk* a következő: minden a *vizsgálat tárgyát képező* beruházási javaslatot *ugyanazon típusú jellemző adatokkal* terjesztjük elő. Tipikus példa erre: a konvertibilis export áru-alapok fejlesztését szolgáló beruházási javaslatok elbírálási rendszere.

A közvetlenül szolgáltatott nyers beruházás-jellemzők számottevő része eleve nem ad határozott információt a beruházás „jóságáról". (Pl. beruházási költség, átfutási idő stb.) Célszerű a jellemzők induló rendszerét alkalmasan átalakítani, s a beruházásokat már határozottabban „véleményező" mutatók egy csoportjával dolgozni. Szeretnénk felhívni a figyelmet e mutatók képzésével kapcsolatos néhány megállapításra. Természetes igény — ha nem akarunk feleslegesen információt veszteni —, hogy *valamennyi* induló adat szerepeljen a képzésben. Érdemes figyelni a gyakorlat során egyébként is alkalmazott mutatók szerepeltetésére. Világosan látni kell azonban a következőket. Már akár 15–20 induló jellemzőből is igen sok származtatott mutató képezhető, ezek azonban — bármilyen sokan vannak — *semmivel sem tartalmaznak több információt*, mint az induló rendszer. Ha ehhez még hozzávesszük azt a gyakori tapasztalatot, hogy egy 15–20 gazdasági jellemzőből álló adatrendszer a legtöbbször *már önmagában erősen redundáns*, akkor világos, hogy nem kell erősen szaporítani a származtatott mutatók körét. Egyébként ezt a gondolatmenetet a konkrét számítások is teljes egészében igazolták.

A másik fontos és a külső szemlélő számára közvetlenül talán nem nyilvánvaló szempont a következő; a származtatott mutatók közötti *belső kapcsolatok* miatt az egyes beruházások végső megítélése, a *csoportosítás elvégzése* szemszögéből elég messzemenően *közömbös*, hogy valójában *mely alakban* végezzük az induló adatokból a származtatott mutatók képzését.

Munkánk lépései ezután a következő fázisokat tartalmazzák:

1. A képzett mutatók összességét téve elemzés tárgyává, a lehető *legkevesebb mutatóra vezetjük vissza* a döntés előkészítését. E lépés célja nem elsősorban a mutatók darabszámának csökkentése, hiszen a csoportosítás több jellemző esetében is elvégezhető. A lényeges a *belső kapcsolatok kiszűrése*, s így a különböző jellemzők halmozódó figyelembevételének megakadályozása. A *további vizsgálatokat tehát az információ lényegét egymástól elkülönítve hordozó jellemzőkre akarjuk építeni.*
2. A következő lépés a *csoportosítás* kialakítása. Hívjuk fel a figyelmet ismét a probléma lényegére. Ha egy döntéshozónak 40 alternatíva közül kell valahányat kiválasztani, s minden változatot 10 mutató jellemez, hogyan biztosítható, hogy — csak a legegyszerűbb esetet tekintve, — *minden lépésben minden információt valóban egyenlően figyelembe vegyen?* (Ha még a súlyozásban is különbség van, a szálak még jobban összekuszálódnak!) Ha meg nem így történik a döntés, akkor a tervezett jellemzők szereplése nem jár együtt az elvárható és kívánatos következményekkel, és a *döntés-előkészítés* sokszor igen *nagy volumenű* munkája egyfajta *látszattevékenység* szintjén konzerválódik.

Kérjük az olvasót, ne vesse most szemünkre, hogy nem vagyunk tisztában a döntésekhez időnként kapcsolódó objektív nehézségekkel, *szükségszerűségek-*

kel. Véleményünk szerint is ilyenkor vagy nincs helye latolgatásnak, vagy a vizsgálatok szerepe legfeljebb az egyébként vitán felül álló tény-helyzet jobb áttekintésében nyilvánulhat meg. Meggyőződésünk azonban, hogy a fejlődés általános folyamatában egyre gyakoribb lesz a *komplex hatékonyság kvantitatív mérlegelésére támaszkodó döntéselőkészítés.* A megfelelő eljárások kialakításához szeretnénk ezúton mi is hozzájárulni.

Célunk a javaslatok *olyan csoportokba* sorolása, *melyek az összes információ szimultán figyelembevételével relatíve homogénebb együtteseket alkotnak.* A döntéshozó e csoportok szem előtt tartásával közvetlenül látja – ha egy javaslat megvalósítása mellett dönt –, melyek azok a további javaslatok, amelyek a vizsgálatban levő adatok *összessége* szerint, a kiválasztott esettel egy kategóriába sorolhatók.

3. Közbevetőleg jegyezzük meg, hogy igen érdekes és tanulságos meggondolásokra juthatunk a következőképpen. *Változtassuk* meg több lépésben a figyelembe vett *jellemzők számát.* Tehát végezzünk csoportosítást először könnyebben áttekinthető *kevesebb,* majd *egyre több* jellemzővel. Így lehet felderíteni – a változás mértékéből – az egyes *jellemzők szerepét,* az eredmények *érzékenységét.* Egy szinttel mélyebben láthatunk be – az elvégzett kvantitatív elemzések útján – a gazdasági problémák természetébe.
4. A munka további része az egyes *javaslatok,* illetve a *kategóriák értékelésével foglalkozik.* Vizsgáljuk tehát, hogy mit lehet mondani az egyes kialakított csoportokról, *vannak-e* egyértelműen *jobb kategóriák, rossz csoportosulások.* A kérdés így is megfogalmazható: mely mutató kombinációk alapján alakíthatók ki jellegzetesen szétváló javaslat együttesek.

Ezek a gondolatmenetek igen jól *gépesíthetők,* az értékeléshez szükséges rajzok is elkészíthetők gépi úton. Ha valamely szervezet egyszer beépít a döntéselőkészítés folyamatába egy ilyen típusú csoportosító, sorbarendező eljárást, akkor a gyakorlat által igényelt rendszerezésben és változatokban a *döntéshozó* elé tárható a rendelkezésre álló *konkrét információk* által rögzített helyzetkép. A döntés ezeknek és *minden esetleges további tényezőnek* a figyelembevételével már a döntéshozó joga és feladata.

1. Az adathalmazról

Az V. ötéves terv 45 milliárd forint kedvezményes beruházási hitelkeretet tartalékol konvertálható export árualapokat növelő kapacitások létesítésére. Annak érdekében, hogy a népgazdasági szempontból előnyösebb, hatékonyabb fejlesztési célok kerüljenek kielégítésre a Magyar Nemzeti Bank pályázatot hirdetett. A hitelkérelmek elbírálásához egységes elv szerint meghatározott információkat kellett a beruházó vállalatoknak szolgáltatni. Az elfogadás legfontosabb feltételei a következők:

– a fejlesztés eredményeként megjelenő termék minden piacon tartósan értékesíthető legyen,
– a fejlesztés teljes költsége (beruházás és forgóeszköz bővítés együtt) legfeljebb 5 éven belül térüljön meg a netto devizahozamból,
– az összes befektetett állóeszközhöz viszonyított vállalati eredmény (eszközarányos nyereség) érje el a hitelpolitikai irányelvekben rögzített minimális 15%-os szintet,

– a fejlesztés minél rövidebb idő alatt valósuljon meg és lehetőleg már 1980-ban teljes kapacitással termeljen,
– a devizakitermelési mutató értéke az átlagosnál kedvezőbb legyen.

Bár a 45 milliárd Ft az V. ötéves terv összes beruházásainak csak alig 5%-a, kedvezőnek kell ítélnünk, hogy a vállalatok közel 300 hitelkérelmet dolgoztak ki, a versenyt alapvetően hatékonysági követelmények alapján bírálták el.

A pályázati feltételek szerint minden egyes beruházásra 19 jellemző adat megadása kötelező. Ezek az adatok az értékelő feldolgozás szempontjából nem tekinthetők egyenrangúaknak – erre a későbbiekben visszatérünk. Ettől eltekintve is egy olyan többváltozós csoportosítási, rangsorolási problémával állunk szemben, amit módszertani segítség nélkül eredményesen, kellő objektivitással nem lehet megoldani.

A rendelkezésünkre álló adathalmaz 1977. évi kezdésre javasolt 32 beruházásra vonatkozik, mindegyikre 19 jellemző adatot tartalmaz az induló tábla. A Bank által kért adatokat eredeti formájukban nem célszerű összehasonlító vizsgálatra felhasználni. Értékelő feldolgozás a „nyers" adatrendszer alkalmas átalakítását igényli, ezért az *eredeti adatokból leszármaztatott hatékonyság típusú mutatórendszer* alapján végeztük a feldolgozást. A mutatórendszer kialakításánál a következő szempontokat vettük figyelembe:

– a rendelkezésre álló információtartalom minél nagyobb hányadát hasznosítsuk,
– az egyes mutatók önálló közgazdasági tartalommal bírjanak és minden esetben a mutatószám nagyobb értéke jelentse a kedvezőbb körülményt, teljesüljön a monotonitás. (Szükség esetén ezt olyan transzformáció segítségével értük el, amely az alkalmazott módszer szempontjából a helyzeten nem módosít.)
– amennyiben lehetséges, az alap (eredeti) adatrendszerben szereplő hatékonyságtípusú mutatókat eredeti formájukban tartsuk meg.

Ilyen módon 16 hatékonyságtípusú mutatót állítottunk össze (l. a függelékben). Ezek jelentős része valamilyen szempont szerinti parciális hatékonyságot fejez ki, amelyek a beruházás hatékonyságát, ill. az export hatékonyságot mérik, mivel a pályázati előírás szempontjai is ezek voltak. Természetesen elképzelhető az alkalmazási tapasztalatok alapján ezen mutatók módosítása is. (Amint a bevezetőben is említettük a hangsúly az információtartalom hordozásán van, s ezt alkalmasan képzett különböző mutató kombinációk – megfelelő sűrítés után – egyaránt biztosíthatják!)

A kialakított mutatórendszer elemei sem feltélenül azonos fontosságúak, teljesen stabil számbavételükhöz objektív súlyrendszerre lenne szükség, amivel természetesen nem rendelkezünk. De ha meg is adnánk valamilyen súlyt minden egyes mutatóhoz, a tényleges figyelembevétel csak akkor történik e szerint, ha a mutatók függetlenek. Általános tapasztalatunk gazdasági adatrendszerekre vonatkozóan, és ebben az esetben is ez áll fenn, hogy adataink között jelentős mértékű sztochasztikus kapcsolatok állnak fenn, ami az adatrendszer nagymértékű sűrítését, a mutatók számának csökkentését teszi lehetővé. A gyakorlatban sok esetben nem súlyozunk, de az előbbiekből következik, hogy a ténylegesen azonos súlyozás megvalósítása nagyon nehéz. Valójában tehát a legtöbb esetben érvényesül valamilyen súlyozás, csak nem ismerjük a súlyokat. E probléma általános megoldása természetesen nem egy-

szerű feladat, s tanulmányunkban nem vállalkozunk még részmegoldásra sem. Azt azonban megjegyezzük, hogy a probléma széles körben felmerül, és a legtöbb esetben egyszerűen nem vesznek róla tudomást.

E vizsgálatban a probléma feloldására törekszünk olyan módon, hogy a kialakított hatékonysági mutatókra elvégeztünk egy faktoranalitikus elemzést – amely korrelálatlan faktorokat állít elő – és a *16 mutatót* az információtartalom 85%-os megtartása mellett *7-re* csökkentettük. Ebből az is következik, hogy ha feltételezzük, hogy indokolt a 19 eredeti jellemző megadása az egyes beruházások megfelelő leírásához, akkor ezek információ mennyiségét hordozó egyszerű szerkezetű hányadostípusú mutatóknak csak valamilyen *rendszere* – nem pedig egyetlen mutatószám – képes megfelelően jellemezni egy beruházást. (A továbbiakban az adatrendszer, változók, jellemzők, megjelölés mindig a hatékonysági mutatókra utal.)

2. A csoportosítás, rangsorolás főbb lépései

A különböző fontosságú mutatók között faktoranalízis segítségével válogattunk. A kiválasztott 7 mutató a következő:

1. a beruházás megtérülése a nettó deviza hozamból,
2. 100 Ft eszközre jutó vállalati nyereség,
3. deviza kitermelési mutató,
4. a beruházási költségre jutó többlettermelés,
5. az egységnyi hitelre jutó nettó devizahozam,
6. a tőkés import gép megtérülése a nettó devizahozamból,
7. az egységnyi építésre jutó nettó devizahozam.

A Bank alapvetően 3 feltétel teljesülését írta elő a beruházási javaslat elfogadásához, az előbbi felsorolásban az 1 – 3. pont alattiakat, ezekre normatív követelményeket támasztott. Ez a 3 változó tehát szerepel a faktor elemzéssel kiválasztott 7 változó között is. (Az összes információnak kb. 50%-át tartalmazzák.)

A vizsgálatokat a stabilitás növelése érdekében a bevezetésben adott séma szerint több változatban végeztük el.

3. Értékelés a bank kritériumai szerint

A Bank elsődlegesen 3 feltételt írt elő a beruházási javaslatok elfogadásához, de nyilvánvalóan törekedett a többi információ felhasználására is a döntésnél. Hogy hagyományos eszközökkel már 3 változó esetében is csak részlegesen oldható meg ez a feladat, annak bemutatására végezzük el a megfelelő számításokat.

A változók és az előírt korlátok:

1. A beruházás megtérülése a nettó deviza hozamból legfeljebb 5 év lehet.
2. A deviza kitermelési mutató értéke legyen kedvezőbb az előírt értéknél (megadott korlát 40 Ft/$).
3. A 100 Ft eszköz ráfordításra jutó vállalati nyereség legalább 15 Ft legyen.

A 3 korlátozó feltétel megadása mellett *változónként külön-külön* rangsorolhatók a beruházások és megadhatók az elfogadási és elutasítási tartományok.

Mindezek alapján a következő megállapításokat tehetjük:

– egyértelműen elfogadhatjuk azokat a javaslatokat, amelyek mind a 3 kritériumnak eleget tesznek (*kurzív számok*), ez számszerűen 13 beruházást jelent,
– elvethetjük az elutasítási tartományok közös részét (*félkövér*) az 5-ös sorszámú beruházást.

Egyértelmű döntést tehát csak 14 esetben hozhatunk és ez alig több mint az összes lehetőség 40%-a. Nyilvánvalóan nem lehetünk elégedettek ezzel a valójában elég durva értékeléssel. Mit csináljunk ha mégis dönteni kell a maradék 18 beruházásról is, amelyek közül egyeseket egy, másokat két feltétel nem teljesülése miatt nem fogadunk el.

1. táblázat

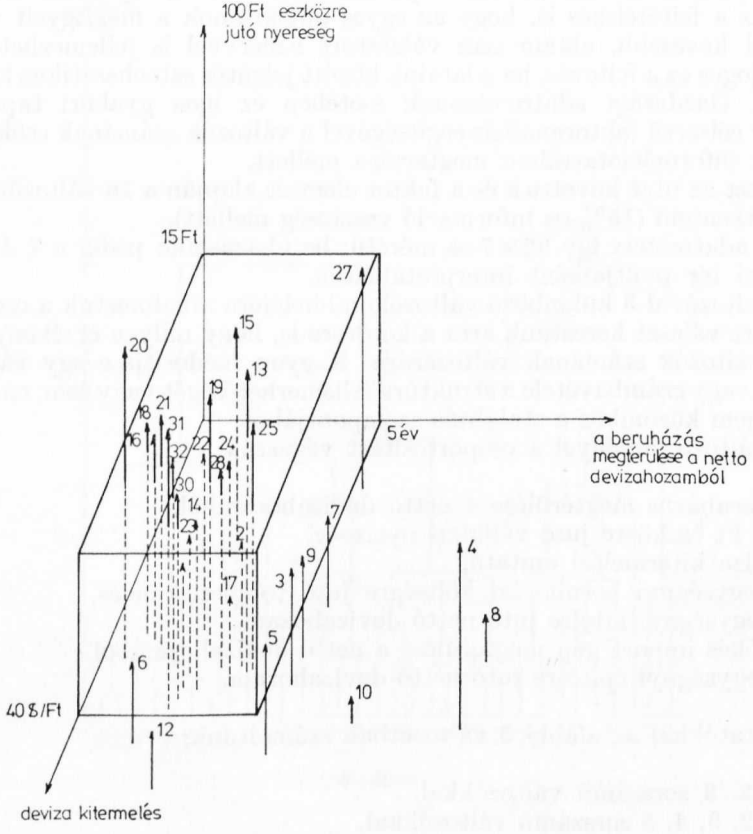
A három szóbanforgó mutató szerint külön-külön elkészített rangsorolás

Beruházások rendje	Mutatók		
	megtérülés a nettó deviza hozamból 1.	deviza kitermelési mutató 2.	100 Ft eszközre jutó nyereség 3.
1	*20*	*13*	*20*
2	14	15	31
3	*21*	27	27
4	*18*	14	*16*
5	*30*	7	*19*
6	*32*	*21*	*25*
7	6	24	*18*
8	*13*	*19*	*21*
9	28	28	*30*
10	*15*	*30*	*32*
11	23	9	*2*
12	19	*25*	*15*
13	24	11	13
14	*16*	26	4
15	22	8	*29*
16	*31*	22	22
17	*29*	*18*	26
18	12	*29*	14
19	17	17	1
20	1	*20*	23
21	*2*	3	3
22	*25*	23	**5**
23	26	*32*	24
24	9	4	17
25	7	2	2
26	3	*31*	28
27	**5**	16	7
28	27	10	12
29	11	6	6
30	10	**5**	8
31	8	1	10
32	4	12	11

elutasítási tartomány (1. oszlop)
elutasítási tartomány (2. oszlop)
elutasítási tartomány (3. oszlop)

Szemléltetés kedvéért ábrázoljuk az egyes beruházásokat a 3 dimenziós tér pontjaiként, adott 3 jellemzőjük mint koordináta értékek alapján. Ha még berajzoljuk a korlátozó feltételek értékeit is, akkor egy olyan felül nyitott hasábot kapunk, amelybe tartozó pontok (beruházások) egyértelműen elfogadhatók a döntési szabály szerint, vagyis mind a 3 feltételnek eleget tesznek. (2. ábra.) Nehezíti a problémát, ha megfigyeljük, hogy milyen jelentős különbség van szigorúság szempontjából az egyes kritériumok között. A deviza kitermelési kritérium csak 4 beruházást utasít el, a 100 Ft eszközre jutó nyereség előírásnak 17 javaslat (több mint fele a 32-nek) nem felel meg. Ésszerűnek látszik az a megfontolás — feltételezve, hogy indokolt e 3 kritérium alapján dönteni — hogy kedvezőbb egy olyan beruházási javaslat elfogadása, amely bizonyos mértékig mind a 3 feltételt egyidejűleg teljesíti, mint pl. egy olyan javaslaté, amely 2 feltételnek magasan eleget tesz ugyan, de a 3-ik követelménynek csak nagyon kis mértékben.

A továbbiakban a jellemzők *együttes* figyelembe vétele alapján lehetővé válik a teljes beruházási halmaz értékelése az előbbi szempontoknak megfelelően, *tetszőleges számú mutató alapján*. Igen jelentős eredmény a megbizhatóság szempontjából az is, hogy a javaslatok csoportokba sorolhatók.



2. ábra

4. Cluster analízis alkalmazása, az eredmények értékelése

A főbb lépések, melyek az alkalmazásoknál mindig felmerülnek, a következők:

a) a változók kiválasztása,
b) az algoritmus kiválasztása,
c) a távolságfogalom megválasztása (az adatrendszer transzformációja),
d) a csoportok számának eldöntése.

Induló adatmátrixunk a hatékonyság típusú változók bevezetésével 32×16 méretű. A 32 beruházási javaslat tehát a 16 dimenziós állapot tér egy-egy pontjának tekinthető. Elméleti megfontolások és jelentős mennyiségű számítástechnikai kísérlet után hierarchikus technikával, euklideszi távolság alkalmazásával dolgoztunk tovább.

Tudjuk, hogy bizonyos esetekben, ha az adatrendszer struktúrája nehezen felismerhető és nem elég karakterisztikus az adatok csoportosulása, előfordulhat, hogy a választott algoritmus nem ismeri fel világosan a valódi struktúrát, hanem – mint amikor rossz szögből nézünk egy kompozíciót – túl bonyolult képet sugall a rendszerről. Ez ellen a veszély ellen többféle módon is védekezhetünk, például az ún. „zavaró" pontok kiszűrésével. Gyakran ésszerűnek látszik az a feltételezés is, hogy az egyes objektumok a megfigyelt változók számánál kevesebb, alkalmasan választott ismérvvel is jellemezhetők. Különösen jogos ez a feltevés, ha adataink között jelentős sztochasztikus kapcsolat áll fenn. Gazdasági adatrendszerek esetében ez igen gyakori tapasztalat. Ilyenkor célszerű faktoranalízis segítségével a változók számának csökkentése, az adott információtartalom megtartása mellett.

Mi is ezt az utat követtük és a faktor elemzés alapján a 16 változót sikerült 7-re csökkenteni (15%-os információ veszteség mellett).

Az új adatmátrix így 32×7-es méretű, beruházásaink pedig a 7 dimenziós euklideszi tér pontjaiként interpretálhatók.

E 7 változóval 3 különböző változókombinációra alkalmaztuk a csoportosítást, mert választ kerestünk arra a kérdésre is, hogy milyen érzékeny a módszer a változók számának változására. Nagyon módosítja-e egy változó elhagyása vagy számbavétele a struktúra felismerhetőségét vagy sem, ez nyilvánvalóan nem közömbös a stabilitás szempontjából.

A 7 változó, amellyel a csoportosítást végeztük:

1. a beruházás megtérülése a nettó devizahozamból,
2. 100 Ft eszközre jutó vállalati nyereség,
3. deviza kitermelési mutató,
4. az egységnyi beruházási költségre jutó többlettermelés,
5. az egységnyi hitelre jutó nettó devizahozam,
6. a tőkés import gép megtérülése a nettó devizahozamból,
7. az egységnyi építésre jutó nettó devizahozam.

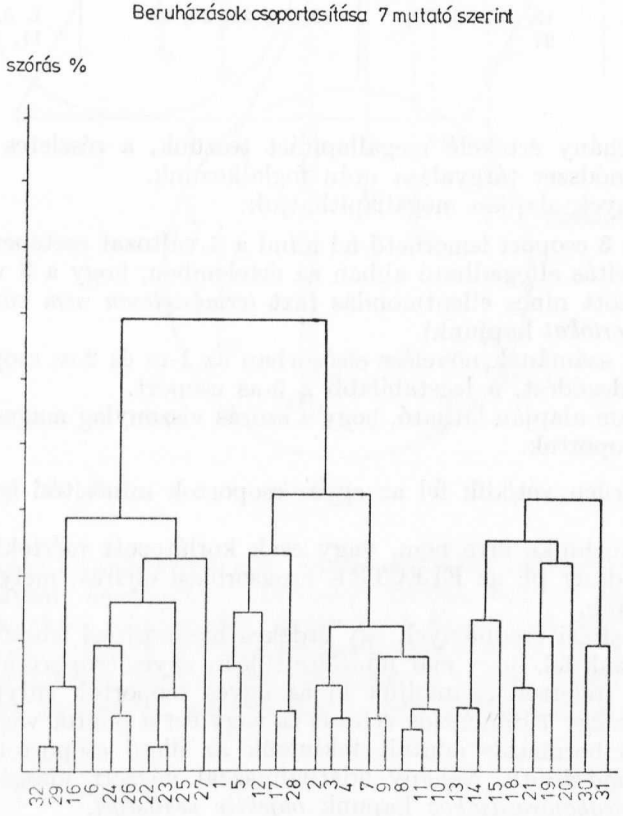
A mutatókkal az alábbi 3 változatban számoltunk:

a) 1, 2, 3 sorszámú változókkal,
b) 1, 2, 3, 4, 5 sorszámú változókkal,
c) 1, 2, 3, 4, 5, 6, 7 sorszámú változókkal.

Hipotézisünk az volt, hogy az osztályozáshoz adatrendszerünket nem kell transzformálni. Ezt egyrészt arra alapoztuk, hogy változóink hatékonyság típusú mutatószámok, másrészt a faktoranalízis eredményeként kapott faktorok közel azonos súlyúak voltak.

Vizsgálatunkban hierarchikus algoritmust alkalmaztunk. Alapelve az objektumok fokozatos egyesítése. Az algoritmus kezdetben minden objektumot egy-egy clusternek tekint, majd az egymáshoz „legközelebb állókat" egyesíti, ezután kiszámolja az új cluster „koordinátáit" és minden további lépésben a clusterek számát eggyel csökkenti, amíg végül minden eredeti megfigyelés egyetlen clusterbe kerül. Előnye, hogy a folyamatot, a csoportok homogenitásának mértékét és a clusterek számának alakulását lépésenként figyelemmel kísérhetjük. Hátránya a viszonylag nagy gépidőigény.

A clusterek számának meghatározásánál figyelembe kell venni, hogy azonos objektum szám mellett az egyes csoportokon belüli homogenitás monoton (nem egyenletesen) csökken a clusterek számának növelésével. Ezt mutatja be a 3. csoportosítási változatra a 3. sz. ábrán látható dendogram,[1] ahol



3. ábra

[1] Gépi úton közvetlenül rajzoltatható.

a vízszintes tengelyen a beruházások sorszáma (azonosítója) olvasható le, a függőleges tengelyen pedig az egyes kapcsolódási szintekhez tartozó homogenitást jelző szórásértékek szerepelnek. A dendogram szemléletes képet ad az adatrendszerben felismerhető csoportok számáról is.

A számítás eredményeként kapott csoportokat a 3 változatra a 2. táblázatban foglaljuk össze.

2. táblázat

Változatok	1. csoport	2. csoport	3. csoport
3 változó alapján	13, 14, 15, 16, 20, 27, 31	2, 3, 4, 18, 19, 21, 22, 23, 25, 26, 29, 30, 32	1, 5, 6, 7, 8, 9, 10, 11, 12, 17, 24, 28
5 változó alapján	13, 14, 15, 16, 18, 19, 20, 21, 27, 30, 32	2, 4, 22, 23, 25, 26, 29, 31	1, 3, 5, 6, 7, 8, 9, 10, 11, 12, 17, 24, 28
7 változó alapján	13, 14, 15, 18, 19, 20, 21, 27, 30, 31	6, 16, 22, 23, 24, 25, 26, 29, 32	1, 2, 3, 4, 5, 7, 8, 9, 10, 11, 12, 17, 28
A három változat közös része	13, 14, 15, 20, 27	22, 23, 25, 26, 29	1, 5, 7, 8, 9, 10, 11, 12, 17, 28

Itt csak néhány értékelő megállapítást teszünk, a részletes elemzéssel a rangsorolási módszer tárgyalása után foglalkozunk.

Az eredmények alapján megállapíthatjuk:

1. Alapvetően 3 csoport ismerhető fel mind a 3 változat esetében.
2. A csoportosítás elfogadható abban az értelemben, hogy a 3 változat eredményei között nincs ellentmondás (azt *természetesen nem várhatjuk*, hogy *azonos csoportokat* kapjunk).
3. A változók számának növelése elsősorban az 1-es és 2-es csoportok között okoz átrendeződést, a legstabilabb a 3-as csoport.
4. A dendogram alapján látható, hogy a szórás viszonylag magas, elég heterogének a csoportok.

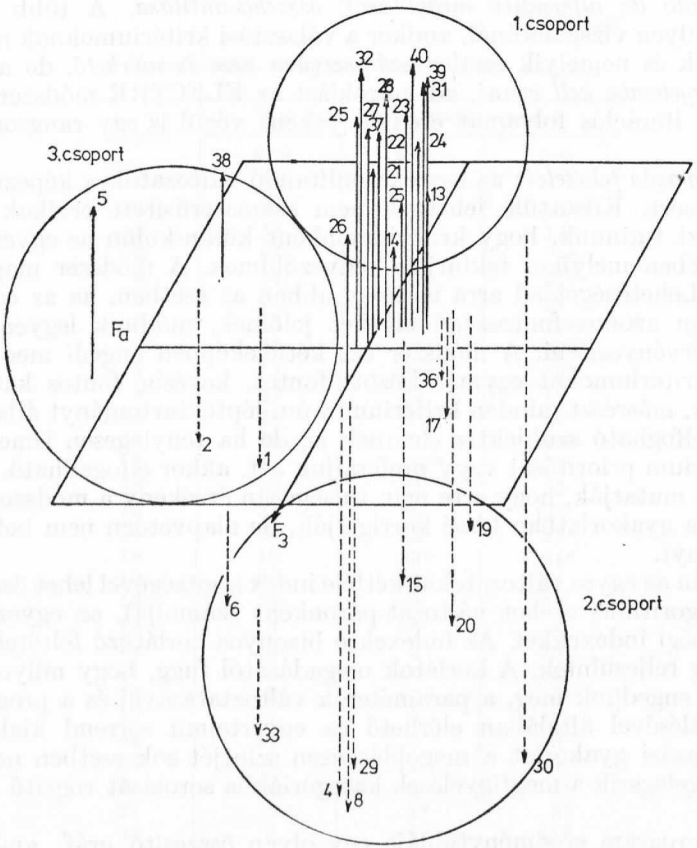
Értelemszerűen vetődik fel az egyes csoportok minősítési igényének gondolata.

A cluster technika erre nem, vagy csak korlátozott mértékben alkalmas, megfelelő módszer pl. az ELECTRE rangsorolási eljárás, melyet röviden ismertetni fogunk.

A csoportosítási eredmények egy érdekes hasznosítási lehetősége a következő. Tételezzük fel, hogy már minősítettük az egyes csoportokat jó – közepes – rossz ítélettel. Számoljuk ki az egyes csoportok súlypontját (a mi esetünkben ez egy 7 dimenziós vektor) és vagy ezt a pontot vagy a hozzá legközelebb álló beruházás adatait tekintsük az illető csoportot reprezentáló *standard beruházásnak*. Néhány adathalmazzal végzett vizsgálat alapján *a normatívák meghatározásához* kapunk *objektív segítséget*.

A szemléltetés érdekében átmenetileg eltekinthetünk az eredeti változóink által rögzített tértől. Megkísérelhetjük faktoranalízissel, maximum 3 faktorba tömöríteni mutatóink információtartalmának jelentős részét. A 3 faktor által

kifeszített „állapottérben" megfigyeléseink clusterei esetleg jól elkülönülő pontfelhőként szemléletesen jelenhetnek meg. A 4. ábrán egyik ilyen menet közben készült munka-ábránkat mutatjuk be. A vázolt gondolatmenet mind a strukturális összefüggések felismerésére, mind az előző eredmények egyfajta



4. ábra

ellenőrzésére igen jól felhasználható. (A mondottakat természetesen nem szabad mechanikusan hasznosítani. Lényeges kérdés, hogy a grafikusan még ábrázolható 3 faktor alkalmazása elegendő-e? Egyébként zavarok származhatnak az eredeti változók tere és a faktortér kapcsolatából is. A két rendszerből — tapasztalataink szerint elég tág határok között adódó — *egybevágó* következtetések viszont jól megerősíthetik elért alakfelismerési eredményeinket.).

5. Az ELECTRE többváltozós rangsorolási módszer alkalmazása

Az ELECTRE[2] a többváltozós döntéselemzés egy rangsoroló módszere. Segítségével *több különböző döntési kritérium egyidejű figyelembe vétele mellett megvalósítható az alternatív megoldások összehasonlítása.* A több dimenziós problémák ilyen vizsgálatánál, amikor a választási kritériumoknak nincs közös mérőszámuk és némelyik esetleg *számszerűen nem is mérhető*, de mégis *valamennyit figyelembe kell venni*, ad megoldást az ELECTRE módszer. Az algoritmus egy iterációs folyamat eredményeként végül is egy rangsort határoz meg.

Az alkalmazás feltételei: az összehasonlítandó változatokra képezni kell egy kritérium sort. Közöttük lehetnek nem számszerűsített értékek is, ekkor elegendő azt tudnunk, hogy kritériumonként külön-külön az egyes változatpárok esetében melyiket tekintjük kedvezőbbnek. A módszer maga nagyon rugalmas. Lehetőséget ad arra is, hogy abban az esetben, ha az egyes kritériumok nem azonos fontosságú elveket jelölnek, módunk legyen súlyozási rendszert érvényesíteni. A módszer ezt kétféleképpen engedi meg. Egyrészt az egyes kritériumokat egymás között fontos, kevésbé fontos kategóriákba sorolhatjuk, másrészt minden kritériumra ún. léptéktartományt állapíthatunk meg. Ez felfogható szubjektív elemnek is, de ha ténylegesen ismerjük valamely kritérium prioritását vagy preferáljuk azt, akkor elfogadható. Tapasztalataink azt mutatják, hogy erre nem túlságosan érzékeny a módszer; a súlyozási elvek a gyakorlatban kissé korrigálják, de alapvetően nem befolyásolják az eredményt.

Ezek után az egyes változatokat kétféle index segítségével lehet összehasonlítani (az algoritmus ezeket változat-páronként számolja), az egyezőségi és a különbözőségi indexekkel. Az indexekre bizonyos korlátozó feltételeknek kell egyidejűleg teljesülniök. A korlátok megadásától függ, hogy milyen mértékű szigorítást engedünk meg, a paraméterek változtatásával és a program többszöri ismétlésével általában elérhető az egyértelmű sorrend kialakítása is. Az alkalmazási gyakorlat a megoldás ezen szintjét sok esetben nem igényli, hanem megelégszik a megfigyelések kategóriákba sorolását rögzítő eredménnyel.

A gépi program eredménytáblája egy olyan összesítő gráf, amelyből felírható a rangsor. A gráf egyben szemléletes képet is ad az értékelésről. Csúcspontjaiban az egyes alternatívák szerepelnek, a gráf irányítása pedig olyan, hogy az a *legkedvezőbb változat, amelyikbe a legtöbb nyíl mutat.*

A vizsgálatainkban szereplő beruházási javaslatok esetében is meghatároztuk a rangsort ahhoz a 3 mutatókombinációhoz, amelyekre a csoportosítást is elvégeztük.

Az eredményt a 3. táblázatban foglaljuk össze. Az egymást követő csillaggal jelölt beruházások a rangsorban azonos helyen állnak. Pl. a 7 db mutató alapján történő rangsorolás esetében a 13, 14, 20, 21 sorszámú beruházások, amelyek a táblázatban az 1, 2, 3, 4-ik helyen szerepelnek valójában azonos rangot jelentenek, a további finomítást e szempontból nem tartottuk jelentősnek.

[2] A módszert Franciaországban dolgozták ki, elnevezése az „Elimination et Choix Trandisant le Realité" kezdőbetűből képzett rövidítés.

Az eredmények alapján látható, hogy a 3 mutató alapján történő rangsorolás elég jelentősen eltér az 5 és 7 mutató alapján kapott eredményektől. Az utóbbiak szerinti rangsorok már jó egyezést mutatnak, említésre méltó különbség csak a 6-os sorszámú beruházásnál áll fenn. (Az ábra az áttekinthetőség

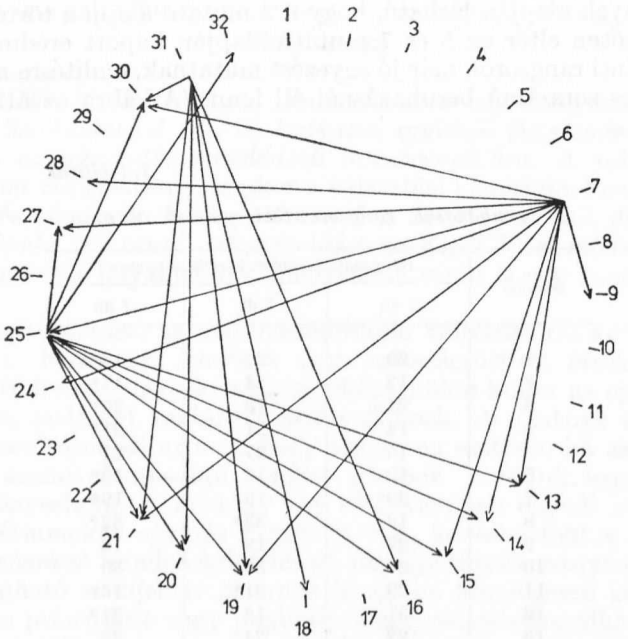
3. táblázat

A beruházások rangsora több változó alapján

Sorszám	Hány mutató alapján készült a rangsor		
	3 db	5 db	7 db
1	20	20	13*
2	13	14	14*
3	21	13	20*
4	14	21	21*
5	15	27	27
6	19	15	15*
7	30	19*	19*
8	18	32*	32*
9	27	18	18
10	25	30*	30
11	29	31*	16*
12	31	16	31*
13	22	24	22
14	28	25	25
15	24	22*	23*
16	16	23*	24*
17	26	28*	28
18	32	29	29
19	2	26	6
20	3	2*	26
21	23	4*	2
22	9	17	4
23	7	9	17
24	17	7	9
25	1	11*	7
26	4	3*	1*
27	5	1*	3*
28	6	8	11*
29	8	6	8
30	10	12	5
31	11	5	12
32	12	10	10

kedvéért csak a 7-es, 25-ös és 31-es sorszámú beruházásokról mutatja meg, hogy mely beruházásoknál minősülnek rosszabbnak.)

Az 5. ábrán bemutatjuk a 7 mutató alapján kapott rangsorhoz tartozó gráfot. A csúcspontokban a beruházás sorszáma szerepel, a nyíl a kedvezőbb irányba mutat, az *ábráról* tehát *leolvasható*, hogy valamely beruházás *melyik* és *hány másik beruházásnál kedvezőbb*.



5. ábra

6. Összehasonlító értékelés, következtetések

A különböző módszerekkel végzett számítási eredményeket a 4. táblázatban foglaljuk össze. A jelzőszámok mindenütt a beruházás sorszámára utalnak, az 5 és 7 mutató alapján képzett rangsorokban az egymás alatt szereplő számok (beruházások) azonos helyen állnak.

1. Megállapíthatjuk, hogy a *bank három kritériuma* szerinti feldolgozás és ugyanezen 3 mutatóra végzett *csoportosítási eredmények között ellentmondás nincs.* Észre kell venni, hogy a megadott normatívák viszonylag *alacsony „mércét"* jelentenek, mert több beruházási javaslat fogadható el (13 db) és csak egy utasítandó el ezek alapján. Az *együttes figyelembe vétel lényegesen szigorúbb, 7 beruházást fogad el és 12-t elutasít.*
2. A cluster analízis minden mutató kombinációra alapvetően *3 csoportot eredményez,* ez azt jelenti, hogy az adatokban 3 relatíve homogén „sűrűsödési" tartomány különíthető el.
3. A többváltozós rangsorolás eredményei alapján a kapott *csoportok* minősíthetők, 3 csoport esetén kézenfekvő a *„jó"* – *„közepes"* – *„rossz"* minősítés.
4. A csoportosítás jelentősége, hogy a viszonylag *homogén csoportok elemei egységesebben kedvező vagy kedvezőtlen eredményt nyújtanak hatékonyság szempontjából.* Ez a felismerés a döntéshozó számára bizonyos *rugalmasságot*

Összesítő táblázat

A bank három kritériuma alapján (nem cluster felfogásban)

2, 13, 15, 16, 18, 19 20, 21, 25, 29, 30, 31, 32	1, 3, 4, 6, 7, 8, 9, 10, 11, 12, 14, 17, 22, 23, 24, 26, 27, 28	5
elfogadási halmaz (13 db)	nincs egyértelmű döntési lehetőség (18 db)	elutasítási halmaz

3 mutató alapján:

rangsor: *20, 13, 21, 14, 15, 19, 16, 18, 27, 25, 29, 31, 22, 28, 24, 30, 26, 32, 2, 3, 23, 9, 7, 17, 1, 4, 5, 6, 8, 10, 11, 12*

13, 14, 15, 16, 20, 27, 31	2, 3, 4, 18, 19, 21, 22 23, 25, 26, 29, 30, 32	1, 5, 6, 7, 8, 9 10, 11, 12, 17, 24, 28
1. csoport (7 db) „jó"	2. csoport (13 db) „közepes"	3. csoport (12 db) „rossz"

5 mutató alapján

rangsor: *20, 14, 13, 21, 27, 15, 19, 18, 30, 16, 24, 25, 22, 26, 2, 17, 7, 9, 11, 6, 8, 12, 5, 10*
32, 31, 23, 4, 3
28, 1
29,

13, 14, 15, 16, 18, 19, 20, 21, 27, 30, 32	2, 4, 22, 23, 25, 26, 29, 31	1, 3, 5, 6, 7, 8, 9, 10 11, 12, 17, 24, 28
1. csoport (11 db) „jó"	2. csop. (8 db) „közepes"	3. csop. (13 db) „rossz"

7 mutató alapján

rangsor: *13, 27, 15, 18, 30, 16, 22, 25, 23, 28, 29, 6, 26, 2, 4, 17, 9, 7, 1, 8, 5, 12, 10*
14, 19, 31, 24, 3,
20, 32, 11,
21,

13, 14, 15, 18, 19, 20, 21 27, 30, 31, 32	6, 16, 22, 23, 24, 25, 26 29	1, 2, 3, 4, 5, 7, 8, 9 10, 11, 12, 17, 28
1. csop. (11 db) „jó"	2. csop. (8 db) „közepes"	3. csop. (13 db) „rossz"

is jelent. Pl. a „jó" csoporton belül, ha az egyes beruházások különböző célt valósítanak meg, akkor lényegében azonos hatékonysági eredmények mellett valamelyik beruházási célt preferálhatja.

5. Az ELECTRE módszer egy jellemző tulajdonsága, hogy nem garantálja, hogy a sorrendben az i és $i + 1$-ik helyen álló beruházások eltérésének mértéke megegyezik a k és $k + 1$-ik helyen állók eltérésével. Ez adta az ötletet a csoportosítási és rangsorolási eredmények összekapcsolásához. Jogosan feltételezhető, hogy ahol a minősített szomszédos csoportok különválnak, ott a rangsorban ugrás, nagyobb eltérés van. Ilyen elgondolás alapján a mi esetünkben a 3 csoportnak megfelelően a rangsor 3 szeletre vágható, és az egyes szeletek megfeleltethetők egy-egy csoportnak.

Ez egy további finomítási lehetőséget is ad *a csoporton belüli rangsorolásra* és felhasználható annak eldöntésére is, hogy a rendelkezésre álló pénzeszközök erejéig milyen sorrendben fogadjuk el a beruházási javaslatokat.

(Beérkezett: 1977. július 18-án.)

Függelék: Az alapadatokból képzett hatékonyság típusú mutatószámok: 1. Többlettermelés/Beruházási költség, 2. Tőkés export/Beruházási költség 3. Többlettermelés/Építés, 4. Tőkés export/Építés, 5. Nettó devizahozam/Építés, 6. Többlettermelés/Tőkés import gép, 7. Tőkés export/Tőkés import gép, 8. Többlettermelés/Hitel, 9. Tőkés export/Hitel, 10. Nettó devizahozam/Hitel, 11. Többlettermelés/Deviza ráfordítás, 12. Tőkés export/Deviza ráfordítás, 13. Beruházás megtérülése a nettó devizahozamból, 14. Tőkés imp. gép megtérülése a nettó devizahozamból, 15. 100 Ft eszközre jutó nyereség, 16. Deviza kitermelési mutató.

GROUPING AND RANKING OF INVESTMENT PROPOSALS

The paper treats the problem in two major steps. Firstly, the grouping of investment proposals aimed at the enlargement of the supply of export goods on convertible currency markets is examined. The problems of choosing and homogenizing the indicators, eliminating their interconnectedness and fist of all the determination of their number, respectively, are raised. Answers to these questions are sought for by economic considerations, the transformation of indicators, reiterated factor analysis and finally to the problem of grouping by cluster analysis.

In the second step an attempt is made to evaluate the emerging clusters by a briefly outlined procedure called "ELECTRE". In the course of these examinations both the individuals and the clusters are ranked, whereby and ordering within the clusters can also be made.

Another point of interest of the examinations is that the questions are answered on the basis of criteria drawn directly from the practice, and then, relying on the methodology outlined, by the application of three, five and seven variables, respectively (together with ranking). The results are compared and evaluated. Finally, a proposal is submitted on the judicious extension of our train of thoughts to a wider sphere.

ГРУППИРОВКА И УПОРЯДОЧЕНИЕ ПРЕДЛОЖЕНИЙ ПО КАПИТАЛЬНЫМ ВЛОЖЕНИЯМ

В данной работе выдвигаемая проблема рассматривается по двум основным этапам. Во-первых, рассматриваются возможности группировки предложений по капитальным вложениям ,расширяющим товарные фонды по конвертированному экспорту. Возникает

проблема выбора показателей, их однородности, отсева их связей друг с другом и, главным образом, их количества. На эти вопросы ответ ищется на основе экономических соображений, трансформации показателей, повторных обследований с помощью факторного анализа и, в заключении, группировка с помощую кластерного анализа.

На втором этапа в отношении оценки сложившихся кластеров предпринимается попытка применить кратко изложенный метод «Electre». В результате этих исследований происходит группировка как индивидуумов, так и кластеров и, таким образом, становится возможым установление очередности и в рамках самого кластера.

Проведенные исследования являются интересными еще и потому, что на выдвигаемые вопросы ответ дается на основании критериев непосредственной практики и, в последующем, на основании указанной методики используются три, пять и, далее, семь переменных (с указанием очередности) и получаемые результаты сравниваются, оцениваются. В заключении вносится предложение и в отношении более широкого распространения приводимых соображений и использования их в соответствии с заложенной в них идеей.

Chikán Attila

Vállalati vélemények a tartalékolási magatartásról

A „Szocialista vállalat" kutatási főirány keretében 1975-ben kutatásokat kezdtünk az erőforrástartalékoknak a vállalati gazdálkodásban betöltött szerepével, a vállalati tartalékolási magatartással kapcsolatban. A kutatás elvi alapjait összefoglaló tanulmányunkban a tartalékok fogalmát úgy határoztuk meg, mint a vállalatnál adott időpontban jelenlevő erőforrások kihasználatlan termelőképességét. Az ily módon értelmezett tartalékok jellemző formáinak a kapacitástartalékot, a munkaerőtartalékot és a készleteket tekintettük. Kimondottuk, hogy ezen tartalékok jelenléte a vállalat rugalmas gazdálkodásának, a vállalat, mint rendszer adaptív magatartásának szükséges feltétele. Megfogalmaztuk ezen erőforrástartalékok néhány olyan fő vonását, amelyek az elméleti vizsgálatok alapján jelenlegi hazai gazdasági, gazdaságpolitikai helyzetünkben jellemzőnek tekinthetők.

Az elméleti kutatással párhuzamosan empirikus felmérést készítettünk vállalataink tartalékolási magatartásának vizsgálatára. A felmérés elvi alapját a fent jelzett tanulmányban kifejtett gondolatok adták. A vizsgálat kikérdezéses kérdőíves felmérés formájában zajlott le, 34 vállalat 134 vezetőjét kérdeztük meg. A vállalatok (amelyek közül 15 gépipari, 7 vegyipari, 4 kohászati, 4 textilipari, 4 egyéb könnyűipari) a legfontosabb ismérveket tekintve megfelelően reprezentálják a magyar ipar egészét. Az egyes vállalatoknál négy-négy vezetőt kérdeztünk meg: a közgazdasági, az anyagellátási, a munkaerőgazdálkodási és a termelésirányítási terület vezetőit – összhangban a tartalékok fentiekben jelzett típusaival.

A kérdőív konstrukciója olyan volt, hogy a kérdések egy részét, amelyek a vállalati tartalékolási magatartás általános vonásaival foglalkoztak, valamennyi megkérdezettnek feltettük, míg az egyes szakterületeket (készletgazdálkodás, munkaerőgazdálkodás, termelésirányítás) érintő speciális kérdéseket mindig a szakterület vezetője és a közgazdasági terület vezetője válaszolta meg. (A „vezető" konkrét beosztása természetesen vállalatonként eltérő volt: osztályvezetőtől vezérigazgatóhelyettesig.)

Jelen cikkünk keretében természetesen nincs mód a felmérés valamennyi részeredményének ismertetésére. (A kérdőíven szereplő 35 kérdéscsoportból mindössze négyet elemzünk e cikkben – persze ezek a legfontosabbak közül valók.) Amit ehelyütt kiemelünk: a felmérés során kapott vélemények egy

[1] Chikán A.: Tartalékok a vállalati rendszerben, MKKE soksz. 1976. Rövidítve megjelent a Vezetéstudomány c. folyóiratban (1977/5) „A vállalati erőforrástartalékok néhány kérdése" címmel.

jól körülhatárolható részhalmaza matematikai és közgazdasági jellemzőit tekintve alkalmas a sokváltozós statisztikai módszereknek a felhasználására a feldolgozásban. Kétségtelen, hogy az ilyen, véleménykérésen alapuló adatok elemzése nem tekinthető ezen módszerek „klasszikus" alkalmazási területének, s ez különös óvatosságra int az eredmények elemzésében. Úgy véljük azonban, hogy az itt ismertetendő eredmények is hozzájárulhatnak annak bizonyításához, hogy nem kell, sőt nem szabad megállni a hasonló jellegű közgazdasági, szociológiai és más társadalomtudományi vizsgálatokban sem azon az elemi statisztikai feldolgozási szinten, ami sajnos még ma is sok esetben jellemzőnek tekinthető.[2]

A sokváltozós statisztikai vizsgálattal nyert eredmények esetünkben sem hoztak forradalmian újszerűt, de jónéhány másként meg nem szerezhető információhoz juttattak és megnyugtatóan alátámasztják a minta strukturálására tett kísérleteinket. Érdekes és fontos tény továbbá az, hogy a statisztikai elemzés több lényeges ponton értékelést ad magára a közgazdasági kérdésfeltevésre, mintegy visszacsatolásszerűen utal arra: jól ragadtuk-e meg az egyes kérdéseket, s egymáshoz való kapcsolatukat. Ez ugyan az adott felmérés szempontjából ex post elemzés, s bár kétségkívül hasznosítható, nagyobb jelentősége mégis a további hasonló elemzések, felmérések előkészítése szempontjából lehet.

Az alábbiakban tömören ismertetjük az elemzés néhány olyan eredményét, amelyek megítélésünk szerint reprezentálják a sokváltozós statisztikai módszerek alkalmazásával hasonló felmérések esetén nyerhető eredményeket.

Az egész felmérés egyik legfontosabb kérdésköre a vállalati tartalékolási magatartásra ható külső és belső tényezők vizsgálata volt. Az elemzés célját a kérdőív több kérdéscsoportja egymással összefüggésben szolgálja.

Ezek közül faktor- és clusterelemzéssel az alábbi két kérdéscsoportot dolgoztuk fel:

1. A vállalati tartalékolási magatartásra a mai hazai viszonyok között általánosan ható tényezők.
2. Az egyes erőforrások (kapacitás, munkaerő, készlet) tartalékolására vezető okok a konkrét vállalati szituációban.

Az első csoportba tartozó kérdéseket valamennyi megkérdezettnek feltettük, míg a második csoport kérdéseinél mindig az adott szakterület vezetője és a közgazdasági vezető válaszolt (ez a minta tehát fele akkora). Valamennyi kérdésnél arra kértük az illetékes vezetőket, hogy 0-tól 9-ig adjanak súlyt az egyes felsorolt tényezőknek: 0-t akkor adjon, ha véleménye szerint az adott tényező semmilyen hatást nem gyakorol a tartalékolásra, 9-et pedig akkor, ha a tényezőt meghatározó erejűnek tartja.

[2] Itt említjük meg, hogy a felmérés, illetve feldolgozás csoportos munka eredménye. A kérdőív kialakításában, s a kérdezés lebonyolításában az MKKE Rajk László Szakkollégiumának számos akkori hallgatója működött közre. A feldolgozási szempontok összeállításában *Fazekas Károly*, a Közgazdaságtudományi Intézet munkatársa, a számítógépes futtatásokkal kapcsolatos módszertani kérdésekben *Füstös László*, a Szociológiai Intézet munkatársa, a matematikai-statisztikai elemzésekben *Meszéna György*, az MKKE Matematikai és Számítástudományi Intézetének osztályvezetője volt segítségünkre. Maga a feldolgozás egyébként az MTA SZTAKI CDC gépén történt.

Az alábbiakban röviden ismertetjük a faktor- és clusterelemzés útján nyert eredményeket.[3] A módszertani kérdésekkel részletesen nem foglalkozunk, mivel munkánk e területen nem ment túl a már ismert módszerek alkalmazásán. Jelen cikkel célunk az alkalmazás lehetőségeinek illusztrálása.

1. A vállalatok tartalékolási magatartására ható tényezők vizsgálata

A vizsgálatot a kérdőív azon kérdéscsoportjának alapján végeztük, amelybe azok a kérdések tartoztak, amelyek a mai hazai feltételek mellett általában ható tényezők súlyára vonatkoztak. Az adott kérdéscsoportban az előzetes elvi megfontolások alapján tizenhét tényezőt soroltunk fel. Ezen tényezőket – a vélemények megoszlásának átlagával és szórásával – a következő táblázat tartalmazza:

1. táblázat

A vállalati tartalékok képzésére, illetve kialakulására vezető okok

Sorszám	Megnevezés	Átlag	Szórás
1.	A termékek iránti szükséglet ismeretlen ingadozásai	3,64	2,52
2.	Beszerzési nehézségek	5,02	2,47
3.	Pénzügyi okok	3,20	2,27
4.	A központi szabályozás módosulásai	3,60	2,41
5.	Munkaerőhelyzet	5,16	2,79
6.	A bérszabályozás rendje	4,00	2,70
7.	Finanszírozási, hitelezési rendszer	4,05	2,46
8.	A felügyeleti hatóságok elvárásai, utasításai	3,02	2,03
9.	Egyéb külső okok	2,46	3,01
10.	Irányítási, vezetési hiányosságok	3,63	2,18
11.	Manőverezési lehetőség biztosítása	3,41	2,43
12.	Vállalatpolitikai célkitűzések	4,22	2,32
13.	Törekvés a termelés folyamatosságának biztosítására	5,19	2,37
14.	Vállalaton belüli termelési kooperáció	3,67	2,34
15.	Végrehajtási fegyelmezetlenség	4,05	2,30
16.	Törekvés a vállalatgazdálkodás részterületeinek rugalmas összehangolására	3,93	2,27
17.	Egyéb belső okok	1,00	2,12

A táblázatból látható, hogy az átlagok alapján a válaszadók a tartalékolási politikára ható legfőbb tényezőknek

– a termelés folyamatosságára való törekvést,
– a munkaerőhelyzetet és
– a beszerzési nehézségeket

[3] Hasonló típusú elemzéseket végeztünk egyrészt a kérdőív más kérdésköreire, másrészt az itt feldolgozott kérdéseknél a válaszadók strukturálására. Cikkünkben a tematikai egységesség kedvéért azonban csak a jelzett, véleményünk szerint legérdekesebb kérdéskörben nyert elemzésekkel foglalkozunk. Egyéb eredményeinket a kérdőív teljes feldolgozásával együtt később publikáljuk.

tekintették, míg a legkisebb hatásúnak – az egyéb okoktól eltekintve – a felügyeleti hatóságok elvárásait, a pénzügyi okokat és a manőverezési lehetőség biztosítását tartották.

A sokváltozós statisztikai módszerek alkalmazása a vizsgált tényezőket közgazdaságilag jól értékelhető módon strukturálta. A következő vizsgálatokat végeztük el:

- faktoranalízis és cluster analízis valamennyi válaszadóra, valamennyi tényező figyelembevételével,
- faktoranalízis és cluster analízis valamennyi válaszadóra, különválasztva a tartalékolásra ható belső és külső tényezőket,
- faktoranalízis valamennyi tényező figyelembevételével, a válaszadókat szakterületenként szétválasztva.

A különböző gépi futtatások után az elemzéshez ténylegesen hét faktoranalízis és három clusteranalízis eredményeit használtuk fel. Az alábbiakban az elemzések során nyert fő eredményeket foglaljuk össze.

A tizenhét tényezőt és valamennyi válaszadót felölelő faktor- és cluster elemzés a tényezőket igen jól strukturálta. A kétféle elemzés egymást e szempontból alátámasztotta.

A faktoranalízis eredményeként a tizenhét tényezőt öt csoportba soroltuk. A besorolást a mellékelt 2. táblázat mutatja, ahol a táblázat belsejében a rotált faktor-mátrix elemei állnak, az egyes faktorokban döntő súlyt képviselő tényezőkre vonatkozóan.

A táblázatból látható, hogy a változók meglehetősen jól beilleszkednek az öt faktorba. (A táblázatban valamennyi 0,4-nél magasabb korrelációs együtthatót feltüntettük.) A két pénzügyi tartalmú változó (3. és 7.), valamint a manő-

2. táblázat

A tartalékalakulás tényezőit sűrítő faktorok

Faktorok	I.	II.	III.	IV.	V.
Tényezők	A faktorok által képviselt információ a teljes információ %-ában				
	0,30304	0,40280	0,47744	0,54918	0.60906
1.					—0,69567
2.			0,61714		
3.		—0,47632	0,44059		
4.		—0,60381			
5.		—0,64157			
6.		—0,75786			
7.		—0,50459			—0,52351
8.	0,53376				
9.			0,82587		
10.				0,79996	
11.				0,43538	—0,48207
12.					—0,60419
13.	0,68933				
14.	0,74431				
15.				0,79679	
16.	0,81537				
17.					

verezési lehetőséget okként megjelölő 11. változó szerepel viszonylag alacsonyabb korrelációs kapcsolattal két-két faktorban, míg a 17. (egyéb belső okok) nem került elég szoros kapcsolatba egyetlen faktorral sem. (A legmagasabb korrelációs szinttel – 0,39 – a második faktorhoz kapcsolódik, ez nem értelmezhető.)

Még többet mond azonban, hogy az egyes faktoroknak minden különösebb erőszakoltság nélkül saját közgazdasági tartalmat tulajdoníthatunk.

Az első faktor a vállalaton belüli kapcsolatrendszer, elvárások három változóját tartalmazza: a részterületek rugalmas összekapcsolására irányuló törekvést, a belső kooperációt, valamint a termelés folyamatosságának biztosítását. Ugyanezen faktorba került, de jóval alacsonyabb korrelációs szinten a 8. tényező (felügyeleti hatóságok elvárásai) – ami azt tanúsíthatja, hogy akik a vállalaton belüli tényezőknek nagy súlyt adtak, azok hajlamosak arra, hogy kívülről utasítást várjanak, illetve attól tartsanak – legalábbis az átlagnál erősebben.

A második faktor döntően a szabályozórendszer által közvetített külső tényezőket tartalmazza. Az egyes változók korrelációs együtthatói alacsonyabbak, mint az első faktor esetében. Ezt a faktort a kifelé való alkalmazkodás faktorának nevezhetjük, s döntően a szabályozásról, konkrétan a munkaerő- és bérszabályozásról, illetve a pénzügyi szabályozásról kialakult vélemény befolyásolja.

A harmadik faktor a szívásos piaci helyzet hatásait tükrözi. Az egyéb külső okok között elsősorban ezt jelölték meg a válaszadók, s a beszerzési nehézségek is erre utalnak. Feltételezhető, hogy a pénzügyi okok megjelenése összhangban áll a készlettartásra korlátozólag ható forgóeszköz finanszírozás kérdésével.

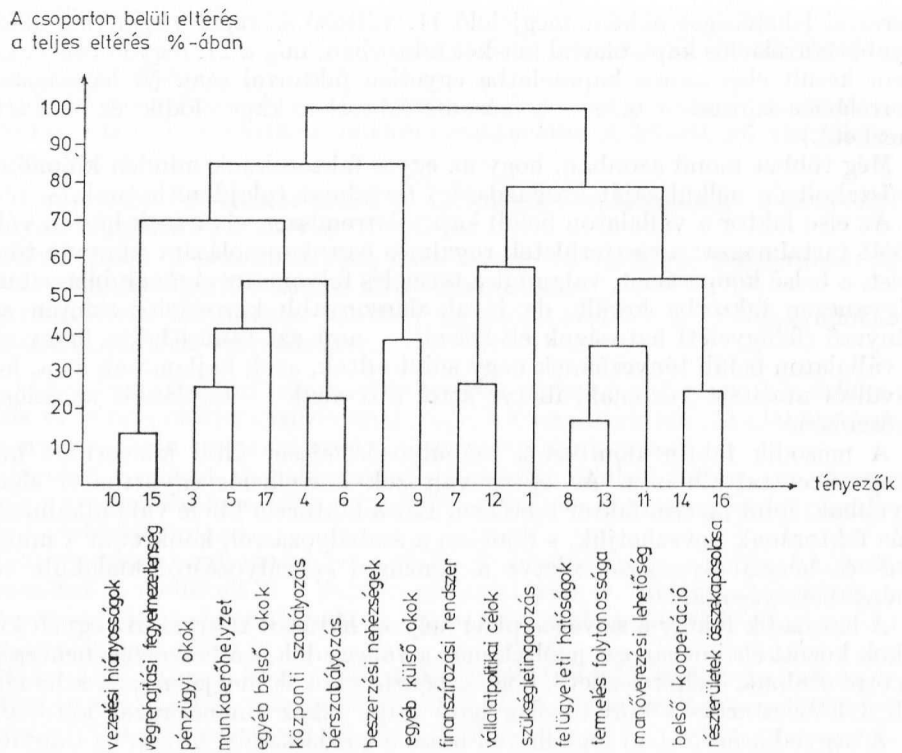
A negyedik faktorban foglalhatók össze a gazdálkodási hibák: az irányítási-vezetési hiányosságok és a végrehajtási fegyelmezetlenség. Az, hogy a manőverezési lehetőség is szerepel itt (igaz, hogy lényegesen alacsonyabb korrelációs együtthatóval), alighanem értelmezési probléma következménye.

Végül az ötödik faktorba kerültek – sajátos és elgondolkodtató módon – a vállalati tartalékolási magatartás tudatosságának kifejezésére hivatott tényezők: a szükséglet ingadozásához való alkalmazkodás, a vállalatpolitikai célkitűzések, a piaci manőverezés lehetőségeinek kihasználására való törekvés. A finanszírozási-hitelezési rendszer itt is megjelenik, ami tartalmi okokkal nem magyarázható.

Az egyes faktorokhoz tartozó sajátértékeknek a százalékos elemzése azt mutatja, hogy a belső okok meghatározó jellegűek (az első faktor az összes eltérések 30,3%-át magyarázza), míg a további faktorok nagyjából azonos súlyúak, s egyenként nem jelentősek (rendre: 9,97; 7,45; 7,17; 5,99%-ot képviselnek). A kérdéskör bonyolultságát tekintve az öt faktor által adott 60,9%-os összsúly kielégítőnek fogadható el.

A változók és megfigyelések azonos halmazára cluster analízist is végeztünk. A hierarchikus clusterezéshez felhasznált távolságfogalom az egyes változók faktorokban betöltött súlyának összehasonlításán alapult, az eredmények tehát alkalmasak a faktorok vizsgálatánál tett megállapítások ellenőrzésére. Tekintettel arra, hogy a faktoranalízis eredménye – mint láttuk – az értelmezés szempontjából kedvező volt, ezt az ellenőrzést igen fontosnak tartjuk a közgazdasági következtetések levonása szempontjából.

A cluster analízis eredményét az 1. ábra szemlélteti. Mint az ábrából látható, a nyert clusterek szoros rokonságot mutatnak a korábban meghatározott



1. ábra. A tartalékolási magatartásra ható tényezők hierarchikus clusterezése

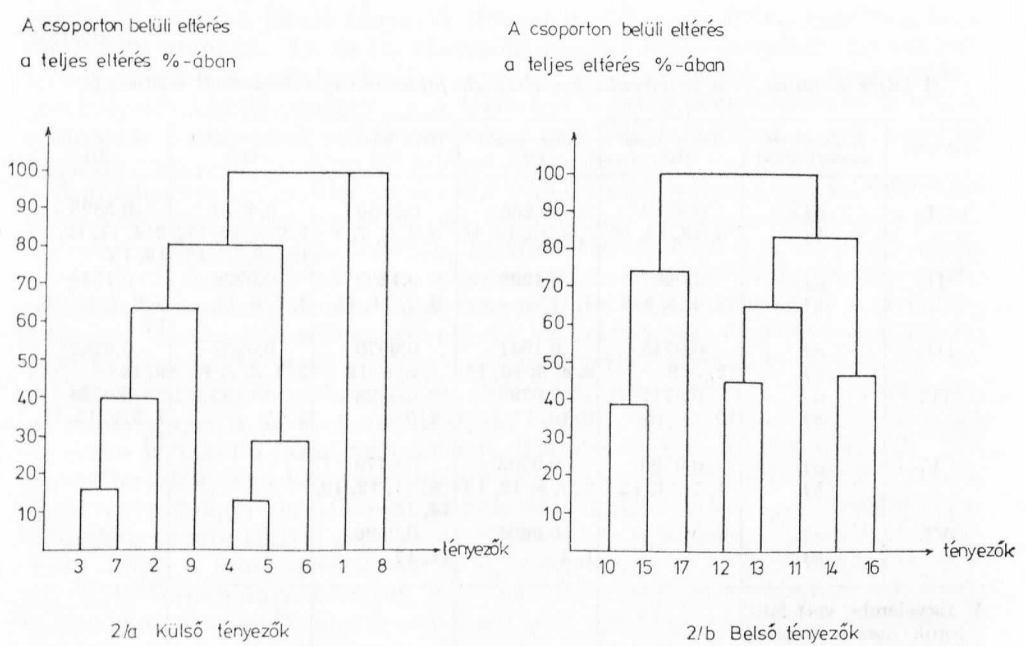
faktorokkal. Az első clusterbe – igen alacsony összekapcsolódási szinten, s más clusterektől karakterisztikusan elválva – a negyedik faktor két fő tényezője, a vezetési hiányosságok és a végrehajtási fegyelmezetlenség került. A második cluster a szabályozó rendszer tényezőit tartalmazza (a beékelődött „egyéb belső okok" szerepe itt nem magyarázható), ez lényegében a második faktor. A harmadik cluster a harmadik faktor két fő tényezőjét tartalmazza, s az ellátási nehézségekre utal. A negyedik cluster az ötödik faktor tényezőire épül, míg – viszonylag szorosabb kapcsolódással – az ötödik cluster megfeleltethető az első faktornak, amely a vállalati belső kapcsolódások rugalmasságára való törekvést fejezi ki.

A tényezőknek a faktoranalízissel és cluster analízissel végzett csoportosításában három tényező foglal el eltérő helyet: a 3, 7, és 11, tehát a pénzügyi okok, a finanszírozási rendszer és a manőverezési lehetőségek. Ennek magyarázata (a statisztikai értékelés törvényszerű esetlegessége mellett) véleményünk szerint ott keresendő, hogy – miként erre más elemzések is utalnak – a kérdőívben ezeknek a tényezőknek a definiálása sikerült talán legkevésbé. A harmadik és hetedik változó (pénzügyi okok – finanszírozási, hitelezési rendszer) egymáshoz való viszonyát és a 11. változónak (manőverezési lehetőség biztosítása) a vállalatpolitikai célkitűzésekkel való kapcsolatát nem definiáltuk sikeresen – azaz, a tényezőkről saját fejünkben kialakított képet nem sikerült a válaszadók felé közvetíteni.

Ez a negatív eredmény egyébként, véleményünk szerint általánosítható módon, a sokváltozós statisztikai módszerek alkalmazásának egy fontos „melléktermé ke". Az összefüggések keresése közben ezek a módszerek törvényszerűen rávezetnek a kérdésfeltevésben rejlő átfedésekre, inkonzisztenciára. Ennek a ténynek a társadalomtudományokban való alkalmazások esetén különös jelentősége van, hiszen itt a kérdések ritkán definiálhatók teljes egzaktsággal.

Az eredmények ellenőrzésére még három további clusterezést végeztünk. Egyszer elhagytuk a páronkénti korrelációkat tekintve feltűnően független első változót. A tényezők csoportosítása lényegében nem változott, ami arra az érdekes következtetésre vezet, hogy a szükségletingadozásra való felkészülést a vállalatok nem tartják lényeges, az egyéb okokkal összefüggő tényezőnek a tartalékolási magatartás szempontjából. Ez a megállapítás összhangban van a kérdőív más feldolgozási szempontjai alapján nyert eredményekkel csakúgy, mint a közvetlen gyakorlati tapasztalatokkal, és az eladói piac létével hozható kapcsolatba.

Elvégeztük továbbá a clusterezést a tartalékolásra ható külső és belső tényezők szétválasztásával, azaz külön a külső és külön a belső tényezőkre. Az eredmények — amelyeket a 2. ábra illusztrál — alátámasztják a minta egészére tett megállapításainkat; itt ismét lényegében azok a tényezőkombinációk jelennek meg, mint a teljes mintában. A külső tényezőket illetően feltűnő az 1. és 8. változó teljes különválása valamennyi többi tényezőtől. Mint már említettük, ezek a változók a teljes mintában is labilisan viselkedtek. A belső tényezők is hasonló csoportokba álltak össze, mint az eredeti vizsgálatban:



2. ábra. A tartalékolási magatartásra ható tényezők hierarchikus clusterezése, külön a külső és belső tényezőkre (A tényezők sorszámozása azonos az 1. ábrával)

ha magas kapcsolódási szinten is, de az eredeti első és negyedik faktor elemei szétválaszthatók. A két részre bontott mintára is elvégeztük egyébként a faktoranalízist, amelynek eredményei ismét igazolták, hogy a változók ily módon történő strukturálódása lényegében stabil.

Következő vizsgálatkörünkben az eredeti 134 megfigyelést négyfelé bontottuk. Arra voltunk kíváncsiak, hogy a már említett négy szakterület vezetőinek véleménye mennyiben konzisztens egymás között. Ebben a körben cluster analízist nem végeztünk, hiszen egyrészt a 34 megfigyelés a 17 dimenziós térben aligha adhat értelmezhető konfigurációt, másrészt korábbi elemzéseink azt a feltevést támasztották alá, hogy a faktorok és a clusterek egymással jól azonosíthatóak. Így e kérdéskör vizsgálatára négy faktoranalízist végeztünk, amelynek eredményeit (a teljes mintára nyert eredményekkel együtt) a 3. táblázat tartalmazza. A táblázatban azok a faktorok szerepelnek, amelyekhez tartozó sajátértékek legalább 1,0000 nagyságúak, a táblázat belsejében pedig az a) sorban az egyes faktorok által magyarázott eltérések találhatók az összes eltérés százalékában, a b) sorban pedig azok a tényezők vannak, amelyeknek korrelációs együtthatója az adott faktorra vonatkozólag legalább 0,4.

A táblázatból az alábbi főbb következtetések adódnak:

Az egyes szakterületeken nyert faktorok egymástól is, az egész mintától is eltérő struktúrát mutatnak. Ez azt jelenti, hogy a tartalékalakulásra ható tényezőket az egyes szakterületek vezetői egymástól eltérő módon ítélik meg. Az egész mintáról az áttekinthető kép csak a vélemények együtteséből alakul ki. Valamennyi szakterületre igaz ugyanakkor, hogy a figyelembe vett faktorok az összes eltéréseknek nagyobb százalékát magyarázzák, azaz — végsősoron —

3. táblázat

A teljes mintára és a szakterületekre elvégzett faktoranalízis összesített eredménye

Faktorok	Szakterületek (megfigyelések)	Teljes minta (134)	Közg. vezető (34)	Term. irányítás (34)	Munkaerőgazd. (34)	Készletgazd. (32)
I.	a)	0,3030	0,256	0,3086	0,4670	0,3893
	b)	8, 13, 14, 16	6, 8, 13, 14, 16	3, 4, 6, 7, 8	1, 2, 8, 10, 11, 12, 13, 15, 16	2, 4, 11, 12, 13 14, 15
II.	a)	0,0997	0,1290	0,1200	0,0926	0,1356
	b)	3, 4, 5, 6, 7	5, 11	3, 7, 11, 15	3, 7, 9, 14	5, 6, 7, 10, 15, 17
III.	a)	0,0745	0,1041	0,0970	0,0867	0,0787
	b)	2, 3, 9	2, 4, 6, 10, 15	1, 5, 6, 12	3, 4, 5, 6, 17	9, 14
IV.	a)	0,0717	0,0798	0,0893	0,0760	0,0724
	b)	10, 11, 15	9, 10, 17	2, 9	3, 17	1, 3, 8, 13, 14, 16
V.	a)	0,0599	0,0703	0,0779		
	b)	1, 7, 11, 12	3, 7, 8, 12, 13	8, 11, 12, 13, 14, 16		
VI.	a)		0,0664	0,0596		
	b)		1, 4	14, 17		
A figyelembe vett faktorok összes részesedése az eltérések magyarázatából		0,6091	0,7050	0,7524	0,7222	0,6761

kisebbek a véleményeltérések az egyes szakterületen belül, mint valamennyi szakterület egyesítésekor. Ez az eredmény a közvetlen belátással összhangban levőnek tűnik.

A közgazdasági területen a legnagyobb a tényezők szóródása, az első faktor súlya itt a legkisebb. Az egyes faktorokhoz való kapcsolódást jelző korrelációs együtthatók relatíve alacsonyak, sok tényező jelenik meg nagyjából azonos együtthatóval több faktorban. (Ugyanakkor ez az a terület, amelynek első faktora lényegében megegyezik a minta egészére számított első faktorral.) Az egyes faktorok – az elsőt kivéve – nem értelmezhetők olyan jól, mint a minta egészére vonatkozólag. Feltűnő, hogy a naturális szemlélettel, illetve a gazdálkodás naturális vonatkozásaival összefüggő tényezők teljesen egyértelmű prioritást kapnak a gazdálkodás egészét illető, szintetizáló jellegű tényezőkkel szemben. Ezek csak az utolsó két faktorban foglalnak helyet, szerepük tehát elenyészőnek tekinthető. Ez lényegében összhangban áll a teljes mintából nyert – s megítélésünk szerint közgazdasági szempontból kedvezőtlen – képpel, de igencsak elgondolkodtató, hogy a szemléleti probléma épp a közgazdasági terület vezetőinél jelentkezik a legerősebben.

A termelésirányítás vezetőinek véleményében a figyelembe vett faktorok az eltérések nagyobb részét (75% felett) magyarázzák meg, mint a többiekben, s az elemzés itt is sajátos vonásokat tár fel. Az első faktorba épp a legkevésbé naturális vonatkozású tényezők kerültek, amit sokkal inkább vártunk volna a közgazdasági terület vezetőitől. Egyéb elemzésekkel együtt arra következtethetünk, hogy a termelésirányítás szinte kivétel nélkül műszaki képzettségű vezetői erről az oldalról érzik leginkább bizonytalanul magukat, s ezért az innen érkező hatásokkal szemben szeretnének elsősorban védettek lenni. Igen érdekes, hogy ugyanakkor ez az egyetlen terület, ahol az összevont vizsgálatban kiemelkedő szerepet játszó tényezők (főként a vállalaton belüli rugalmas kapcsolatokra utaló 13., 14. és 16. tényezők) ennyire hátra szorultak. Ez véleményünk szerint a vállalatoknál a gyakorlatban uralkodó termeléscentrikus gondolkodás következménye – a termelési vezetők természetesnek tekintik a termelési szempontok elsődlegességét, s nem látják értelmét annak, hogy az említett tényezők hatására tartalékot képezzenek.

A munkaerőgazdálkodási terület kiemelendő sajátossága, hogy az első faktor súlya a többihez, s más területek első faktoraihoz képest igen nagy, és ez a faktor viszonylag magas korrelációs együtthatók mellett sok tényezőt sűrít. Ezek között megtalálhatók: a) a teljes mintában is az első faktorban szerepelt, a belső rugalmassággal összefüggő tényezők (ez teljesen ellentétes a termelésirányítás vezetőinek véleményével), b) az értékesítési és beszerzési piaccal való kapcsolatokra utaló tényezők (ez az egyetlen vizsgálat, ahol az első tényező, a szükségletingadozás ilyen előkelő helyen szerepel), és c) a vállalatpolitikai tényezők is. Azt mondhatjuk, hogy a négy szakterület közül talán ez hozta leginkább össze valamennyi, általunk fontosnak ítélt tényezőt. (Ez persze megítélésbeni bizonytalanságra is utalhat.) Figyelemre méltó egyrészt, hogy a munkaügyi vezetők nem tudtak mit kezdeni a pénzügyi okokkal (gyakorlatilag azonos súllyal szerepel az elsőt kivéve valamennyi faktorban), másrészt, hogy a munkaerőhelyzet és a bérszabályozás csak a harmadik (kis súlyú) faktorban kapott helyet. Ez utóbbi tény egyaránt jelentheti azt, hogy a munkaügyi vezetők mai hazai munkaerő- és bérgazdálkodási feltételeink mellett szükségtelennek, vagy azt, hogy lehetetlennek tartják a tartalékképzést.

A tartalékolásról szóló elméleti tanulmányunkban a készletezésről azt mondottuk, hogy véleményünk szerint ez az a gazdálkodási terület, amely a vállalati rugalmassággal a legszorosabb kapcsolatba hozható. Ezt a megállapításunkat a faktoranalízissel nyert eredmények alátámasztják: ez az egyetlen terület, ahol a rugalmassághoz kapcsolódó változók ilyen szépen összejönnek, s ráadásul az első faktorban. (A vállalatpolitikai célkitűzések, a manőverezési lehetőségek, s a szabályozó rendszerhez való alkalmazkodás sehol másutt nem kerül össze egy faktorba.) Ugyanakkor a szükségletekhez való alkalmazkodás itt is csak az utolsó faktorban szerepel – a rugalmasság tehát ezáltal passzívnak, defenzívnek tűnik. Érdekes továbbá, hogy a teljes minta első faktorával összefoglalt négy változó itt együttesen az utolsó faktorba került – azaz, a készletgazdálkodás vezetői – hasonlóan a termelésirányítás vezetőihez, de nyilvánvalóan más okokból – szintén nem tulajdonítanak nagy jelentőséget a belső kapcsolatok rugalmasságának. (Talán arról van itt szó, hogy a készletezést a vállalatok inkább a külvilággal szemben használják pufferként, s belső szerepét a termelés által determináltnak tekintik.)

*

A fentiekben ismertetett vizsgálatok mellett még további elemzéseket is végeztünk (ezekre elszórtan utaltunk is), a terjedelmi korlátok miatt részletesen nem foglalkozunk velük. Ehelyett inkább arra térünk még ki, hogy miként konkretizálódnak az egyes ható tényezők a vállalatok tényleges gazdálkodásáról szóló véleményekben. Cikkünk második része ezzel foglalkozik.

2. A vállalatok erőforrástartalékolásához vezető okok szakterületenkénti vizsgálata

Ez a kérdéskör, bár logikáját tekintve hasonló kérdésekre keres választ, mint az előzőek, más konkrét megfogalmazást igényel. Ebből, valaminta kérdőív egészének struktúrájából következően, szemben az előző kérdéscsoport vállalatgazdasági nézőpontjával, itt az üzemgazdasági szintű megközelítést alkalmaztuk. Ily módon tartalmilag hasonló, de konkrét megfogalmazásaiban eltérő kérdéseket tettünk fel a három erőforrástartalékról. Valamennyi kérdésre az adott szakterület vezetője és a közgazdasági vezető válaszolt.

A faktoranalízist alapvetően a tényezők (a kölcsönös kapcsolatokat is kifejező) rangsorolására, a cluster analízist a csoportosításra használtuk fel, módszertanilag az előző fejezettel analóg módon. Az alábbiakban az egyes szakterületekre vonatkozó eredményeket ismertetjük, s ezután teszünk néhány általános megállapítást.

a) *Termelésirányítás*

A kapacitástartalékokra vonatkozólag 10 tényezőt soroltunk fel – ezekre (csakúgy, mint valamennyi további kérdésre) ismét 0-tól 9-ig kellett súlyokat adni. A faktorelemzés eredményeképpen a tíz tényezőt a 4. táblázatban rangsoroljuk.

A rangosorolás alapjául a 10×10-es rotált faktor-mátrix szolgált. Ebben a mátrixban ugyanis esetünkben – a közgazdasági feladatoknál ritkaságszámbamenő tisztasággal – jól elválaszthatók az egyes tényezők, s igen

magas korrelációs együtthatók mellett azonosíthatók egy-egy faktorral. (A legalacsonyabb korrelációs együttható 0,84, a legmagasabb 0,99.) Tudatában vagyunk annak, hogy ez az eredmény messzemenő következtetésekre nem alkalmas, felhasználható azonban az egyes tényezők vélt fontosságának jellemzésére a kapacitástartalékok képzésében. Egyértelmű ugyanis, hogy az okok között nagyobb jelentőséget tulajdonítanak az első faktorokkal azonosítható tényezőknek, mint az utolsóknak. A szemléletes összehasonlítás céljából a táblázatban feltüntetjük az egyszerű megoszlás átlagai alapján kialakítható sorrendet is.

4. táblázat

Milyen okok késztetik a vállalatot kapacitástartalékok tartására?

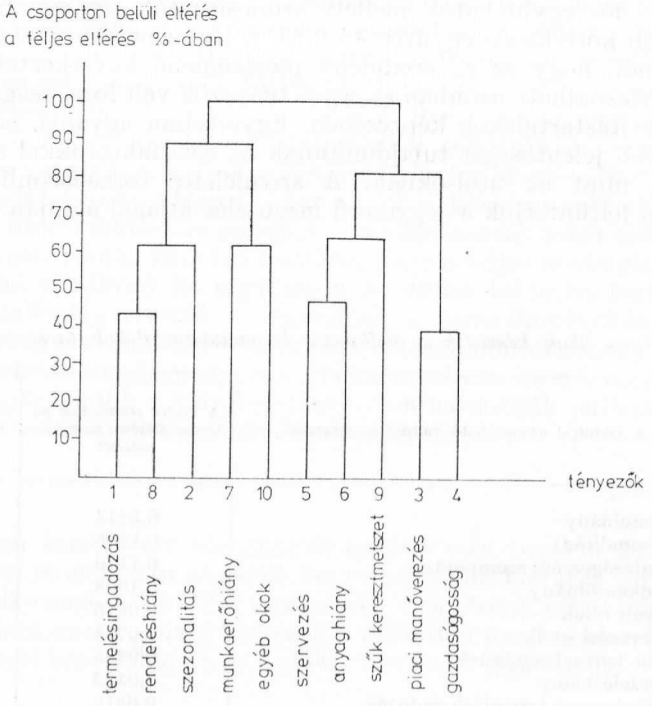
A faktor sorszáma	A faktorral azonosítható változó megnevezése	A faktor részesedése az összes eltérések magyarázatából	A faktorral azonosított változónak az egyszerű átlag alapján elfoglalt helye
1.	Anyaghiány	0,3414	2
2.	Szezonalitás	0,1405	9
3.	Gazdaságossági szempontok	0,1170	5
4.	Munkaerőhiány	0,1008	1
5.	Egyéb okok	0,0815	10
6.	Szervezési okok	0,0734	3
7.	Szűk keresztmetszetek	0,0485	6
8.	Rendeléshiány	0,0433	8
9.	Szükségszerű termelésingadozás	0,0318	4
10.	Piaci manőverezési lehetőségek	0,0218	7

A táblázatból látható, hogy a válaszadók döntő okként az anyagellátási nehézségeket nevezik meg, s ezután nagyjából azonos súllyal szerepel a szezonalitás, a gazdaságossági szempont és a munkaerőhiány. Az egyéb tényezők fokozatosan csökkenő és kis súllyal szerepelnek. Az első négy tényező közül három kényszerítő jellegű (s észre kell vennünk, hogy a két másik erőforrás hiánya szerepel itt), de pozitívan kell értékelnünk, hogy a gazdaságossági szempontok ilyen előkelő helyen szerepelnek. Feltűnő viszont, hogy a piaci szempontok az utolsó helyre kerültek. A faktoranalízis erősen átrendezte az egyszerű átlagok alapján kialakítható sorrendet.

A cluster analízis eredményét a mellékelt 3. ábra illusztrálja. A 10 változó négy jól értelmezhető clustert képez: az elsőben a termeléssel közvetlenül összefüggő három változó szerepel, a második a munkaerőhiány mellett az egyéb okokat tartalmazza, a harmadik a termelési folyamat lebonyolíthatóságának feltételeiként értelmezhető tényezőkből áll, s végül a negyedik a vállalatgazdasági szintű tényezőket foglalja össze.

b) *Munkaerőgazdálkodás*

Erre a szakterületre ugyanazt az elemzést végeztük el, mint az előzőre, de itt 11 változót soroltunk fel. Ezek közül nyolc pontosan megegyezett a termelésirányítás vezetőinek feltett kérdésekkel, a többi három tartalmilag hasonló: az adott erőforrás korlátozottságára, s a többi erőforrással való kapcsolatára utalt. A rotált faktor-mátrixban itt, bár kissé alacsonyabb, de még



3. ábra. A vizsgált vállalatok kapacitástartalékaira ható tényezők

mindig magas korrelációs szinten azonosíthatók az egyes faktorok a változókkal. (A legalacsonyabb korrelációs együttható 0,82, a legmagasabb pedig 0,983.) A változóknak a faktorelemzés alapján kialakított sorrendjét, a faktorsúlyokat és az egyszerű átlag alapján képzett sorrendet az 5. táblázat mutatja.

A táblázatból az a vélemény olvasható ki, hogy a munkaerő tartalékolására egy tényező, a termelés szükségszerű ingadozása, döntő hatással van. Ez is kényszerítő tényező, csakúgy mint a termelésirányítást befolyásoló fő szempontok, de lényegesen eltérő tartalmú: a működés egészét érintő külső, objektív tényező, a vállalatvezetés befolyásától jórészt független. (Szemben a termelésirányítással, ahol a tartalékolásra ható fő tényezők között az egyéb erőforrásokkal való ellátottság szerepelt, ami elsősorban vállalati döntési szféra.) Tehát a vállalatok vezetői azt állítják, hogy saját belső céljaik megvalósítása érdekében munkaerőt nem tartalékolnak. A népgazdasági munkaerőhelyzet (amely itt mint külső erőforráskorlát szerepel) második helye ugyancsak ezt jelzi.

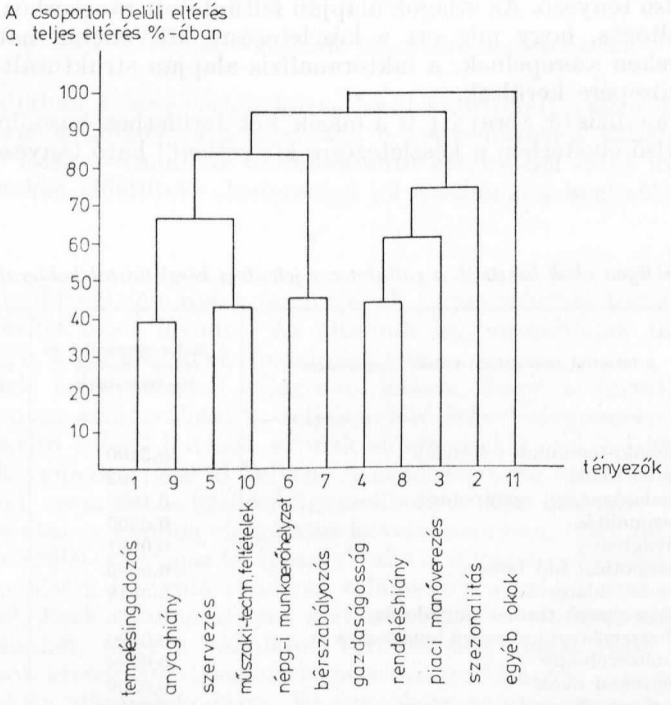
A többi tényező súlya kicsi. A rangsorban itt is érdekes a gazdaságossági szempontok előkelő helyezése, és feltűnő, de magyarázható, hogy az anyaghiány, amelyet a kapacitástartalékolás első számú okaként értékeltek, itt a sor végére került. Csak a kapacitáskorlátra utaló műszaki-technikai feltételeket előzi meg. Tehát úgy vélik, hogy az erőforráskombinálás belső tényezői nem játszanak szerepet a munkaerőtartalékolásban, ami ismét az előző bekezdésben tett megállapítást erősíti.

5. táblázat

Milyen okok késztetik a vállalatot munkaerőtartalékolásra?

A faktor sorszáma	A faktorral azonosítható változó megnevezése	A faktor részesedése az összes eltérések magyarázatából	A faktorral azonosított változónak az egyszerű átlag alapján elfoglalt helye
1.	Szükségszerű termelésingadozás	0,4269	6
2.	Népgazdasági munkaerőhelyzet	0,1178	2
3.	Egyéb okok	0,0988	11
4.	Gazdaságossági szempontok	0,0939	8
5.	Szervezési okok	0,0648	1
6.	Szezonalitás	0,0499	10
7.	Bérszabályozási rendszer	0,0418	3
8.	Piaci manőverezési lehetőségek	0,0383	7
9.	Rendeléshiány	0,0300	9
10.	Anyaghiány	0,0228	4
11.	Műszaki-technikai feltételek	0,0151	5

A cluster analízis (4. ábra) itt is jól értékelhető eredményeket hozott. Három clustert és két különálló tényezőt találunk az előző vizsgálathoz hasonló 70% körüli szinten összekapcsolódva. Az első cluster a vállalati működés technikai feltételeire utaló változókat tartalmazza, a második a munkaerőgazdálkodás



4. ábra. A vizsgált vállalatok munkaerőtartalékaira ható tényezők

központi szabályozásával összefüggő két változót, míg a harmadik a vállalatgazdasági szintű változókat foglalja össze. Külön áll a szezonalitást és az egyéb okokat jelző tényező.

c) *Készletgazdálkodás*

Az előzőekhez hasonló logikával végeztünk elemzéseket a készletezési szakterületre is. Itt 12 tényezőt tüntettünk fel, amelyek között szerepelt ugyanaz a nyolc, amelyre az előző két szakterületen is rákérdeztünk, s volt további egy-egy közös tényező egyenként az előző két szakterülettel. A rotált faktormátrix itt is jól strukturálta a változókat, s talán itt a legmagasabbak az egyes tényezőket az egyes faktorokhoz rendelő korrelációs együtthatók. (A legalacsonyabb 0,858, a legmagasabb 0,976.) A 6. táblázat tünteti fel a faktoranalízis adta rangsorolást.

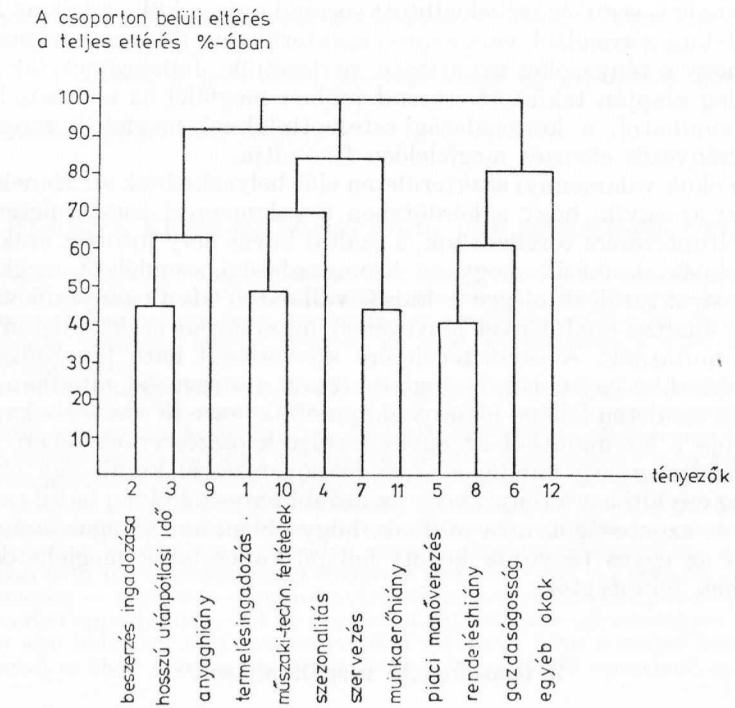
A faktoranalízis alapján megállapított sorrend itt különbözik legélesebben az egyszerű átlagszámítással nyert sorrendtől. Ennek értékelésére még visszatérünk. Az eredmény azt mutatja, hogy a kapacitások és a készletek tartalékolását komplementer gazdálkodási aktusként értékelik. Míg a kapacitástartalékolásra vezető okok között a készletnek, mint erőforrásnak korlátozottsága volt az első, addig itt a fő oknak a termelés technikai feltételeit tartják, amelyek az állóeszközkorlát megjelenítői. Ez az eredmény híven fejezi ki a gazdálkodás tényleges gyakorlatában megjelenő adottságokat, a vállalataink működésében uralkodó jelenlegi helyzetet. Az itt kialakult sorrend közös a másik két erőforrásnál tapasztalttal abban, hogy az egyéb okok és a gazdaságossági szempontok elől helyezkednek el, a piaci manőverezés lehetősége pedig itt is az utolsó tényező. Az átlagok alapján feltüntetett rangsorhoz képest igen lényeges változás, hogy míg ott a készletezésre közvetlenül ható tényezők az első helyeken szerepelnek, a faktoranalízis alapján strukturált halmazban a mezőny közepére kerültek.

A cluster analízis (5. ábra) itt is a másik két területhez hasonló struktúrát tár fel. Az első clusterben a készletezésre közvetlenül ható tényezők szerepel-

6. táblázat

Milyen okok késztetik a vállalatot a jelenlegi készlettartalékolásra?

A faktor sorszáma	A faktorral azonosítható változó megnevezése	A faktor részesedése az összes eltérések magyarázatából	A faktorral azonosított változónak az egyszerű átlag alapján elfoglalt helye
1.	Műszaki-technikai feltételek	0,3400	9
2.	Egyéb okok	0,1192	12
3.	Gazdaságossági szempontok	0,1068	5
4.	Szezonalitás	0,0900	10
5.	Anyaghiány	0,0681	3
6.	Utánpótlási idő hossza	0,0589	1
7.	Rendeléshiány	0,0548	11
8.	Szükségszerű termelésingadozás	0,0426	6
9.	A beszerzés szükségszerű ingadozása	0,0390	2
10.	Munkaerőhiány	0,0324	7
11.	Szervezési okok	0,0269	4
12.	Piaci manőverezési lehetőség	0,02134	8



5. ábra. A vizsgált vállalatok készleteire ható tényezők

nek, a másodikban a termeléshez közvetlenül kapcsolódó változók, a harmadikban a munkaerőhiány és a szervezés, a negyedikben a vállalatgazdasági szempontok. Ezek – mint az összehasonlító elemzésből látni fogjuk – az egyéb területeken előállított clusterekkel jól összhangba hozhatók.

*

A három szakterületen nyert eredmények összevetéséből további érdekes következtetéseket lehet levonni. Az általunk legfontosabbnak tartott megállapításokat a következőkben foglaljuk össze:

A sorrendek összevetéséből világosan látszik, hogy a tartalékolás motívumai az egyes szakterületeken teljesen eltérőek. Valamennyi szakterületen a kényszerítő jellegű hatások vannak előtérben, de ezek közül különböző konkrét okok vannak a vezető helyen. A gazdaságosság előtérbe kerülése az üzemgazdasági szemléletű kérdésfeltevésekben a közvetlenebb érdekeltség megjelenésére utal – gondos vizsgálatot követel azonban, hogy mit is értenek az egyes szakterületeken „gazdaságosság" alatt. A piaci manőverezési lehetőségek kihasználására irányuló törekvés valamennyi részterületen a sor végén helyezkedik el. Ezek az eredmények alátámasztják az elméleti anyagunkban szereplő tételünket, hogy a vállalatok tartalékolási magatartása döntően a negatív hatások kivédésére, „passzív rugalmasságra" irányul, s nem a környezethez való aktív alkalmazkodásra. Ez egyébként a jelen tanulmányunk első részében nyert eredményekkel is alátámasztható.

A faktoranalízis segítségével előállított sorrend erősen különbözik az átlagok alapján kialakult sorrendtől, valamennyi szakterületen. Ez annak jelentőségére mutat rá, hogy a tényezőket együttesen mérlegeljük. Jellemzőnek tekinthető, hogy az átlag alapján tekintett sorrend jobban megfelel az első közelítésben várható, mondhatni, a közgazdasági sztereotipiáknak megfelelő rangsornak, s ezt a sokténye zős elemzés megfelelően finomítja.

Az egyéb okok valamennyi szakterületen elől helyezkednek el. Ennek két fő okát látjuk: az egyik, hogy a kérdőívben a valamennyi szakterületen közös tényezők feltüntetésére törekedtünk, s ezáltal kevés hely jutott a szakterületi specialitásoknak, a másik, hogy az üzemgazdasági szemléletű megközelítés miatt nem strukturáltuk eléggé a külső, vállalaton kívüli tényezőket.

A cluster analízis eredményei lényegében mindhárom szakterületen azonos struktúrát mutatnak. A saját területére közvetlenül ható tényezőket valamennyi szakterület egy clusterbe foglalta össze, a szervezést mindhárom esetben egy más területen fellépő hiánnyal kapcsolták össze (a szervezés kapcsolatteremtő ereje e szempontból az egyes területek összeszervezésében jelenhet meg). A gazdaságosság mindig a piaci manőverezéssel került egy clusterbe, ami némileg enyhíti azt a tényt, hogy az utóbbi tényezőt olyan hátul rangsorolták. Ezek az azonosságok arra utalnak, hogy ebben az üzemgazdasági szintű elemzésben az egyes tényezők között fellépő kapcsolatok meglehetősen törvényszerűnek tekinthetők.

3. Összefoglaló megállapítások

Ha a vállalatpolitikai szemléletű, a tényezők általános hatására rákérdező vizsgálatunk eredményeit összevetjük az egyes vállalatok különböző erőforrás-tartalékainak képzésére ható, üzemgazdasági szemléletű eredményekkel, a válaszok lényegében azonos tendenciákat tükröznek. Röviden összefoglalva ezeket a következőkben fogalmazhatjuk meg:

1. A megkérdezett szakemberek szerint a vállati erőforrás-tartalékokat döntően a naturális vonatkozású tényezők határozzák meg.
2. A cluster analízissel nyert struktúrák azt támasztják alá, hogy az egyes erőforrásokra vonatkozó különböző korlátozásokat nem annyira önmagukban, hanem sokkal inkább az erőforrások szükséges kombinációival összefüggésben tartják jelentősnek.
3. A vállalati tartalékolási magatartást a megkérdezettek passzív alkalmazkodásként jellemzik; a kedvezőtlen lehetőségektől való félelem összehasonlíthatatlanul nagyobb szerepet játszik a tartalékolásban, mint a vállalkozói készség.
4. A sokváltozós statisztikai módszerek alkalmazása azt igazolja, hogy a megkérdezett vállalati vezetők meglehetősen homogén halmazt alkotnak a tartalékolásra vonatkozó véleményüket illetően, s a véleményalkotásuk általában konzekvens.

Befejezésül hangsúlyozzuk: tisztában vagyunk azzal, hogy az alkalmazott módszerek alapján nyert eredmények — épp a módszerek és a statisztikai megfigyelések lényegéből adódóan — nem abszolútizálhatók. (Ezt a cikkben azzal is igyekeztünk érzékeltetni, hogy nem egy alkalommal használtunk feltételes módot megfogalmazásainkban.) Úgy véljük azonban, hogy elemzéseink

igazolják: a sokváltozós statisztikai módszereknek a közgazdasági jelenségek vizsgálatában való felhasználása igen sok, más módon hozzá nem férhető információt adhat, s jelentősen hozzájárulhat az elemzések elmélyítéséhez.

(Beérkezett: 1977. nov. 10-én.)

RESERVE KEEPING BEHAVIOUR: THE MANAGERS' OPINION

The paper is based on the evaluation of a sample survey interviewing 134 managers. The survey was aimed at the examination of enterprise behaviour in keeping reserves of resources. Reserves have been interpreted as a condition of flexible firm management and capacity reserves, labour reserves as well as inventories were considered as characteristic forms.

A part of the survey data is suited for processing by multivariate statistical methods. In the present study the results of factor and cluster analyses are presented on factors which influence reserve keeping either generally or by types of reserves. Analyses show that the factors can be classified well and the results of factor and cluster analyses mostly support each other thus providing possibilities for drawing several important conclusions. Among them we point out that from among the factors influencing the development of enterprise reserves those in physical terms are usually of more importance than those in value terms; restrictions concerning various resource reserves have a part only in connection with the combination of resources and that – in the opinion of the interviewed managers – reserve keeping behaviour is motivated decisively not by making best out of market opportunities, but by the fear of unfavourable circumstances. The examination has also indicated that the interviewed managers form a rather homogeneous set with regard to their opinion and their opinion formation is consistent, as a rule.

МНЕНИЕ ПРЕДПРИЯТИЙ О СОЗДАНИИ РЕЗЕРВОВ

Данная работа основана на подведении итогов опроса 134 руководителей предприятий посредством вопросника. Цель опроса заключалась в изучении поведения предприятий в отношении накопления ресурсов. Ресурсы рассматривались в качестве предпосылки гибкого хозяйствования предприятия и в качестве их характерных форм указывались резервные мощности, трудовие реасурсы и запасы.

Часть данных исследования пригодна для применения статистических методов со многими переменными. В настоящей работе излагаются результаты факторного и кластерного анализа, касающегося факторов, влияющих на формирование резервов вообще, а также на отдельные типы резервов. Анализ показывает, что факторы пригодны для создания структуры и результаты проведенного факторного и кластерного анализа во многом подтверждают друг друга и позволяют делать много важных выводов. Из их числа следует указать на то, что из факторов, влияющих на формирование резервов предприятий, большее значение имеют натуральные, чем те, которые по своему характеру являются стоимостными: ограничения, касающиеся различных резервов играют определенную роль лишь в связи с комбинацией различного рода резервов; по мнению опрошенных руководителей предприятий формирования ресурсов мотивируется в большей части не столько использованием рыночных возможностей, сколько опасениями, связанными с неблагоприятными возможностями. Исследование еще подчеркивает и то, что мнение опрошенных руководителей предприятий является довольно однородным множеством и в создании своих мнений они довольно последовательны.

S. Benedikt Vera — Vári Anna

Egyes clusteranalízis-eljárások és gazdasági alkalmazásuk

Bonyolult rendszerekkel kapcsolatos modellezési feladatok megoldásában matematikai, számítástechnikai és humán oldalról hasonló problémák (a feladat egyszerűsítésére, az implicit összefüggések feltárására stb.) merülnek fel. A cikk első része e problémákat általában tárgyalja, majd ismerteti az ezek megoldását segítő statisztikai osztályozó eljárásokat. Végezetül bemutatjuk az eljárások alkalmazási lehetőségeit strukturális döntések szimulációs modellezésében.

A modellezési feladatok által felvetett problémák

Komplex rendszerek *modelljének megalkotásához* gyakran nélkülözhetetlen a rendszer leíró változói közötti kapcsolatok feltárása, a változók struktúrába rendezése, számuk redukálása. Ugyancsak felvetődhet az igény a rendszer objektumainak (alrendszereinek) leíró változóik alapján történő tipologizálására és összevonására is.

A változók és objektumok strukturálása és összevonása nem az egyetlen módja a rendszerelemzés megkönnyítésének. Másik módja a kvantitatív skálákon mért változóknak (illetve változócsoportoknak) kvalitatív változókká történő transzformációja. A kvalitatívvá transzformált változók és a rendszer egyéb – kvantitatív vagy kvalitatív – változói között általában jóval egyszerűbben kezelhető statisztikai, illetve logikai összefüggések állapíthatók meg, mint ez utóbbiak és az eredeti kvantitatív változók (változócsoportok) között.

A modell-változók vagy objektumok összevonása, valamint a kvalitatív skálák alkalmazása a rendszerelemzés matematikai és számítástechnikai problémáinak enyhítésén túl a modellek alkalmazásával kapcsolatos *emberi döntések* hatékonyságát is növeli. Döntéselméleti kísérletek bizonyították, hogy a döntéshozóknak egy-egy szituáció felismerésében és megítélésében nyújtott teljesítménye a szituációt leíró változók számának, pontosabban a változók lehetséges értékkombinációi számának növekedésével rohamosan csökken (*Miller*, 1967; *Edwards, Philips*, 1966; *Tversky, Kahneman*, 1972).

A következőkben néhány, a modell-alkotás és a modell-használat során felmerülő tipikus döntési feladaton mutatjuk be a fenti elvek érvényesítési lehetőségeit.

Kvantitatív skáláknak leképezésekor kvalitatív skálára az osztályok határainak kijelölésével tulajdonképpen meghatározzuk, hogy a kvantitatív változók értékeinek különbsége mely tartományok között lényegi, és mely tartományo-

kon belül hanyagolható el az adott célú elemzés szempontjából. Ez igen hasznos lehet pl. a *modellek érvényességének vizsgálatánál*, amikor a modell által előállított egyes adatokat a valóságos rendszer tényleges adataival kell összehasonlítanunk. A modell érvényességének kritériuma az, hogy a modellezett rendszer és a modell megfelelő adatai adott pontossági korláton belül megegyezzenek. A kívánt pontosság megadása körül felvetődő problémákat a vizsgált változók kvalitatív leképzésével, azaz a lényeges és kevésbé lényeges különbségek előzetes szétválasztásával jórészt kiküszöbölhetjük.

A kvalitatív skálák alkalmazásának említett előnye fontos lehet azokban az esetekben is, amikor nem egyetlen jól definiált feltételrendszer és célfüggvény alapján keressük a legjobb döntési alternatívát, hanem több *döntési politikát* akarunk összevetni az ezek eredményességét kifejező ún. cél-változókra gyakorolt hatásuk alapján. (Ez a helyzet általában a heurisztikus eljárásoknál.) A döntéshozó a több cél-változónak egyetlen kvalitatív preferenciaskálára való leképezésével értékítéleteit, a cél-változóknak az értékelés szempontjából ekvivalens tartományait határozza meg. Ez nagymértékben megkönnyíti a döntési politikák értékelését és összehasonlítását.

Másrészről a döntési politikákat reprezentáló kvantitatív változókat – az ún. tényezőket – is átalakíthatjuk kvalitatív változókká. A tényezők vagy a cél-változók számának csökkentése, illetve ezek kvalitatív transzformációja megkönnyíti annak megállapítását, hogy a döntési politika tényezői milyen hatást gyakorolnak a modell működésére *(modell-érzékenységvizsgálat)*, s ezáltal elősegíti a legeredményesebb politikai megtalálását.

A leíró változók vektorai és az osztályok közötti kapcsolat (osztályhatárok vagy osztályba tartozási függvény) explicit megadása ugyancsak nehéz döntési problémát vethet fel, különösen sokváltozós terek esetében. Sok esetben alkalmazható az a módszer, hogy a döntéshozók implicit módon, azaz összetartozó kvantitatív-kvalitatív értékeket (ún. tanítási mintákat) adnak meg és ebből számítógépi úton megfelelő statisztikai eljárások állítják elő a minta alapján legvalószínűbb összefüggéseket. Ezen eljárásoknak természetesen alkalmasaknak kell lenniök a döntéshozó inkonzisztens döntéseinek kiszűrésére is. Így lehetővé válik, hogy az ember is fokozatosan tanuljon a gépi eljárások által szolgáltatott eredményekből.

A döntések megfelelő előkészítése, az egyidejűleg figyelembe veendő változók számának és típusának helyes meghatározása, az emberi tanulás biztosítása különösen a folyamatos emberi beavatkozást igénylő, interaktív eljárások (pl. számítógépes szimuláció) alkalmazása esetén lényeges.

Statisztikai osztályozó eljárások

A fentiekben felsorolt problémák megoldására többek között a statisztikai osztályozó (alakfelismerő) eljárások alkalmasak.

Ezek egyik csoportját a *tanító nélkül osztályozó* (cluster analízis) eljárások képezik, amelyek az objektumok tulajdonságait leíró változók értékei alapján az objektumok osztályozására vonatkozó hipotéziseket generálnak. A másik csoportba a *tanítóval működő osztályozó* eljárások tartoznak, amelyek az elemeikkel megadott osztályokból (tanítási mintákból) indulnak ki. Az eljárások feladata, hogy a leíró változók és az osztályokat reprezentáló kvalitatív változók között összefüggéseket határozzanak meg, s egyben lehetővé tegyék

újabb objektumok osztályba sorolását. Mindkét eljárástípusnál az az objektumok összevonásának, az objektumok és osztályok egymáshozrendelésének kritériuma valamely geometriailag vagy statisztikusan értelmezhető távolság (pl. euklideszi távolság, osztályon belüli négyzetes eltérés stb.) minimalizálása.

A clusteranalízis-eljárásokat az általuk igényelt kiindulási információ mennyisége és jellege alapján két nagy csoportba sorolhatjuk.

A hierarchikus eljárások az objektumok eloszlására, illetve az osztályok számára vonatkozóan semmiféle előzetes információt nem igényelnek. Az objektumok összevonásával a közöttük levő összefüggések hierarchiáját határozzuk meg. *A nem hierarchikus eljárások* adott (rögzített) vagy az algoritmusok által iteratív módon változtatható számú osztályt képeznek az objektumokból. Az osztályok számára vonatkozó információt tehát előre meg kell adni számukra.

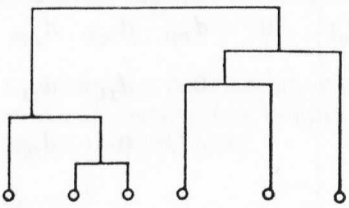
A modellezés során – mint azt példánkból is látni fogjuk – célszerű a különböző mennyiségű kiindulási információt igénylő hierarchikus és nem hierarchikus cluster analízist, valamint a tanítóval működő eljárásokat kombináltan, egymásra építve alkalmazni.

A következőkben részletesen ismertetjük a vizsgálatainkban alkalmazott cluste ranalízis eljárásokat. Ezek programját FORTRAN nyelven a CDC 3300 számítógépre készítettük el, s a közeljövőben elkészül egy, az R20 számítógépen futtatható változatuk is. A vizsgálatainkban ugyancsak felhasznált tanítóval működő – ún. potenciálfüggvényes – osztályozó eljárásra vonatkozóan l. pl. *Andrews*, 1972.

A Wishart-féle hierarchikus clusteranalízis-eljárás

A hierarchikus cluster analízis az egyes objektumokat rangsorolja, ún. hierarchiát állapít meg közöttük. Az egyes objektumok közötti hierarchikus összefüggést fával vagy más néven dendogrammal szokták ábrázolni (1. sz. ábra). Az ábrán látható fán a körök egy-egy objektumot jelentenek, az összekötő vonalak az objektumok közötti kapcsolatokat. A dendogram függőleges tengelye az egyesítési szintek megadására szolgál. A hierarchikus clusteranalízis fogalmát az alábbiakban határozzuk meg.

Legyen $E = \{x_1, x_2, \ldots, x_n\}$ az objektumok halmaza, amelyet kiindulásként tekintsünk $P_1, \ldots P_n$ egyelemű clusterek halmazának. Válasszuk ki közülük azt a P_p és P_q osztályt, amelyek valamilyen értelemben a legköze-



1. ábra

lebb vannak egymáshoz, s ezeket vonjuk össze egy osztályba. Az így kapott osztályhalmaznak már csak $n-1$ eleme lesz:

$$P_1, P_2 \ldots, (P_p, P_q), \ldots, P_n$$

Ismételjük meg ezt az eljárást. Így az osztályhalmazoknak egy olyan sorozatát kapjuk, amelynek a továbbiakban $n-2$ eleme, majd $n-3$ stb. eleme lesz. Végül egyetlen osztályt kapunk, amely az eredeti n osztályt tartalmazza.

Kérdés, hogy mely osztályokat nevezzük legközelebbieknek. Ehhez definiálni kell két objektum távolságát és ennek alapján meg kell határozni, hogy mit értünk két osztály távolságán. Érthetjük ezalatt pl. egymáshoz legközelebbi, illetve egymástól legtávolabbi elemeik, vagy középpontjaik (centroidjaik) távolságát stb., a különböző hierarchikus módszerek lényegében ebben különböznek egymástól.

Legyen az osztályok közötti távolságfüggvény

$$d: E^* \times E^* \to R^-,$$

ahol E^* az E részhalmazainak halmaza, és R^+ a nem negatív valós számok halmaza.

Jelöljük a P_i és P_j osztályok távolságát, azaz $d(P_i, P_j)$-t d_{ij}-vel. A kezdeti n számú egyelemű osztályra így egy $n \times n$-es D_0 távolság-mátrixot nyerünk:

$$D_0 = \begin{array}{c|cccc} & P_1 & P_2 & P_3 & \ldots P_n \\ \hline P_1 & 0 & d_{12} & d_{13} & d_{1n} \\ P_2 & & 0 & d_{23} & d_{2n} \\ \vdots & & & 0 & d_{3n} \\ P_n & & & & 0. \end{array}$$

Tegyük fel, hogy P_p és P_q vannak egymáshoz a legközelebb. Akkor a P_p és P_q összevonása után egy új $(n-1)\times(n-1)$ dimenziós távolság-mátrix elemeit kell meghatározni.

$$D_1 = \begin{array}{c|ccccc} & (P_p, P_q) & P_1 & P_2 & P_3 & \ldots P_n \\ \hline (P_p, P_q) & 0 & d_{pq1} & d_{pq2} & d_{pq3} & \ldots d_{pqn} \\ P_1 & & 0 & d_{12} & d_{13} & \ldots d_{1n} \\ P_2 & & & 0 & d_{23} & \ldots d_{2n} \\ \vdots & & & & & \vdots \\ P_n & & & & & 0. \end{array}$$

A D_1 mátrixnak $n-2$ sora azonos a D_0 mátrix megfelelő sorával csak egy sorát kell újra kiszámítani. Ha azonban meg tudnánk adni a D_i, $i = 1 \ldots n-1$ számolására egy olyan transzformációs formulát, amely nem az eredeti objektumok, hanem csak az előző mátrix adatait használja, akkor ez az eljárás legegyszerűsödne.

Wishart adott meg egy olyan rekurzív formulát, amelynek segítségével hat különböző hierarchikus módzsert lehet megoldani (*Wishart*, 1969). A módszerek egymástól a d függvény definíciójában különböznek. Az objektumok között értelmezett távolság valamennyi módszernél az euklideszi távolság. Ha a P_p osztályt egyesítjük a P_q osztállyal, akkor az így kapott új P_r osztály távolságát a többi P_t ($t = 1, \ldots n$; $t \neq p$, $t \neq 0$) osztálytól is ki kell számítani. Így a távolság mátrix is megváltozik. Az új távolság-mátrixot a következő transzformációs formula segítségével számítjuk:

$$d_{ir}^2 = \alpha_p d_{ip}^2 + \alpha_q d_{iq}^2 + \beta d_{pq}^2 + \gamma \mid d_{ip}^2 - d_{iq}^2 \mid, \tag{1}$$

ahol α_p, α_q β és γ paraméterek, és $P_r = P_p \cup P_q$.

Jelöljük k_i-vel az i-edik osztály elemeinek számát. Az α_p, α_q, β és γ paraméterek különböző megválasztásával a következő hat módszert kapjuk:

A) *Legtávolabbi szomszéd módszer*

Két osztály közötti távolságot akkor tekintjük minimálisnak, ha az összevonásukkal nyert osztály legtávolabbi objektumai közötti távolság minimális. Ebben az esetben a paraméterek:

$$\alpha_p = \alpha_q = \frac{1}{2}; \quad \beta = 0; \quad \gamma = \frac{1}{2}.$$

B) *Legközelebbi szomszéd módszer*

A két osztály közötti távolság akkor minimális, ha az összevonással kapott osztály legközelebbi objektumai közötti távolság minimális. A paraméterek:

$$\alpha_p = \alpha_q = \frac{1}{2}; \quad \beta = 0; \quad \gamma = -\frac{1}{2}.$$

C) *Centroid módszer*

Két osztály közötti távolságot a centroidjaik közötti távolsággal definiáljuk. Olyan osztályokat von tehát össze a módszer, amelyek centroidja közötti távolság minimális.

A transzformációs formula paraméterei:

$$\alpha_p = \frac{k_p}{k_r}; \quad \alpha_p = \frac{k_p}{k_r}; \quad \beta = \alpha_p \alpha_q; \quad \gamma = 0.$$

D) *Medián módszer*

Az osztályok közötti távolságot a mediánjaik közötti távolsággal definiáljuk. Az osztályozásnál azokat az osztályokat vonjuk össze, amelyek mediánjai közötti távolság minimális. A paraméterek:

$$\alpha_p = \alpha_q = \frac{1}{2}; \quad \beta = -\frac{1}{4}; \quad \gamma = 0.$$

E) *Csoportátlag módszer*

A módszer a két osztály közötti távolságot a két osztály elemei közötti átlagos távolsággal definiálja és az osztályozásnál azokat az osztályokat vonja össze, ahol az átlagos távolság minimális.

A paraméterek:

$$\alpha_p = \frac{k_p}{k_r}; \quad \alpha_q = \frac{k_q}{k_r}; \quad \beta = \gamma = 0.$$

F) *Ward módszere*

Az osztályozásnál olyan új osztályok létrehozására törekszünk, amelyekben négyzetes hibának az összevonás által eredményezett növekedése minimális.

A paraméterek:

$$\alpha_p = \frac{k_t + k_p}{k_t + k_r}; \quad \alpha_q = \frac{k_t + k_q}{k_t + k_r}; \quad \beta = \frac{-k_t}{k_t + k_r}; \quad \gamma = 0.$$

A Diday-féle hierarchikus clusteranalízis-eljárás

A nem hierarchikus eljárások lényege, hogy az objektumok egy kezdeti osztályozásából, vagy az osztályok valamely feltételezett jellemzőiből kiindulva iteratív módon változtatják az objektumok besorolását, mindaddig, amíg valamilyen szempont szerint jobb osztályozást nyernek.

Az eljárásoknak az objektumokat előre rögzített, vagy az eljárás során iteratív módon meghatározott K számú osztályba kell sorolniok. Ettől függően megkülönböztetünk rögzített számú osztályt feltételező algoritmusokat (pl. *Forgy*, 1966; *McQueen*, 1967 stb.), és e változó osztályszámú algoritmusokat (pl. *Ball–Hall*, 1965). Az előbbi csoportba tartozik az alábbiakban ismertetendő Diday-féle eljárás is (*Diday, Govaert*, 1974).

Legyen:

E az R^q (q dimenziós euklideszi tér) részhalmaza, az objektumokat reprezentáló vektorok halmaza,

E^* az E részhalmazainak halmaza,

P_K^* az E K-osztályú particióinak halmaza,

$$P \in P_K^* \Longleftrightarrow P = (P_1, P_2, \ldots, P_K), \qquad P_i \in E^*,$$
$$P_i \cap P_j = \Phi \quad \text{ha} \quad i \neq j, \qquad UP_i = E.$$

L^* az ún. „magok" tere. A magok az E részhalmazaihoz rendelt jellemzők,

L_K^* az L^*-ból képzett K-elemű sorozatok halmaza,

$$L \in L_K^* \Longleftrightarrow L = (\lambda_1, \lambda_2, \ldots, \lambda_K), \lambda_i \in L^*,$$

D az objektumok és magok közötti értelmezett távolságfüggvény,

$$D : E \times L^* \to R^+,$$

ahol R^+ a nem negatív valós számok halmaza.

R a magok és az objektumokból képzett halmazok között értelmezett távolságfüggvény

$$R : L^* \times E^* \to R^+$$

W az R által szolgáltatott távolságoknak egy-egy K-partícióra történő összegezése, egyúttal a minimalizálandó célfüggvény,

$$W : L_K^* \times P_K^* \to R^+,$$

$$W(L, P) = \sum_{i=1}^{K} R(\lambda_i, P_i).$$

Definiáljuk továbbá az alábbi leképezéseket:

f: $L_K^* \to P_K^*$,

olyan leképezés, amelynek eredménye olyan új P partíció:

$$L = (\lambda_1, \ldots, \lambda_K); f(L) = P = (P_1, \ldots, P_K),$$

amelyre:

$$P_i = \{x \in E \mid D(x, \lambda_i) \leq D(x, \{\lambda_j), \forall_j \neq i\}.$$

g: $P_K^* \to L_K^*$

olyan leképezés, amelynek eredménye olyan új K-elemű L mag-halmaz:

$$P = (P_1, \ldots, P_K); g(P) = L = (\lambda_1, \ldots, \lambda_K),$$

amelyre:

$$R(\lambda_i, P_i) = \min_{\lambda \in L^*} R(\lambda, P_i).$$

Legyen adott P^0 kezdeti K-partició, vagy L^0 kezdeti K-elemű mag-halmaz.
Az f és g leképezések végrehajtásával rekurzív módon előállíthatók az alábbi sorozatok:

$$\begin{aligned} L^n &= g(P^n), \\ P^{n+1} &= f(L^n), \\ w_n &= W(L^n, P^n). \end{aligned}$$

Bizonyítható, hogy amennyiben

$$W[L, f(L)] \leq W(L, P), \qquad L \in L_K^*, P \in P_K^*, \tag{2}$$

akkor a w_n sorozat monoton csökken és véges számú lépés után *lokális* minimumot ér el.
A (2) feltétel teljesüléséhez elégséges, hogy

$$R(\lambda_i, P_i) = \sum_{x \in p_i} D(x, \lambda_i) \tag{3}$$

A Diday-féle algoritmus P^0-ból vagy L^0-ból kiindulva az f és g leképezések sorozatát hajtja végre, mindaddig, amíg a W értéke tovább már nem csökken. Tetszőleges D távolságfüggvény alkalmazható, R-nek pedig a (3) feltételt kell teljesítenie, így az algoritmus

$$W(L, P) = \sum_{i=1}^{K} \sum_{x \in P_i} D(x, \lambda_i)$$

típusú célfüggvény minimalizálására alkalmas.

A következőkben a Diday-féle algoritmus három speciális esetét vizsgáljuk meg:

A) Legyenek magok az osztályok középérték-vektorai:

$$\lambda_i = \mu_i = \frac{1}{N_i} \sum_{x \in P_i} x,$$

és alkalmazzuk távolságfüggvényként az euklideszi távolságot:

$$D(, \lambda_i) = (x - \mu_i)^T (x - \mu_i).$$

Ennek megfelelően a célfüggvény az osztályon belüli négyzetes eltérések összege, azaz:

$$W(L, P) = \sum_{i=1}^{K} \sum_{x \in P_i} (x - \mu_i)^T (x - \mu_i).$$

Megjegyezzük, hogy a Diday-féle algoritmus ezen speciális esete megegyezik a Forgy által kidolgozott eljárással (*Forgy*, 1966).

B) Válasszuk magnak egy-egy osztály középérték-vektorát és kovariancia mátrixát:

$$\lambda_i = (\mu_i, V_i)$$

$$V_i = \frac{1}{N_i} \sum_{x \in Pi} (x - \mu_i) (x - \mu_i)^T.$$

Alkalmazzuk távolságfüggvényként a *Mahalanobis*-távolságot:

$$D(x, \lambda_i) = (\det V_i)^{1/q} (x - \mu_i)^T V_i^{-1} (x - \mu_i).$$

Ekkor:

$$W(L, P) = \sum_{i=1}^{K} (\det V_i)^{1/q} \sum_{x \in P_i} (x - \mu_i)^T V_i^{-1} (x - \mu_i),$$

azaz a célfüggvény az osztályok inercia-főtengelyeire transzformált négyzetes eltérések összege.

C) Tegyük fel, hogy az objektumok K számú ismert típusú eloszlás keverékéből származnak, azaz:

$$F(x) = \sum_{i=1}^{K} p_i \varphi(\lambda_i, x),$$

ahol $F(x)$ a keverékeloszlás sűrűségfüggvénye, $\varphi(\lambda_i, x)$ az i-edik komponens sűrűségfüggvénye, p_i az i-edik komponens a priori valószínűsége.

Válasszuk magnak az osztályok sűrűségfüggvényeinek λ_i paramétereit, a távolságfüggvényt pedig definiáljuk egy-egy objektumnak egy-egy osztályba való tartozása a posteriori valószínűségének függvényeként – (az osztályok a priori valószínűségeit egyenlőnek tekintve) – az alábbi módon:

$$D(x, \lambda_i) = \log [C/\varphi(\lambda_i, x)],$$

ahol C egy 1-nél nagyobb konstans.
Ekkor:

$$W(L, P) = C' - \log \prod_{i=1}^{K} \prod_{x \in P_i} \varphi(\lambda_i, x).$$

Az így használt Diday-féle algoritmus tehát az objektumoknak a hozzájuk rendelt osztályokba tartozásának – a teljes objektumhalmazra számított – együttes valószínűségét igyekszik maximalizálni.

Speciálisan Gauss-eloszlást feltételezve, $\lambda_i = (\mu_i, V_i)$-t választva, ahol μ_i az i-edik osztály középértékvektora, V_i az i-edik osztály kovariancia mátrixa:

$$D(x, \lambda_i) = C + \frac{1}{2} [\log(\det V_i) + (x - \mu_i)^T V_i^{-1} (x - \mu_i)]$$

$$W(L, P) = C' + \frac{1}{2} \left[\sum_{i=1}^{K} N_i \log(\det V_i) + \sum_{i=1}^{K} \sum_{x \in P_i} (x - \mu_i) V_i^{-1} (x - \mu_i) \right].$$

Alkalmazási példák

1. Vizsgálat

A vizsgálat célja az volt, hogy egy ágazati modell, a magyar szénhidrogénipar strukturális döntéseinek vizsgálatára készült szimulációs modell (*Vári, Kelemen*, 1974) eredményeit elemezzük és meghatározunk egy kellőképpen eredményes döntési politikát.

A modell a szénhidrogénipari vertikum termelési, tárolási és értékesítési folyamatait írta le. A strukturális döntési politikák a rendszer termelő-, tároló- és szállítókapacitásait érintő döntésekből (pl. beruházások, átcsoportosítások stb.) tevődtek össze, az eredményességet pedig – többek között – a hazai igények kielégítettségi fokával mértük. Így a vizsgált tényezők az említett kapacitások idősorai, a cél-változók pedig a rendszer által kibocsátott legfontosabb termékek kínálatai és keresleti idősorai voltak.

Mivel az egyes termékek termelőkapacitásai nem változtathatók meg egymástól függetlenül (ez a szénhidrogénipari technológia sajátos ága), a szállító- és tárolóeszközök pedig a különböző termékfajták között bizonyos mértékig átcsoportosíthatók, így szükséges volt a termékek adatainak együttes vizsgálata. A 10 legfontosabb terméknek 13 évre, azaz 52 negyedévre vonatkozó kínálati és keresleti adatainak együttes áttekintése, illetőleg ennyi adatra nézve az érzékenységvizsgálat elvégzése nehéz feladatot adott volna, ezért előzleg cluster analízisnek vetettük alá az eredményeket.

A szénhidrogénipari termékek nagy részénél (üzemanyagok, fűtőanyagok) a keresleti idősorok szezonális ingadozásokat tartalmaznak, így a kereslet-kínálat viszony évközi alakulása is jelentős eltéréseket mutathat a különböző anyagoknál és időszakokban. Az osztályozandó vektorokat ezért egy-egy termék egy-egy évre vonatkozó negyedéves bontású kínálat-kereslet hányadosaiból alakítottuk ki. Az így nyert 130 db négyegyelemű vektorból először hierarchikus osztályozással egy megfelelő induló osztályozást képeztünk ki. A hierarchikus elemzést a *Wishart*-féle eljárással (a *Ward*-módszerrel) elvégezve, azon vektorokból alakítottunk ki osztályokat, amelyek viszonylag korán kapcsolódtak össze és összevonásuk más osztályokkal a négyzetes eltérések összegét viszonylag nagymértékben megnövelte volna.

Az így nyert osztályozásból kiindulva a Diday-féle nem hierarchikus algoritmussal néhány iteráció után olyan új osztályozást nyertünk, amelyhez tartozó összes négyzetes eltérés az indulási értéknek kb. 2/3 része volt. Mind a hierarchikus, mind a nem hierarchikus osztályozásnál euklideszi távolsági mértéket alkalmaztunk.)

A kialakult 17 osztály mindegyike az alábbi 4 főbb viselkedéstípus valamelyikét reprezentálta:

a) a kínálat jól követi a keresletet,
b) hiányok és többletek váltják egymást,
c) állandó hiány,
d) állandó túlkínálat.

A modell összefüggéseinek ismeretében általánosságban megállapítható, hogy a b) típusú osztályoknál a tárolási és szállítási kapacitások növelése, a c) típusúaknál az adott termék (esetleg ennek nyersanyagai) termelési kapacitásainak növelése, a d) típusúaknál az exportlehetőségek bővítése eredményezheti a kereslet-kínálati viszonyok javulását. A felsorolt típusokon belüli osztályok a hiányok és többletek mértékében, vagy fázisviszonyaiban különböztek egymástól. Ezek további vizsgálata alapján feltárhatók a termelő-, tároló- és szállítókapacitások időszakos átcsoportosítási lehetőségei is.

A modell összefüggéseinek ismeretén túl érzékenységi vizsgálatra is szükség volt ahhoz, hogy a strukturális változtatások arányait és mértékét megfelelően határozhassuk meg. Az érzékenységi vizsgálat során azt kellett feltárni, hogy az egyes kínálat-keresleti vektorok mozgása milyen törvényszerűségeket követ a strukturális döntések függvényében. Első lépésben a vektoroknak az osztályok közötti mozgását vizsgáltuk meg, ezután került sor – a kritikus termékeknél és időszakokban – a finomabb kvantitatív összefüggések feltárására.

Példánkban tehát a clusteranalízis az eredmények áttekintését, a lényegi összefüggések megragadását, s az érzékenységi vizsgálat hatékony elvégzését segítette elő.

2. *Vizsgálat*

Célunk az volt, hogy egy beruházási szimulációs játék (*Rabár, Kelemen*, 1975) modelljét, a döntési lehetőségeket didaktikai szempontból helyesen alakítsuk ki, és hogy lehetővé tegyük a játékosok teljesítményének értékelését gépi úton.

A modell több vállalatnak és a központi gazdaságirányításnak a tevékenységét fogta át. A játékosok egy-egy vállalatot képviseltek. Feladatuk az volt, hogy vállalatuk helyzetének ismeretében beruházási döntéseket hozzanak. Döntéseik következményei alapján meg kellett tanulniok a döntések és a vállalat helyzetére gyakorolt hatásuk összefüggéseit.

A beruházási döntéseket az alábbi paraméterekkel jellemeztük:

- a beruházás költsége;
- a beruházás nyereséghozama;
- a beruházás átfutási ideje;
- a beruházás célja (pl. pótlás, munkaerőhelyettesítés stb.).

A fenti jellemzők adott lehetséges kombinációihoz tartozó 144 döntésből, a tanulásnak és az összefüggések meghatározásának megkönnyítése érdekében, a Wishart-féle hierarchikus eljárással 7 osztályt képeztünk. A képződött osztályokba az alábbi típusú beruházási döntéseket soroltuk:

a) pótlás vagy munkaerőhelyettesítés,
b) kisebb felújító jellegű beruházás,
c) közepes beruházás,
d) kisebb rekonstrukció,
e) kisebb rekonstrukció és kisebb felújítás,
f) kisebb rekonstrukció és közepes beruházás,
g) közepes vagy nagy rekonstrukció.

A fenti osztályokat kvalitatív (rendezési) skálán tudtuk elhelyezni, oly módon, hogy a hozzájuk rendelt rangszámok az egyes osztályokba tartozó döntések horderejének sorrendi viszonyait tükrözték. Ily módon pl. a pótlás-munkaerőhelyettesítés az 1, a közepes vagy nagy rekonstrukció a 7 rangszámot kapta.

Az osztályozással elértük, hogy a játékosnak egyidejűleg csak 7 beruházási típus között kellett döntenie, majd további megfontolások figyelembevételével a választott típuson belül egyetlen alternatívát kiválasztania.

Következő feladatunk az volt, hogy a modell vizsgálata és a játékosok döntéseinek automatikus értékelése céljából explicitté tegyük a döntési helyzet és a helyes döntés közötti összefüggéseket.

A döntéseket a clusteranalízissel kialakított osztályok (döntéstípusok) rangszámaival jellemeztük. A döntési helyzetet leíró változók (a vállalatot jellemző állóeszköz-forgóeszköz arány, nyereség-eszköz arány, eszköz-bér arány, nyereség, fejlesztési alap, központi hitel és támogatás, a korábban megkezdett beruházások terhei stb.) több különböző kombinációját használtuk fel vizsgálatainkban. E kvantitatív változók és a döntéseket reprezentáló kvalitatív változók közötti kapcsolat meghatározására egy tanítóval működő osztályozó eljárást, az ún. potenciálfüggvényes algoritmust (*Andrews*, 1972) alkalmaztuk. Tanítási mintaként felhasználtuk a lejátszott játékok azon összetartozó döntési helyzet-döntési párjait, amelyeknél a döntések kedvező hatásúnak bizonyultak.

Az eljárás meghatározta a döntési helyzet és a jó döntés között a legvalószínűbb függvénykapcsolatot. Így választ kaptunk arra, hogy melyek a döntési helyzetnek a döntések meghozatalánál elsődlegesen figyelembe veendő paraméterei és hogy milyen értelemben és milyen súllyal kell ezeket figyelembe venni.

A döntési helyzetet leíró változók közül az alábbiak bizonyultak lényegesnek a döntések szempontjából:

– az adott évben képződő nyereség;
– a fejlesztési alap + az adott évre esedékes központi hitel és támogatás;
– a korábban megkezdett beruházások aktuális költségei;
– a lekötött eszközök és a bérköltség hányadosa.

A legnagyobb (pozitív) súllyal a nyereség és a fejlesztési alap + hitel + támogatás szerepeltek, az eszköz-bérköltség hányados és a megkezdett beruházások terhei kisebb, de nem elhanyagolható szerepet játszottak a meghozott döntésekben. (Az utóbbi változó természetesen negatív súllyal szerepelt.)

A fentiekben meghatározott összefüggés módot nyújtott arra, hogy segítségével

– ellenőrizzük, hogy modellünk összefüggései megfelelnek-e a valóságos összefüggéseknek (érvényesség vizsgálat);
– felülvizsgáljuk, hogy modellünk eléggé érzékeny-e a döntések közötti különbségekre, alkalmas-e a lényegi összefüggések megtanítására, a játék egyéb véletlen komponensei nem fedik-e el ezeket az összefüggéseket;
– automatikusan értékeljük egy újabb játék játékosainak teljesítményét, összahasonlítva az általuk hozott döntéseket az elméleti összefüggés által meghatározott döntésekkel.

Láttuk, hogy a döntések kvalitatív skálára való leképezése olyan globális vizsgálatokat tett lehetővé, amelyekre enélkül nem lett volna módunk. Természetesen az így nyert információk csak a döntések osztályaira vonatkoznak. A következő lépés az egy-egy osztályba tartozó döntések közötti finomabb különbségek vizsgálata, amely az egy-egy osztályba tartozó elemek kis száma miatt aránylag egyszerű eszközökkel elvégezhető. A cluster analízis tehát esetünkben mind a döntési, mind az elemzési és értékelési feladatok dekompozícióját elősegítette.

Végül fölhívjuk a figyelmet néhány, az alkalmazások során felmerült problémára. Az egyik – a 2. vizsgálatnál – adódott abból, hogy mind a döntési helyzeteket, mind a döntéseket különféle mértékegységekben mért paraméterekből álló vektorokkal jellemeztük. Ezek együttes kezelése csak úgy volt lehetséges, hogy a vektorokat valamennyi dimenzió szerint normáltuk.

A második problémát az egymástól statisztikusan nem független változók kezelése adta. A mintavektorok halmazára korrelációanalízist végeztünk, így a 2. vizsgálatban szereplő, a döntési helyzeteket leíró változók között számottevő korrelációt találtunk. Az analízis eredményei alapján alakítottuk ki azokat a különböző, korrelált változókat nem tartalmazó változókombinációkat, amelyekből képzett vektorokra a potenciálfüggvényes osztályozó eljárást elvégeztük. Esetünkben a változók közötti kapcsolatok aránylag egyszerűek voltak, így a változók számának csökkentéséhez nem volt szükség gépi eljárások (pl. faktoranalízis) alkalmazására.

A Diday-féle eljárás alkalmazása során merültek fel az osztályok számának előzetes meghatározásával kapcsolatos problémák, mivel az eljárás előre meghatározott számú osztállyal dolgozik. Természetesen minél több osztályt engedünk meg, annál kisebbé tehető a négyzetes eltérések összege. Ugyanakkor általában nem célszerű túlságosan sok kevéselemű osztály képzése sem. A probléma megoldására beépíthető olyan algoritmus, amely a nagy elemszámú,

ill. nagy szórású osztályok szétbontásával, és a kis elemszámú, egymáshoz közeli osztályok egyesítésével iteratív módon alakítja ki a végső osztályozást. Ilyen algoritmust tartalmaz pl. az ISODATA eljárás (*Ball–Hall*, 1965). Olyan algoritmus azonban, amely automatikusan optimális megoldásra vezet, nem ismeretes. Sőt magának az optimalitási kritériumnak a megadása is problematikus. Ráadásul a fenti típusú algoritmusok sok előzetes információ megadását igénylik (pl. maximálisan és minimálisan megengedett osztályelemszám stb.), és túlságosan mechanikusak is, ezért alkalmazásukat nem láttuk célszerűnek. Ehelyett a Diday-eljárást több különböző számú osztály esetére végrehajtottuk úgy, hogy az induló osztályokat a korábbi osztályozások során kialakult osztályok szétbontása, ill. egyesítése révén nyertük. Az eredményeket megfelelő mutatók (pl. az osztályok szórásainak átlaga) segítségével összehasonlítva, kiválasztottuk a legkedvezőbb megoldást.

Az előzőekből kitűnik, hogy az osztályozó eljárások – heurisztikus jellegüknél fogva – nem alkalmazhatók mechanikusan, megkövetelik a folyamatos emberi beavatkozást, az eredmények állandó elemzését és értékelését.

(Beérkezett: 1977. ápr. 12-én.)

IRODALOMJEGYZÉK

1. Andrews, P.: Introduction to Mathematical Techniques in Pattern Recognition. Wiley-Interscience. New York, 1972.
2. Ball, G. H. – Hall, D. J.: ISODATA, a Novel Method of Data Analysis and Pattern Classification Techn. Report, Stanford Research Inst. Menlo Park. California, 1965.
3. Diday, E. – Govaert, G.: Apprentissage et Mesures de Ressemblances Adaptatives. IRIA, Rapport de Recherche, 1974. No. 89.
4. Edwards, W. – Phillips, L. D.: Conservatism in a Simple Probability Inference Task. Journal of Experimental Psychology, 1966.
5. Forgy, E. W.: Classification so as to Relate to Outside Variables. Final Rep. Conf. Cluster Analysis of Multivariate Data, Washington, 1966.
6. MacQueen, J. B.: Some Methods for Classifications and Analysis of Multivariate Observations. Proc. Symp. Math. Statist. and Probability, 5th. Berkeley, University of California Press, 1967.
7. Millen, G. A.: The Magical Number Seven, Plus or Minus Two in: The Psychology of Communication. Penguin Books, 1967.
8. Rabár, F. – Kelemen, K.: A központi és a vállalati beruházási politika szimulációja. A Számítógépes Rendszerszimuláció Szimpozium előadása, 1975.
9. Tversky, A. – Kahneman, D. K.: Subjective Probability. A Judgement of Representativeness. Cognitive Psychology, 1972.
10. Vári, A. – Kelemen, K.: Az OKGT strukturális döntéseinek vizsgálatára készült szimulációs modell formális és verbális leírása. INFELOR tanulmány, 1974.
11. Wishart, D.: An Algorithm for Hierarchical Classifications. Biometrics, 1969.

ECONOMIC APPLICATION OF CLUSTER ANALYSIS PROCEDURES

The article is aimed at presenting the application possibilities of statistical classification procedures – first of all cluster analysis – for the reduction of the complexity of economic system-modelling as well as at reviewing two less-known cluster analysis procedures.

In the first part some complexity problems are dealt with which emerge in the different phases of system modelling (model formation, validity test of the model, sensitivity analysis, evaluation of decision rules, revealing and learning of connections, etc.). A way to their solution is the application of statistical procedures.

In the second part the main types of statistical classification procedures and the theoretical possibilities of their application are briefly reviewed. *Wishart*'s hierarchic and *Diday*'s non-hierarchic cluster analysis procedures are discussed in detail. As a matter of fact the former is suitable for carrying out six different hierarchic methods and within the latter we can vary the distance function and the objective function (expressing the anality of the classification) within wide limits.

In the third part examples are presented for the application of classification procedures in economic modelling. In one of the examples we present the results of a simulation model, which was set up for the examination of structural decisions and sensitivity analysis. Another example deals with an investment simulation game, where the above procedures facilitated the model formation, the evaluation of results and learning.

Finally, we offer potential solutions to problems often arising in the course of the application of procedures.

ИСПОЛЬЗОВАНИЕ МЕТОДОВ КЛАСТЕРНОГО АНАЛИЗА ПРИ МОДЕЛИРОВАНИИ ЭКОНОМИЧЕСКИХ СИСТЕМ

Цель данной статьи заключается в том, чтобы показать возможности применения методов статистической классификации, в первую очередь кластерного анализа, в отношении ограничения комплексности экономического системного моделирования, а также изложить два относительно малоизвестных метода кластерного анализа.

В первой части затрагиваются проблемы комплексности, возникающие на различных фазах системного моделирования (разработка модели, изучение действенности модели, изучение чувствительности, оценка политики принятия решения, выявление взаимосвязей и их изучение и т. д.), одним из возможностей решения которых является использование статистических методов.

Во второй части дается краткий обзор основных типов методов статистической классификации и принципиальных возможностей их использования. Детально рассматриваются такие методы кластерного анализа как иерархический метод Вишарта и неиерархический метод Дидея. Первый из них, по существу, пригоден для применения шести различных иерархических методов, а во втором случае функция расстояния и функция, выражающая правильность классификации, могут изменяться в довольно широких пределах.

В третьей части приводятся примеры по использованию методов классификации в экономическом моделлировании. В одном из примеров излагается использование результатов анализа и исследования чувствительности симуляционной модели, составленной для изучения структурных решений. В другом примере описывается применение симуляционной игры по капитальным вложениям, когда указанные выше методы направлены на облегчение разработки модели, оценки результатов, а также и учебы.

В заключении, указываются проблемы, часто возникающие в ходе применения этих методов и излагаются некоторые возможности их решения.

FUTÓ PÉTER

A cluster analízis egy új modellje és algoritmusa

1. A cluster analízis gráf és hipergráf modelljei

A cluster analízis alapfeladata az, hogy objektumok és jellemzőik bonyolult rendszerének struktúráját feltárja; az objektumokat – előzetes ismeretek, tapasztalatok nélkül – kizárólag a jellemzőikből adódó kapcsolataik alapján természetes csoportokba ún. clusterekbe sorolja úgy, hogy az egymáshoz hasonló objektumok azonos clusterekbe, az egymáshoz kevésbé hasonló objektumok különböző clusterekbe kerüljenek.

A clusteranalízis feladatkörébe tartozik a jellemzők csoportosítása is, az általuk jellemzett objektumokból adódó kapcsolatok felhasználásával. A bonyolult rendszerek struktúrájának feltárása nagy jelentőségű és gyakori feladat az orvostudományban, biológiában, a közgazdaságtanban, a mérnöki tudományokban, az információ-tudományban és még sok más területen.

Tehát a cluster analízis modelljeinek, eljárásainak felhasználási területe igen széles. Ezt tükrözi terminológiájának heterogenitása is. A biológusok, orvosok numerikus taxonomiáról, a mérnökök tanító nélküli tanuló algoritmusokról, a statisztikusok, információtudományi szakemberek cluster analízisről, az operációkutatók partícionálási feladatról beszélnek, de valamennyien ugyanazon probléma megoldására, egy bonyolult rendszer struktúrájának feltárására törekszenek. A különböző felhasználási területeken alkalmazott modellek és eljárások összefonódásával a hetvenes évek elejére a cluster analízisnek két fő ága alakult ki: a matematikai-statisztikai és az információtudományi cluster elemzés.

A matematikai-statisztikai clusteranalízis a vizsgált rendszer valamennyi objektumát egy adott rendezett jellemző halmaz felhasználásával írja le. Tehát adott az objektumoknak a $D = \{d_1, \ldots d_m\}$ halmaza, és az objektumok mindegyikén megfigyelhető jellemzők $C = (C_1, \ldots C_p)$ vektora.

A jellemzők mérési skálájuk alapján négy fő csoportba sorolhatók.

Jelölje a C jellemzőnek a d_j, ill. d_k objektumra jellemző értékét $C(j)$, ill. $C(k)$.

Ha C nominális skálájú, akkor csak azt tudjuk, hogy $C(j) = C(k)$, vagy $C(j) \neq C(k)$ (pl. szem színe, születési hely).

Ha C ordinális skálájú, akkor azt tudjuk, hogy $C(j) = C(k)$ vagy $C(j) > C(k)$ vagy $C(j) < C(k)$ (pl. indiai kasztrendszerben elfoglalt hely).

Ha C intervallum skálájú, akkor ismerjük $C(j) - C(k)$ értékét (pl. hőmérséklet °C-ban).

Ha C hányados skálájú, akkor ismerjük $C(j)/C(k)$ értékét (pl. hőmérséklet °K-ban).

Hányados, ill. intervallum skálájú változók esetén nyilván nem mellékes a mértékegység megválasztása sem.

A matematikai-statisztikai cluster elemzés alapvető problémája a különböző skálájú jellemzők értékeinek felhasználásával az objektum-párokra hasonlósági mérőszám konstruálása. A probléma igen bonyolult, és sikeres megoldása a rendszer-struktúra feltárásának szükséges feltétele. Nem véletlen, hogy a cluster analízissel foglalkozó cikkek nagy hányada a hasonlósági mérőszám meghatározására szolgáló heurisztikus eljárások konstruálásával és összehasonlításával foglalkozik, szinte háttérbe szorítva a cluster kereső algoritmusok kidolgozását is (Anderberg [2]). Az információtudományi cluster elemzésnél adott az objektumoknak (általában dokumentumoknak) a $D = \{d_1, \ldots d_m\}$ halmaza, és az objektumok jellemzésére (indexelésére, leírására) használt tárgyszavaknak a $K = \{k_1, \ldots k_n\}$ halmaza.

Az objektumok jellemzésére a K halmaznak egy-egy általában nem rögzített elemszárú részhalmazát használják.

Az objektumok közötti hasonlóság és a hasonlósági mérőszám az objektumokat jellemző tárgyszó részhalmazok megegyező és egymástól különböző elemei számának valamilyen függvényén alapul.

A hasonlósági mérőszám konstruálása itt is igen bonyolult és döntő jelentőségű probléma. Ezt tükrözi a gyakorlatban használt mérőszámok széles skálája is.

A cluster elemzés mindkét ágában használt hasonlósági mérőszámok $[s(d_i, d_j)]$ két objektum között értelmezettek és a legtöbb esetben a következő tulajdonságokkal rendelkeznek:

1. $0 \leq s(d_i, d_j) \leq 1 \qquad (i, j = 1 \ldots m)$
2. $s(d_i, d_i) = 1 \qquad (i = 1 \ldots m)$
3. $s(d_i, d_j) = s(d_j, d_i) \qquad (i, j = 1 \ldots m)$

Ennek alapján megszerkeszthető a cluster analízis súlyozott gráf modellje, amelyen értelmezhető a gyakorlatban használt valamennyi cluster kereső algoritmus.

1. Modell

Adott a $D = \{d_1, \ldots d_m\}$ objektum halmaz, és valamennyi objektumpár esetén a hasonlósági mérőszámuk: $s(d_i, d_j)$ $(i, j = 1 \ldots m)$. Legyen $X = \{x_1, \ldots X_m\}$ egy gráf pontjainak a halmaza. Modelezze az x_i pont a d_i objektumot. Ha $d_i \neq d_j$ és $s(d_i, d_j) > 0$, akkor a gráf x_j és x_i pontja között irányítatlan élt (E_k) húzunk, amelyhez súlyként hozzárendeljük a $v(E_k) = s(d_i, d_j)$ értéket. Legyen $\varepsilon = \{E_1, \ldots E_m\}$. Tehát a $G = (X; \varepsilon)$ gráf és az élein értelmezett $v(E_j) > 0$ $(j = 1 \ldots n)$ súlyfüggvény modellezi az objektumokat és az objektumpárok hasonlósági mérőszámait.

A cluster kereső eljárásokat két nagy csoportba sorolhatjuk:

- hierarchikus eljárások,
- nem hierarchikus eljárások.

A hierarchikus eljárások a $G = (X; \varepsilon)$ gráf pontjai közül olyan részhalmazokat jelölnek ki, amelyek vagy egymást tartalmazzák, vagy diszjunktak. Minden esetben a kijelölt részhalmazok között lesznek a gráf pontjai is. Ezeket

a részhalmazokat a tartalmazási reláció felhasználásával hierarchiába rendezhetjük. (Innen adódik az eljárások neve is.) A hierarchikus eljárásokat két további csoportba sorolhatjuk:

– agglomeratív jellegű eljárások,
– nem agglomeratív jellegű eljárások.

Az agglomeratív jellegű eljárások a gráf pontjainak fokozatos összekapcsolásával alkotják a clustereket (pl. legközelebbi szomszéd módszere, centroid módszer, Ward-módszer) (*Sibson* [27], *Gower* [15], *Ward* [31]).

A nem agglomeratív jellegű eljárások a gráf fokozatos szétdarabolásával határozzák meg a clustereket. Ezek hatékony particionálási eljárások hiánya miatt egyenlőre nem nagyon elterjedtek, pl. Edwards és Cavalli-Sforza (*Anderberg* [2]).

A nem hierarchikus eljárások a $G = (X; \varepsilon)$ gráf pontjai közül olyan részhalmazokat (clustereket) jelölnek ki, amelyek diszjunktak (tehát a tartalmazás nem megengedett). Általában nem megkötés az, hogy a gráf valamennyi pontja legyen eleme legalább egy részhalmaznak.

A nem hierarchikus eljárásokat további két csoportba oszthatjuk:

– nem strukturális jellegű kritérium alapján osztályozó eljárások,
– strukturális kritérium alapján osztályozó eljárások.

A nem strukturális jellegű kritérium alapján osztályozó eljárásoknál általában előre adott a kívánt cluster szám, és valamilyen célfüggvény (pl. csoportokon belüli szórásnégyzetek összege), mint vezérfonal felhasználásával keresik a gráf pontjainak jobb elosztását (*MacQueen* [24], *Forgy* [10]).

A strukturális kritérium alapján működő eljárások nagy hányadánál súlyozott élű $G = (X; \varepsilon)$ gráfból küszöbszintek (pl. $t = 0{,}1; 0{,}3$) bevezetésével olyan $G = (X; \varepsilon_t)$ gráfokat állítanak elő, amelynél az x_i és x_j pontok között akkor húzódik él (amelyhez már nem rendelnek súlyt), ha $s(d_i, d_j) > t$. A $G = (X; \varepsilon_t)$ gráf(ok) komponenseit (*Auguston, Minker* [3]), vagy maximális teljes részgráfjait (*Osteen* [26]), vagy az ezekből képzett struktúrákat tekintik clustereknek. A strukturális kritériumok alapján működő eljárások közé sorolhatók a gráfelmélet particionálási módszerei is (*Lawler* [22]).

A cluster kereső eljárásokkal szemben a következő észrevételek tehetők:

– Több elem hasonlóságát csak elempárok hasonlóságával képesek kifejezni.
– Az algoritmusok futtatása előtt semmit, vagy csak igen keveset tudnak mondani az eredményként adódó clusterek tulajdonságairól. (Nincs explicit cluster definíció.)
– Az eljárások között nincs igazán hatékony, bizonyíthatóan az optimumhoz konvergáló algoritmus.

Jelenleg is széles körű nemzetközi kutatómunka folyik, amelynek célja a gyakorlatban jól használható egzakt cluster definíció megszerkesztése, valamint hatékony és konvergens cluster kereső algoritmusok kidolgozása. Ez a dolgozat a fenti kutatómunkát szeretné előbbre vinni a cluster elemzés hiper-

gráf modelljeinek megszerkesztésével, strukturális kritériumon alapuló cluster definíció, és olyan polinommal fedhető lépésszámú, konvergens, hierarchikus cluster eljárás kidolgozásával, amelynek alkalmazásával elkerülhető a hasonlósági mátrix megalkotása.

A szerzőt a hipergráf modellek megalkotására az információtudományi cluster elemzés gráf modelljének és eljárásainak kritikája sarkallta, míg a hipergráf kvázi-komponense fogalmának megalkotásához – ami az új cluster definíció és eljárás alapköve – *Lawler* [22] cikke adta az ötletet, aki *Luccio* és *Sami* [23] eredményeinek felhasználásával észrevette, hogy vannak a hipergráfoknak olyan ponthalmazai, amelyek a minimális két részre vágás során nem vágódnak el. Lawler erre a felismerésre alapozva hatékony heurisztikus eljárásokat dolgozott ki hipergráfok több részre vágására, de nem foglalkozott mélyebben az el nem vágódó ponthalmazok tulajdonságaival.

Az információtudományi cluster elemzésben két objektumot akkor tartanak hasonlónak, ha a jellemzésükre (indexelésükre) használt deszkriptorok (tárgyszavak) közül legalább egy közös.

Ez a bináris reláció nyilván szimmetrikus, reflexív, de nem tranzitív, tehát tolerancia reláció (*Srejder* [29]). Ez a reláció nyilván egy olyan több elemű relációból származik, amelynél minden egyes deszkriptor kapcsolatot létesít azon (nem feltétlenül kettő) objektumok között, amelyek jellemzésére az adott deszkriptort felhasználták.

Ezzel teljesen analóg módon értelmezhető egy több elemű reláció a deszkriptorok között is úgy, hogy minden egyes objektum kapcsolatot létesít a jellemzésére használt deszkriptorok között.

A matematikai-statisztikai cluster elemzésben az objektumok közötti hasonlóság fogalmát általában az objektumokat jellemző vektorok között definiált valamilyen távolság fogalmából vezetik le. Tehát a hasonlósági reláció itt is bináris, mégpedig szimmetrikus, reflexív és általában nem tranzitív (tolerancia) reláció. Ez a tolerancia reláció is nyilván egy olyan több elemű relációból származik, ahol az objektumok közötti hasonlóság alapja az, hogy egy vagy több jellemzőjük értéke nagyon közeli, vagy megegyezik.

Ezen alapul az a gondolat, hogy az objektumok jellemzésére szolgáló mátrix cellái értékeinek felhasználásával itt is deszkriptorokat alakítsunk ki. Például deszkriptor lehet az, hogy egy adott jellemző értéke egy adott intervallumba esik. A feladat természetétől függően deszkriptorokat definiálhatunk úgy is, hogy több jellemző értékét szorítjuk határok közé.

A deszkriptorok definiálásánál nem kikötés az, hogy segítségükkel az objektumoknak egy osztályozását hozzuk létre; tehát megengedhetjük azt is, hogy például a „súly 1" deszkriptor a 10 kp és a 15 kp közötti súlyú objektumok jellemzésére szolgáljon, míg a „súly-hossz" deszkriptor a 14 kp és a 16 kp közötti súlyú és a 1 m és a 2 m közötti hosszúságú objektumok jellemzésére szolgáljon.

Természetesen csak olyan deszkriptorokat definiálunk, amelyek legalább egy objektum jellemzésére szolgálnak, és a definiált deszkriptorok között található minden egyes objektumhoz legalább egy, amely az illető objektum jellemzésére szolgál.

Bár jelentősége nem olyan nagy mint az információtudományi cluster elemzésben, de itt is definiálható a deszkriptorok között egy több elemű reláció

úgy, hogy minden egyes objektum kapcsolatot létesít a jellemzésére használt deszkriptorok között.

A fentiek alapján mind az információtudományi, mind a matematikai-statisztikai cluster elemzésben jól használható a következő két hipergráf modell.

2. *Modell*

Adott a $D = \{d_1, \ldots d_m\}$ objektum halmaz. Jelölje a d_j objektum jellemzésére használt deszkriptorok halmazát $E_j (j = 1, \ldots m)$. Mivel az objektumok jellemzésére legalább egy, de véges sok deszkriptort használnak, ezért

$$1 \leq | E_j | \leq K \quad (j = 1, \ldots m). \tag{1}$$

Az objektum halmaz objektumainak jellemzésére szolgáló deszkriptorok halmazát jelölje: X

$$X = \bigcup_{j=1}^{m} E_j \tag{2}$$

Ha ε jelöli az objektumok jellemzésére szolgáló deszkriptor halmazok osztályát: $\varepsilon = \{E_1, \ldots E_m\}$, akkor $H = (X; \varepsilon)$ hipergráf, ugyanis (vö. 1. Definíció).

(I) $X = \{x_1, \ldots x_n\}$ véges halmaz (1), (2) miatt,
(II) $E_j \neq \emptyset (j = 1, \ldots m)$ (1) miatt,
(III) $\bigcup_{j=1}^{m} E_j =$ (2) miatt.

A $H = (X, \varepsilon)$ hipergráf pontjai tehát deszkriptorok, élei pedig az egy-egy objektum jellemzésére szolgáló deszkriptor halmazok.

3. *Modell*

A $H = (X; \varepsilon)$ hipergráf duálisa a $H^* = (E; X_1, \ldots X_n)$ hipergráf, amelynek pontjai $(e_1, \ldots e_m)$ a $H = (X; \varepsilon)$ hipergráf élet $(E_1, \ldots E_m)$ reprezentálják, élei pedig $(X_1, \ldots X_n)$ a $H = (X; \varepsilon)$ hipergráf pontjainak felelnek meg a következő értelemben:

$$X_i = \{e_j \mid j \leq m, \; x_i \in E_j\}. \tag{3}$$

$H^* = (E; \mathfrak{X})$ valóban hipergráf, ugyanis

(I′) $E = \{e_1, \ldots e_m\}$ véges halmaz (I) miatt,
(II′) $X_i \neq \emptyset (i = 1, \ldots n)$ (III) és (3) miatt.
(III′) $\bigcup_{i=1}^{m} X_i = E$ (II) és (3) miatt.

A $H^* = (E; \mathfrak{X})$ hipergráf pontjai objektumok, élei pedig az egyes deszkriptorok által meghatározott olyan objektum halmazok, amelyek objektumai jellemzésére az adott deszkriptort felhasználtuk.

A modellezés mindkét modell esetén az éleken értelmezett pozitív súlyfüggvény $v(E_j) > 0$ $(j = 1 \ldots m)$, illetve $u(X_i) > 0$ $(i = 1, \ldots n)$ bevezetésével finomítható. Ha finomításra nincs szükség, vagy nem lehetséges, akkor is bevezetünk egy súlyfüggvényt, mégpedig úgy, hogy minden egyes élhez az 1 súlyt rendeljük hozzá.

A dolgozat 2. része azt a strukturális kritériumon alapuló cluster definíció ismerteti, amely a cluster elemzés mindhárom fent említett modelljére sikerrel alkalmazható.

2. Matematikai alapok. A hipergráf kvázi-komponensének fogalmán alapuló új cluster definíció

A dolgozatban szereplő – hasonlósági mérőszámot nem használó – cluster eljárás a csoportosítandó objektumok és a jellemzésükre használt deszkriptorok kapcsolatait feltáró hipergráf modelleken, valamint a cluster definícióként alkalmazott kvázi-komponens fogalmán alapul.

A dolgozatnak ez a része tartalmazza azokat a matematikai alapokat, amelyek a kvázi-komponens definíciójához és a cluster elemzés szempontjából fontos tulajdonságainak leírásához szükségesek. A kvázi-komponens definícióját követik azok a megjegyzések, lemmák, tételek, amelyek alátámasztják a kvázi-komponens fogalom alkalmazhatóságát a cluster elemzésben. A dolgozatban csak azokat a tételeket bizonyítjuk, amelyek részletes bizonyítása a *Futó* [13] dolgozatban nem szerepel.

2.1. Definíció: Adott az $X = \{x_1, \ldots, x_n\}$ véges halmaz és $\varepsilon = \{E_1, \ldots, E_m\}$ az X halmaz részhalmazainak osztálya.

A $H = (X;\varepsilon)$ pár *hipergráf*, ha

$$E_j \neq \emptyset \qquad (j = 1, \ldots m), \tag{1}$$

$$\bigcup_{j=1}^{m} E_j = X. \tag{2}$$

Az X halmaz elemeit pontoknak, az ε halmaz elemeit éleknek nevezzük.

Ha $X = \emptyset$, akkor a hipergráfot üresnek nevezzük.

Értelmezzük a hipergráf ponthalmazai és élhalmazai között az $\mathcal{E}$: $2^X \to 2^\varepsilon$ és az $\mathcal{H}$: $2^\varepsilon \to 2^X$ leképezéseket:

2.2. Definíció: Tetszőleges S ponthalmaz $(S \subseteq X)$ esetén

$$\mathcal{E}(S) = \{E_j | E_j \in \varepsilon,\ \exists x_i \in S\colon\ x_i \in E_j\}.$$

2.3. Definíció: Tetszőleges $\mathcal{F}$ élhalmaz $(\mathcal{F} \subseteq \varepsilon)$ esetén

$$\mathcal{H}(\mathcal{F}) = \{x_i | x_i \in X,\ \exists E_j \in \mathcal{F}\colon\ x_i \in E_j\}.$$

2.1. Megjegyzés: Egyszerűen belátható, hogy tetszőleges S ponthalmaz $(S \subseteq X)$ és $\mathcal{F}$ élhalmaz $(\mathcal{F} \subseteq \varepsilon)$ választása esetén $S \subseteq \mathcal{H}\big(\mathcal{E}(S)\big)$ és $\mathcal{F} \subseteq \mathcal{E}\big(\mathcal{H}(\mathcal{F})\big)$, tehát az $\mathcal{E}$: $2^X \to 2^\varepsilon$ és az $\mathcal{H}$: $2^\varepsilon \to 2^X$ leképezések egymásnak nem inverzei.

Új fogalmak bevezetését, tételek egyszerűbb bizonyítását és a kvázi-komponensek meghatározására szolgáló algoritmus gyorsítását teszi lehetővé az $\mathscr{E}'\colon 2^X \otimes 2^X \to 2^\varepsilon$ leképezés, amely az $\mathscr{E}\colon 2^X \to 2^\varepsilon$ leképezés általánosítása.

2.4. Definíció: Tetszőleges S és T ponthalmazok $(S \subseteq X)$, $(T \subseteq X)$ esetén

$$\mathscr{E}'(S|T) = \{E_j | E_j \in \varepsilon;\ \exists x_i \in S\colon x_i \in E_j,\ E_j \subseteq T\}.$$

2.2. Megjegyzés: Egyszerű számolással bizonyíthatók az $\mathscr{E}'\colon 2^X \otimes 2^X \to 2^\varepsilon$ leképezés következő tulajdonságai, amelyeket a továbbiakban gyakran felhasználunk:

Ha $S \subseteq X$ és $T = X$, akkor $\mathscr{E}'(S|T) = \mathscr{E}(S)$.
Ha $T \subseteq S \subseteq X$, akkor $\mathscr{E}'(S|T) = \mathscr{E}'(T|T)$.
Ha $S \subset S' \subseteq X$ és $T \subseteq X$, akkor $\mathscr{E}'(S|T) \subseteq \mathscr{E}'(S'|T)$.
Ha $S \subseteq X$ és $T \subset T' \subseteq X$, akkor $\mathscr{E}'(S|T) \subseteq \mathscr{E}'(S|T')$.

A gráfelméleti szakirodalomban széleskörűen használt az élhalmaz által generált rész-hipergráf és a ponthalmaz által generált alhipergráf fogalma.

2.5. Definíció: A $H = (X; \varepsilon)$ hipergráfnak az $\mathscr{F}$ élhalmaz $(\mathscr{F} \subseteq \varepsilon)$ által *generált rész-hipergráfja:* $H = \big(\mathscr{H}(\mathscr{F}); \mathscr{F}\big)$.

2.6. Definíció: A $H = (X; \varepsilon)$ hipergráfnak az S ponthalmaz $(S \subseteq X)$ által *generált alhipergráfja:* $H = (S; \varepsilon_S)$, ahol

$$\varepsilon_S = \{E_i \cap S | E_i \in \mathscr{E}(S)\}.$$

A generált rész-hipergráf vagy a generált alhipergráf fogalom alkalmazása további munkánkat indokolatlanul elbonyolítaná. Ugyanis a kvázi-komponensek tulajdonságainak leírásához, és a megkeresésükre szolgáló eljáráshoz is olyan rész-hipergráf fogalom szükséges, amelyet ponthalmaz segítségével definiálunk. A rész-hipergráfnak mindazokat és csak azokat az éleket kell tartalmaznia, amelyek a definiáló ponthalmaznak részei.
Ezeket a követelményeket elégíti ki a következő definíció.

2.7. Definíció: A $H = (X; \varepsilon)$ hipergráfnak az S ponthalmaz $(S \subseteq X)$ felhasználásával *kifeszített rész-hipergráfja:*

$$H = \big(\mathscr{H}(\mathscr{E}'(S|S);\ \mathscr{E}'(S|S)\big).$$

Nyilvánvaló, hogy az $(S;\ \mathscr{E}'(S|S)$ pár nem lett volna jó definíció, ugyanis az S halmaznak lehet olyan pontja, amelyet egyetlen él sem tartalmaz. Ezzel szemben $H = \big(\mathscr{H}(\mathscr{E}'(S|S));\ \mathscr{E}'(S|S)\big)$ valóban hipergráf (nincs üres éle és izolált pontja), de ponthalmaza nem feltétlenül egyezik meg S-sel.

2.3. Megjegyzés: Könnyen belátható, hogy $S \subseteq X$ esetén $\mathscr{H}\big(\mathscr{E}'(S|S)\big) \subseteq S$.
A továbbiakban jelöljük a $H = \big(\mathscr{H}(\mathscr{E}'(S|S));\ \mathscr{E}'(S|S)\big)$ hipergráfot röviden H_S-sel.

2.4. Megjegyzés: A H_S kifeszített rész-hipergráf megegyezik az $\mathscr{E}'(S|S)$ élhalmaz által generált rész-hipergráffal.

Mivel a dolgozat következő részeiben csak ponthalmazok által kifeszített rész-hipergráfokkal dolgozunk, ezért ezeket a továbbiakban röviden csak rész hipergráfoknak nevezzük.

2.8. Definíció: A H_S rész-hipergráfban a K ponthalmaz $[K \subseteq \mathscr{H}(\mathscr{E}'(S|S))]$ *hipergráfot kifeszítő* (vagy röviden kifeszítő), ha $\mathscr{H}\big(\mathscr{E}'(K|K)\big) = K$.

Az elnevezést indokolja az, hogy a kifeszítő ponthalmaz megegyezik a felhasználásával kifeszített rész-hipergráf ponthalmazával, azaz $K = \mathscr{H}\big(\mathscr{E}'(K|K)\big)$ miatt $H_K = \big(K;\ \mathscr{E}'(K|K)\big)$.

2.1. Lemma: A H_S rész hipergráfban a K ponthalmaz $\big(K \subseteq \mathscr{H}(\mathscr{E}'(S|S))\big)$, akkor és csak akkor kifeszítő, ha $\exists \mathscr{F}\big(\mathscr{F} \subseteq \mathscr{E}'(S|S)\big)$ élhalmaz, amelyre $K = \mathscr{H}(\mathscr{F})$.

2.5. Megjegyzés: A 2.1. Lemma egyszerű következménye az, hogy egy K ponthalmaz $(K \subseteq X)$ vagy kifeszítő minden olyan H_S rész-hipergráfban, amelyre $K = \mathscr{H}\big(\mathscr{E}'(S|S)\big)$, vagy egyikben sem kifeszítő.

Ez indokolja, hogy a továbbiakban csak kifeszítő ponthalmazról fogunk beszélni (nem tesszük hozzá, hogy a $H = (X;\ \varepsilon)$ hipergráf melyik rész-hipergráfjában), és a K-val jelölt ponthalmazok mindig kifeszítők lesznek.

2.2. Lemma: Ha $S \subseteq X$, akkor $\exists K$ kifeszítő ponthalmaz $\big(K = \mathscr{H}(\mathscr{E}'(S|S))\big)$, amelyre $H_S = H_K$. Az $\mathscr{H}\big(\mathscr{E}'(S|S)\big)$ halmaz az S által tartalmazott maximális kifeszítő ponthalmaz.

2.6. Megjegyzés: A 2.2. Lemma egyszerű következménye, hogy az általánosság megszorítása nélkül feltehetjük bármelyik kifeszített rész-hipergráfról, hogy az egy kifeszítő ponthalmaz felhasználásával keletkezett.

Hasonlóan az $\mathscr{H}\big(\mathscr{E}'(S|S)\big)$ és az S halmaz közötti reláció elemzéséhez (amely a kifeszítő halmaz fogalmának bevezetését eredményezte), $K \supseteq S$ esetén az $\mathscr{H}\big(\mathscr{E}'(S|K)\big)$ és az S halmaz kapcsolatának vizsgálata vezet el a komponens fogalmához. Jelöljük az S halmaz $(S \subseteq X)$ valódi részhalmazainak osztályát $\mathscr{S}_S$-sel:

$$\mathscr{S}_S = \{T|T \neq \emptyset,\ T \subset S\} = 2^S - \{S\} - \{\emptyset\}$$

2.9. Definíció: A H_K rész-hipergráfban a P ponthalmaz $(O \neq P \subseteq K)$ *komponens*, ha

(1) $$\mathscr{H}\big(\mathscr{E}'(P|K)\big) = P,$$

(2) $$T \in \mathscr{S}_P \text{ esetén } \mathscr{H}\big(\mathscr{E}'(T|K)\big) \supset T$$

2.7. Megjegyzés: Ha a H_K rész-hipergráfban a P ponthalmaz komponens, akkor kifeszítő. $[\mathscr{E}'(P|K) = \mathscr{F}$ választással a 2.1. Lemma alkalmazásával adódik.]

2.8. Megjegyzés: A 2.7. Megjegyzés miatt $\mathscr{H}\big(\mathscr{E}'(P|P')\big) = P$. Ezért a 2.2. Lemma alkalmazásával egyszerűen belátható, hogy ha a P halmaz kompo-

nense a H_K rész-hipergráfnak azaz $\mathcal{H}(\mathcal{E}'(P|K)) = P$, akkor $P \subseteq K' \subset K$ esetén $\mathcal{H}(\mathcal{E}'(P|K')) = P$, azaz P komponense a $H_{K'}$ rész-hipergráfnak is.

2.10. Definíció: A H_K rész-hipergráf összefüggő, ha $T \in \mathcal{S}_K$ esetén

$$\mathcal{H}(\mathcal{E}'(T|K)) \supset T.$$

A komponens és az összefüggő rész-hipergráf definíciója már mutatja, hogy hasznos volt a hipergráf ponthalmazai és élhalmazai között értelmezett leképzések bevezetése. Ugyanis a fenti definíciók nyilvánvalóan megegyeznek a szakirodalomban használt definíciókkal, amelyek az út fogalom bevezetése miatt bonyolultabbak.

A kvázi-komponens definíciójához és a kvázi-komponenseket meghatározó eljáráshoz egyaránt szükséges a vágás most következő definíciója.

2.11. Definíció: A H_K rész-hipergráfnak a T *ponthalmaz* ($T \subseteq K$) *által generált vágása:* [jelölése $C_K(T)$].

$$C_K(T) = \mathcal{E}'(T|K) \cap \mathcal{E}'((K - T)|K).$$

Nyilvánvaló, hogy a T és a $K-T$ halmazok által generált vágások megegyeznek, és hogy az üres halmaz és a K által generált vágás mindig az üres halmaz.

2.3. Lemma: $C_K(T) = \mathcal{E}'(T|K) - \mathcal{E}'(T|T)$.

2.4. Lemma: Legyen T ($T \subseteq K$) tetszőleges ponthalmaza a H_K rész-hipergráfnak.
$\mathcal{H}(\mathcal{E}'(T|K)) = T$ akkor és csak akkor, ha $C_K(T) = \emptyset$.

2.5. Lemma: A P ponthalmaz ($P \subseteq K$) akkor és csak akkor komponense a H_K rész-hipergráfnak, ha

(1′) $$C_K(P) = \emptyset,$$

(2′) $$T \in \mathcal{S}_P \text{ esetén } C_K(T) \supset \emptyset.$$

A komponens itt közölt alternatív definíciójának általánosításán alapul a kvázi-komponens definíciója. További vizsgálatainkhoz értelmezzük a hipergráf élein a $v(E_j) > 0$ ($j = 1, \ldots m$) pozitív függvényt. Terjesszük ki a függvény értelmezési tartományát élhalmazokra is. Vezessük be a $w: 2^{\varepsilon} \to R^+$ leképezést, amelynek révén még diszjunkt élhalmazokat is össze tudunk hasonlítani.

2.12. Definíció: Tetszőleges $\mathcal{F}$ élhalmaz ($\mathcal{F} \subseteq \varepsilon$) esetén $w(\mathcal{F}) = \sum_{E_j \in \mathcal{F}} v(E_j)$.

2.9. Megjegyzés: Egyszerű számolással adódnak a $w: 2^{\varepsilon} \to R^+$ függvény következő tulajdonságai, amelyeket a továbbiakban gyakran felhasználunk:

Ha $\mathcal{F} \subseteq \mathcal{G} \subseteq \varepsilon$, akkor $w(\mathcal{F}) < w(\mathcal{G})$.
Ha $\mathcal{F} \subseteq \mathcal{G} \subseteq \varepsilon$, akkor $w(\mathcal{G} - \mathcal{F}) = w(\mathcal{G}) - w(\mathcal{F})$.
Ha $\mathcal{F} \subseteq \varepsilon, \mathcal{G} \subseteq \varepsilon$, akkor $w(\mathcal{F} \cup \mathcal{G}) = w(\mathcal{F}) + w(\mathcal{G}) - w(\mathcal{F} \cap \mathcal{G})$.

Az élhalmazokon értelmezett függvény módot ad a vágások értékének definiálására is.

2.13. Definíció: A H_K rész-hipergráf T ponthalmaza ($T \subseteq K$) által generált vágásának *értéke* [jelölése: $\overline{w}_K(T)$]:

$$\overline{w}_K(T) = w[C_K(T)] = w[\mathcal{E}'(T|K) \cap \mathcal{E}'\big((K - T)|K\big)].$$

A kvázi-komponens definíciója a komponens utolsó definíciójának általánosítása.

2.14. Definíció: A H_K rész-hipergráfban a Q ponthalmaz ($\emptyset \neq Q \subseteq K$) *kvázi-komponens*, ha bármely $T \in \mathcal{S}_Q$ választása esetén $\overline{w}_K(Q) < \overline{w}_K(T)$.

2.10. Megjegyzés: A kvázi-komponens fogalom valóban a komponens fogalmának általánosítása, ugyanis 2.5. Lemma felhasználásával triviális, hogy a H_K rész-hipergráf valamennyi komponense egyben kvázi-komponense is.

2.11. Megjegyzés: A H_K rész hipergráfban a Q ponthalmaz ($Q \subseteq K$) kvázi-komponens, ha $|Q| = 1$, ugyanis ez esetben $\mathcal{S}_Q = \emptyset$.

Az egy elemű kvázi-komponenseket a továbbiakban *triviális* kvázi-komponenseknek nevezzük.

2.6. Lemma: Ha a H_K rész-hipergráfban a Q ponthalmaz nem triviális kvázi-komponens ($|Q| \geq 2$, $Q \subseteq K$), akkor Q kifeszítő.

A 2.6. Lemma egyszerű, de a továbbiak során gyakran felhasznált következményét ismerteti a következő megjegyzés.

2.12. Megjegyzés: Ha a Q ponthalmaz ($Q \subseteq K$) nem triviális kvázi-komponens H_K-ban, akkor legalább egy élt tartalmaz, és $x_i \in Q$ esetén $\exists E_j \in \mathcal{E}'(Q|Q)$, amelyre $x_i \in E_j$ (Ugyanis a Q ponthalmaz kifeszítő.)

A következő tétel a kvázi-komponenseknek a cluster elemzés szempontjából nagyon fontos tulajdonságát fejezi ki. Azt mondja ki, hogy a nem-triviális kvázi-komponens bármely valódi részhalmaza „erősebben kapcsolódik" a kvázi-komponens többi részéhez, mint a kvázi-komponens teljes környezetéhez.

2.7. Tétel: A H_K rész-hipergráfban a Q ponthalmaz ($Q \subseteq K$) akkor és csak akkor nem-triviális kvázi-komponens, ha bármely $T \in \mathcal{S}_Q$ ponthalmaz választása esetén:

$$w[\mathcal{E}'(T|Q)] > w[\mathcal{E}'\big(T|(K-(Q-T))\big)].$$

2.13. Megjegyzés: A 2.7. Tétel tovább nem élesíthető, ugyanis $T = Q$ választása esetén a 2.2. és 2.9. Megjegyzések felhasználásával triviálisan adódik, hogy $w[\mathcal{E}'(Q|Q)] \leq w[\mathcal{E}'(Q|K)]$.

2.14. Megjegyzés. A 2.7. Tétel felhasználásával egyszerű számolással adódik a kvázi-komponensek egyik fontos tulajdonsága: Ha a Q halmaz kvázi-komponense a H_K rész-hipergráfnak, akkor $Q \subseteq K' \subset K$ esetén kvázi-komponense a $H_{K'}$ rész-hipergráfnak is.

Ugyanis a triviális kvázi-komponensekre az állítás nyilvánvaló, a nem-triviálisra pedig a 2.2. és 2.9. Megjegyzést felhasználva a következő adódik:

$$w\,\mathcal{E}'(T|Q)] > w[\mathcal{E}'(T|(K-(Q-T)))] \geq w[\mathcal{E}'(T|(K'-(Q-T)))].$$

2.15. Megjegyzés: Ha Q nem-triviális kvázi-komponense a H_K rész-hipergráfnak, akkor a H_Q rész-hipergráf összefüggő. Ugyanis a 2.6. Lemma alapján Q kifeszítő. 2.14. Megjegyzést $Q = K'$-re alkalmazva $w[\mathcal{E}'(T|Q)] > w[\mathcal{E}'(T|T)]$ adódik tetszőleges $T \in \mathcal{S}_Q$ esetén. Ez pedig a 2.4. Lemma alapján pontosan azt jelenti, hogy H_Q összefüggő.

A most következő tétel alapvető jelentőségű a hipergráf összes kvázi-komponense meghatározására szolgáló hatékony algoritmus konstruálásához.

2.8. Tétel: Legyen $K(|K|) \geq 2$) a $H = (X; \varepsilon)$ hipergráfnak egy kifeszítő ponthalmaza, és S olyan ponthalmaz, amelyre teljesül az, hogy $S \subseteq K$ és $|S| \geq 2$. Legyen T^* az a ponthalmaz ($T^* \in \mathcal{S}_S = \{T|T \neq \emptyset; T \subset S\}$), amelyre teljesül az, hogy bármely $T \in \mathcal{S}_S$ esetén $\overline{w}_K(T^*) \leq \overline{w}_K(T)$.

Legyen $Q(Q \subset S)$ tetszőleges kvázi-komponense a $H = (X; \varepsilon)$ hipergráfnak. Ekkor $Q \subseteq T^*$ vagy $Q \subseteq S - T^*$ teljesül.

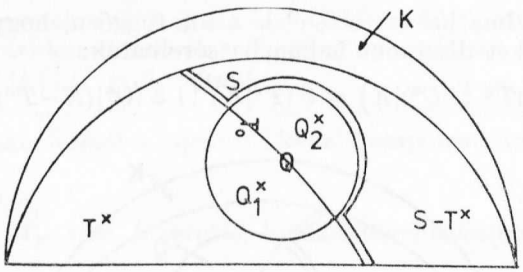
Biz.: Ha Q triviális kvázi-komponens, akkor a tétel állítása triviális.

Tegyük fel, hogy létezik olyan $Q(Q \subset S)$ nem-triviális kvázi-komponense a $H = (X; \varepsilon)$ hipergráfnak, amelyre $Q_1^* = Q \cap T^* \neq \emptyset$ és $Q_2^* = Q \cap (S - T^*) \neq \emptyset$.

Mivel $Q \subset S$, ezért $Q \supseteq T^*$ és $Q \supseteq S - T^*$ egyszerre nem teljesülhet.

1. Eset: Tegyük fel, hogy $Q \supseteq S - T^*$. Legyen $T^0 = T^* \cup Q = T^* \cup Q_2^*$. $T^* \subseteq T^0$ miatt $T^0 \neq \emptyset$ és $Q \supseteq S - T^*$ miatt $T^0 \subset S$, tehát $T^0 \in \mathcal{S}_S$.

Számítsuk ki $\overline{w}_K(T^0)$ értékét $\overline{w}_K(T^*)$ függvényében: $\overline{w}_K(T^0) = w[\mathcal{E}'(T^0|K)] - w[\mathcal{E}'(T^0|T^0)]$.



1. ábra

Az $\mathcal{E}'(T^0|K)$ halmazba tartozó élek attól függően, hogy tartalmaznak-e T^*-beli pontot, két diszjunkt halmazba sorolhatók:

$$\mathcal{E}'((T^* \cup Q_2^*)|K) = \mathcal{E}'(T^*|K) \cup \mathcal{E}'(Q^*|(K - T^*)).$$

Ebből a halmaz egyenlőségből a következő skalár egyenlőség adódik:

$$(1) \qquad w[\mathcal{E}'((T^* \cup Q_2^*)|K)] = w[\mathcal{E}'(T^*|K)] + w[\mathcal{E}'(Q^*|(K-T^*))].$$

Az $\mathcal{E}'(T^0|T^0)$ halmazba tartozó élek attól függően, hogy tartalmaznak-e Q_2^*-beli pontot két diszjunkt halmazba sorolhatók:

$$\mathcal{E}'((T^* \cup Q_2^*)|(T^* \cup Q_2^*)) = \mathcal{E}'(T^*|T^*) \cup \mathcal{E}'(Q_2^*|(T^* \cup Q_2^*)).$$

Ebből a halmaz egyenlőségből a következő skalár egyenlőség adódik.

$$(2) \qquad w[\mathcal{E}'(T^0|T^0)] = w[\mathcal{E}'(T^*|T^*)] + w[\mathcal{E}'(Q_2^*|(T^* \cup Q_2^*))].$$

A (2) egyenlőséget az (1) egyenlőségből kivonva kapjuk, hogy $\overline{w}_K(T^0) = \overline{w}_K(T^*) + w[\mathcal{E}'(Q_2^*|(K-T^*))] - w[\mathcal{E}'(Q_2^*|(T^* \cup Q_2^*))]$.
$\overline{w}_K(T^*) \leq \overline{w}_K(T^0)$ miatt:
$w[\mathcal{E}'(Q_2^*|(K-T^*))] \geq w[\mathcal{E}'(Q_2^*|(T^* \cup Q_2^*))]$
$K-T^* \subseteq K-(Q-Q_2^*)$ és $Q \subseteq T^* \cup Q_2^*$ felhasználásával.

(3) $w[\mathcal{E}'(Q_2^*|(K-(Q-Q_2^*)))] \geq w[\mathcal{E}'(Q_2^*|Q)]$ adódik. Mivel $Q_1^* \neq \emptyset$ és $Q_2^* \neq \emptyset$, tehát $Q_2^* \in \mathcal{S}_Q$.
Az 5. Tétel következménye miatt Q kvázi-komponense a $H_K = (K; \mathcal{E}'(K|K))$ rész-hipergráfnak is. A 2.7. Tétel miatt ekkor Q_2^*-re teljesülnie kell a következő egyenlőtlenségnek:

$$(4) \qquad w[\mathcal{E}'(Q_2^*|(K-(Q-Q_2^*)))] < w[\mathcal{E}'(Q_2^*|Q)].$$

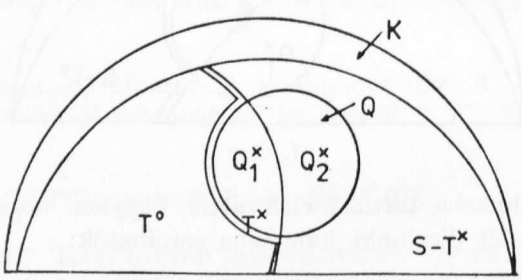
amely ellentmond a (3) egyenlőtlenségnek.

2. Eset: Tegyük fel, hogy $Q \not\supseteq T^*$. Legyen $T^0 = T^* - Q = T^* - Q_1^*$, azaz $T^* = T^0 \cup Q_1^*$. $T^0 \subseteq T^*$ miatt $T^0 \subset S$, és $Q \not\supseteq T^*$ miatt $T^0 \neq \emptyset$, tehát $T^0 \in \mathcal{S}_S$.
Számítsuk ki $\overline{w}_K(T^*)$ értékét $\overline{w}_K(T^0)$ függvényében:

$$\overline{w}_K(T^*) = w[\mathcal{E}'(T^*|K)] - w[\mathcal{E}'(T^*|T^*)].$$

Az $\mathcal{E}'(T^*|K)$ halmazba tartozó élek attól függően, hogy tartalmaznak-e T^0-beli pontot két diszjunkt halmazba sorolhatók:

$$\mathcal{E}'((T^0 \cup Q_1^*)|K) = \mathcal{E}'(T^0|K) \cup \mathcal{E}'(Q_1^*|(K-T^0)).$$



2. ábra

Ebből a halmaz egyenlőségből a következő skalár egyenlőség adódik:

$$w[\mathcal{E}'(T^*|K)] = w[\mathcal{E}'(T^0|K)] + w[\mathcal{E}'\big(Q_1^*|(K-T^0)\big)]. \tag{5}$$

Az $\mathcal{E}'(T^*|T^*)$ halmazba tartozó élek attól függően, hogy tartalmaznak-e Q_1^*-beli pontot két diszjunkt halmazba sorolhatók:

$$\mathcal{E}'\big((T^0 \cup Q_1^*)|(T^0 \cup Q_1^*)\big) = \mathcal{E}'(T^0|T^0) \cup \mathcal{E}'\big(Q_1^*|(T^0 \cup Q_1^*)\big)$$

Ebből a halmaz egyenlőségből a következő skalár egyenlőség adódik:

$$w[\mathcal{E}'(T^*|T^*)] = w[\mathcal{E}'(T^0|T^0)] + w[\mathcal{E}'\big(Q_1^*|(T^0 \cup Q_1^*)\big)]. \tag{6}$$

A (6) egyenlőséget az (5) egyenlőségből kivonva kapjuk, hogy $\overline{w}_K(T^*) = \overline{w}_K(T^0) + w[\mathcal{E}'\big(Q_1^*|(K-T^0)\big)] - w[\mathcal{E}'\big(Q_1^*|(T^0 \cup Q_1^*)\big)]$.

$\overline{w}_K(T^*) \leq \overline{w}_K(T^0)$ miatt

$w[\mathcal{E}'\big(Q_1^*|(K-T^0)\big)] \leq w[\mathcal{E}'\big(Q_1^*|(T^0 \cup Q_1^*)\big)]$.

$Q \subseteq K-T^0$ és $T^0 \cup Q_1^* \subseteq K-(Q-Q_1^*)$ felhasználásával.

$$w[\mathcal{E}'(Q_1^*|Q)] \leq w[\mathcal{E}'\big(Q_1^*|(K-(Q-Q_1^*))\big)] \text{ adódik.} \tag{7}$$

Mivel $Q_1^* \neq \emptyset$ és $Q_2^* \neq \emptyset$, tehát $Q_1^* \in \mathcal{S}_Q$.

Az 5. Tétel következménye miatt Q kvázi-komponense a $H_K = \big(K; \mathcal{E}'(K|K)\big)$ rész-hipergráfnak is. A 2.7. Tétel miatt ekkor Q_1^*-ra teljesülnie kell a következő egyenlőtlenségnek:

$$w[\mathcal{E}'(Q_1^*|Q)] > w[\mathcal{E}'\big(Q_1^*|(K-(Q-Q_1^*))\big)], \tag{8}$$

amely ellentmond a (7) egyenlőtlenségnek. Q.E.D.

A következő tétel a hipergráf kvázi-komponensei közötti relációra mutat rá. Azt fejezi ki, jogy a kvázi-komponensek vagy tartalmazzák egymást vagy diszjunktak.

2.11. Tétel: Legyenek a $Q(Q \subseteq K)$ és a $Q'(Q' \subseteq K)$ ponthalmazok egymástól különböző kvázi-komponensek a H_K rész-hipergráfban. Ekkor vagy $Q \subset Q'$, $Q' \subset Q$, vagy $Q \cap Q' = \emptyset$ teljesül.

A most következő tétel a hipergráf kvázi-komponenseinek számára ad felső korlátot.

2.12. Tétel: A H_K rész hipergráf kvázi-komponenseinek száma legfeljebb $2|K| - 1$.

A tétel állítása egyébként a fa struktúrájú partíciók jól ismert tulajdonsága.

A következő két lemma már nem a kvázi-komponensek tulajdonságainak feltárására szolgál, hanem a dolgozat 3. részében szereplő algoritmus szerkesztéséhez szükséges.

2.13. Lemma: Legyen adott a $H = (X; \varepsilon)$ hipergráfnak a $K(|K| \geq 2, K \subseteq X)$ kifeszítő ponthalmaza és az S halmaz, amelyre teljesül az, hogy $|S| \geq 2$ és $S \subseteq K$. Legyen $T^* \in \mathfrak{H}_S$ az a ponthalmaz, amelyre teljesül az, hogy $\overline{w}_K(T^*) \leq \overline{w}_K(T)$ tetszőleges $T \in \mathfrak{H}_S$ halmaz választása esetén.

Az S halmaz akkor és csak akkor kvázi-komponens a H_K rész-hipergráfban, ha $\overline{w}_K(S) < \overline{w}_K(T^*)$.

2.14. Lemma: Legyen adott a $H = (X; \varepsilon)$ hipergráfnak a K $(K \subseteq X, |K| \geq 2)$ kifeszítő ponthalmaza és az S ponthalmaz, amelyre teljesül az, hogy $S \subseteq X$ és $|S| \geq 2$. Legyen $T^* \in \mathfrak{H}_S$ az a ponthalmaz, amelyre teljesül az, hogy $\overline{w}_K(T^*) \leq \overline{w}_K(T)$ tetszőleges $T \in \mathfrak{H}_S$ halmaz választása esetén.

Ha $\mathcal{H}(\mathcal{E}'(T^*|T^*)) = \emptyset$, akkor T^* csak triviális kvázi-komponenseket tartalmaz.

Ha $\mathcal{H}(\mathcal{E}'(T^*|T^*)) \neq \emptyset$ és a P ponthalmaz komponense a H_{T^*} rész-hipergráfnak, akkor

$$C_K(T^*) = C_K(\mathcal{H}(\mathcal{E}'(T^*|T^*))) = C_K(P),$$

azaz

$$w_K(T^*) = w_K(\mathcal{H}(\mathcal{E}'(T^*|T^*))) = w_K(P).$$

A kvázi-komponens definíciója és bizonyított tulajdonságai lehetővé teszik a dolgozat első részében említett három modell bármelyikére sikerrel alkalmazható új *cluster definíció* bevezetését:

2.15. Definíció: Az objektumok clusterjei az 1. Modell gráfjának, ill. a 3. Modell hipergráfjának kvázi-komponensei, a deszkriptorok clusterjei a 2. Modell hipergráfjának kvázi-komponensei.

Az így definiált clusterek legfontosabb tulajdonságai a következők:

1. Ha a Q halmaz clustere az R objektum vagy deszkriptor halmaznak, akkor clustere R minden olyan részhalmazának is, amely tartalmazza a Q halmazt. (2.14. Megjegyzés.)
2. Ha a Q halmaz clustere az R halmaznak, akkor bármely $T \in \mathfrak{H}_Q$ részhalmaz erősebben kapcsolódik a Q clusterhez, mint annak környezetéhez (2.7. Tétel).
3. A clusterek vagy diszjunktak vagy tartalmazzák egymást (2.11. Tétel).
4. Az R halmaz clustereinek száma kisebb, mint a rendszer elemeinek számának kétszerese (2.12. Tétel).

A kvázi-komponensek megkeresésére szolgáló eljáráson alapuló új cluster technika

A hipergráf összes kvázi-komponensének meghatározására szolgáló eljárásnak két alapvető rutinja van.

R1: Egy hipergráf komponenseinek meghatározására szolgáló rutin.
R2: Egy hipergráf „minimális két részre vágása" meghatározására szolgáló rutin.

Az R1 rutin feladata a következő:

Adott a $H = (X; \varepsilon)$ hipergráf, és az élein értelmezett $v(E_j) > 0$ $(j = 1, \ldots m)$

függvény. Adott továbbá a $K (K \subseteq X)$ kifeszítő ponthalmaz, és az általa kifeszített $H_K = (K; \mathcal{E}'(K|K))$ rész-hipergráf.

Határozzuk meg a H_K rész-hipergráf összes komponensét, azaz azokat a $P (0 \neq P \subseteq K \subseteq X)$ ponthalmazokat, amelyekre teljesül az, hogy $\overline{w}_K(P) = 0$, és bármely $T \in \mathcal{S}_K = \{T | \emptyset \neq T, T \subset K\}$ esetén $\overline{w}_K(T) > 0$.

A feladat elvégzésére az irodalomban több algoritmus is ismert. Elterjedtek az indexelési technikán alapuló (*Klafszky* [17]) és az ekvivalencia osztályt generáló (*Knuth* [18]) eljárások is.

Az R1 rutin az indexelési technikán alapul és $O(|K| \cdot |\mathcal{E}'(K|K)|)$ lépésben határozza meg a $H_K = \left(K; \mathcal{E}'(K|K)\right)$ hipergráf összes komponensét.

Az R2 rutin feladata a következő:

Adott a $H = (X; \varepsilon)$ hipergráf ($|X| \geq 2$), és az élein értelmezett $v(E_j) > 0$ $(j = 1, \ldots m)$ függvény. Adott S ($|S| \geq 2$ és $S \subseteq X$) ponthalmaz. Határozzuk meg azt a $T^* \in \mathcal{S}_S = \{T | T \neq \emptyset\, T \subset S\}$ ponthalmazt, amelyre teljesül az, hogy $w(T^*) \leq w(T)$ bármely $T \in \mathcal{S}_S$ esetén.

A feladat elvégzése céljából az irodalomban ismertetett eljárások a hipergráf feladatot gráf problémára vezetik vissza, majd Ford – Fulkerson algoritmusával keresik a megoldást (*Lawler* [22]). Ezzel szemben az R2 rutin közvetlenül hipergráfon dolgozik és a maximális folyam problémánál egyszerűbb kereslet-kínálat feladat megoldásán alapul. $O(|X|^3 \cdot |\varepsilon|^2)$ lépésben határozza meg a $H = (X; \varepsilon)$ hipergráf minimális vágását. (*Futó* [13].)

Elemezzük eredeti problémánkat; egy hipergráf összes kvázi-komponensének meghatározására szolgáló eljárás megszerkesztését:

Ha a hipergráf nem összefüggő, akkor kvázi-komponensei a komponensei által kifeszített összefüggő rész-hipergráfoknak is kvázi-komponensei, és az összefüggő rész-hipergráfok kvázi-komponenseinek meghatározásával az eredeti hipergráf valamennyi kvázi-komponensét megkapjuk. (2.14. Megjegyzés, 2.15. Megjegyzés, Futó [13].)

Tehát elegendő olyan algoritmust konstruálni, amely egy összefüggő hipergráf kvázi-komponenseit keresi meg.

Feladat

Adott a $H = (X; \varepsilon)$ *összefüggő* hipergráf és $v(E_j) > 0$ $(i = 1, \ldots m)$ a hipergráf élein értelmezett pozitív függvény. Határozzuk meg a $H = (X; \varepsilon)$ hipergráf összes kvázi-komponensét, azaz azokat a $Q \subseteq X$ ponthalmazokat, amelyekre teljesül az, hogy bármely T esetén $(\emptyset \neq T \subset Q)$, $\overline{w}(Q) < \overline{w}(T)$.

Algoritmus

Legyen $\mathcal{K}_j = \{K_j^{(l_j, h_j)}, \ldots, K_{\nu_j}^{(l_{\nu j}, h_{\nu j})}\}$,

az eljárás j-edik lépése előtt azon $K_i^{(l_i, h_i)}$ ponthalmazok rendezett osztálya, melynek elemeiről a j-edik és a további lépések során kell eldönteni, hogy kvázi-komponensek-e.

Legyen $Q_j = \{Q_1, Q_2, \ldots Q_{tj}\}$,

az eljárás j-edik lépése előtt ismert vagy meghatározott kvázi-komponensek halmaza.

0. lépés

$$\mathcal{H}_0 = \{X\},\ \mathcal{Q}_0 = \{\{x_1\}, \ldots \{x_n\}\}.$$

Ha $|X| = 1$, akkor a 2.11. Tétel értelmében $H = (X; \varepsilon)$-nek csak 1 kvázi-komponense van, amelyet azonban már a 0. lépés előtt ismertünk, tehát az eljárás véget ért. Tehát

$$\mathcal{Q}_1 = \mathcal{Q}_0 = \{X\},\ \mathcal{H}_1 = \emptyset.$$

Ha $|X| \geq 2$, akkor mivel a $H = (X; \varepsilon)$ hipergráf összefüggő, ezért $|X| \geq 2$ miatt az X halmaz nem-triviális kvázi-komponens, tehát $X \in \mathcal{Q}_1$, és így $\mathcal{Q}_1 = \{[x_1], \ldots \{x_n\}, X\}$.

Ha $|X| = 2$, akkor az algoritmus véget ér, azaz $\mathcal{H}_1 = \emptyset$, mivel a $H = (X; \varepsilon)$ összes kvázi komponensét meghatároztuk, hiszen a 2.11. Tétel szerint ezek száma legfeljebb 3, és $Q_1 = \{X, \{x_1\}, \{x_2\}\}$.

Ha $|X| > 2$, akkor az R2 rutin segítségével $S_0 = X$ választás mellett meghatározzuk azt a $T_0^* \in \mathcal{S}_{S_0}$ halmazt, amelyre teljesül az, hogy bármely $T \in \mathcal{S}_{S_0}$ halmaz választása esetén $\overline{w}(T_0^*) \leq \overline{w}(T)$.

Jelölje a továbbiakban $\overline{w}(T_0^*)$-ot w_{01} és $\overline{w}(S_0 - T_0^*)$-ot w_{02}. A 2.11. Definíció miatt $w_{01} = w_{02}$, ezért w_{02} is már ismert.

A $H = (X; \varepsilon)$ hipergráf valamennyi X-től különböző kvázi-komponensét vagy T_0^* vagy $X - T_0^*$ tartalmazza (2.8. Tétel).

A nem triviális kvázi-komponenseket (amelyek megkeresése a célunk, ugyanis a triviálisak már a 0. lépés előtt ismertek voltak) a 2.11. Megjegyzés miatt $\mathcal{H}(\mathcal{E}'(T_0^*|T_0^*))$ vagy $\mathcal{H}(\mathcal{E}'((S_0 - T_0^*)|(S_0 - T_0^*)))$ is tartalmazza.

Ha $\mathcal{H}(\mathcal{E}'((S_0 - T_0^*)|(S_0 - T_0^*))) = \emptyset$ és $\mathcal{H}(\mathcal{E}'(T_0^*|T_0^*)) = \emptyset$, akkor T_0^* és $S_0 - T$ csak triviális kvázi-komponenseker tartalmaz, amelyek már a 0. lépés előtt ismertek voltak. Tehát az alogirtmus véget ér. $\mathcal{H}_1 = \emptyset$.

Ha $\mathcal{H}(\mathcal{E}'(T_0^*|T_0^*)) \neq \emptyset$ vagy $\mathcal{H}(\mathcal{E}'((S_0 - T_0^*)|(S_0 - T_0^*))) \neq \emptyset$, akkor a $H = (X; \varepsilon)$ hipergráf nem triviális kvázi-komponensei $H_{T_0^*}$ vagy a $H_{S_0 - T_0^*}$ rész-hipergráfoknak is kvázi-komponensei, sőt ezeket $H_{T_0^*}$ vagy $H_{S_0 - T_0^*}$ komponensei is tartalmazzák (2.14. Megjegyzés).

Határozzuk meg az R1 rutin segítségével $H_{T_0^*}$ és $H_{S_0 - T_0^*}$ komponenseit. (Megjegyzés: az R2 rutin alkalmazása esetén $H_{T_0^*}$ összefüggő lesz.) Jelölje ezeket $K_1^{(0,h_1)}, \ldots K_{v_1}^{(0,h_{v_1})}$.

A felső indexben első helyen álló 0 arra utal, hogy ezek a komponensek a 0. lépésben keletkeztek. A $h_i = 1$ érték esetén a $K_i^{(0,h_i)}$ komponenst T_0^*, a $h_i = 2$ érték esetén a $K_i^{(0,h_i)}$ komponenst az $X - T_0^*$ halmaz tartalmazza.

A 2.14. Lemma következtében $\overline{w}(T_0^*) = \overline{w}(X - T_0^*) = w_{01} = w_{02} = \overline{w}(K_1^{(0,h_1)}) = \ldots = \overline{w}(K_{v_1}^{(0,h_{v_1})})$.

Legyen $\mathcal{H}_1 = \{K_1^{(0,h_1)} \ldots K_{v_1}^{(0,h_{v_1})}\}$.

j. lépés:

($j \geq 1$)

$$\mathcal{H}_j = \{K_j^{(l_j,h_j)} \ldots K_{v_j}^{(l_{vj},h_{vj})}\},$$

$$\mathcal{Q}_j = \{Q_1, Q_2, \ldots Q_{tj}\}.$$

Ha $\mathcal{H}_j = \emptyset$, akkor az eljárás már a $j-1$. lépésben véget ért, ugyanis nincs több olyan ponthalmaz, amely kvázi-komponens lehet.

Ez esetben a Q_j halmaz tartalmazza a $H = (X; \varepsilon)$ hipergráf valamennyi kvázi-komponensét.

Ha $\mathfrak{K}_j \neq \emptyset$, akkor válasszuk ki a $\mathfrak{K}_j$ rendezett halmaz soron következő elemét: $K_j^{(l_j,h_j)}$-t.

Ha $|K_j^{(l_j,h_j)}| = 1$, akkor $K_j^{(l_j,h_j)}$ triviális kvázi-komponens, tehát már a 0. lépés előtt ismert volt. $K_j^{(l_j,h_j)}$ nyilván más kvázi-komponenst már nem tartalmazhat.

Tehát ez esetben: $\mathfrak{K}_{j+1} = \mathfrak{K}_j - \{K_j^{(l_j,h_j)}\}$, $\mathfrak{Q}_{j+1} = \mathfrak{Q}_j$.

A $K_j^{(l_j,h_j)}$ halmaz által generált vágás értékét már keletkezésekor az l_j. lépésben meghatároztuk ($l_j < j$). Ugyanis $h_j = 1$ esetben $K_j^{(l_j,h_j)} \subseteq T_{lj}^*$, $h_j \geq 2$ esetben $K_j^{(l_j,h_j)} \subseteq S_{l_j} - T_{lj}^*$ tartalmazási relációk állnak fenn, és a 2.13. Lemma miatt $\overline{w}(K_j^{(l_j,hj)}) = w_{l_j,h_j}$. A w_{l_j,h_j} értékét már az l_j. lépésben kiszámoltuk.

Ha $|K_j^{(l_j,h_j)}| = 2$, akkor meghatározzuk $K_j^{(l_j,h_j)}$ mindkét elemére (legyenek ezek x_{j1} és x_{j2}) a $\overline{w}(x_{j1})$ és $\overline{w}(x_{j2})$ értékeket. A kvázi-komponens definíciója szerint:

- ha $w_{l_j,h_j} < \overline{w}(x_{j1})$ és $w_{l_j,h_j} < \overline{w}(x_{j2})$, akkor $K_j^{(l_j,h_j)}$ kvázi-komponens, tehát $\mathfrak{Q}_{j+1} = \mathfrak{Q}_j \cup \{K_j^{(l_j,h_j)}\}$ és $\mathfrak{K}_{j+1} = \mathfrak{K}_j - K_j^{(l_j,h_j)}$.
- míg, ha a fenti két egyenlőtlenségnek legalább az egyike nem igaz, akkor $K_j^{(l_j,h_j)}$ nem kvázi-komponens, tehát $\mathfrak{Q}_{j+1} = \mathfrak{Q}_j$ és $\mathfrak{K}_{j+1} = \mathfrak{K}_j - \{K_j^{(l_j,h_j)}\}$.

Ha $K_j^{(l_j,h_j)} > 2$, akkor $S_j = K_j^{(l_j,h_j)}$ választás mellett az R2 rutin segítségével meghatározzuk azt a $T_j^* \in \mathfrak{S}_{sj}$ halmazt, amelyre teljesül az, hogy bármely $T \in \mathfrak{S}_{sj}$ esetén $\overline{w}(T_j^*) \leq \overline{w}(T)$.

Jelöljük a továbbiakban $\overline{w}(T_j^*)$-ot w_{j1}-gyel.

Azt, hogy $S_j = K_j^{(l_j,h_j)}$ kvázi-komponens volt-e, az l_j. lépésben kiszámított w_{l_j,h_j} és a j. lépésben meghatározott w_{j1} összehasonlításával döntjük el.

Ha $w_{l_j,h_j} < w_{j1}$, akkor a $K_j^{(l_j,h_j)}$ halmaz kvázi-komponens és így $\mathfrak{Q}_{j+1} = \mathfrak{Q}_j \cup K_j^{(l_j,h_j)}$.

Ha $w_{l_j,h_j} \geq w_{j1}$, akkor a 2.13. Lemma miatt a $K_j^{(l_j,h_j)}$ halmaz nem kvázi-komponens és így $\mathfrak{Q}_{j+1} = \mathfrak{Q}_j$.

A $H = (X; \varepsilon)$ hipergráf összes – $S_j = K_j^{(l_j,h_j)}$ által valódi részként tartalmazott – kvázi-komponensét vagy T_j^* vagy $S_j - T_j^*$ tartalmazza.

Ezek közül a nem-triviálisakat a 2.14. Megjegyzés miatt

$$\mathfrak{K}\big(\mathfrak{S}'((S_j - T_j^*)|(S_j - T_j^*))\big) \text{ vagy } \mathfrak{K}\big(\mathfrak{S}'(T_j^*|T_j^*)\big)$$

is tartalmazza.

Ha $\mathfrak{K}\big(\mathfrak{S}'((S_j - T_j^*)|(S_j - T_j^*))\big) = \emptyset$ és $\mathfrak{K}\big(\mathfrak{S}'(T_j^*|T_j^*)\big) = \emptyset$, akkor $S_j - T_j^*$ és T_j^* csak triviális kvázi-komponenseket tartalmaz (2.14. Lemma), amelyek már a 0. lépés előtt ismertek voltak.

Így ez esetben $\mathfrak{K}_{j+1} = \mathfrak{K}_j - \{K_j^{(l_j,h_j)}\}$.

Ha $\mathfrak{K}\big(\mathfrak{S}'((S_l - T_j^*)|(S_j - T_j^*))\big) \neq \emptyset$ vagy $\mathfrak{K}\big(\mathfrak{S}'(T_j^*|T_j^*)\big) \neq \emptyset$, akkor a $H = (X; \varepsilon)$ hipergráfnak az S_j által tartalmazott nem-triviális kvázi-komponensei a $H_{T_j^*}$ vagy a $H_{S_j - T_j^*}$ rész-hipergráfnak is kvázi-komponensei, sőt ezeket $H_{S_j - T_j^*}$ vagy $H_{T_j^*}$ komponensei is tartalmazzák (2.14. Megjegyzés).

Határozzuk meg az R1 rutin segítségével H_{T^*} és $H_{S_j - T_j^*}$ komponenseit. Jelölje $H_{T_j^*}$ komponenseit $K_{v_j+1}^{(j,1)}, \ldots K_{v_j+i_j}^{(j,1)}$.

A 2.14. Lemma miatt $w_{j1} = \overline{w}(K_{v_j+1}^{(j,1)}) = \ldots = \overline{w}(K_{v_j+i_j}^{(j,1)})$.

Jelölje $H_{S_j - T_j^*}$ komponenseit $K_{v_j+i_j+1}^{(j,2)}, \ldots K_{v_j+r_j}^{(j,r_j-i_j+1)})$, $(v_j + r_j = v_{j+1})$ $(r_j - i_j + 1 = h_{v_{j+1}})$. A $H = (X; \varepsilon)$ hipergráfnak a $H_{S_j-T_j^*}$ rész-hipergráf komponensei által generált vágásainak az értékeit meghatározzuk. Jelölje ezeket rendre $w_{j,2}, \ldots w_{j,r_j-i_j+1}$.

Általában $w_{j,1} \neq w_{j,2} \neq \ldots \neq w_{j,r_j-i_j+1}$.

Legyen ez esetben

$$\mathcal{K}_{j+1} = \mathcal{K}_j - \{K_j^{(l_j,h_j)}\} \cup \{K_{v_j+1}^{(j,1)}, \ldots K_{v_{j+1}}^{(j,r_j-i_j+1)}\} = \{K_{j+1}^{(l_{j+1},h_{j+1})}, \ldots K_{v_{j+1}}^{(j,h_{v_{j+1}})}\}.$$

3.1. Tétel: A Feladat megoldására, azaz az összefüggő $H = (X; \varepsilon)$ hipergráf összes kvázi-komponensének meghatározására szolgáló algoritmus legfeljebb $|X| - 1$ számú lépésben véget ér, ha $|X| \geq 2$, és az utolsó $r \cdot (r \leq |X| - 2)$ lépésben kapott Q_{r+1} halmaz tartalmazza a $H = (X; \varepsilon)$ hipergráf összes kvázi-komponensét.

Biz. 1: $|X| = n$ szerinti teljes indukcióval

a) $n = 2$: Ekkor az eljárás már a 0. lépésben véget ért ($\mathcal{K}_1 = \emptyset$), tehát $r = 0 = n - 2 = 0$ teljesül.

b) $n > 2$: Tehát $|X| \geq 3$. Végezzük el az algoritmus 0. lépését.

Ha $\mathcal{H}\big(\mathcal{E}'(T_0^*|T_0^*)\big) = \emptyset$ és $\mathcal{H}\big(\mathcal{E}'(X - T_0^*)|(X - T_0^*)\big) = \emptyset$, akkor az eljárás már a 0. lépésben véget ér, tehát $r = 0 \leq n - 2 \geq 1$.

Ha $\mathcal{H}\big(\mathcal{E}'(T_0^*|T_0^*)\big) \neq \emptyset$ vagy $\mathcal{H}\big(\mathcal{E}'((X - T_0^*)|(X - T_0^*))\big) \neq \emptyset$ akkor jelölje a kapott komponensek $(K_1^{(0,h_1)}, \ldots K_{v_1}^{(0,h_{v_1})})$ elemszámát $n_1, n_2, \ldots n_{v_0}$. Nyilván $n \geq \sum_{i=1}^{v_0} n_i$.

Tudjuk, hogy az X által tartalmazott kvázi-komponensek, a fenti komponensek által kifeszített rész-hipergráfoknak is kvázi-komponensei (2.14. Megjegyzés).

Indukciós feltevésünk értelmében a komponensek által kifeszített hipergráfok kvázi-komponenseit legfeljebb $n_1 - 1, \ldots n_v - 1$ számú lépésben határozhatjuk meg.

Tehát az algoritmus lépésszáma legfeljebb $\sum_{i=1}^{v_0} (n_i - 1) + 1$, ugyanis a 0. lépést már elvégeztük.

Ha $\sum_{i=1}^{v_0} n_i = n$, akkor $v_0 \geq 2$, mivel ez esetben $\mathcal{H}\big(\mathcal{E}'(T_0^*|T_0^*)\big) \neq \emptyset$ és $\mathcal{H}\big(\mathcal{E}'((X - T_0^*)|(X - T_0^*))\big) \neq \emptyset$ kellett legyen.

Ha $v_0 = 1$, akkor $\sum_{i=1}^{1} n_i = n_1 < n$, mivel ez esetben vagy $\mathcal{H}\big(\mathcal{E}'(T_0^*|T_0^*)\big) = \emptyset$, vagy $\mathcal{H}\big(\mathcal{E}'((X - T_0^*)|(X - T_0^*))\big) = \emptyset$ teljesült. Tehát

$$\sum_{i=1}^{v_0} (n_i - 1) + 1 = \sum_{i=1}^{v_0} n_i - v_0 + 1 \geq 2 + 1 = n - 1.$$

Tehát az algoritmus lépésszáma legfeljebb $n - 1$.

Biz. 2: Legyen a Q $(\emptyset \neq Q \subseteq X)$ halmaz a $H = (X; \varepsilon)$ hipergráfnak tetszőleges kvázi-komponense.

Ha Q triviális kvázi-komponens, akkor már $\mathcal{Q}_0$ is tartalmazta és $\mathcal{Q}_j \subseteq \mathcal{Q}_{j+1}$ $(j = 0, \ldots r)$ miatt $Q \in \mathcal{Q}_{r+1}$ is teljesül.

Ha Q nem-triviális kvázi-komponens és $Q = X$ akkor már az algoritmus 0. lépésében ismert, azaz $X \in \mathcal{Q}_1$ és így $\mathcal{Q}_j \subseteq \mathcal{Q}_{j+1}$, $(j = 0, \ldots r)$ miatt $X \subset \mathcal{Q}_{r+1}$.

Ha Q nem-triviális kvázi-komponens, és $Q \neq X$, akkor jelölje $K_j^{(l_j,h_j)}$ a $\bigcup_{j=0}^{r} \mathcal{H}_j$ elemei közül azt a legszűkebb halmazt, amely Q-t tartalmazza.

Ilyen legszűkebb halmaz biztos van, mert $X \subset \mathcal{H}_0$ miatt $X \subset \bigcup_{j=0}^{r} \mathcal{H}_j$.

Jelölje $\mathcal{H}_j$ azt a rendezett halmazt, amelynek $K_j^{(l_j,h_j)}$ az első eleme.

Ha $Q \subset K_j^{(l_j,h_j)}$, akkor a 2.14. Megjegyzés és a 2.8. Tétel miatt az eljárás j. lépésében keletkező komponensek valamelyike tartalmazza Q-t. Ez viszont ellentmond annak, hogy $K_j^{(l_j,h_j)}$ volt az a legszűkebb eleme $\bigcup_{j=0}^{r} \mathcal{H}_j$-nek, amely Q-t tartalmazta.

Tehát $Q = K_j^{(l_j,h_j)}$. Ez pedig azt jelenti, hogy a j. lépésben $K_j^{(l_j,h_j)} \in \mathcal{Q}_{j+1}$ lesz. De $\mathcal{Q}_j \geq \mathcal{Q}_{j+1}$ $(j = 0, \ldots r)$ miatt $Q \in \mathcal{Q}_{r+1}$. Q.E.D.

Mivel az R1 rutin lépésszáma $O(|X| \cdot |\varepsilon|)$ és az R2 rutin lépésszáma $O(|X|^3 \cdot |\varepsilon|^2)$, ezért a 2.12. Tétel alapján állíthatjuk, hogy a hipergráf összes kvázi-komponensének meghatározására szolgáló eljárás lépésszáma $O(|X|^4 \cdot |\varepsilon|^2)$.

A kvázi-komponensek meghatározására szolgáló eljárás a 2. részben bevezetett cluster definíció alapján egy új *hierarchikus cluster technika* lesz, amelynek legfontosabb jellemzői a következők:

1. Az R (objektum vagy deszkriptor) halmaz minden egyes eleme legalább egy clusternek is eleme (2.11. Megjegyzés).
2. Az eljárás konvergens (3.1. Tétel).
3. Az eljárás lépésszáma felülről becsülhető az objektumok és a deszkriptorok számának polinom alakú függvényével (3.1. Tétel).

A cluster elemzés hipergráf és gráf modelljeinek felhasználásával jelenleg folyamatban van az új cluster technika programozása FORTRAN nyelven R20 és TPA/i számítógépekre.

4. Az új cluster modellek és technika alkalmazási lehetőségei

A dolgozat első részében bevezett gráf és hipergráf modellek alkalmasak bonyolult rendszerek szerkezetének leírására. A hipergráf modellek alkalmazása különösen az információtudományi cluster analízis területén nagy jelentőségű. A probléma teljesen kézenfekvő modelljéül szolgálnak, és lehetővé teszik a hasonlósági mérőszámok definiálásának, kiszámolásának elkerülését.

A matematikai-statisztikai cluster elemzés problémáinak modellezésére mind a gráf, mind a hipergráf modellek alkalmasak. Itt azonban már nem kerülhető el vagy a hasonlósági mérőszámok, vagy a deszkriptorok definiálása, amely óhatatlanul a modell torzulására vezet.

A hipergráf vagy gráf kvázi-komponensének fogalmán alapuló cluster definíció és eljárás mindhárom modellre alkalmazható, tehát a klasszikus gráf

modell esetén is egy új lehetőséget nyújt a rendszer struktúrájának feltárására. Igazi jelentősége azonban a hipergráf modellek felhasználásánál látható. Lehetővé teszi olyan objektum rendszerek szerkezetének felderítését is, amelynél az objektumok jellemzésére tárgyszavakat és számadatokat is használnak. A cluster analízissel foglalkozó szakemberek által jól ismert tény, hogy a vizsgálandó rendszerek nagy hányada ilyen tulajdonságú.

Például az orvostudomány területén a csopotosítási probléma megoldása kulcs szerepet játszik a differenciál diagnosztikában, ahol a betegségeket a tünetek és a vizsgálatok eredményei alapján kell csoportosítani, és az analitikus epidemiológiában, ahol a betegségek csoportosítását külső és belső környezeti hatások alapján kell elvégezni. Nyilvánvaló, hogy a tüneteket vagy környezeti hatásokat leíró tárgyszavak kódolása, majd a hasonlósági mérőszámok kidolgozása igen erős torzulásokat eredményez. Hasonló a helyzet a közgazdasági és információtudományi problémák nagy hányadánál is.

A következőkben részletesebben bemutatjuk az új cluster modellek és technika alkalmazási lehetőségeit a kutatásirányítás területén. A választást indokolja egyrészt az, hogy a kutatásirányítás az utóbbi években a figyelem középpontjába került, másrészt, de nem utolsósorban az, hogy a szerző ezen a területen dolgozik, és az itt felmerült problémák sarkallták a cluster analízis mélyebb megismerésére és új módszer kidolgozására. Az Építéstudományi Intézetben dolgozó szűkebb kollektíva közel egy évtizedes kutató és fejlesztő munkájának eredménye – a tudományelmélet, információtudomány és az operációkutatás módszereinek és eredményeinek együttes felhasználásán alapuló LOGEL (*log*ikai *el*járások, vagy angolul: *log*ical mod*el*) tematikai kutatásirányítási módszer, amelynek lényege vázlatosan a következőkben foglalható össze.

1. A koordinált indexelés alapelvének felhasználásával *tárgyszavak* (deszkriptorok) halmazaira képzi le a vizsgált kutatási programok, témák, témajavaslatok tartalmát, módszerét, célját (esetleg ráfordításait, várható eredményeit is). Ezek a tárgyszavak lehetőség szerint egy ellenőrzött, szinonimáktól (különböző szóképek, azonos jelentés) és homonimáktól (azonos szókép, több jelentés) mentes, állandóan bővülő – általában hierarchikus szerkezetű – tárgyszórendszer, az ún. tezaurusz elemei. (A tezaurusz építése sok esetben a kutatási témák tárgyszavazásával párhuzamosan folyik.)
2. A kutatási témák, programok, vagy tudományos tételek, hipotézisek, és az ezeket a különböző szempontok szerint leíró tárgyszavak, valamint mindezek rendszerei közötti kapcsolatokat a LOGEL elmélet logikai modeljeinek felhasználásával írja le, amelyek irányított vagy irányítatlan gráfok, illetve hipergráfok.
3. A különböző kutatásirányítási problémák megoldását a logikai modellek bizonyos részhalmazainak különböző szempontú és mélységű *elemzésével* teszi lehetővé. Az elemzés révén, amely gráfelméleti és matematikai-statisztikai módszerekkel történik, az irányítandó kutatás rendszerről átfogó kép nyerhető.

   A *gráfelméleti* elemzés célja a logikai modellek elemei kapcsolatainak feltárása, míg az elemek eloszlásfüggvényeinek vizsgálata a LOGEL módszer fontos részét képező *deszkriptorstatisztika* feladata.

4. A kutatásirányítási *akciók modellezését* a logikai modellként szolgáló gráfok strukturális változtatásával (pl. bővítésével, szűkítésével, vágásával) segíti elő.

   A gyakorlatban elérni kívánt megoldásoknak a gráfokon strukturális optimumkritériumokat feleltetünk meg. Ezek tényleges eltérését *operációkutatási* algoritmusok teszik lehetővé.
5. A logikai modellek szerkesztését, elemzését, az operációkutatási algoritmusok futtatását az irányítandó kutatási programok vagy témák nagyobb száma esetén a LOGEL módszer számítógépes programrendszere teszi lehetővé, amelynek programjai FORTAN nyelven eddig CDC-3300, SIEMENS 4004/45 és TPA/i számítógépekre dolgozták ki. A fentiekből nyilvánvalóan adódik, hogy a cluster elemzés hipergráf modelljei és az új cluster technika eredményesen alkalmazható a kutatásirányítás területén.

A kutatás helyzetelemzése során lehetővé teszi mind a kutatási témarendszer, mind az indexelésükre felhasznált tárgyszavak clusterjeinek meghatározásával a rendszerek tematikai gócpontjainak és periferikus területeinek meghatározását, ill. extrapolációs módszer felhasználásával a tematikai centrumok elhelyezkedésének előrejelzését.

A kutatások koordinálása során a kutatási célprogramok, irányprogramok kidolgozásánál nyilván ezek magvai a helyzetelemzésnél meghatározott clusterek lesznek.

Az új cluster technika másodlagos alkalmazása az hogy az R2 minimális vágási rutin jól felhasználható kutatási témák optimális csoportosítására, célprogramok, irányprogramok konkrét kialakítására is.

A LOGEL módszeren alapul jelenleg az Építéstudományi Intézet, az ÉVM és az országos számítástechnikai kutatási célprogram kutatásirányítási rendszere. Az új cluster technika számítógépes programjainak elkészülte után 1978 elejétől megkezdjük annak folyamatos alkalmazását mindhárom területen. A számítástechnikai és alkalmazási tapasztalatokról külön cikkben számolunk be 1978 végén.

(Beérkezett: 1977. július 5-én.)

IRODALOMJEGYZÉK

1. Adamson, G. W. – Boreham, J.: The Use of an Association Measure Based on Character Structure to Identify Semantically Related Pairs of Words and Document Titles; Inform. Stor. Retr., Vol. 10. (253 – 260), 1974.
2. Anderberg, M. R.: Cluster Analysis for Applications; Academic Press, 1973.
3. Augustson, J. G. – Minker, J.: An Analysis of Some Graph Theoretical Cluster Techniques; Journal of A.C.M., Vol. 17. (571 – 588), 1970.
4. Balas, E. – Padberg, M.: On the Set Covering Problem: II. An Algorithm for Set Partitioning; Op. Res., No. 23. (74 – 90), 1975.
5. Benedikt, V. – Kelemen, K. – Pintér, Zs. – Vári, P.-né: Cluster analízis és lényegkiemelő eljárás-rendszer terve; SZÁMKI, 1976.
6. Berge, C.: Graphs and Hypergraphs; North Holland/American Elsevier, 1973.
7. Boulton, P. M. – Wallace, C. S.: An Information Measure for Single Link Classification; Comp. Journ., Vol. 18. (236 – 238), 1975.
8. Diday, E. – Schroeder, A.: A New Approach in Mixed Distributions Detection; IRIA, Rapport de Recherche, No. 52, 1974.
9. Edmonds, J. – Karp, R.: Theoretical Improvements in Algorithmic Efficiency for Network Flow Problems; Journal of A.C.M., Vol. 19. (248 – 264), 1972.
10. Forgy, E. W.: Classification so as to Relate to Outside Variables; Final Report,

Conf. Cluster Analysis of Multivariate Data (13.01 – 13.12), Cath. Univ. America, 1966.
11. Fritz, J.: Tanuló algoritmusok alkalmazása az alakfelismerésben; MTA MKI, 1975.
12. Futó, P.: Computer Aided Management of Industrial Research – the Method LOGEL; 12. Progr. Op. Res. (353 – 371), North Holland, 1976.
13. Futó, P.: Hipergráf elméleten alapuló új cluster definíció és technika; Alk. Mat. Lapok (Sajtó alatt).
14. Futóné Szántó, Zs.: Számítástechnika az egészségügyért. ESZTIK, Budapest, 1976.
15. Gower, J. C.: A Comparison of Some Methods of Cluster Analysis; Biometrika, Vol. 23 (623 – 637), 1967.
16. Johnson, S. C.: Hierarchical Clustering Schemes; Psychometr., Vol. 32. (241 – 254), 1976.
17. Klafszky, E.: Hálózati folyamok; Bólyai J. Mat. Társ., 1969.
18. Knuth, D. E.: The Art of Computer Programming; Vol. 1. Fundamental Algorithms, Addison – Wesley, 1968.
19. Kovács, L. B.: A diszkrét programozás kombinatorikus módszerei; Bólyai J. Mat. Társ., 1969.
20. Kunszt, Gy.: A tudományos kutatás logikai modellezése és tematikai irányítása; Akadémiai Kiadó, 1975.
21. Lawler, E. L.: Cutsets and Partitions of Hypergraphs; Networks, Vol. 3. (275 – 286), 1973.
22. Lawler, E. L.: Algorithms, Graphs and Complexity; Networks, Vol. 5. (89 – 92), 1975.
23. Luccio, F. – Sami, M.: On the Decomposition of Networks into Minimally Interconnected Networks; IEEE Trans. Circuit Theory, CT 16. (184 – 188), 1969.
24. MacQueen, J. B.: Some Methods for Classification and Analysis of Multivariate Observations; Proc. Symp. Math. Stat. and Prob., Vol. 1. (281 – 297), 1967.
25. Mulligan, G. B. – Corneil, P. G.: Corrections to Bierstone's Algorithm for Generating Clique; Journal of A.C.M., Vol. 19. (244 – 247), 1972.
26. Osteen, R. E.: Clique Detection Algorithms Based on Line Addition and Line Removal; SIAM Journal Appl. Math., Vol. 26. (126 – 135), 1974.
27. Sibson, R.: SLINK – An Optimally Efficient Algorithm for the Single-link Cluster Method; Comp. Journ., Vol. 16. (30 – 34), 1973.
28. Sparck-Jones, K.: Automatic Indexing "74"; Comp. Lab., Univ. of Cambridge, 1974.
29. Srejder, Ju. A.: Egyenlőség, hasonlóság, rendezés; Gondolat, 1975.
30. Tanimoto, T. T.: An Elementary Mathematical Theory of Classification and Prediction; IBM, 1958.
31. Ward, J. H.: Hierarchical Grouping to Optimize an Objective Funtion; Journ. Amer. Statist. Assoc., Vol. 58. (236 – 244), 1963.
32. Wishart, D.: An Algorithm for Hierarchical Classifications; Biometr., Vol. 22. (165 – 170), 1969.

A NEW MODEL AND ALGORITHM OF CLUSTER ANALYSIS

The first part of the paper presents the hypergraph model of cluster analysis, which enables us to eliminate the clumsy procedure of constructing the similarity matrix. The presentation of the new, hypergraph-based definition and its characteristics (part two) is followed by the detailed description of a new non-agglomerative, hierarchic cluster algorithm (part three). The fourth part of the paper deals with the application possibilities of the new cluster technique.

НОВАЯ МОДЕЛЬ КЛАСТЕРНОГО АНАЛИЗА И ЕЕ АЛГОРИТМ

В первой части работы рассматривается кластерный анализ на модели гиперграфа, посредством использования которой можно обойти несколько сложный метод разработки матрицы подобия. После описания понятия кластера, базирующегося на модели гиперграфа и его свойств (вторая часть) следует детальное изложение нового и неагломеративного по своему характеру иерархического кластерного алгоритма (третья часть). В четвертой части работы рассматриваются возможности применения новой кластерной техники.

KÖNYVEKRŐL

ANDERBERG, M. R.: *Cluster analysis for applications.* New York – London, 1973. Academic Press, 359 p.

Az olvasóban a könyv kiadásának dátumát olvasva felvetődhet a kérdés, hogy a SZIGMA cluster analízisnek szentelt különszámában miért ezt a már 5 éve megjelent könyvet ismertetjük.

A könyv választását a következők indokolják: A matematikai-statisztikai cluster analízis területén a 70-es évek elejéig elért eredményeket Anderberg munkája nemcsak összefoglalja, hanem nagyon alaposan és didaktikusan rendszerezi is. A könyv a cluster elemzés hazai művelőinek szinte egyöntetű véleménye szerint alapvető fontosságú.

A később megjelent cluster analízis témájú könyvek (pl. *J. A. Hartigan:* Clustering Algorithms, John Wiley and Sons, 1975) már csak egy-egy fontosabb részterülettel foglalkoznak behatóan.

Ugyanakkor a hetvenes évek elejétől igen gyors fejlődésnek induló információtudományi cluster elemzésről összefoglaló jellegű könyv még nem jelent meg, csak cikkek és szélesebb területet is felölelő tanulmányok (pl. *K. Sparck Jones:* Automatic Indexing '74. University of Cambridge, 1974).

Anderberg könyve 10 fejezetet és 8 függeléket tartalmaz.

Az első két fejezet általánosan ismerteti a cluster analízis témáját, felhasználási területeit és kapcsolatait más tudományágakkal. A cluster analízis alapfeladata az, hogy objektumok bonyolult rendszerének struktúráját feltárja, az objektumokat – előzetes tapasztalatok nélkül – kizárólag jellemzőiből adódó kapcsolataik alapján úgy csoportosítsa, hogy az egymáshoz hasonló objektumok egy csoportba (clusterbe) az egymáshoz kevéssé hasonló objektumok különböző csoportokba kerüljenek. Tágabb értelemben a cluster elemzés feladata a jellemzők csoportosítása is.

A szerző élesen meghúzza a határvonalat az osztályzási eljárások és a cluster elemzés módszerei között azzal, hogy a klasszifikálásnál a struktúra már ismert, és csak egy vagy több kategóriát kell besorolni, míg a cluster elemzés elsődleges célja egy „természetes" struktúra megalkotása.

A struktúra feltárás problémája nagy jelentőségű a biológiában, földtudományban, orvostudományban, társadalomtudományban, mérnöki tudományokban, információtudományban és végül, de nem utolsó sorban az operációkutatásban. Ebből is látszik, hogy a cluster elemzés módszereinek felhasználási területe milyen széleskörű.

A harmadik fejezet az objektumok jellemzőinek (a továbbiakban változóknak) csoportosításával foglalkozik. Az értelmezési tartomány számossága szerint megkülönböztet folytonos, diszkrét és – a diszkréten belül még – bináris változókat. Különösen jelentős a változók csoportosítása a mérési skála alapján. Jelöljön A és B két tetszőleges objektumot, x_A, ill. x_B az X változónak az A, ill. a B objektumra jellemző „értékét".

Ha X normális skálájú változó, akkor csak azt tudjuk, hogy $x_A = x_B$ vagy $x_A \neq x_B$. Ha X ordinális skálájú változó, akkor azt tudjuk, hogy $x_A = x_B$ vagy $x_A > x_B$ vagy $x_A < x_B$. Ha X intervallum skálájú változó, akkor ismerjük $x_A - x_B$ értékét. Ha X hányados skálájú változó, akkor ismerjük $x_A - x_B$ értékét. Ha X hányados skálájú változó, akkor ismerjük x_A/x_B értékét is. Ebben a fejezetben található skála konverziók részletes ismertetése is.

A negyedik fejezet a változók közötti asszociációs mértékekkel foglalkozik. Külön tárgyalja a hányados és intervallum skálájú (kvantitatív jellegű) változók, valamint a nominális és ordinális skálájú (kvalitatív jellegű) változók között használatos mértékeket. Részletesen elemzi és összehasonlítja a bináris változók asszociációs mértékeit (pl. *Dice, Tanimoto* mértéke).

Az ötödik fejezet az objektumok közötti asszociációs mértékeket tárgyalja. A kvantitatív jellegű változók esetén a metrikán alapuló mértékek fontosságát emeli ki. A kvalitatív jellegű változók esetén valószínűségi jellegű mértékek alkalmazása is elterjedt, ezért ezeket is részletesen tárgyalja. Külön foglalkozik a szerző a bináris változók esetén használatos mértékekkel. A fejezet végén értékes tanácsokat ad arra az esetre, ha az összehasonlítandó objektumok jellemzői között kvalitatív és kvantitatív jellegű változók egyaránt szerepelnek.

A következő két fejezet a cluster eljárásokkal foglalkozik. A hierarchikus cluster eljárások (hatodik fejezet) zömmel az asszociációs mérték segítségével képzett hasonlósági mátrix felhasználásán alapulnak, amely i-edik sorának j-edik eleme megadja az i-edik és a j-edik objektum, ill. változó közötti asszociációs mértéket. Az eljárások célja olyan cluster hierarchia (fa) megalkotása, amelynél két clusternek vagy nincs közös eleme, vagy az egyik tartalmazza a másikat. (Általában clusternek tekintik az egyes objektumokat és a teljes vizsgálandó objektum halmazt is.) A fejezet elsősorban az agglomeratív jellegű hierarchikus cluster eljárásokat tárgyalja (az eljárás során mindig kisebb clusterek egyesítéséből képezzük a nagyobb clustert). A fordított irányú (particionáláson alapuló) eljárások közül csak néhányat említ meg. Az agglomeratív jellegű eljárásokat aszerint csoportosítja, hogy a számítógép központi memóriájában a kiindulási adatokat vagy a hasonlósági mátrixot tárolják. Részletesen tárgyalja a széleskörűen alkalmazott módszereket (legközelebbi szomszéd, legtávolabbi szomszéd, *Ward*, *Wishart* módszerét és a középpont szerint osztályozó eljárásokat).

A hetedik fejezet az ún. nem hierarchikus cluster eljárások ismertetését tartalmazza, itt nem egy cluster hierarchia megalkotása a cél, hanem az objektumok particionálása előre megadott számú vagy az eljárás közben adódó számú csoportba (clusterbe). A fejezet elején a szerző részletesen elemzi a kialakítandó clusterek magjainak meghatározására szolgáló heurisztikus algoritmusokat. Ezt követi *Forgy*, *MacQueen*, és *Wishart* eljárásának, valamint az ISODATA módszerének ismertetése.

A nyolcadik és kilencedik fejezet azokat a technikákat és stratégiákat tartalmazza, amelyek alkalmazásával a cluster analízis hatékonysága növelhető. A hierarchikus osztályozás segédeszközeinek vázlatos ismertetése mellett részletesen tárgyalja a szekvenciális jellegű cluster eljárásokat, amelyek zömmel heurisztikus módszerek, de nagy méretű problémák megoldásához szinte nélkülözhetetlenek. Vázolja a több cluster eljárás párhuzamos alkalmazásából eredő előnyöket és esetleges hátrányokat (pl. költség) is.

Vizsgálja a cluster analízis felhasználási területeit a matematikai statisztikán belül, és a külső kritériumokat is figyelembe vevő cluster eljárások megalkotásának általános irányelveit.

A tizedik fejezet a cluster eljárások összehasonlításának módszereivel foglalkozik. Kiemeli, hogy általában nem található legjobb eljárás egy adott feladat megoldására. Kísérletet tesz a hierarchikus eljárások és a particionáláson alapuló eljárások közötti hasonlóság mérésére. Felsorolja a problémák legfontosabb jellemzőit (pl. objektumok, változók száma, típusa, változók súlyozása, clusterrel szemben támasztott követelmények) és a megoldási módszerek fő paramétereit (pl. az eredmények struktúrája, gépidő igény, memória igény, a kiindulási állapot megválasztásának hatása a clusterek struktúrájára).

Az A függelék elméleti jellegű. A nominális skálájú változók közötti asszociációs mérték megalkotásával foglalkozik. A könyv B, C, D, E, F és G függeléke FORTAN nyelven írt számítógépes programokat tartalmaz, amelyek géptől függetlenek és valóban egyszerűen felhasználhatók. A B függelékben a skála konverziók elvégzését a C és D függelékben az asszociációs mértékek meghatározását elősegítő programok szerepelnek. Az E és F függelék a leggyakrabban használt cluster eljárások programjait, a G függelék az eljárások eredményeinek interpretálását elősegítő programokat tartalmazza. A H függelék a számítógépes programok kapcsolatait mutatja be.

A könyvet több mint 150 tételes referencia lista és tárgymutató zárja.

FUTÓ PÉTER

RUBIN, J. – FRIEMAN, H. P.: (Cluster analízis és taxonomikus rendszer az adatok csoportosítására és osztályozására) *A Cluster Analisis and Taxonomy System for Grouping and Classifying Data.* IBM Contributed program library. August 1967.

Az IBM gondozásában 1967-ben megjelent könyv első részében betekintést nyújt a cluster analízis problémafelvetésébe. Ismertet egy, a súlypontok módszerén alapuló nem hierarchikus cluster modellt, illetve az adatok hasonlósági mértékének vizsgálatán alapuló osztályozási rendszer elméletét és a módszerek gyakorlati meg-

valósításának elvét. A második részben a mellékelt programcsomag használatát ismerteti. A függelékekben a gyakorlati tapasztalatokat összegezi és javaslatokat ad a különböző alkalmazási területek felhasználói részére.

Ebben az ismertetőben a cluster modell elméletének és gyakorlati megvalósításának elvét írjuk le.

Matematikai leírás

A felhasznált cluster definíció megköveteli, hogy a clusterek diszjunktak legyenek és minden elem tartozzék valamely clusterbe. Az analízis eredményétől megkívánjuk, hogy optimális legyen, vagy ha nem lehetséges az optimalizáció, akkor erről felvilágosítást adjon. Ezért célszerű egy $\mathcal{F}$ cluster függvény bevezetése, amely jellemző minden egyes felbontásra.

A cluster analízis problémája az $\mathcal{F}$ függvény maximumának/minimumának meghatározására.

A vizsgált elemek legyenek $\xi_1, \xi_2, \ldots, \xi_n$ p-dimenziós valószínűségi változók, amelyeknek létezik közös sűrűségfüggvényük. Tegyük fel, hogy előre adott a clusterek száma és a függvény olyan egyszerű szerkezetű, hogy $f(x) = k$, ha $x \in G_i$ és $f(x) = 0$, ha $x \notin G_i$ $(i = 1, 2, \ldots g)$, ahol G_i az i-edik cluster és g a clusterek száma.

A mérések eredményét mátrix alakban ábrázoljuk. X mátrix sorai $P_i = (x_{i,1}, x_{i,2}, \ldots x_{i,p})$ $(i = 1,2, \ldots, n)$ és P_i legyen a p dimenziós euklideszi-tér egy pontja. Az általánosság megszorítása nélkül feltehető, hogy az n pont tömegközéppontja az origó. Ekkor az n pont totális szórás mátrixa $T = X^T X = \sum_{i=1}^{n} P_i^T P_i$. ($T$ a transzponálás jele). Jelölje $n_1, n_2, \ldots, n_g$ az objektumok számát az egyes csoportokban úgy, hogy $n_1 + n_2 + \ldots + n_g = n$. Ezután definiáljuk a csoportok szórásmátrixát a C_k tömegközépponttal

$$W_k = \sum_{l=1}^{n_k} (P_{kl} - C_k)^T (P_{kl} - C_k)$$

A csoportok összegezésében a szórásmátrix legyen $W = \sum_{k=1}^{g} W_k$ és a csoportok közötti szórásmátrixot definiálja $B = \sum_{k=1}^{g} n_k C_k^T C_k$. Ezek után vizsgáljuk az (1) $T = W + B$ mátrix egyenletet. Mivel a feldolgozás kezdetén az adatok mértékéről és méretéről semmilyen kikötést nem tettünk, a felbontást jellemző clusterfüggvényt (továbbiakban kritériumot) úgy kell választani, hogy az invariáns legyen a pontok közötti nem szinguláris lineáris transzformációkkal szemben.

Belátható, hogy $p = 1$ esetében az (1) egyenletből kapott $\frac{T}{W} = 1 + \frac{B}{W}$ egyenletben $\frac{B}{W}$ a fenti feltételnek megfelelő kritériumot ad.

$p > 1$ esetben ha $p \leqq n - g$ akkor W pozitív definit szimmetrikus mátrix, így létezik inverze. Ekkor képezzük a Mahalanobis-távolságot a következő módon

$$md_{ij} = (P_i - P_j)\, W^{-1} (P_i - P_j)^T$$
$$(i, j = 1,2, \ldots, n)$$

Ez a távolság invariáns a pontok közötti nem szinguláris lineáris transzformációkkal szemben. Így kaptunk egy, a felbontáshoz rendelhető megfelelő kritériumot. Ebből adódik a módszer heurisztikus jellege; ha ismerjük a felbontást, megfelelő metrikát tudunk meghatározni, ha ismerjük a megfelelő metrikát, segítségével eljuthatunk az optimális felbontáshoz. Feltételezzük, hogy a számítások során eljutunk a legjobb felbontásig.

Az eljárást a következőképpen fogalmazzuk meg. Az (1) egyenletből a determinánsok szorzástételének felhasználásával képezzük a $\frac{|T|}{|W|} = |I + W^{-1} B|$ egyenletet. A bal oldal egy skalárfüggvény, ezt a $\frac{|T|}{|W|}$-t maximalizáljuk, elfogadva azt a felbontást, amelyre a $\frac{|T|}{|W|}$ a legnagyobb, majd az ehhez tartozó W-t használjuk a Mahalanobis-különbség meghatározására.

Mivel módszereink heurisztikusak, szükséges más kritériumok meghatározása is az ellenőrzés érdekében. Vezessük be a $\mathrm{Tr}A$ jelölést, amely a mátrixhoz olyan konstansot rendel, mely jellemzi a mátrix elemeinek egymáshoz való viszonyát. (Bizonyos feltételek mellett a $\mathrm{Tr}A = \Sigma a_{ii}$.) Vezessük be az (1) egyenlet alapján a Hotteling-féle trace-kritériumot. Belátható, hogy a $\mathrm{Tr}\,(W^{-1}B)$ megfelelő kritériumot ad, ahol $\mathrm{Tr}\,(W^{-1}B) = \Sigma \lambda_i$, ahol λ_i-k a $W^{-1}B$ sajátértékei, azaz a $|B - \lambda W| = 0$ egyenlet megoldásai. Ezen λ_i-k segítségével kifejezhetjük a $\frac{|T|}{|W|}$ hányadost is, melyre igaz $\frac{|T|}{|W|} = \prod_i (1 + \lambda_i)$.

A feldolgozás során a $\log(|T|/|W|)$-t fogjuk vizsgálni. A különböző feldolgozások során kitűnt, hogy nem dönthetünk egyik kritérium javára sem. A választást

az éppen adott feladat határozza meg, illetve az ismertetett programcsomag automatikusan választja ki a legmegfelelőbbet.

A cluster analízis gyakorlati megvalósítása

A clusterizáláshoz javasolt kritériumok mindegyike felhasználja az összes lehetséges g csoportra való felbontást. Ezek száma még alacsony elemszám esetén is igen nagy, ezért fontos olyan módszer kidolgozása, amely ha nem is az összes, de legalább annyi lehetőséget végigvizsgál, amelyekből már megfelelő következtetést levonhatunk. Használjuk a már sok esetben bevált, ún. hegymászó eljárást.

Mivel ez a rendszer valamilyen kezdeti felbontást feltételez az objektumok halmazán, hatékonysága nagyon függ a kiindulás jóságától.

Először legyen a kezdeti felbontás valamilyen véletlen felbontás, majd alkalmazzuk az ún. „gyorsított menet"-et. Válaszszuk ki a kezdeti felbontás valamely csoportját és minden elemét vigyük az elemhez legközelebb levő csoport súlypontjának közelébe (a mérték legyen az éppen adott felbontáshoz tartozó Mahalanobis-mérték, vagy az általános eukledeszi mérték). Minden esetben számoljuk ki az egyszeri páronkénti kritériumot ($\mathrm{Tr}\, W_k = \frac{1}{n_k}\left(\sum_{l,m=1}^{n_k}(P_{lk}-P_{mk})(P_{lk}-P_{mk})\right)$, $\mathrm{Tr}\, W = \sum_{k=1}^{g}\mathrm{Tr}\, W_k$ és ha ez kisebb, mint az előző esetben, hagyjuk az elemet az új helyén.

Másik jól használható eljárás a kezdeti felbontáshoz az ún. újrajelölési mód. Jelöljön valamilyen kiinduló felbontást Q és minden egyes objektumot abba a Q-hoz tartozó csoportba soroljunk, amelynek súlypontjához a legközelebb van (mértékül az eukledeszi mértéket használjuk). Vizsgáljuk az új felosztáshoz tartozó valamelyik kritériumot. Az eljárást addig folytatjuk, amíg vagy nem kaptunk új felbontást, vagy az új felbontáshoz gyengébb érték tartozik.

A hegymászó eljárás. Tegyük fel, hogy valamilyen módszerrel adott egy kezdeti felbontás. Vegyük az egyik objektumot és mozgassuk el csoportról csoportra. Ha egy mozgatás során a választott kritérium jobb értéket szolgáltat, akkor az objektumot az új csoportba tartozónak tekintjük és folytatjuk a mozgatását. Ha az új felbontásnál a kritérium ugyanazt az értéket szolgáltatja, meghagyjuk az eredeti felbontást. Ha a kezdeti felbontásban az adott csoportok száma kisebb, mint az a szám, amit a felhasználó megkívánt, akkor az üres halmazba való mozgatást is megvizsgáljuk. Így az egyes objektumok elvándorolnak a „jobb" csoport felé. Ezután vesszük a következő objektumot és így tovább. Az összes objektum egyszeri tologatása lesz egy hegymászó menet. Mivel véges halmaz mozgatásainak száma is véges, véges ismételt hegymászó eljárás után eljutunk a legjobb felbontásig.

Mivel a vázolt eljárás igen sok számítást igényel a cluster-analízis eredményét egyszeri hegymászó menet után kapjuk. A gyakorlati feladatoknál ez általában elegendő, de természetesen mód van a program paraméterezésének segítségével további pontos számolásokra. A hatékonyságot inkább a kezdeti felbontás jobb megadásával, illetve jobb adatelőkészítéssel lehet fokozni.

Az IBM felhasználói programkönyvtárában közreadott rendszer IBM alapú, így (ESZR gépen is) 128K, 256K, illetve 512K vagy nagyobb központi memóriájú gépen futtatható. A terjesztett mágnesszalagon az IBM OS konvekcióknak megfelelő object modul, illetve az eredeti forrásprogram is rögzítve van. A program Assembler és FORTRAN nyelven megírt modulokból áll. A program szolgálatatásai igen sokrétűek. Az adatokat a felhasználó által megadott FORTRAN formátum szerint kell megadni. A kezdeti felbontás is megadható, de kívánságra a program véletlen, vagy gyorsított, vagy újrajelöléses módon is képezhet kiinduló felbontást. Szabályozni lehet, hogy a hegymászó eljárást hányszor ismételje meg és azt, hogy a feldolgozás során a Mahalanobis ($\mathrm{Tr}\,(W^{-1}B)$), a Wilks – Lambda $\left(\log\frac{|T|}{|W|}\right)$ kritériumot, vagy ezek közelítését használja. Végül mód van arra, hogy a feldolgozást megszakítva, más időpontban a korábbi eredményeket felhasználva pontosítsuk az eredményeket.

A programrendszer a KSH Számítástechnikai Igazgatóságán instalálva van, s ott használatáról gyakorlati tapasztalatokkal is rendelkeznek.

CSICSMAN JÓZSEF

TUDOMÁNYOS ÉLET

A VII. Magyar Operációkutatási Konferencia

(Válogatott észrevételek.)

A hetedik – jubileumi – Magyar Operációkutatási Konferencia abban a tekintetben tökéletesen megegyezik elődeivel, hogy méltatása kemény dió az erre vállalkozók számára. Minden szempontból heterogén: a tartalom, a módszer, az előadói modor és a témaválasztási indíték annyiféle, ahány előadás elhangzik. Ha a témakörökről azt mondjuk, hogy jobbára közgazdaságiak, alig mondtunk valamit, hiszen maga a konferencia lehet a legjobb bizonyíték arra, hogy ez a kategória mennyi – szinte diszjunkt – önálló világot takar. A felvonultatott módszerek a mennyiségi vonásokat teljesen nélkülöző abszolút verbálistól a matematikai formulák dömpingjéig, az előadói stílus az elsősorban didaktikustól a legfeljebb utolsó sorban didaktikusig, végül a témaválasztási indíték a kizárólagosan módszer-orientálttól a kizárólagosan eredmény-orientáltig olyan széles spektrumokat tár elénk, hogy az elhangzottaknak valamiféle közös nevezőre hozása, összevetése vagy éppen a kiemelésre érdemesség szempontjából való értékelése megvalósíthatatlan feladatot jelent, kivált akkor, ha a szakmailag indokolatlan elfogultság bűnébe nem akarunk beleesni. A kiemelendők kiválasztása különösen kényes feladat, mivel az egy-egy előadás mögött álló kutatói munkának nyilván nagyobb súllyal kellene latba esnie, mint az azt közlő előadói teljesítménynek, akár az előadásra felkészülés munkáját is hozzászámítva. Viszont a közönségnek ehhez a többhónapos, vagy éves, kemény munkához jószerivel egyetlen csatornája a jobban vagy gyengébben sikerült előadás, amely esetenként az előadók rutinjától, orgánumától, szuggesztivitásától függően arat vagy nem arat tetszést, illetőleg értést. Az elfogult véleményalkotás veszélyét az is fokozza, hogy mindenkinek az tetszik a legjobban, amihez ért, aminél nem álmos és – nem mellékes – aminél jelen volt.

A felsorolt szempontoktól indíttatva arra az elhatározásra jutottunk, hogy megpróbáljuk ezt az ismertetést anélkül megejteni, hogy egyetlen előadó nevét vagy előadás címét explicit módon megemlítenénk. Mivel mindenféle dologról mindenféle módon volt szó, úgy mint az előző alkalmakkor, de ugyanakkor mindegyik előadás – ezt joggal állíthatjuk róluk – megint hozott valami újat és többé vagy kevésbbé tágította a horizontot, úgy találtuk, hogy arra érdemes a figyelmet felhívni, ami ezen a konferencián minőségileg, témakör fölötti értelemben új volt, vagy mint ilyet vettük észre. Ezekből a kiemelt újdonságokból persze rá lehet ismerni a vonatkozó előadásokra és előadókra, de a válogatást így indirektebbnek és természetesebbnek érezzük.

Előzőleg azonban hadd hívjuk fel a figyelmet egy nem új, de nagyon pozitív vonásra. Ez a konferencia legalább annyira megérdemelte volna „Operációkutatás a gyakorlatban" címet, mint a közvetlen elődje. Alig találkoztunk olyan előadással, amely teljesen elvi síkon maradt volna vagy gyakorlatilag érdektelen méretű mintafeladatok megoldásait produkálta volna. A legtöbb előadás olyan eredményről szólt, amely nemcsak hogy a reális alkalmazás küszöbén állt, de már alkalmazási tapasztalatokról számolhatott be.

Nézzük hát az újdonságokat! Még nem minőségi, csak számszerű újdonság a faktor- és clusteranalízis alkalmazásának előretörése. Meghonosodóban van ez az eddig sem ismeretlen, de sokkal ritkábban használt technika.

Ugyan nem előzmény nélküli jelenség, de ezen a konferencián vált olyan mérvűvé, hogy felfigyeljünk rá: a makroökonomiai problémák teljesen kitöltötték a konferencia legnagyobb volumenű, leglátogatottabb szekcióját, az „A" szekciót. Ez különösen akkor érdekes, ha meggondoljuk, hogy az operációkutatás tárgyköre kialakulásakor mikroproblémákra korlátozódott és a klasszikus tankönyvek például a Leontyev-analízist nem is említik. Az a benyomásunk tehát, mintha a makroökonómia 1977-re kinőtte volna az operációkutatás kereteit. Még markánsabban jelentkezett ugyanez az ökonometriára

nézve, amellyel kapcsolatban az „A" szekció egyik elnöke maga nyilatkozott úgy, hogy az már megérett az önálló fellépésre.

Általában jól sikerült, ez nem új, de az „A" szekcióban feltűnően szerencsés volt az előadások csoportosítása. A többnyire függetlenül előállt témák úgy hatottak, mint egy előre szervezett és szisztematikusan felépített kurzus elemei: egymást készítették elő, egymás mondanivalóját magyarázták és mélyítették el. Ezt a nagyfokú összhangot a konferencia szervező bizottsága aligha érhette volna el, ha a spontán témakínálat összetétele a kezére nem játszott volna. Ez a körülmény viszont ismét azt a feltevést látszik igazolni, hogy az „A" szekció témaköre, a makroökonómia és ökonometria nálunk is rendelkezik már az önálló tudományágakra jellemző egységes kutatási célkitűzéssel és a realizáláshoz szükséges módszertani apparátussal.

Minőségi és lényeges újdonságnak tekintjük azt az irányzatot, amely az információhiányt, illetőleg annak figyelembevételét igyekszik algoritmizálni. Nem az adatok sztochasztikus bizonytalanságának a kezeléséről van szó, tehát nem a determinisztikus adatoknak valószínűségi változókkal való helyettesítéséről, hanem az olyan problémákkal való korrekt bánásmódról, amelyben például a valószínűségi változó ismeretlen, szukcesszív mérési vagy statisztikai adatok hiányoznak, vagy bizonyos szóbajöhető megoldások ellen nehezen megfogalmazható, kvalitatív jellegű kifogás merül fel, azaz minőségi információhiány esete áll fenn. A használható megodás szóbanforgó feltételeinek hiányát úgy szokás – végülis nagyon érthetően – kezelni, hogy a modellező vagy a matematikus nem kevés szellemi befektetés árán elkészíti a megbizható és hiánytalan adatokkal jól működő algoritmust és útjára bocsátja. A többi a munkamegosztásból kifolyólag nem az ő gondja. Ha mégis az, akkor a hiányzó adatokat feltételezi, interpolálja, vagy „meghasalja", vagyis előteremti valahonnan, hogy a rendszer formálisan mégis csak működjék.

A modell-módszer szerkesztés és az információellátás két különböző terület, amely legfeljebb kényszerházasságra lép. Szerves egység teremtődik azonban közöttük, ha az algoritmus az információhiányra eleve ráfigyel, és vagy megengedett módon áthidalja, vagy exakt módon kimutatja a hiányból vagy pontatlanságból adódó eredménytorzulás mértékét. Szemléletben, technikában és távlatokban egyaránt nagyon ígéretes irányzatról van tehát szó.

Figyelemreméltó volt két, a játékszabályok értelmében itt megnevezésre szintén nem kerülő előadás témája, a számítógéphálózatok modellezése operációkutatási eszközökkel, valamint a számítógépi programok optimális ütemű megszakítása és kimentése (archiválása) a géphibák következtében megismétlendő futtatási idők minimalizálása céljából. Ezek a témák ugyanis azt jelzik, hogy a számítógép az operációkutatás *eszközéből* az operációkutatás *tárgyává* lépett elő! A kutatási eredmény természetesen (többek között) az operációkutatásnak *további* eszköze. Képzeljük el, ha ez iterálni kezd. . . Az a benyomásunk, hogy a csak írásbelileg közreadott előadások rendszere nem szerencsés, mert a hallgatóság ezeket egyszerűen figyelmen kívül hagyja. Az írásbeli dolgozatok egyik vitája azzal a meglepetéssel szolgált, hogy a jelenlevők, akik a terítéken levő dolgozatokat tökéletes közönnyel fogadták, mivel feltehetően el sem olvasták őket, heves vitában kifejezésre jutó élénk érdeklődést tanúsítottak a dolgozatok témája iránt, mihelyt a szerzők mégis rászánták magukat szóbeli ismertetésükre.

Természetesen önkényes és szubjektív módon, továbbá szűkmarkúan válogattunk, de válogatásunk szempontja minden bizonnyal objektív és helytálló. Nem hallottunk, nem vettünk észre és nem értettünk meg mindent. Mások mást emeltek volna ki, alighanem ugyanennyire önkényesen. Rengeteg tiszteletreméltó erőfeszítést éreztünk majdnem minden meghallgatott előadás mögött. De mindegyikük megemlítése éppen annyira nem vált volna hasznukra, mint amennyire nem érinti őket említetlenül hagyásuk sem.

A konferencia hangulatát befelhőzte egy tragikus hír: *Weitz Tamás* és felesége a konferenciára utaztukban autóbalesetet szenvedtek és a helyszínen meghaltak. Mind a ketten előadást tartottak volna. A szerencsétlenül járt házaspárt sokan ismerték, de ez a borzalmas esemény az összes résztvevőt mélységesen megrázta. A szakmai foglalkozás és az oldottabb programok közben újból és újból eszünkbe jutott sorsuk. Nem tudtunk, nem is lehetett fölötte napirendre térni!

ZSELLÉR GYULA

A szerkesztőség kiegészítő megjegyzése

A VII. Magyar Operációkutatási Konferenciáról szóló beszámoló is említi a konferencia heterogenitását. Ez ismert és ismétlődő sajátossága a konferenciáknak: így válik a konferencia a hazai operációkutatók széles körű, országos seregszemléjévé, ahová mindenki

eljöhet és témájáról beszélhet is. Kellenek viszont szűkebb szakértői szinten rendezett „kis konferenciák" is, ahol egy-egy részterület művelői találkoznak. 1978-ban két ilyen konferencia is lesz, egy a hosszútávú (és esetleg középtávú) tervezésről és egy az operációkutatásról a mezőgazdaságban.

Az Ökonometriai Társaság Európai Konferenciája (ESEM '77)

Az Ökonometriai Társaság 1977. évi Európai Konferenciáját Bécsben rendezték meg szeptember 6. és 9. között. A résztvevők száma mintegy 400 fő, többnyire Európából. A benyújtott előadások száma 240 körül volt. A konferencián a plenáris üléseken kívül párhuzamosan 4 – 5 szekcióülés is folyt.

A konferencia *nem csupán ökonometriai*, hanem *általános közgazdasági, gazdaságelméleti és matematikai-közgazdasági kérdésekkel is foglalkozott*. A nyugateurópai és amerikai előadások elsősorban a még mindig divatos *egyensúlyelmélet* alapján álltak. Mind elméleti fejtegetéseik, mind pedig modeljeik az egyensúly és egyensúlytalanság kérdéseit állították középpontba; ezt az irányzatot jól tükrözte a konferencia egyik plenáris ülése, ahol az előadó (*L. W. McKenzie*, Rochester) a kompetetív egyensúly létezésének problémáit és feltételeit fejtette ki. Ugyancsak általános tendenciaként értékelhető az is, hogy meglepően sok előadás foglalkozott a társadalmi, politikai tervezés és prognóziskészítés módszereivel és eddigi eredményeivel. Ugyanakkor – feltehetően a világgazdasági folyamatok áttekinthetetlensége következtében – feltűnően kevés előadó vállalkozott külgazdasági modellek, illetve számítások ismertetésére.

Ami a modellek és általában az ökonometria módszertani vonatkozásait illeti, úgy tűnik, hogy újabb kiemelkedő elméleti eredményekről nem számoltak be; *a módszertani előadások általában* a meglevő, *ismert módszerek speciális* körülmények közt való *alkalmazásával foglalkoztak*. Említésre méltónak tartjuk, hogy az előadások a korábbiaknál nagyobb súlyt helyeztek arra, hogy a gyakorlat számára hasznosítható kutatásokról adjanak áttekintést. *Az elmélet és gyakorlat közeledésére* utalt egy sor olyan előadás, amelyek a meglevő adottságokhoz igazodó, elméletileg talán kevésbé elegáns, de feltétlen hasznos eljárásokat ismertettek és nyilvánvalóan mem lehet véletlen az sem, hogy a konferencia másik fő plenáris előadása (*C. W. Simc*, Minneapolis) is az elmélet és a gyakorlat kérdéseivel foglalkozott.

Tekintettel a konferencián szereplő témák sokrétűségére és nagy számára, az ismertetésben nem térhetünk ki minden részletre. A témákat a hazai kutatási irányoknak megfelelő csoportosításban tekintjük át: 1. tervezési modellek; 2. ökonometriai modellek és módszertani kérdések; 3. fogyasztási modellek; 4. a szabályozás modellezése; 5. egyéb.

*A tervezési modelek*kel 3 szekcióülés foglalkozott. Sajnos ezeken a szekciókon viszonylag kevés előadást tartottak az e témakörben kétségkívül járatosabb szocialista országok kutatói; a nyugati előadások pedig a téma kívülről való kezelésével és az aktuális problémák ismeretének hiányában aligha járultak hozzá a tervezési modellek számunkra is hasznosítható továbbfejlesztéséhez. *D. Conn* (Ohio) két előadást is tartott, melyekben a szocialista gazdaságok felépítését, információáramlását, döntési mechanizmusát és ösztönzőrendszerét vizsgálta. *M. Deleau* (Paris) a gazdasági szabályozás egyes matematikai vonatkozásaival foglalkozott. *M. Desai* (Heverlee) a tervezési ciklusok vizsgálatára két differenciálegyenletből álló rendszert konstruált, mely segítségével a gazdaság stabil, illetve instabil növekedési szakaszait elemezte. *V. Simunek* (Kent) csehszlovák származású amerikai professzor egy – a szocialista tervezés tapasztalatait és módszereit adaptáló – óriásmodellt ismertetett. *G. Forbrig* (Rostock) az NDK-ban alkalmazott hatékonysági számítások egyes kérdéseiről tartott előadást. *Kádas K*, pedig a szállítás iránti igények ökonometriai vizsgálatairól.

Az ökonometriai modellek közül több előadás foglalkozott a *rövidtávú* modellekkel kapcsolatos specifikációs, tesztelési és becslési kérdésekkel. *W. Maciejewski* (Varsó) a lengyel gazdaság KP-3K negyedéves modelljét ismertette, amely 17 egyenletből áll. A modellt 1967 – 75-ös idősorok alapján számszerűsítették. Vizsgálni akarták, hogy az 1971-ben bevezetett gazdasági reformok milyen irányú és nagyságrendű változásokat okoztak a legfontosabb változók közötti kapcsolatokban. *G. Gelauff* (Rotterdam) előadásában éppen az olyan negyedéves modellek becslési problémáival foglalkozott, amelyeknél hiányos a negyedéves adatbázis. Miután ismertette – meglehetősen jól

rendszerezve – azokat a módszereket, melyekkel a hiányos negyedéves adatbázis alapján számszerűsített rövidtávú modellek paraméterbecslése elvégezhető, a különböző módszerekkel készített paraméterbecslések kisminta tulajdonságait hasonlította össze Monte Carlo módszerrel. *H. Erlat* (Ankara) ugyancsak a hiányos megfigyelésekkel rendelkező negyedéves modellek becslési problémáival foglalkozott előadásában. Az ismertetett módszer a magyarázó változókhoz rendelt dummy változók segítségével és az általánosított legkisebb négyzetek módszerének felhasználásával végzi el a hiányos negyedéves adatbázison becsült szimultán lineáris regressziós modellek becslését.

A kifejezetten matematikai-statisztikai *módszertani előadások* közül kettőt említünk meg. *N. E. Savin* (Cambridge) a regressziós együtthatók közötti lineáris összefüggések tesztelésével foglalkozott. Általánosította Scheffé eljárását, s így egy olyan módszert kapott, amely úgy viszonylik az ökonometriai modelleknél használatos Wald teszthez, mint az eredeti Scheffé eljárás az *F* próbához: ha a Wald-teszt elveti a lineáris összefüggések hipotézisét, az általánosított Scheffé eljárás azt is megmondja, hogy az összefüggések közül melyik vezetett a hipotézis elvetéséhez. A korábban említett *A. Deaton* (Bristol) előadása egy próbát mutatott be annak eldöntésére, hogy a változók között lineáris, vagy logaritmikus összefüggés áll-e fenn. Ez a kérdés ekvivalens azzal a problémával, hogy a hiba additív, avagy multiplikatív. Az előadás a statisztikai próba számítási kérdéseivel is foglalkozott.

A fogyasztás modellezésével kapcsolatos kérdésekről három szekcióülésen volt szó. A dinamikus struktúrák modellezésével foglalkozó ülésen *J. E. H. Davidson* (Warwick) tartott előadást a fogyasztói kiadás és a jövedelem kapcsolatát leíró ökonometriai modellekről. Három különböző modellt dolgoztak ki az Egyesült Királyság fogyasztására vonatkozóan, amelyek nagyon különböző jövedelemrugalmasságokat eredményeztek. A *keresletelemzéssel* és a *fogyasztói magatartás vizsgálatával* egy négy előadásból álló szekció foglalkozott. Ezen előadások közül az egyik elméleti jellegű volt; *D. Weiserbs* (Louvain) előadása egy konkrét preferenciafüggvényt és az abból levezetett keresleti egyenleteket tárgyalta. Két előadása Hollandiára, illetve Ausztriára jellemző fogyasztói magatartást vizsgált. *D. H. Saks* (East Lansing) előadása a családi jövedelemelosztást és az iskoláztatási költségek alakulását elemezte azzal a gyakorlati céllal, hogy megállapítsa, hogy a tandíj és ösztöndíjrendszer milyen mértékben egyenlíti ki a jövedelemdifferenciákat. Egy másik szekció a hasznossági függvény segítségével meghatározott *keresleti modellekkel* foglalkozott. Itt a legérdekesebb előadás *A. Deaton*nak (Bristol) a költségfüggvény duálisáról az ún. transzformációs függvényekről tartott előadása volt.

A gazdasági szabályozás modellezési kérdéseivel foglalkozó előadások közül kiemeljük *K. W. Clements* (Chicago) előadását, mely egy nyílt gazdaság többszektoros egyensúlyi modelljének specifikációját és becslését ismertette. A modell lényegében azt vizsgálta, hogy bizonyos gazdaságpolitikai változók (növekedési ütem, hitelállomány, adók, árfolyamok stb.) hogyan befolyásolnak olyan fontos gazdasági folyamatokat mint pl. a kereskedelmi mérleg alakulása. *A. H. Hallett* előadása azzal a kérdéssel foglalkozott, hogy hogyan deríthető ki egy több időszakot átölelő optimális politikáról, hogy érzékeny-e parametrikus változásokra, illetve exogén információkra. Az előadás a probléma megoldására alkalmas szabályozáselméleti módszereket ismertetett. *G. Tintner* (Bécs) és szerzőtársai az osztrák gazdaság dinamikus lineáris modelljét mutatták be. Ennek kapcsán az egyensúlyi helyzetet és a stabilitást két szempontból vizsgálták. Egyrészt foglalkoztak azzal a kérdéssel, hogy a gazdasági rendszerben bekövetkezett zavarok után az egyensúlyi helyzet milyen eszközökkel és milyen idő alatt állítható vissza, másrészt azt vizsgálták, hogy a stabilitás mennyire érzékeny a különféle külső változókra.

Az *egyéb* csoportba tartozó témák közül néhány olyan előadást emelünk ki, melyek az optimális gazdasági növekedéssel, a termelési függvények elemzésével, valamint a pénzügyi politikai céljait szolgáló modellekkel foglalkoztak. *Az optimális növekedést* az előadók az egyensúlyelmélettel összefüggésben vizsgálták. *J. M. Harwick* (Ontario) modelljében az optimális növekedést a volumen növekvő hozadékának feltételezése mellett írja le, ez lehetőséget nyújt a beruházások optimális időbeli szabályozásának és egy optimális megtakarítási szabálynak a megfogalmazására. A növekvő hozadék feltételezése *J. Cremer* (Paris) modelljének is központi kérdése, ugyanakkor ez a modell több, egymástól függő termék iránti kereslet kielégítésének hosszútávú tervezésére alkalmas.

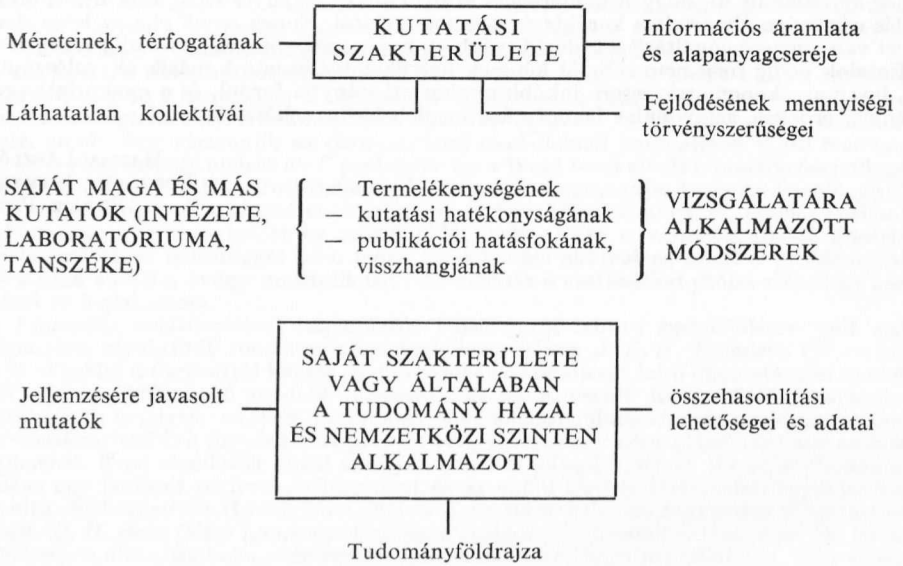
Kifejezetten a *termelési függvények* elemzésével és számszerűsítésével leginkább *T. Pray* (New York) és szerzőtársa foglalkozott, akik dolgozatukban a megtestesült és meg nem testesült technikai változás egyidejű számszerűsítését kísérelték meg vállalati adatok alapján. *A. Cukierman* professzor (New York) modelljében a gazdaságban a kereslet vagy a kínálat oldalán végbemenő véletlen megrázkódtatások hatását vizsgálta.

A *pénzügyi politika* céljait szolgáló modellek ismertetésével szintén több előadás foglalkozott. Ezek közül *M. M. G. Fase* (Amsterdam) és szerzőtársa dolgozatát emelhetjük ki, akik a bankkölcsönök iránti kereslet és a kölcsönzési ráta meghatározást tűzték ki célul.

Az előadások meghallgatása és a résztvevőkkel való beszélgetések alapján az az általános kép alakult ki, hogy a konferencia színvonala és az egyes előadások iránti érdeklődés némiképp elmaradt a korábbi évek színvonalától. Ennek egyik oka az lehet, hogy most van „generációváltás" a tudományágban, így a „nagy öregek" már alig szerepeltek, a fiatalok pedig még nem mindig képesek helyüket betölteni. A másik ok valószínűleg az, hogy az ökonometria egyre inkább gyakorlati irányba fordul, és a gyakorlati eredmények érdekes, színvonalas bemutatása meglehetősen nehéz feladat.

Hamza Lászlóné
Losonczy Istvánné
Subicz Péter

Önt bizonyára érdekli



mert tudja, hogy

A TUDOMÁNY MAI FEJLŐDÉSI SZAKASZÁBAN, A KUTATÁSBAN CSAKIS ÚGY

– LEHET EREDMÉNYES
– TERVEZHETI MUNKÁJÁT
– VÁLASZTHAT SAJÁT MAGA, INTÉZETE, ORSZÁGA SZEMPONTJÁBÓL LEGMEGFELELŐBB TÉMÁT, HA

a fenti kérdésekre megtalálja a választ

Hol? Hol? Hol? Hol? Hol?

Scientometrics

An International Journal for All Quantitative Aspects of the Science of Science and Science Policy

című

angol nyelvű nemzetközi folyóiratban

AKADÉMIAI KIADÓ, Budapest kiadja ELSEVIER PUBLISHING CO., Amsterdam

Saját eredménye e területen — ÖNNEK TALÁN VAN — Véleménye, megjegyzése

Megírt vagy megirandó közleménye

A FOLYÓIRAT EZEKET ÉRDEKLŐDÉSSEL VÁRJA

Cím: Dr. J. VLACHY Kankovského 1241 180 00 Praha 8 CSSR

vagy Dr. T. BRAUN Eötvös Loránd Tudományegyetem
1443 Budapest, P.F. 123, Magyarország

1828—1978

MEGJELENT
AZ AKADÉMIAI KÖNYVKIADÁS
150. ÉVÉBEN

A kiadásért felel az Akadémiai Kiadó igazgatója.
Műszaki szerkesztő: Sándor István
A kézirat nyomdába érkezett: 1978. I. 26. — Terjedelem: 10,85 (A/5) ív
78.5458 Akadémiai Nyomda, Budapest — Felelős vezető: Bernát György

CONTENTS

BOOK REVIEWS

SCIENTIFIC LIFE

СОДЕРЖАНИЕ

О КНИГАХ

НАУЧНАЯ ЖИЗНЬ

Ára: 12,— Ft
Előfizetés egy évre: 40,— Ft

INDEX: 26793
ISSN 0039—8128

TARTALOM

KÖNYVEKRŐL

TUDOMÁNYOS ÉLET

AKADÉMIAI KIADÓ, BUDAPEST