10.12663/PSYHUNG.3.2015.1.5

KLASSZIFIKÁCIÓS MÓDSZEREK MUTATÓI

Takács Szabolcs[1], Makrai Balázs[2], Vargha András[3]
Károli Gáspár Református Egyetem, Pszichológiai Intézet
1tanársegéd, 2 hallgató, 3tanár

Kivonat

Számos kutatásban előfordulhat, hogy valamilyen technika segítségével tipizálnunk, klasszifikálnunk kell az eseteinket. Ennek egyik bevett formája a klaszterelemzés. A klaszterezés során felmerülő egyik legfontosabb kérdés az, hogy mennyire jó a klaszterezés eredményeként kapott klaszterstruktúra. Ennek eldöntésére egy eljárást igyekszünk bemutatni – továbbá érzékeltetni, hogy e módszerben rejlő döntéshozatal közel sem statisztikai/matematikai feladat, így az eljárás alkalmazói nem vonhatják ki magukat egy-egy klaszterezés szakmaiságának, helytállóságának megítéléséből, a felelős döntéshozatalból.

Kulcsszavak: kockázatvállalás ▪ klasszifikáció, klaszteranalízis ▪ hatékonyság ▪ ROPstat ▪ SPSS©

Abstract

In this study we demonstrated examples for measuring the goodness of a classification method. In the beginning the definitions of distances are given for the reader, then a short summary clarifies the differences between hierarchical and non-hierarchical classification – focusing on the non-hierarchical k-means clustering method. Utilizing the results of a risk-connected research, we made a three-dimensional classification in SPSS© with the help of the Xie–Beni index and evaluated the output. Later we repeated the run but now in the statistical program ROPstat©; that time we used the Silhouette index, the point-biserial correlation coefficient and the EESS-percentage. The output resulted in a rather similar consequence with the ascertainment that the simultaneous usage of more classification indexes let us to gain a more precise picture about the goodness of the final cluster structure.

Keywords: attitude toward risk ▪ classification ▪ cluster analysis ▪ efficiency ▪ ROPstat ▪ SPSS©

BEVEZETŐ

Egy statisztikai vizsgálat általános célját úgy fogalmazhatjuk meg: az általunk vizsgált populációról információkat szerezni. Ezt két jól elkülöníthető irányból tehetjük meg: egyik oldalról vizsgálatunk tárgyát képezhetik azok a véletlen változók, jelenségek, melyekkel magát a populációt tudjuk jellemezni – a másik oldalról pedig magát a populációt helyezzük a középpontba. Ezen utóbbi vizsgálatokat összefoglalóan személyorientált módszereknek nevezhetjük. A személyorientált megközelítés holisztikus, dinamikus szemlélet (Magnusson és Allen, 1983), amelyben a hangsúly azon vizsgálati személyek típusokba sorolásán van, akik sok változó tekintetében hasonló együttes holisztikus mintázatot mutatnak (Bergman, Magnusson és El-Khouri, 2003).

Jelen dolgozatunkban a klaszterelemzést, mint az egyik leggyakrabban alkalmazott személyorientált vizsgálati módszert szeretnénk bemutatni – illetve az eljárás néhány adekvációs mutatóját fogjuk ismertetni. A klaszterelemzés célja elsősorban az, hogy a populáció egyedei között olyan csoportokat hozzunk létre, melyek sok változó együttes mintázataiban egymástól a lehető legjobban különböznek (közöttük nagy távolságok legyenek), míg egy-egy csoporton (klaszteren) belül lehetőleg ne találjunk nagy különbségeket (a csoportok, klaszterek legyenek homogének). Látható tehát, hogy itt nem a változók közötti kapcsolat feltárása az elsődleges cél (ami a változóorientált módszerek jellemzője), hanem olyan személyeket csoportosítunk egybe, ahol a személyen belüli, több változó által kirajzolt együttes mintázat hasonló.

A személyiségpszichológiában kurrens kutatási irány köszönhető a főként klaszteranalízist alkalmazó személyorientált szemléletnek, miszerint nem a különböző változók menti (interindividuális) különbségekre, hanem a személyen belüli (intraindividuális) tipikus holisztikus mintázatokra, a személyiségjellemzők konfigurációira érdemes fókuszálni (Asendorpf, 2002). John és mtsai (2013) mindezek figyelembevételével a klaszteralapú Big Five-személyiségprototípusok kapcsolatát vizsgálták patológiás mutatókkal.

A klaszteranalízis számos további kurrens pszichológiai kutatásban is sikerrel alkalmazott eljárásnak bizonyult, Marton, Surányi, Farkas és Egri (2014) fogyatékkal élőknek a fizikai, pszichoszociális és kognitív mutatóin végzett klaszterelemzéssel azonosították a jól, ill. rosszul funkcionálás különböző mintázatait. Ugyanígy klaszterelemzéssel sikerült feltárni tehetséges serdülők lelki problémamintázatait (Bagdy, Mirnics és Kövi, 2014; Bagdy, Kövi és Mirnics, 2014).

A klaszteranalízis alapvetően egyszerű eljárás – és talán épp az egyszerűsége adja a nehézségét is. Számos területen jól alkalmazható, könnyen átlátható, azonban annak eldöntése, hogy egy adott változószett alapján az egyedek besorolása mennyire jó, használható, már lényegesen nehezebb kérdés. A klaszteranalízisnek orvosi, biológiai vagy geológiai alkalmazásai ugyanúgy vannak, mint pszichológiaiak vagy nyelvészetiek – a módszerek kellően

általánosak és könnyen alkalmazhatók. Ezen széles spektrum miatt is fontos, hogy a módszer jóságát, megbízhatóságát valamilyen eszköztárral mérni tudjuk.

A klaszterezés egyik kulcsmomentuma az, hogy akár az egyedek, akár a klaszterek között miként mérjük a távolságokat. Ez látszólag könnyű feladat – pedig korántsem az. Mutatunk néhány példát arra, hogy milyen távolságokat definiálhatunk akár az egyedek, akár a klaszterek között.

Egyedek közötti távolságok értelmezését a legkönnyebben egy konkrét példán tudjuk felvázolni. Tegyük fel, hogy mindössze két változót mérünk, és a két alanyunk értékei legyenek (1; 5) és (4; 9). Definiálunk közöttük négy különböző távolságot:

1) *Euklideszi távolság*: ez a két pontot összekötő szakasz hossza:

$$\sqrt{(1-4)^2+(5-9)^2}=5$$

2) *Euklideszi távolság négyzete*: ez a két pontot összekötő szakasz hosszának négyzete – igen gyakori, hogy ezt alkalmazzuk az 1-es pontban tárgyalt távolság helyett, jobb matematikai tulajdonságai miatt:

$$(1-4)^2+(5-9)^2=25$$

3) *Manhattan-távolság*: két pont között úgy számolható, mintha a koordináta-rendszer rácsvonalán sétálva szeretnénk eljutni egyik pontból a másik pontba:

$$|1-4|+|5-9|=7$$

4) *Maximumtávolság*: a két pont koordinátái közötti legnagyobb különbséget mérjük:

$$max\{|1-4|;|5-9|\}=4$$

Összefoglalva:

Távolság fajtája	Pontpár	Távolság mértéke
Euklideszi távolság	(1; 5) és (4; 9)	5
Euklideszi távolság négyzete	(1; 5) és (4; 9)	25
Manhattan-távolság	(1; 5) és (4; 9)	7
Maximumtávolság	(1; 5) és (4; 9)	4

1. táblázat. Az adott pontpárok távolságai különböző távolságmértékek mellett

Megfigyelhető, hogy ugyanazon pontpárokat alkalmazva más és más távolságot kapunk két pont között.[1] Az egyedeket ilyen távolság alapján vonjuk össze klaszterekbe.

[1] Elmondható, hogy a távolságok szimmetrikusak, azaz az (1; 5) és (4; 9) pontok között ugyanaz a távolság, mint ha a (4; 9) és (1; 5) pontpárok között vennénk fel. Ez egyértelműnek látszik, pedig

Kérdés az is, hogy két klaszter között miként mérjük a távolságot. Erre vonatkozóan illusztrációs céllal három lehetséges metódust vázolunk:

1) A két klaszter közötti távolságot az mondja meg, hogy mekkora a két klaszter egyedei közötti legkisebb távolság.
2) *Átlagos távolság*: a két klaszter távolságát úgy kapjuk, hogy átlagoljuk a két klaszter egyedei között páronként kiszámított távolságokat.
3) *Ward-módszer*: eszerint az a két klaszter van a legközelebb egymáshoz, amelyek összevonása során a legkisebb mértékben növekszik a pontok közötti négyzetes eltérés (azaz amely két klaszter „leginkább hasonlít egymáshoz" a többi klaszterhez képest).

Mind a klaszterek közötti távolságokról (illetve más megközelítésben összevonási metódusokról), mind az egyedek közötti távolságokról számos helyen olvashatunk: például Wilson és Ritter könyvében, melyben a különböző távolságok kiszámításának számítógépes eljárásait is ismertetik (Wilson, Ritter, 2000) vagy magyar nyelven akár a Móri Tamás és Székely Gábor által szerkesztett könyvből (Móri, Székely, 1986).

Ezek után az alábbi feladatot kívánjuk megoldani: tegyük fel, hogy egy adatállományon elvégeztünk egy klaszterelemzést, melynek során az egyedeket besoroltuk például 3 klaszterbe. Mitől lesz ez jobb, mint egy olyan klaszterezés, amelyben 2 vagy 4 klasztert hozunk létre? Jobb-e ez a 3 klaszteres modell, mint egy másik 3 klaszteres, ahol azonban más távolságot definiálunk akár a klaszterek, akár az egyedek között?

HIERARCHIKUS ÉS NEM HIERARCHIKUS KLASZTEREZÉSI ELJÁRÁSOK

A klaszterezési eljárások egyik lehetséges felosztása az ún. hierarchikus és nem hierarchikus szempont szerinti besorolás. Lényeges különbség a két eljárás között, hogy a klaszterek száma a hierarchikus módszerekben nincs előre meghatározva, míg a nem hierarchikus osztályozásoknál előre meghatározott számú klaszterbe sorolódnak az esetek.

nem az: gondoljunk arra, hogy ha például repülőgéppel utazunk, akkor két város között a repülőút akár még időben sem feltétlenül azonos, tehát „az út hossza A és B város között" nem ugyanaz, ha A-ból utazunk B-be, mint ha B-ből utazunk A-ba. Ezekre az esetekre nem szeretnénk kitérni, de például közgazdasági, marketinges alkalmazásairól Van den Poel és társai több dolgozatot is jegyeznek (Prinzie & Van den Poel, 2006a, 2006b, 2007).

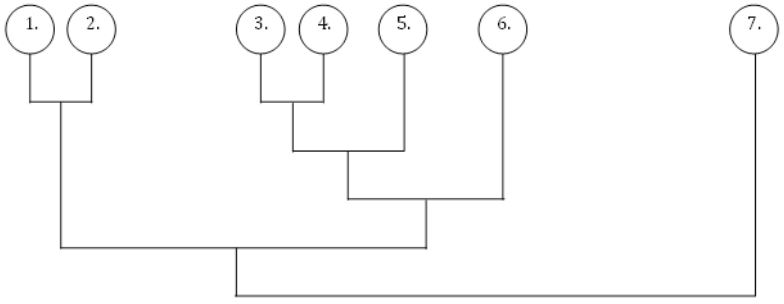
Hierarchikus klaszterezés

1) A hierarchikus eljárásban a klaszterek folyamatos egyesítésével vagy szétbontásával alakítjuk a csoportok számát, így jutva el a két lehetséges végállapothoz, tehát itt klasszifikációsorozatokat kapunk. A két eljárástípus: minden egyes elem különállva képez egy klasztert, majd lépésenkénti összevonások sorozatával létrejön egyetlen nagy – minden elemet tartalmazó – klaszter;
2) a vizsgálni kívánt esetek összessége kezdetben egyetlen klaszterben tömörül, majd a felosztás végére minden elem külön képez egy-egy klasztert.

Az előbbi típust – módszerének megfelelően – **összevonó** (agglomerative), az utóbbit pedig **felosztó** (divisive) eljárásnak nevezzük. Megjegyezzük, hogy az ismertebb szoftverek (IBM-SPSS©, továbbiakban csak SPSS©, R, SAS, ROPstat) mind az első, összevonó eljárást alkalmazzák.

A klaszterek összevonására/szeparálására több eljárás is ismert: ezek lényegében azonosak azzal, ahogy a klaszterek közötti távolságot mérjük. Hogy éppen melyik módszer szerint, azt a megfelelő program futtatása során általában opcionálisan tudjuk beállítani.

Jogosan merül fel a kérdés, hogy a fenti ciklusok valamelyikének futtatásánál hol érdemes megállni (ha nem megyünk el a végállomásig), tehát melyik az a klaszterezési állapot, amely már jól reprezentálja egy adott adathalmaz csoportosulási tulajdonságait. Ez részint az elemzést végző személy saját döntése, ugyanis különböző céloknak és szempontoknak megfelelően több jó megoldás is létezhet. Azonban a számszerű mérhetőség itt is, mint a statisztika valamennyi területén, alapvető elvárás. E téren a tanulmány további részében ismertetett adekvációs mutatószámok fognak segíteni.



1. ábra. Összevonó hierarchikus klaszterezési eljárást szemléltető faábra (dendogram)

Nem hierarchikus klaszterezés

A nem hierarchikus eljárások esetében a klaszterezési folyamat a kívánt klaszterszám megadásával kezdődik. Alábbiakban a sokféle lehetséges algoritmus közül most egyetlen módszert, a legszélesebb körben elterjedt, ún. **k-központú** (k-means) eljárást ismertetjük. Ennek egyik első leírása MacQueen 1967-es tanulmánya (MacQueen, 1967).

K-központú klaszteranalízis

A k-központú algoritmus során első lépésként szükséges a létrehozni kívánt klaszterek számának megadása *(k)*. Ez egyben (hacsak valamilyen egyéb szempont nem indokol egy konkrét értéket) a feladat nehézsége, hiszen sokdimenziós adatállomány esetén még a grafikus ábrázolás lehetősége sem áll fenn, tehát előre, mintegy „vakon" kell a klaszterszámot meghatározni. A folyamat a következő:

(1) A megadott *k* paraméter segítségével az algoritmus létrehoz *k* darab klasztert úgy, hogy véletlenszerűen meghatározza a középpontjait.
(2) Minden elemet a legközelebb[2] eső klaszterközépponthoz rendel, így egyúttal megtörténik a klaszterbe sorolás, tehát létrejönnek a klaszterek.
(3) Az algoritmus meghatározza a klaszterek új középpontjait, tehát a véletlenszerűen létrehozott értékek helyett most már a klaszterbe tartozó elemek által számítható középpontokat.
(4) Ismét meghatározza az eljárás valamennyi elem távolságát az új középpontoktól. Amennyiben egy elem egy másik klaszterközéphez közelebb került azok újraszámítása révén, akkor másik klaszterbe sorolódik át. Az elemek fennmaradó része a helyén marad.
(5) Az algoritmus ismételte újraszámolja a klaszterközéppontokat, majd újból átsorolja az elemeket. Ezt a folyamatot hívjuk iterációnak, amelyet addig ismételünk, amíg (a) nem sorolhatóak át az elemek tovább, következésképpen mindegyik a megfelelő klaszterbe került, vagy (b) a program eléri az előre meghatározott iterációszámot (ezt annak érdekében célszerű megtenni, hogy az eljárás során ne jöjjön létre végtelen ciklus).

A fenti algoritmusból egyértelműen látszik, hogy a k-központú eljárás célja, hogy az egyes klasztereken belüli variancia a lehető legkisebb legyen. Ez formálisan a következőt jelenti:

Egy adathalmazt (X_N) kívánunk adott számú *(k)* klaszterre *(C)* bontani.

[2] A távolságok meghatározásáról lásd a Bevezetőben tárgyaltakat.

$$C = \{C_1, C_2, \cdots, C_k\}$$

Egy klaszter jóságát annak belső négyzetösszege *(SS)* méri. Az adathalmazunk valamennyi elemének *(x)* távolságát meghatározzuk az aktuálisan létrehozott klaszterközéppontoktól (M_i). Ezeknek az eltéréseknek a négyzetét klaszterek szerint összegezzük, a cél pedig, hogy az összes klaszteren belül képezett négyzetösszeg *(SS(S))* minimális legyen. Ekkor kap szerepet az iterációs folyamat.

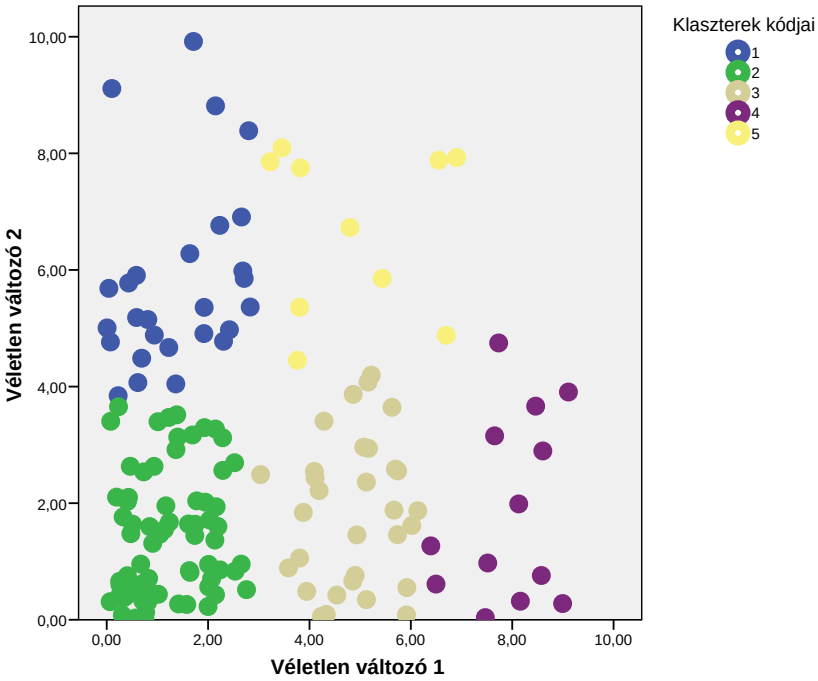
$$min\, SS(C) \quad vagy\ másképpen \quad min \sum_{i=1}^{k} SS(C_i)$$

Az esetek besorolása, tehát az analízis lefuttatása utáni végeredmény az újabb és újabb lefuttatások után változhat, ahogy a kezdeti lépésben létrehozott klaszterek és klaszterközéppontok is változnak minden eljárásban. Ennek tükrében tehát csak a kiindulási paraméterek és olyan tényezők állandók, amelyek az eljárás során adhatók meg (pl. hányszor iteráljon az algoritmus). Eme tulajdonsága és a nagy mintákra való hatékony alkalmazhatósága miatt az egyik leggyakrabban használt eljárás.

ADEKVÁCIÓS MUTATÓK A KLASZTERANALÍZISBEN

Tételezzük fel, hogy valamilyen módszerrel eljutottunk odáig, hogy van egy kész klaszterrendszerünk. Ekkor annak eldöntésére, hogy a klaszterezés mennyire jó, több mutatót is alkothatunk. Például az SPSS©[3] segédanyagaiban arra vonatkozóan nem találunk semmifajta érdemi információt, hogy egy általunk elkészített klaszterezés mennyire elfogadható. A ROPstat[4] viszont mérni tudja egy sor klaszteradekvációs mutató (EESS%, klaszterhomogenitási együtthatók, Silhouette-mutató, pontbiszeriális korreláció) segítségével valamely klaszterstruktúra jóságát.

[3] www-01.ibm.com/software/analytics/spss
[4] Lásd www.ropstat.com, illetve Vargha (2007) és Vargha (2008).



2. ábra. K-központú klaszteranalízis eredménye, N=150, k=5, iterációk száma 10.

Az R©[5] programcsomagban számos mutatóval dolgozhatunk, mely mutatók teljes leírását megtalálhatjuk a programhoz tartozó dokumentációkban. Dolgozatunkban azt szeretnénk bemutatni, hogy egy általunk kiválasztott mutató segítségével miként tudjuk eldönteni, mennyire jó a klaszterezési eljárásunk. A következő alfejezetben ismertetünk egy olyan mutatót, melynek segítségével ezt a döntést akár magunk is meg tudjuk hozni.

Egy kiválasztott klaszterezési mutató ismertetése

A Xie–Beni-index

A Xie–Beni-index megalkotói 1991-ben írták dolgozatukat a klaszterezés validitását mérő mutatójukról (Xie, Beni, 1991), bár e mutatót elsősorban nem a klasszikus klaszteranalízisben, hanem az úgynevezett fuzzy klaszterezés[6] során

[5] www.r-project.org

[6] A fuzzy klaszterezés esetén az egyedek bizonyos valószínűséggel tartoznak csak egy-egy klaszterhez, azaz lényegében minden pontnak (egyednek) adott a valószínűsége, hogy egy-egy adott klasz-

szokták előszeretettel alkalmazni. Azért választottuk ezt a mutatót, mert az SPSS© segítségével néhány beállítással, illetve egy-két számítással magunk is kiszámíthatjuk értékét. Az erre vonatkozó Syntax állományt az első melléklet tartalmazza.

Az index kiszámításához vegyük alapul a teljes négyzetes hibát (azaz a megfigyelési egységek négyzetes eltérését az adott, saját klaszterük középpontjától, súlypontjától – majd tekintsük ezek összességét):

$$W_K = \sum_{k=1}^{K} \sum_{i \in I_k} \left(X_i^{\{k\}} - M^{\{k\}} \right)^2,$$

majd ezt a mennyiséget átlagoljuk:

$$W = \frac{W_K}{N}.$$

Vegyük észre, hogy e fenti mennyiség lényegében a korábban ismertetett euklideszi távolságok négyzeteivel egyezik meg (hiszen 1-1 klaszterben a klaszterek középpontjaitól vett euklideszi négyzetes távolságot összegzi, majd átlagolja). A négyzetes távolságok a klasztereken belül a belső variancia kiszámítására szolgálnak – tehát a fenti átlagolás nem más, mint a mintában a saját klaszterközépponttól való átlagos távolság.

Tekintsük továbbá az adott klaszterek távolságait (négyzetes euklideszi távolságként definiálva):

$$D_{k,k\prime} = d(M_k, M_{k\prime}).$$

Vegyük most $D = \min\{D_{k,k'}\}$ az egymáshoz legközelebbi két klaszter távolságaként.

Ekkor a Xie–Beni-indexet az alábbi formulával definiáljuk:

$$XB = \frac{W}{D}.$$

Ebben az esetben azt szeretnénk, ha az index minél kisebb lenne, hiszen a *W* érték a belső távolságok átlaga, míg a *D* érték a külső, klaszterek közötti páronkénti távolságok minimuma. Amennyiben a klaszterek homogének, úgy a *W* érték kicsi, illetve jól szeparált rendszer esetén a klaszterek távol kerülnek egy-

terhez tartozzon. Így a klaszterek nem szeparáltak – és lényegében minden pont minden klaszterhez is tartozhat akár adott valószínűséggel.

mástól, tehát a *D* érték nagy – így az *XB* érték szintén alacsony. Az alacsony *XB* érték tehát azt jelzi, hogy homogén, egymástól jól szeparált klasztereket alkottunk – míg magas *XB* érték nem jól elkülönülő, heterogén klaszterstruktúrát sejtet.

SZEMÉLYORIENTÁLT KUTATÁSOK AKTUÁLIS EREDMÉNYEI

A személyorientált elemzésekről elmondható, hogy az informatikai fejlesztések nagy nyertesei közé tartoznak. A jelenkori adatgyűjtések (internetes, ill. nagy mintás felmérések) nehezen tették lehetővé, hogy e nagy mintás elemzések feldolgozásánál az egyedek közötti tipológiákra, csoportokra koncentráljanak, így a feldolgozásokat elsősorban a változók felől értelmezték. A statisztikai programok fejlődése magával hozta az újfajta megközelítések lehetőségét is – ez pedig új eredményeket is hozott magával.

Kiderült, hogy pusztán a változók elemzésén keresztül nem feltétlenül lehetséges az egyedek, személyek közötti különbségek vizsgálata, illetve az ott tetten érhető különbségek szakszerű feltárása (Borsboom, 2003). Szükségszerű volt tehát, hogy a felmérésekben részt vevők közötti, egyedszinten megjelenni látszó különbségek és hasonlóságok feltárására a meglévő módszereket is továbbfejlesszék. Az 2000-es évek elejétől számos, nem hagyományos elemzési eljárás vált könnyen elérhetővé: ezek összehasonlító elemzését mutatja be például Nock munkatársával közös cikkében (Nock és Nielsen, 2006). Dolgozatukban a k-központú elemzést, a valószínűségeken alapuló fuzzy klaszterezést, illetve különböző súlyozott változataikat hasonlítják össze. Teszteléseik során azt tapasztalták, hogy a különböző adatstruktúrákon a különböző módszerek nem egyformán viselkednek, ezért a módszerek iteratív kombinációját tartják egyfajta járható útnak.

Más irányt mutat be a klasszifikációs módszereket érintő fejlesztésekben az az újfajta klasszifikációs eljárás és terület (Surányi, 2011), melyben nem többfajta klaszterelemzést kombinálnak, hanem a klaszterelemzést kombinálják például sűrűsödéselemzéssel és más statisztikai eljárásokkal. Ezen kombinációk arra szolgálnak, hogy akár a szélsőséges eseteket kiszűrjük, akár az ideális klaszterszámot (sűrűsödési helyek segítségével) megkeressük, illetve akár a kialakított klaszterek vizuális megjelenítését elősegítsük. Ilyen elemzések könnyűszerrel elvégezhetők a ROPstat legfrissebb változatának személyorientált menüpontja segítségével (lásd Vargha, Torma és Bergman, 2014, megjelenés előtt).

Dolgozatunkban egy harmadik fejlesztési irányt szerettünk volna bemutatni: az adott klaszterezési eljárásunk jóságát, adekvátságát szeretnénk megmérni, illetve szeretnénk jellemezni. Ehhez számos mutató rendelkezésünkre áll, mint ahogy korábban már bemutattuk. Ezen mutatók együttes értelmezése segíthet, hogy a nekünk leginkább megfelelő döntést hozzuk meg.

SZÁMÍTÁSI TAPASZTALATOK – ESETBEMUTATÁS

Az alkalmazott esetben tehát 3 változó szerint végzünk elemzést (hogy miért csak 3 változót választottunk, annak az interpretálásnál lesz fontos szerepe). Elemzésünkben kérdőíves felmérést alkalmazva arra voltunk kíváncsiak, hogy a vizsgált alanyok milyen valószínűséggel, milyen mértékű kockázattal és milyen haszon reményében vállalnak sporttevékenységeik közben kockázatot. A felméréshez a Dospert kérdőívet alkalmaztuk, melyről bővebben például Radnóti közleményében olvashatunk (Radnóti, 2008). Radnóti dolgozatában a kérdőív magyarországi adaptációját mutatja be, a dolgozatban a kérdőív érvényességének és validitásának vizsgálata is elolvasható.

Vizsgálatunkban a teljes kérdőívet használtuk, most azonban kizárólag a sportra vonatkozó skálákat alkalmazzuk a módszer gyakorlati hátterének bemutatására.[7] Kérdésünk az, hogy a kockázatvállalás gyakorisága (valószínűsége), a kockázat nagyságának mértéke, illetve a kockázat vállalásával elérhető haszon alapján milyen csoportok hozhatók létre – illetve az is kérdés, hogy e csoportok mennyire különülnek el egymástól.

A most alkalmazandó algoritmusrészlet tehát az alábbi:

1) 3 klaszterre k-központú elemzés lefuttatása.
2) Klaszterek középpontjainak és az esetek klaszterközépponttól való távolságának mentése.
3) A Xie–Beni-index meghatározása ezen információk alapján: négyzetes (euklideszi) távolságok átlaga, valamint a klaszterközéppontok páronkénti távolságainak minimuma.

A fenti számítást először az SPSS© programcsomagban mutatjuk be, melynek Syntax állományát a mellékletekben szerepeltetjük, így elemzésünk megismételhető (természetesen például Excel© alkalmazásával is könnyedén elvégezhetjük ezeket a számításokat).

Másodszor, a ROPstat használatával egy azonos célú elemzést mutatunk be, külön kitérve a két programban található különbségekre. A ROPstat használata esetén – miután ott nincs Syntax állomány – a fenti algoritmus lépéseit közvetlenül a menürendszerben találhatjuk meg.

Az SPSS©-ben elvégzett elemzés első lépéséhez használt Syntax állomány az első mellékletben olvasható. A keletkezett klaszterek középpontjai a 2. táblázatban láthatók.

[7] Az adatállomány *.por formátumban elérhető a www.ropstat.com oldalról letölthető demóváltozatban (a c:_vargha\ropstat\dat\demodat mappában). Az ilyen fájlok mind a ROPstat, mind az SPSS© programba egyszerűen beolvashatók.

	1. klaszter	2. klaszter	3. klaszter
Sportolásban való kockázatvállalás, valószínűség	2,51	3,87	1,48
Sportolásban való kockázatvállalás, kockázat mértéke	3,11	2,91	3,76
Sportolásban való kockázatvállalás, elvárt haszon nagysága	2,50	3,15	1,73

2. táblázat. Klaszterközéppontok koordinátái

A klaszterek középpontjai alapján a három klasztert az alábbi módon tudjuk jellemezni:

1) Az első klaszterben azt láthatjuk, hogy közepesen magas mind a kockázatvállalás valószínűsége, mind annak mértéke, mind pedig az elvárt haszon.
2) A második klaszter esetén elmondható, hogy nagy valószínűséggel vállalnak nagy hasznossággal kecsegtető helyzeteket – azonban a kockázat mértékének alacsonynak kell lennie.
3) A harmadik klaszter esetén az igazi veszélykeresőket lehet megtalálni: kis valószínűséggel vállalnak kockázatot, nem igazán érdekli őket a belőle származó haszon – cserébe azonban nagynak kell lennie a kockázat mértékének. Azaz: akkor vállalnak kockázatos helyzeteket, ha azok valóban kockázatosak.

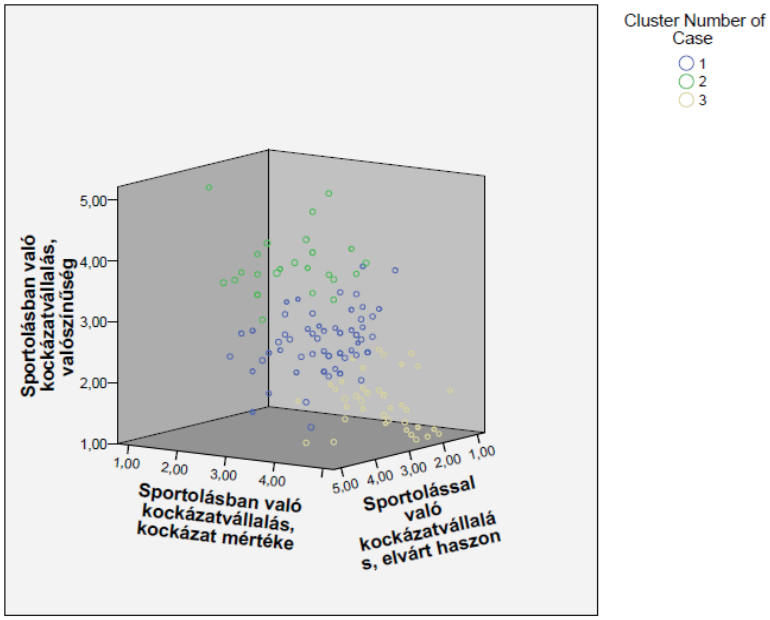
Természetesen e harmadik klaszter egyedei másként is értelmezhetők: miután náluk a kockázat mértéke számít, könnyen megeshet, hogy azok a helyzetek, amelyek mások számára már kockázatosnak tűnnének, számukra egészen hétköznapiak. Így önbevallásos kérdés esetén[8] (Milyen gyakran kerülnek „kockázatos" helyzetekbe?) nem azért alacsony a szintjük, mert valóban ritkán kerülnek ilyenbe, hanem azért, mert az ő esetükben az ingerküszöb máshol van.

Azt azonban, hogy a fenti klaszterek mennyire különülnek el egymástól, illetve milyen mértékben tekinthető jónak a fenti klaszterstruktúra, az SPSS semmilyen módon sem jelzi számunkra.

Általában egy-egy klaszterezés során a magasabb változószám miatt nem tudunk egyetlen 2 vagy 3 dimenziós ábrát szerkeszteni arról, hogy miként is fest klaszterezett pontfelhőnk. Azért mutattuk be az eljárásunkat háromdimenziós (3 változós) esettel, hogy ábrán is érzékeltetni tudjuk a hasonlóságot. Az SPSS©

[8] Fontos kiemelnünk: az önbevallás miatt nemcsak ebben az esetben, hanem mindhárom skála miatt árnyalódik a kockázatvállalásról alkotott kép. Például a kockázatvállalás gyakori volta mellett annak hasznossága is önbevallás-alapú. Így ezt a hasznosságot magasnak fogja önmaga számára értékelni – függetlenül attól, hogy egyébként az adott szituációból származó végkifejlet összességében jelent-e neki bárminemű objektív hasznot.

segítségével grafikusan is meg tudjuk jeleníteni (lásd 3. ábra), amit a számok mutatnak. Az ábra Syntax állományát a második melléklet tartalmazza.



3. ábra. A keletkezett klaszterek 3 dimenziós ábrája

A 3. ábrán megfigyelhető, hogy a zölddel jelzett ponthalmaz (2. klaszter) a kockázatvállalás dimenzió mentén magas tartományba esik, míg a kék (1. klaszter) a közepesbe. A fehér ponthalmaz az „előtérben" van, azaz magas a vállalt kockázat mértéke, azonban ezt ritkán teszi, másként fogalmazva: kis valószínűséggel (ha úgy tetszik, megfontoltan) vállal nagyobb kockázatot.

A Xie–Beni-index kiszámítása az SPSS programcsomag segítségével nem túl bonyolult (bár nem is feltétlenül triviális). Az indexben szereplő D értéket könnyen meghatározhatjuk, hiszen az SPSS a klaszterek középpontjainak távolságát megadja:

	1. klaszter	2. klaszter	3. klaszter
1. klaszter	0	1,528	1,440
2. klaszter	1,528	0	2,914
3. klaszter	1,440	2,914	0

3. táblázat. A klaszterek középpontjainak távolsága

Ezen távolságok legkisebbjének négyzete lesz a képletben szereplő D érték.

Amennyiben beállítottuk (a Syntax állományban) az egyedek távolságát saját klaszterük középpontjától, úgy ezek átlaga adja a képletben szereplő W értéket. Kettőjük hányadosa a keresett index.

	Esetszám	Összeg
Az esetek távolságösszegei a saját klaszter-középpontjuktól	130	112,22259

4. táblázat. Távolságösszegek

Azaz:

$$XB = \frac{\frac{112{,}22}{130}}{1{,}44^2} = 0{,}416.$$

Azonos elemzés elvégzése a ROPstat segítségével

A ROPstat programcsomag használata esetén nincs szükségünk más, külső alkalmazásra, hiszen ez a program automatikusan kiszámít legalább 3 olyan mutatót, melyek segítségével egy klaszterezési eljárás eredménye objektív módon összehasonlítható más klaszterezések eredményeivel. Ráadásul a ROPstat legújabb változata a Xie–Beni-indexet is kiszámítja (a hierarchikus elemzés felületén találjuk meg). Segítségével arra kaphatunk választ, hogy mely klaszterszám esetén van az indexnek lokális minimumértéke – hiszen ez a klaszterszám jó választás lehet a további vizsgálatokhoz. A továbbiakban azonban azt mutatjuk be, hogy egy k-központú elemzésben milyen döntést segítő mutatóink vanak.

Az elemzéseket tehát azonos változószetten végeztük el. A ROPstatban a Xie–Beni mellett több klaszteradekvációs mutatót is találhatunk, ezek az alábbiak:

1) **Silhouette-mutató**: a Silhouette mutatót Rousseeuw publikálta 1987-ben (Rousseeuw, 1987), és a klaszterek egymáshoz való távolságán alapul. Ez a komplex mutató egyben arról is tájékoztat, hogy a klaszterek mennyire homogének (belső egység), illetve arról is, hogy mennyire különböznek egymástól (szeparáltság). A mutató értéke –1 és 1 közé esik; általában elvárható, hogy elfogadható klaszterstruktúra esetén elérje a 0,5-ös értéket.
2) **Pontbiszeriális korreláció**: a pontbiszeriális korreláció nagyon egyszerű elven alapuló korrelációs együttható. Lényegében azt mondjuk, hogy akik egy klaszterbe tartoznak, legyenek közel egymáshoz, míg akik távol vannak egymástól, soroltassanak különböző klaszterekbe. Erről bővebben olvashatunk például Baker és Huber munkájában, akik a hagyomá-

nyos Pearson-féle korreláció helyett a Kendall-féle Gamma mutató segítségével is vizsgálták e mennyiség viselkedését (Baker, 1975).

3) **EESS% (Explained Error Sum of Squares %)**: Ez a mutató hasonlít talán a Xie–Beni-indexre abban a tekintetben, hogy az EESS% esetén a klaszterek belső négyzetes összege és a teljes négyzetes összeg arányát elemezzük. Értelemszerűen minél magasabb ez a mutató, annál inkább igaz az, hogy a klaszterek felállításával az egyedek közötti variancia egyre nagyobb hányada magyarázható, tehát a klaszterezésünk annál jobb. Miután ez a mutató egyfajta megmagyarázott varianciaarányként is értelmezhető, ezért azt is mondhatjuk, hogy a ROPstatban található klaszterezési eljárások[9] egyik legfontosabb adekvációs mutatója (de talán általánosságban is).

Az adatállományon végrehajtott elemzés (Ward-féle módszerrel végrehajtott hierarchikus klaszteranalízis után 3 klaszterre elvégzett relokáció) a ROPstat használatával az alábbi eredményeket[10] hozta a különböző sportolási szokásokat mérő változóink alapján:[10]

EESS%	XieBeni	Pontbisz	Sil.eh.	HCátlag	HCmin-HCmax
55,31	0,418	0,439	0,631	0,604	0,48-0,79

5. táblázat. Adekvációs mutatók a ROPstatban

NEM STANDARDIZÁLT ÁTLAGOK				
Klaszter	gyak.	Val_sp	Val_ko	Val_has
1	31	3,796	2,989	3
2	62	2,349	3,075	2,505
3	37	1,477	3,874	1,658

6. táblázat. Klaszterek tulajdonságai

Megfigyelhető, hogy az adott klasztermintázatok nagyon hasonlóak az SPSS©-ben nyertekhez[11] – legfeljebb a klaszterek számozásában vannak eltérések.

[9] Általános, amolyan ökölszabályként elfogadható, hogy azokat a klaszterstruktúrákat nevezzük jónak, amelyben az ESS% legalább 60-65%-os. Ezért a további alfejezetekben bemutatunk egy olyan megnövelt klaszterszámú esetet, amikor ez a mutató meghaladja az e szabály szerinti 60-65%-os elvárt mértéket.

[10] Fontos megjegyezni, hogy a ROPstatban lehetséges kombinálni a hierarchikus és a k-központú elemzést. Ha például egy hierarchikus klaszterezésben a 3-3 közötti kiírással kezdünk, majd az így kapott 3 klaszterre kérünk relokációt, várhatóan jobb eredményt kapunk, mint a véletlenszerű középpontokból indított esetekben.

[11] Általánosságban eltérések adódnak abból a szempontból, hogy az SPSS© mindenképpen nyers adatokkal számol, míg a ROPstat elemzéseiben lehetőségünk van standardizált adatokkal is számolni.

A mostani 1. klaszter egyedei azok, akik nagy valószínűséggel, alacsony kockázati rátával vállalnak nagy haszon reményében kockázatot (korábban ez volt a 2. klaszter). A mostani elemzés 2. klasztere lényegében mindenben közepes értéken szerepel (korábbi 2. klaszter), míg a 3. klaszter megegyezik a korábbi 3.-kal. Ez utóbbi esetén láthatjuk tehát azokat az egyedeket, akik jellemzően magas kockázatot vállalnak – igen csekély haszon reményében, ám igen kis valószínűséggel, tehát ritkán.

A ROPstat a fenti elemzés eredményének illusztrálására egy olyan táblázatot is elkészít, melyben a standardizált átlagok alacsony (A) és magas (M) szintjét betűk és + jelek mutatják (nagyban megkönnyítve így az értelmezést):

STANDARDIZÁLT ÁTLAGOK MINTÁZATA (M = Magas, A = Alacsony)					
Klaszter	gyak.	Homog.	Val_sp	Val_ko	Val_has
1	31	0,79	M++	.	M
2	62	0,58	.	.	.
3	37	0,48	A+	M	A

7. táblázat. Klaszterek gyors áttekintő táblázata 3 klaszterre, ROPstatban

A fenti táblázat tartalmaz egy új mutatót is: a homogenitási együttható a klaszterek belső távolságainak átlaga, így megfigyelhető, hogy az egyedek a 3. klaszterben „hasonlítanak" a leginkább, míg legkevésbé az 1. klaszter tűnik homogénnek.

Elmondható tehát, hogy a klaszterbe sorolás és a távolságok között pozitív együttjárást találtunk (pontbiszeriális együttható 0,439), valamint azt is, hogy a Silhouette-mutató nagynak tekinthető, a magyarázott varianciaarány 55,3%-os, tehát a klasztereink elkülönülőnek látszanak.

ROPstat-eredmények alapján az 5 klaszteres modell ismertetése

Miután a 3 klaszteres modell esetén nem értük el az elvárható 60-65%-os EESS%-szintet, ezért olyan elemzést is lefuttattunk (most a változóinkat némileg eltérő varianciájú skáláik miatt standardizálva), ahol már elértük az elvárt szintet – tehát lényegesen magasabb magyarázóerejű modellhez[12] jutottunk.

Az 5 klaszteres hierarchikus modell esetén az EESS% értéke 63,41%, mely a relokáció segítségével 66,17%-ra növelhető. A kapott modell Xie–Beni-indexe

[12] A ROPstat esetén reziduálanalízis segítségével kiugró értékeket, illetve extrém értékeket tudunk azonosítani, amelyeket aztán (például feltételes csoportosító változó használatával) ideiglenesen eltávolíthatunk az elemzésből, még tisztább modellre igényt tartva. Továbbá például sűrűsödéselemzéssel is tovább finomíthatjuk a klaszterezési eljárásainkat, de ezek terjedelmi, illetve átláthatósági szempontból lényegesen nehezítenék a mostani értelmezéseinket, ezért eltekintünk alkalmazásuktól.

0,352, pontbiszeriális együtthatója 0,370, míg a Silhouette-mutató értéke 0,619 lett. Látható tehát, hogy a pontbiszeriális együtthatóban és a Silhouette-mutatóban nem tudtunk elérni jelentős emelkedést, azonban a legfontosabb mutatóban, a magyarázott varianciahányadnak is értelmezhető EESS%-ban jelentős, 16%-os emelkedést tapasztaltunk.

Az így nyert klaszterekben a változóátlagok alacsony és magas értékeit összefoglaló táblázat az alábbi:

Klaszter	gyak.	Homog.	Valószínűség	Kockázat	Elvárt haszon
1	25	0,92	(M)	A+	M+
2	24	0,96	M+	.	(M)
3	41	0,57	.	.	(A)
4	18	0,48	A+	M+	A+
5	22	0,57	A	M	.

8. táblázat. Az 5 klaszteres felbontás tulajdonságai

A fenti táblázatban tehát látható, hogy a leghomogénebb csoport ismét a korábbi elemzésben is megjelenő negyedik klaszter volt (kis valószínűséggel, kis haszonnal, de nagy kockázattal járó események választása). Ezenkívül ennek ellenkezője is jelentkezik, a nagy valószínűséggel és nagy haszonnal járó, de kis kockázatokat tartalmazó első klaszter (korábban is megjelenő csoport).

Megállapítható tehát, hogy az elemzés alapvetően a közepes értékekkel bíró csoportot igyekezett további szétválasztással szeparálni. Ebben található egy, az első klaszterhez hasonlító második csoport, amely szintén nagyobb gyakorisággal és nagy haszon reményében vállal kockázatot, de az első csoportra jellemzőnél már nagyobbat is elviselve.

A harmadik klaszter továbbra is a közepes kockázatvállalási hajlandóság csoportjaként azonosítható, míg az eddig nem említett ötödik klaszter tagjai kis valószínűséggel (ritkán), de jelentősebb kockázattal járó, közepes haszonnal kecsegtető eseményeket keresnek.

A ROPstat és az SPSS© eredményeinek összevetése, értelmezése

A ROPstat esetén a Xie–Beni mellett legalább 3 objektív mutató áll rendelkezésünkre annak eldöntésére, hogy a klaszterezésünk elfogadható-e. Mindhárom mennyiség abszolút mutatónak tekinthető abban az értelemben, hogy egy-egy klaszterstruktúráról a 3 mutatót egyszerre figyelve tudunk döntést hozni (nincsen szükség egyéb információra). Ez köszönhető annak, hogy a mutatóink abszolút mutatók (tehát fix tartományban mozognak, a korreláció és a Silhouette-mutató esetén –1 és 1 között, míg az EESS-nél 0 és 100% között).

A ROPstatban a Xie–Beni és a 3 objektív mutató együttes megfigyelése és értékelése biztosítja a klaszterezési eljárásunk jóságát. Azaz: ha alacsony Xie–Beni-mutató mellett kellően magas a pontbiszeriális együttható, és megfelelő Silhouette-mutatóval és elegendően magas magyarázott varianciával rendelkezünk, akkor elfogadhatónak tekintjük a klaszterezési eljárásunk eredményét[13] – míg ha valamely mutató gondot jelez, akkor kénytelenek vagyunk kételyeket megfogalmazni.

Az SPSS© ezzel szemben nem ad nekünk támpontokat. Sőt, az SPSS nem is következetes a távolságokat illetően, hiszen a klaszterek középpontjainak távolságát euklideszi, míg az egyedek klaszterközéppontoktól való távolságát négyzetes euklideszi távolságként adja meg! Tehát ha magunk szeretnénk mutatókat számítani, akkor még előtte azt is meg kell vizsgálnunk, hogy az SPSS© milyen adatokkal szolgálhat.

A Xie–Beni nem abszolút mutató, azaz nincsen felső korlátja (a távolságok miatt nyilván nem lehet negatív). Ez egyben rögtön nehézséget is jelent a tekintetben, hogy egy Xie–Beni-mutató „mennyire" kicsi. Más megközelítésben viszont: ha ugyanazon adatállományon alkalmazva több k-központú elemzést is végrehajtunk, akkor közülük azt érdemes választani, amelynek a Xie–Beni-mutatója a legkisebb.

ÖSSZEGZÉS

Dolgozatunkban bemutattunk két lehetséges módot abból a célból, hogy klaszterezési eljárásunk jóságát mérni tudjuk, illetve megmutattuk ennek egy lehetséges megvalósítását az SPSS© programcsomagban, hiszen e programcsomag nem szolgáltat számunkra összetett mutatórendszert az eljárás jóságának mérésére. Felhívjuk a figyelmet, hogy ezzel szemben például a ROPstat programcsomag olyan mutatókat is kiszámít, melyek segítségével döntéseket tudunk hozni – így ebben az esetben akár kényelmesebb megoldást is nyújthat, mint az SPSS©![14]

A két módszer alkalmazása után a lényeget az alábbiakban foglalhatjuk össze: Ha nincsenek a klaszterezés jóságát mérő mutatóink, akkor több abszolút (vagy akár relatív) mutató *együttes alkalmazását javasoljuk*. A mutatókat együtt kell értelmeznünk, és ha lehetőségünk van rá, mindenképpen több mutató segítségével hozzunk döntéseket. E tekintetben ha választani lehet a két eljárás

[13] Ezért végeztünk el egy olyan elemzést is, melyben 5 klaszter kialakításakor már elértük, illetve meghaladtuk az elvárható 60-65%-os EESS%-ot. Erről az SPSS alkalmazásával nem lett volna információnk!

[14] Miután igazolni tudtuk, hogy az SPSS a különböző távolságok esetén egyáltalán nem következetes, így még az is nehézségekbe ütközhet, hogy egy saját magunk számára kidolgozott fejlesztésben a megfelelő adatokkal dolgozhassunk.

között, akkor a ROPstat megoldása és alkalmazása kényelmesebbnek mutatkozik, hiszen egyetlen eljárásban, 1-2 futtatással több mutatót is nyerünk egyszerre.

Ismertetett példánkban bemutattuk, hogy az abszolút mutatók alkalmazásával a saját adatainkon könnyen tudtunk döntést hozni. Azon indexek esetén, amelyeknek nem voltak szigorú korlátozó tartományai, nehezen tudjuk eldönteni, hogy mennyire jó egy adott klaszterstruktúra. Mindkét eljárás során azonos következtetésre jutottunk (bár az előbbinél ez részben a szerencse dolga is lehetett) – így azt is elmondhatjuk, hogy klaszterezési eljárásunkat teljesen más irányokból megközelítve azonos strukturális eredményekre jutottunk, tehát a mintázatunk stabilnak mondható.

Kiemelnénk azonban azt is, hogy pusztán az SPSS© alkalmazásával nem láttuk volna, hogy a magyarázott varianciahányad alacsonyabb – és egyáltalán nem garantálja semmi, hogy a jobb magyarázóerejű, de hasonló egyéb mutatókkal rendelkező klaszterezésig eljutunk!

Fontos azt is megjegyeznünk, hogy például az R© programcsomag használatakor számos mutatót tudunk egyszerre kiszámítani klaszterezéskor – azonban az R© programcsomag használatához némi programozási jártasság szükséges, valamint a különböző mutatók matematikai leírásának értelmezése, amelynek során a matematikai statisztikában kevésbé járatos olvasók számára komoly gondot jelenthet a számos index közül kiválasztani a megfelelőt.

BIBLIOGRÁFIA

Asendorpf, J. B. (2002a). Editorial: The puzzle of personality types. European Journal of Personality, 16, S1-S5.

Bagdy E., Kövi, Zs., Mirnics, Zs. (2014). A tehetség kibontakozása. Budapest: Helikon Kiadó.

Bagdy E., Mirnics, Zs., Kövi, Zs. (2014). Fény és árnyék. A tehetségerők felszabadítása. Budapest: Matehetsz.

Baker, F. B., Hubert, L. J. (1975). Measuring the power of hierarchical cluster analysis, Journal of the American Statistical Association, 70, 31-38.

Ball, G. H., Hall, D. J. (1965). A Novel method of data analysis and pattern classification, Menlo Park: Stanford research Institute (NTIS No. AD 6996116).

Bergman, L. R., Magnusson, D., & El-Khouri, B. M. (2003). Studying individual development in an interindividual context: A person-oriented approach. Vol. 4 in the series Paths through life (D. Magnusson, Ed.). Mahwah, NJ: Erlbaum.

Borsboom, D., Mellenbergh, G. J., van Heerden, J. (2003). The theoretical status of latent variables, Psychological Review, 110(2), 203-19.

Devi, G. (2014). A survey on distributed data mining and its trend, International Journal of Research in Engineeering & Technology, 2(3), 107-120.

Dunn, J. (1974). Well separeted cluster and optimal fuzzy partitions, Journal of Cybernetics, 4, 95-104.

Forman, G., Zhang, B. (2000). Distributed Data Clustering can be efficient and exact, SIGKDD Explorations, 2(2), 34-38.

Hubert, L, Schultz, J. (1976). Quadratic assigment as a general data-analysis strategy, British Journal of Mathematical and Statistical Psychologie, 29, 190-241.

John, B., Mirnics Zs., Bagdy Gy., Gonda X., Benkő A., Molnár E., Lázáry E., Surányi Zs. (2013). Klaszteralapú Big Five-személyiségprototípusok kapcsolata patológiás mutatókkal. Psychologia Hungarica Caroliensis, 1, 1.

MacQueen, J. B. (1967). Some Methods for Classification and Analysis of Multivariate Observations. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, 281-297. Letöltve: University of California, Los Angeles. University of Maryland Institute for Advanced Computer Studies honlapja. http://www.umiacs.umd.edu/~raghuram/ENEE731/Spectral/kMeans.pdf

Magnusson, D., & Allen, V. (Eds.). (1983). Human development: An interactional perspective. New York, NY: Academic Press.

Marton, K., Surányi, Zs., Farkas, L., Egri, T. (2014). Everyday functions and needs of individuals with disability: A reliability and validity study based on the principles of the ICF. Psychiatria Hungarica, 4.

McClain, J. O., Rao, V.R. (1975). Clustisz: A program to test for the quality of clustering of a set of objects, Journal of Marketing Research, 12, 456-460.

Móri T., Szekély G. (1986). Többváltozós Statisztikai Analízis, Budapest: Műszaki Kiadó.

Nock, R., Nielsen, F. (2006). On Weighted Clustering, IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(8), 1-13.

Prinzie, A., Van den Poel, D. (2006a). Investigating Purchasing Patterns for Financial Services using Markov, MTD and MTDg Models, European Journal of Operation Research, 170(3), 710-734.

Prinzie, A., Van den Poel, D. (2006b). Incorporating sequential information into traditional classification models by using an element/position-sensitive SAM, Decision Support System, 42(2), 508-526.

Prinzie, A., Van den Poel, D. (2007). Predictin home-appliance acquisition sequences: Markov/MTD/MTDg and survival analysis for modeling sequential inforation in NPTB models, Decision Support System, 44(1), 28-45.

Radnóti I. (2008). A kockázatvállalási szándék mérése, ÁVF Tudományos Konferencia 2007. 11. 13., Tudományos Közlemények, Általános Vállalkozási Főiskola, (19)103-114.

Rousseeuw, P. J. (1987). Silhouettes: a grapchical aid to the interpretation and validation of cluster analysis, Journal of Computational and Applied Matheatics, 20, 53-65.

Rohlf, F. J. (1974). Methods of comparing classifications, Annual Review of Ecology and Systematics, 5, 101-113.

Surányi Zs., Babocsay Á., Takács Sz., Vargha A. (2011). Új klasszifikációs módszerek a személyiségpszichológiában, Pszichológia, 31., 317-340.

Vargha A. (2007). Matematikai statisztika pszichológiai, nyelvészeti és biológiai alkalmazásokkal (2. kiadás). Budapest: Pólya Kiadó.

Vargha A. (2008). Új statisztikai módszerekkel új lehetőségek: a ROPstat a pszichológiai kutatások szolgálatában, Pszichológia, 28(1), 81-103.

Vargha, A., Torma, B., Bergman, L. R. (2014). ROPstat: a general statistical package useful for conducting person-oriented analyses. Journal for Person-Oriented Research, 1, (megjelenés előtt).

Wilson, J. N., Ritter, G, X. (2000). Handbook of Computer Vision Algorithms in Image Algebra, (2nd ed.), CRC Press.

Xie, X. L., Beni, G. (1991). A validity measure for fuzzy clustering, IEEE Transactions on Pattern Analysis and Machine Intelligence, 13(4), 841-846.

MELLÉKLETEK

M1: SPSS Syntax a relokációs klaszterezéshez és segítség a Xie–Beni-index meghatározásához

```
*** K-központú elemzés alkalmazása ***

QUICK CLUSTER
 val _ sp val _ ko val _ has
 /MISSING=LISTWISE
 /CRITERIA= CLUSTER(3) MXITER(10) CONVERGE(0)
 /METHOD=KMEANS(NOUPDATE)
 /SAVE CLUSTER DISTANCE
 /PRINT INITIAL ANOVA CLUSTER DISTAN.

*** Távolságok négyzetösszege ***

DESCRIPTIVES
 VARIABLES=QCL _ 2
 /STATISTICS=SUM .
```

M2: SPSS Syntax a 3 dimenziós ábrázoláshoz

```
**** Grafikon, a 3 dimenziós ábrázoláshoz ***
* Chart Builder.
GGRAPH
  /GRAPHDATASET NAME="graphdataset" VARIABLES=val _ ko val _ sp
val _ has QCL _ 1
 MISSING=LISTWISE REPORTMISSING=NO
 /GRAPHSPEC SOURCE=INLINE.
BEGIN GPL
 SOURCE: s=userSource(id(„graphdataset"))
```

```
 DATA: val _ ko=col(source(s), name(„val _ ko”))
 DATA: val _ sp=col(source(s), name(„val _ sp”))
 DATA: val _ has=col(source(s), name(„val _ has”))
 DATA: QCL _ 1=col(source(s), name(„QCL _ 1”), unit.category())
 COORD: rect(dim(1,2,3))
 GUIDE: axis(dim(1), label(„Sportolással való kockázatvállalás,
 elvárt „,
 „haszon”))
 GUIDE: axis(dim(2), label(„Sportolásban való kockázatvállalás,
 kockázat „,
 „mértéke”))
 GUIDE: axis(dim(3), label(„Sportolásban való kockázatvállalás, „,
 „valószínűség”))
 GUIDE: legend(aesthetic(aesthetic.color.exterior), label(„Cluster
 Number „,
 „of Case”))
 SCALE: cat(aesthetic(aesthetic.color.exterior))
  ELEMENT: point(position(val _ has*val _ ko*val _ sp), color.
 exterior(QCL _ 1))
END GPL.
```