KLASZTERELEMZÉSEK SZEMÉLY-ORIENTÁLT PSZICHOLÓGIAI KUTATÁSOKBAN A ROP-R SZOFTVER SEGÍTSÉGÉVEL

Vargha András
KRE BTK Pszichológiai Intézete
ELTE PPK Pszichológiai Intézete
vargha.andras@kre.hu

Összefoglaló

Háttér és célkitűzések: A klaszteranalízis a személy-orientált pszichológiai kutatások kedvelt módszere. Míg a változó-orientált kutatások megrekednek olyan mutatók (pl. átlag, korreláció stb.) vizsgálatának szintjén, ami inkább a változókat, mintsem az egyéneket jellemzi, a személy-orientált megközelítés az egyénnel kapcsolatos folyamatokra fókuszál és azt hangsúlyozza, hogy a személyt jellemző adatokat, változóértékeket feldarabolatlan egységként kell tekinteni és kezelni. A személy-orientált többváltozós statisztika olyan eljárásokra fókuszál, amelyek esetében központi szerepet játszanak az egyének közti kvalitatív jellegű különbségek. Ezek hátterében típusmodellek állnak, amelyek jellemzően klasszifikációs módszerekkel tárhatók fel. Cikkünkben áttekintjük a klaszteranalízis alapfogalmait, majd valódi pszichológiai kutatásból származó adatokon mutatjuk be, hogy hogyan lehet a ROP-R ingyenes többváltozós statisztikai szoftver segítségével hierarchikus, *k*-középpontú nem hierarchikus és modell-alapú klaszterelemzéseket végrehajtani.

Kulcsszavak: személyközpontú többváltozós statisztika, klaszteranalízis, ROP-R.

DOI: 10.17627/ALKPSZICH.2023.4.115

Mi a klaszteranalízis?

Az empirikus pszichológiai vizsgálatok célja, hogy feltárják a lelki élet törvényszerűségeit, a pszichológiai jellemzők közti összefüggéseket, okokat-okozatokat, csoportok közti különbségeket. A statisztika ebben úgy segít, hogy megmondja: hogyan kell a vizsgált pszichológiai populációkból megfelelő mintákat kiválasztani, érvényes módon mérni, adatokat feldolgozni, elemezni és alkalmas statisztikai módszerek segítségével helyes következtetéseket levonni. Bár ezekkel az elemzésekkel többnyire az emberekről szeretnénk többet megtudni, gyakran megrekedünk olyan mutatók (pl. átlag, korreláció stb.) vizsgálatának a szintjén, ami inkább a változókat, mintsem az egyéneket jellemzi.

Ha vizsgálatunk fókuszában a változók közti kapcsolatok, összefüggésük mikéntje, rendszere áll, változó-orientált elemzésekről beszélünk, de ugyanez a helyzet akkor is, ha különböző csoportokat az átlag vagy más középérték szerint hasonlítunk össze (Vargha, 2019, 2020). Ezekben az elemzésekben az a közös, hogy statisztikai modelljeikben egy-egy konkrét személynek sehol se találunk helyet. A személy-orientált megközelítés szerint viszont a személyt jellemző adatokat, változóértékeket feldarabolatlan egységként kell tekinteni és kezelni. (Bergman & Lundh, 2015).

A személy-orientált (személyközpontú) többváltozós statisztika olyan eljárásokra fókuszál, amelyek esetében központi szerepet játszanak az egyének közti kvalitatív jellegű különbségek. Ezek hátterében olyan típusmodellek állnak, amelyek jellemzően klasszifikációs módszerekkel tárhatók fel. A személyiség tipológiai megközelítése nem új, mint például Hippokratész-Galénosz vérmérsékleti tipológiája, amely az első, részletesen kidolgozott, két és félezer éves, ma is jól ismert típustan (vö. Bartha, 1980).

A legismertebb típusfeltáró eljárás a klaszteranalízis (klaszterelemzés). Ez matematikai módszerekkel keresi egy többváltozós minta olyan alcsoportjait, ún. klasztereit, amelyek egymástól határozottan különböznek, de amelyeken belül a vizsgált személyek erősen hasonlítanak egymásra. A típusfeltáráson kívül a klaszteranalízis használható az adatminta takarékos leírására is (adatredukció), melynek során egy nagyobb minta személyeit szakmailag értelmezhető és további elemzésekre alkalmas kategóriákba soroljuk.

A klaszteranalízisben alapvető fontosságú a hasonlóság és a különbözőség fogalma mind a személyek, mind a klaszterek tekintetében. A klaszterstruktúrák megfelelőségének mérésére a statisztika irodalmában számos adekvációs mutató ismeretes.

A KLASZTERANALÍZIS ALAPFOGALMAI

A klaszteranalízis alkalmazási feltételei

Megbízható többváltozós elemzéshez legalább 300-400 személyre van szükség és jó, ha a személyek száma legalább tízszerese a változók számának (vö. Vargha, 2019, p. 80). Speciálisan a klaszteranalízis esetében jó, ha a mintaelemszám az 1000-et is meghaladja, hogy a ritkább, 2-5%-os arányú típusokat képviselő klaszterek is feltárhatók legyenek. A lényeg, hogy adott mintanagyság esetén minden klaszter által képviselt típus kellő számú esettel legyen képviselve.

A mintában ne legyenek outlierek, mivel a kilógó személyek rontják a feltárt klaszterstruktúra homogenitását. Hozzá kell

tenni, hogy nem lehet csupán matematikai eszközök segítségével eldönteni valakiről, hogy outlier-e. Az elemzésből csak azokat szabad kihagyni, akik értékmintázata szakmai szempontok alapján is elfogadhatatlan, műtermék jellegű.

Az input változók száma ne legyen túl nagy, ideális, ha 2-7 közötti. A változószám növekedésével – eléggé általános feltételek mellett – csökken a legtávolabbi és a legközelebbi elempár távolságának várható relatív különbsége, vagyis egyre nehezebb egymástól jól elkülönülő klasztereket találni (vö. Moisl, 2015, pp. 71–77).

Törekedjünk tehát arra, hogy a változók száma minél kisebb legyen, de azért fedjenek le egy releváns területet, amelyen belül a típusokat keressük. Kerüljük a redundanciát! Általában közepes szorosságú kapcsolatban lévő változókat érdemes az elemzésbe bevonni. Legyenek legalább intervallum-skálájúak. A változók pszichometriai reliabilitása legyen magas. Minél nagyobb ugyanis a változók mérési hibája, annál nehezebb egy populációban létező klaszterstruktúrát akár nagy minták segítségével is azonosítani (vö. Vargha & Bergman, 2019). Ha a változók különböző mértékegységűek, szükséges a változókat azonos skálára hozni, például *z*-standardizálással. A *z*-standardizálás során minden változónál szórás léptékű skálára térünk át.

A változók szóródása ne legyen nagyon kicsi, mert a nagyon kis variabilitású változók értékei szorosan, egyetlen centrum körül tömörülnek, így az nem lehet valódi klaszterképző komponens. Az ilyen változókat hagyjuk ki az elemzésből. A változók együttes eloszlása ne legyen normális! Számos többváltozós eljárás (pl. a faktoranalízis) alkalmazási feltétele az elemzett változók többdimenziós normalitása. A klaszteranalízisben ennek éppen a fordítottja igaz. Ha a klaszteranalízis változói többváltozós normális eloszlást követnek, a klaszteranalízis bizonyosan nem vezethet nekünk tetsző megoldásra (vö. Vargha, 2022, p. 62). Persze ha sikerül statisztikailag igazolni a többdimenziós normalitás sérülését (pl. polinomiális regresszióval; vö. Vargha, 2019, 2. fejezet), ez csak esélyt ad egy nem triviális klaszterstruktúrára, de nem garantálja azt.

Klaszteranalízis típusa

A személyeken végezhető legismertebb hagyományos klaszterelemzési módszer a hierarchikus klaszteranalízis (HKA). Ez valójában nem egyetlen klaszteranalízis, hanem olyan többlépéses klasszifikáció sorozat, amelynek minden lépésében vagy összevonunk két klasztert egy közös nagyobb klaszterbe (összevonó, agglomeratív HKA, röviden AHKA), vagy szétbontunk, felosztunk egy klasztert két kisebb klaszterre (osztódó, felosztó HKA, röviden OHKA). Ezek közül az AHKA az elterjedtebb, melynek első lépésében az egymáshoz legközelebbi két személyt vonjuk össze közös klaszterbe (1-elemű klaszternek tekintve őket), majd minden további lépésben az egymáshoz legközelebbi két klasztert. Az összevonást addig folytatjuk, amíg vannak különböző klaszterek, így a végén minden személy egyetlen nagy közös klaszterbe kerül. OHKA esetében a folyamat fordított, itt az első lépésben egyetlen nagy klaszterben van mindenki, az utolsóban pedig mindenki egyedül egy 1-eleműben. Itt a szétbontás alapelve, hogy mindig a legheterogénebb klasztert bontjuk két alklaszterre úgy, hogy azok a lehető legjobban különbözzenek egymástól (lásd alább, illetve Vargha, 2022, 5. fejezet).

A nem hierarchikus klaszteranalízisek közül a legismertebb a *k*-középpontú klaszteranalízis (KKA). Ennek lényege, hogy előre rögzítünk egy *k* klaszterszámot, definiálunk egy random vagy más elemzésből kapott induló klaszterstruktúrát, majd egy többlépéses iterációs folyamat során addig javítjuk a klaszterstruktúrát a személyek egyik klaszterből a másikba átrakásával, amíg el nem érünk egy lehetséges maximumot a klaszterek összhomogenitása tekintetében (vö. Vargha, 2022, 6. fejezet).

Kevésbé ismert, de ma már egyre gyakrabban használt módszer a modell-alapú klaszteranalízis (MKA), mely a többváltozós normális eloszlások keverékeként tekint a minta által képviselt populációra, és ezen komponenseloszlások rendszerével teremt háttér-modellt a feltárandó klaszterstruktúrára (vö. Vargha, 2022, 7. fejezet).

Személyek távolsága

Klaszteranalízissel egy többváltozós minta homogén alcsoportjait keressük, vagyis olyan klasztereket, amelyeken belül a személyek hasonlítanak egymásra. A személyek klaszteranalízisének eredményét alapvetően befolyásolja a személyek közti távolság megválasztása, definiálása. Klaszteranalízisekben a személyek hasonlóságának mérésére számos távolságmérték ismeretes (lásd pl. Füstös et al., 2004, pp. 169–173, illetve Takács et al., 2015). A leggyakrabban használt ilyen mutatókat az *1. táblázat*ban foglaltuk össze.

1. táblázat. A klaszteranalízisekben a személyek hasonlóságának mérésére használt leggyakoribb távolságmértékek

Távolság neve	**Távolság definíciója**
Euklideszi (ED)	Négyzetes eltérések összegének négyzetgyöke
Négyzetes euklideszi (SED)	Négyzetes eltérések összege
Átlagos négyzetes euklideszi (ASED)	Négyzetes eltérések átlaga
Minkowski, rögzített *p* értékkel	A változónkénti eltérések *p*-edik hatványait összegezzük, majd az összegből *p*-edik gyököt vonunk (az ED távolság általánosítása tetszőleges *p* hatványra)
Mannhattan (city-block)	Abszolút eltérések összege (a Minkowski távolság $p = 1$ értékkel)
Canberra	A Manhattan távolság egy súlyozott változata
Csebisev (Maximum)	Változónkénti abszolút eltérések maximuma
Adatsorok Pearson-féle *r* korrelációja	A lehetséges maximális pozitív együttjárástól való elmaradás: $d = 1 - r$

Az ED[1] euklideszi távolság a szokásos síkbeli, illetve térbeli távolság a vizsgált változók ortogonális (egymással páronként derékszöget bezáró) terében. Ha a változóértékek különbségei közül büntetni akarjuk a nagyobbakat (így csökkentve a nagyobb különbségek hatását), akkor nem végezzük el a gyökvonást az euklideszi távolság képzé-

[1] Euclidean Distance rövidítése

sekor, amivel a SED[2] négyzetes euklideszi távolságot kapjuk. Ha ezt még leosztjuk a változók számával, az ASED távolsághoz jutunk. SED és ASED között csak egy konstans osztóban van különbség, így ezek minden klasszifikációs elemzésnél ugyanarra az eredményre vezetnek, de az ASED távolság jobban értelmezhető, mert jelzi az egy változóra eső átlagos négyzetes eltérést.

Ha a személyek hasonlóságának megítélésekor értékmintázataik hasonlósága a döntő, akkor a hasonlóság mérésére jó választás lehet két személy adatsorának a Pearson-korrelációja, ehhez persze legalább három változóra van szükség. Ha viszont a hasonlóságot a változónkénti értékek közeliségével mérjük, akkor jobb az euklideszi távolság valamilyen változata. Személy-orientált kutatásokban a legpreferáltabb személytávolság SED, illetve ASED.

Klaszterek távolsága

A klaszteranalízisben nem csak az fontos, hogy a feltárt klaszterek homogének legyenek, hanem az is, hogy jól elkülönüljenek egymástól. Ehhez pedig az szükséges, hogy a klaszterek kellő távolságra legyenek egymástól. Két klaszter távolsága is olyan fogalom, amelyet matematikailag többféleképpen lehet definiálni, s amely ugyancsak döntő hatással van a feltárt klaszterstruktúrára, különösen AHKA esetén. A legismertebb klasztertávolságokat a *2. táblázat*ban foglaltuk össze.

2. táblázat. Két klaszter távolságának legismertebb mértékei

Magyar elnevezés	Angol elnevezés	Alkalmazott klasztertávolság
1. Minimális távolság (legközelebbi szomszéd)	Nearest neighbor	A két klaszter egymáshoz legközelebbi elemének távolsága
2. Maximális távolság (legtávolabbi szomszéd)	Furthest neighbor	A két klaszter egymástól legtávolabbi elemének távolsága
3. Átlagos távolság	Between-groups distance	A két klaszter elemei közti távolságok átlaga
4. Centroid-távolság	Centroid distance	A két klaszter centroidjának (többdimenziós átlagának) a távolsága

A klaszterstruktúrák jóságának mérése

Minden klaszteranalízis használ valamilyen algoritmust arra, hogy optimalizáljon egy kritériumot a feltárt klaszterstruktúra megfelelőségével kapcsolatban. A klaszterstruktúra jóságának megítélése fontos annak eldöntéséhez is, hogy a feltárt struktúra jobb-e, mint egy másik, amelyet például más távolságmértékkel vagy más módszerrel kapunk, vagy amelynél más a változók vagy a klaszterek száma. Az ilyen jellegű kérdések megválaszolására találták ki a különböző klaszter adekvációs mutatókat (angolul: clustering quality coefficients), amelyeket a továbbiakban QC-ként rövidítünk. A QC-k tehát olyan mutatók, amelyek segítségével egy klasztermodell jósága megítélhető. A QC-ket részletesen ismerteti Vargha (2022, 4.4. alfejezet), ezért most csak a legfontosabbakról szólunk röviden.

[2] Squared Euclidean Distance rövidítése

Egy jó klasztermodelltől elvárjuk, hogy a klaszterek homogének legyenek, amit jól mér például a klaszterbeli személyek egymástól való átlagos ASED távolsága (vö. *1. táblázat*), a HC[3] klaszter homogenitási együttható. Minél kisebb egy klaszter HC-értéke, annál homogénebb a klaszter. Ha a klaszterelemzést standardizált változókkal végezzük, 0,5-nél kisebb HC értékek jellemeznek egy igazán homogén klasztert[4]. Ha a változók ugyanolyan skálán mérnek, de átlaguk és varianciájuk jelentősen eltér egymástól, érdemes a változókat még a klaszterelemzés végrehajtása előtt standardizálni. A teljes klaszterstruktúra homogenitását mérhetjük a HC-értékek átlagával (HCátlag; vö. Vargha et al., 2015; Vargha, 2022, p. 69). Jó struktúra esetén a változók standardizálását feltételezve – saját tapasztalataim alapján – 0,50-nél kisebb HC értékekre, de legalábbis 1-nél érezhetően kisebb HCátlagra számítunk.

Jó klaszterhomogenitási mutató EESS% is, a klaszterek által megmagyarázott varianciaarány, mely a varianciaanalízisből ismert eta-négyzet (η^2) többváltozós általánosítása (vö. Bergman et al., 2003, pp. 113-115; Vargha et al., 2015):

EESS% = 100*(SStotal – SSklaszter)/SStotal.

Itt az SSklaszter, az ún. klasztereken belüli összhiba (röviden: összhiba), a klaszteren belüli – klaszterecentroidtól való – négyzetes eltérések összegének az összege, míg SStotal a teljes mintában a mintacentroidtól való négyzetes eltérések összege (a négyzetes eltéréseket minden esetben változónként kiszámítva, majd a változókra összegezve).

Minél jobban megközelíti egy klaszterstruktúra EESS% értéke a 100-at, annál kisebb az összhiba, és annál jobban magyarázzák a struktúra klaszterei az input változók variabilitását, vagyis az egyének közti eltéréseket. Más szavakkal EESS% azt méri, hogy az egyének közti különbségekért milyen mértékben felelős az, hogy az egyének melyik klaszterbe tartoznak. Egy elfogadható klaszterstruktúra esetén EESS%-nak illik elérnie a 65%-ot, jó struktúra esetén pedig a 70-75%-ot. Az a jó, ha EESS% nagy, de a klaszterek száma viszonylag kevés. Ezt a két ellentétes szempontot kell a klaszterelemzések során összhangba hozni.

Egy klaszterstruktúra jóságának megítélésekor fontos arra is figyelni, hogy a klaszterek mennyire különböznek, mennyire szeparálhatók egymástól. A QC-k közül a Rousseeuw nevéhez fűződő Silhouette-együttható (SC[5]) közkedvelt szeparációs mutató, mely azt méri, hogy a mintabeli személyek mennyivel vannak közelebb saját klasztercentrumukhoz, mint a legközelebbi idegen centrumhoz. Eredeti formulája kicsit bonyolult (vö. Rousseeuw, 1987), de van egy egyszerűbb és hasonlóan értelmezhető változata is (vö. Vargha et al., 2016; Vargha, 2022, p. 73). Egy magas SC érték arra utal, hogy az adott klaszterstruktúrában a személyek általában közelebb vannak saját klaszterük centrumához, mint a legközelebbi idegen klaszter centrumához. Az SC értékek -1 és 1 közé esnek. Negatív értékek esetén a személyek általában távolabb vannak saját klaszterük centrumától, mint a legközelebbi idegen klasztercentrumtól, ami nem szerencsés. Jó struktúra esetén

[3] Homogeneity Coefficient

[4] Ha nem standardizálunk, érdemes HC-t leosztani a változók teljes mintabeli varianciájának átlagával és ezen „standardizált" HC mutató (HCstan) alapján ítéletet mondani a klaszter homogenitásáról.

[5] Silhouette Coefficient

elvárt, hogy SC nagyobb legyen 0,5-nél, a 0,2-nél kisebb SC értékek pedig gyenge klaszterstruktúrára utalnak (vö. Kaufman & Rousseeuw, 1990). SC általában könnyebben ér el kívánatosan magas értéket kevés számú klaszternél, ahol könnyebben különülnek el egymástól a klaszterek.

Szintén hasznos szeparációs mutató az XBmod, a módosított Xie-Beni index (Vargha, 2022, p. 74), mely azt méri, hogy az egymáshoz legközelebbi két klasztercentrum távolsága relatíve mennyivel nagyobb, mint a saját klasztercentrumtól való átlagos távolság. Az XBmod értékei szintén -1 és 1 közé esnek, kiértékelése és értelmezése az SC Silhouette-együtthatóéhoz hasonlóan végezhető el.

A ROP-R-ben rendelkezésre álló HCátlag, EESS% és az XBmod QC mutatók tájékoztatnak egy klaszterstruktúra esetében a Vargha et al. (2016) által megemlített két legfontosabb jellemzőről, a klaszterek homogenitásáról (kohéziójáról) és szeparálhatóságáról, így alkalmasak annak megítélésére, hogy egy ROP-R által feltárt klaszterstruktúra mennyire megfelelő. Természetesen léteznek még más fontos QC mutatók is (pl. a ROPstat szoftverrel kiszámítható klaszter pontbiszeriális együttható, a CLdelta és a GDI24 mutató, vö. Vargha, 2022, 4.4. alfejezet), de ROP-R kifejlesztésekor nem volt cél, hogy a szoftver mindent tudjon a témában, amit más szoftverek is. Az is komoly érték, hogy a ROP-R klaszterező moduljai számos egyedi lehetőséggel, új opcióval (pl. a modell-alapú klaszterelemzéssel) jól kiegészítik a ROPstatban rendelkezésre álló lehetőségeket.

A ROP-R-rel végrehajtható klaszterelemzések szemléltetése egy kötődéskutatás adatain

A ROP-R szoftver

A ROP-R egy R szoftver (R Core Team, 2021) alapú, de ROPstat (vö. Vargha, 2016) keretben használható többváltozós statisztikai programcsomag, letöltését és használatát illetően lásd Vargha & Bánsági (2022), illetve Vargha et al. (2024). A ROP-R moduljai a statisztikai elemzés típusa szerint három csoportba sorolhatók: regressziós elemzések, dimenzió redukciók (főkomponens- és faktorelemzések), illetve klaszteranalízisek, amelyek ROP-R-ben a „Többváltozós_elemzések_R_segítségével" menüponttal futtathatók. A klaszteranalízisek végrehajtására négy modul áll rendelkezésre, amelyeket az alábbi fejezetekben majd részletesen ismertetünk.

Az elemzett minta jellemzői

A klaszteranalízisek szemléltetésére Jantek & Vargha (2016) kötődéskutatásából merítünk adatokat, ahol 336 felnőtt magyar személy (124 férfi és 212 nő) állt rendelkezésre, akik mind heteroszexuális kapcsolatban éltek. A kutatás fő célja az ECR–RS (Experiences in Close Relationships – Relationship Structures) kérdőív (Fraley et al., 2011) magyar populációra való adaptálása volt. Az ECR–RS a kötődést két dimenzió (elkerülés és szorongás) mentén méri négy kötődési személy (anya, apa, romantikus partner és barát) viszonylatában. A 40 tételes kérdőív minden viszonylatban ugyanazt a 10 tételt alkalmazza, amelyek közül az első 6 tétel az elkerülést, az utóbbi 4 pedig a szorongást méri.

A jelen cikk klaszterelemzéseiben az anyával kapcsolatos elkerülés (AnyaElk) és szorongás (AnyaSzor) skáláját használjuk majd input változóként. Ezeket az érintett tételek átlagaiként definiáljuk. Mivel a tételek 7-fokú Likert skálájúak (1 és 7 közötti értékekkel), a két skála értékei is ebbe a tartományba esnek. Megjegyezzük, hogy a szorongás skála definiálásakor Fraley et al. (2011), illetve Jantek & Vargha (2016) ajánlására a tíz anyai tételből az utolsót nem vettük figyelembe.

Fraley et al. (2011) modelljében biztonságos, jó kötődéssel rendelkeznek azok, akik az elkerülés és a szorongás tekintetében egyaránt alacsony szinten vannak, míg a félelemteli, elkerülő kötődésűek mindkét dimenzión magas értékűek. A magas elkerülés – alacsony szorongás kombináció az elutasító–elkerülő, a magas szorongás – alacsony elkerülés kombináció pedig az elárasztott–megszállott típusra jellemző (vö. Jantek & Vargha, 2016, 1. ábra).

Mielőtt belevágnánk a klaszterelemzésekbe, érdemes kiszűrni az outliereket és normalitásvizsgálatot végezni. Az outliereket a ROP-R főkomponens-analízis moduljával, az AnyaElk és AnyaSzor skálán végzett elemzés során azonosítottuk, az „Extrém esetek azonosítása" opció bejelölésével. Ez egyetlen outlier esetre hívta fel a figyelmet[6], amit az adatállományból töröltünk, így minden további elemzést a maradék 335 fős adatmintának azon a 322 fős részén hajtottunk végre, ahol mindkét skála érvényes értékkel rendelkezett.

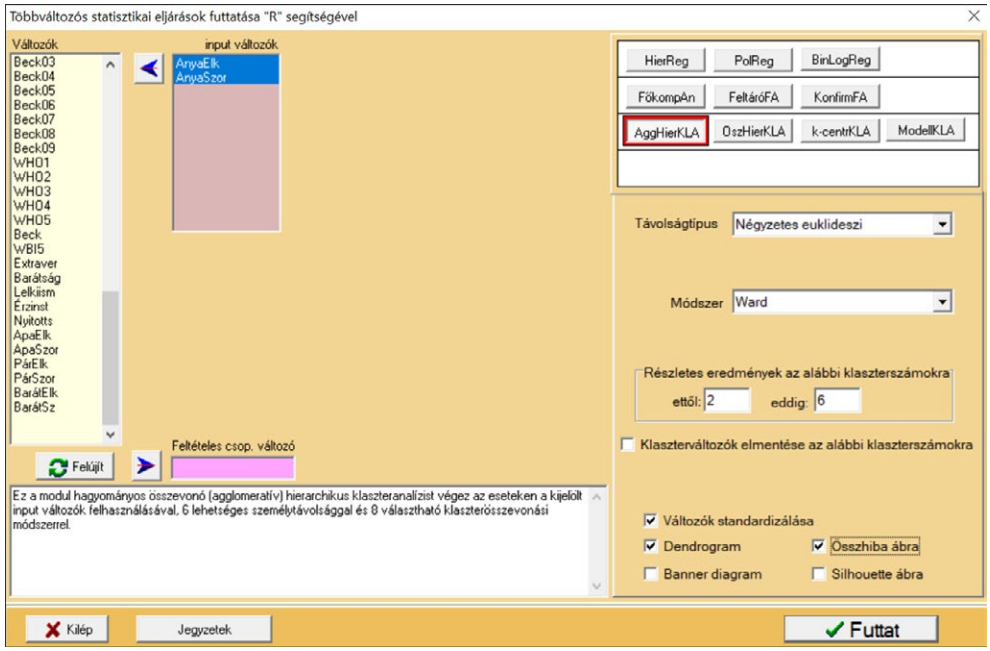
A normalitás vizsgálatára kiszámítottuk a ferdeségi és a csúcsossági együtthatót[7], s ez az AnyaSzor skála esetében jelezte a normalitás súlyos sérülését (ferdeség = 2,45, csúcsosság = 5,69, mindkét esetben $p < 0{,}001$). A többdimenziós normalitás sérülését a ROP-R polinomiális regresszióelemzése jelezte, függő változó = AnyaSzor, független változó = AnyaElk szerepkiosztással. Ez esetben ugyanis a független változónak mind a 3. hatványa ($p = 0{,}020$), mind az 5. hatványa ($p = 0{,}025$) szignifikánsan megemelte a megmagyarázott varianciát, a nemlineáris hatások pedig ellentmondanak a többdimenziós normalitásnak. Az egy- és többdimenziós normalitás ily módon detektált sérülése esélyt adhat egy nem triviális klaszterstruktúra feltárására.

Összevonó hierarchikus klaszteranalízis (AHKA)

Ilyen elemzést ROP-R-ben az összevonó (agglomeratív) hierarchikus klaszteranalízis (röviden AgglHierKLA vagy AHKA) modul segítségével végezhetünk el. Ez a modul AHKA-t végez a mintabeli eseteken a kijelölt változók felhasználásával. AHKA a *stats* (R Core Team, 2021), a *cluster* (Maechler et al., 2022), a *ggplot2* (Wickham, 2016) és a *factoextra* (Kassambara & Mundt, 2020) R-package-et használja fel elemzéseihez. Az AHKA menüablak (lásd *1. ábra*) szolgál az elemzésbe bevonandó változók kiválasztására (input változók ablaka), a személytávolság (Távolságtípus) és az AHKA összevonási módszerének (Módszer) megválasztására, valamint egy sor opció kijelölésére, amelyekhez az alábbi eligazítást fűzzük.

[6] a GMD általánosított Mahalanobis távolság alapján, mellyel minden személy esetén megnézzük, hogy milyen messze esik a teljes minta többdimenziós centrumától (lásd Aggarwal, 2017, illetve https://www.cfholbert.com/blog/outlier_mahalanobis_distance/).

[7] a ROPstat szoftver segítségével



1. ábra. Az AHKA modul menüablaka ROP-R-ben

Input változók kiválasztása

Az AHKA-hoz szükséges változók kiválasztásához a Változók feliratú ablakból kell az elemzéshez szükséges változókat az „input változók" ablakba átküldeni a két ablak közti nyílra rákattintva.

Személytávolság megválasztása

Az AHKA modulban hat személytávolság áll rendelkezésre, az *1. táblázat*ban felsoroltak közül az utolsó sor (Pearson korreláció) kivételével[8] mindegyik. ASED csak azért hiányzik, mert a vele ekvivalens SED jelenléte feleslegessé teszi.

Klaszterösszevonás módszerének megválasztása

Emlékeztetőül az AHKA minden lépésében az egymáshoz legközelebbi két klasztert vonjuk össze közös klaszterbe. A ROP-R AHKA moduljában nyolc klaszterösszevonási módszer áll rendelkezésre. Ezek közül az első négy a *2. táblázat*ban összefoglalt klasztertávolságon alapul, a maradék négy pedig az alábbi.

1. Medián-módszer: ez a centroid módszer olyan variánsa, amely szimmetrikus eloszlású változók esetén hasonló, erősen ferde eloszlások esetén pedig gyakran jobb struktúrához vezet.
2. Ward-féle módszer: ekkor minden lépésben azt a két klasztert egyesítjük, amelyekre vonatkozóan a klaszterstruk-

[8] A Pearson metrika nem elérhető ROP-R-ben, kettőnél több változó esetén sem.

túra összhomogenitását mérő EESS% mutató a legkisebb mértékben csökken.

3. Flexibilis béta: egy viszonylag bonyolult, de néha igen sikeres klasszifikációt eredményező összevonási módszer, melynél egy b (beta) paraméter értékének 0,1 és 1 közötti lehetséges megválasztásával különböző klasztertávolságok állíthatók be (a részletekről lásd Vargha, 2022, p. 86). Például b = 0,2 a Ward-féle módszerhez, egy 1-hez közeli b pedig a minimális távolság módszeréhez hasonló megoldásra vezet.
4. McQuitty[9]: az átlagos távolság módszerének egy olyan variánsa, amelynél a létrehozott új és a régi klaszterek távolságának meghatározásában egyenlővé tesszük az éppen összevont két klaszter esetleg eltérő elemszámának hatását (Vargha, 2022, p. 85).

Megjegyezzük, hogy a centroid-, a medián- és a Ward-féle módszer választása esetén ROP-R mindig a SED személytávolságot használja, bármit is állítunk be.

Pszichológiai kutatásokban a fenti nyolc módszer mindegyike vezethet érdekes eredményre, de személy-orientált vizsgálatokban, típusok feltárására a Ward-módszert tartják a legjobbnak (Bergman et al., 2003).

Segítő ábrák

Az AHKA eredményének értelmezését négy választható ábra segíti az AHKA modulban, amelyeket a menüablak jobb alsó paneljén lehet kijelölni:

- Dendrogram: az AHKA lépésenkénti összevonásait szemléltető fadiagram;
- Banner-ábra: a dendrogram jégcsap-ábrához hasonlító változata;
- Összhiba ábra: a klasztereken belüli összhiba, vagyis SSklaszter lejtődiagramja 1 és 10 klaszterszám között az AHKA személytávolság és összevonási módszer beállításával;
- Silhouette-ábra: a Silhouette-együttható értékének ábrázolása 1 és 10 klaszterszám között az AHKA személytávolság és összevonási módszer beállításával.

Egyéb lehetőségek

Az AHKA menüablakában megadható a klaszterszámok egy övezete, amelyen belül minden megoldásra megkapjuk a HCátlag, EESS%, XBmod QC mutatók értékét, a klaszterek alapstatisztikáit (elemszám, átlag, szórás, minimum, maximum), valamint a standardizált átlagok mintázatát. Külön kérésre a klaszterszámok egy megadott övezetére a klaszterkódot személyenként megadó klaszterváltozók elmenthetők (az adott msw adatfájlhoz illeszthetők) és a menüablakban állítható be az is, hogy standardizáljuk-e az input változókat (alapértelmezés szerint igen).

Egy AHKA elemzés végrehajtása után a c:_vargha\ropstat\aktualis mappában megtaláljuk az elemzéshez elkészített ideiglenes adatfájlt (tmpdat.txt), a kért klaszterszámokhoz tartozó klaszterváltozókkal kiegészített ideiglenes adatfájlt (tmpdat2.txt), a futtatott R-scriptet (AHCA.r), valamint a kért diagramokat jpg vagy pdf fájlban (pl. Dendr1.jpg vagy Banner1.pdf). Ha feltételes csoportosító változót is kijelölünk, akkor minden feltételes csoport elemzése során elkészülnek a kért diagramok, a diag-

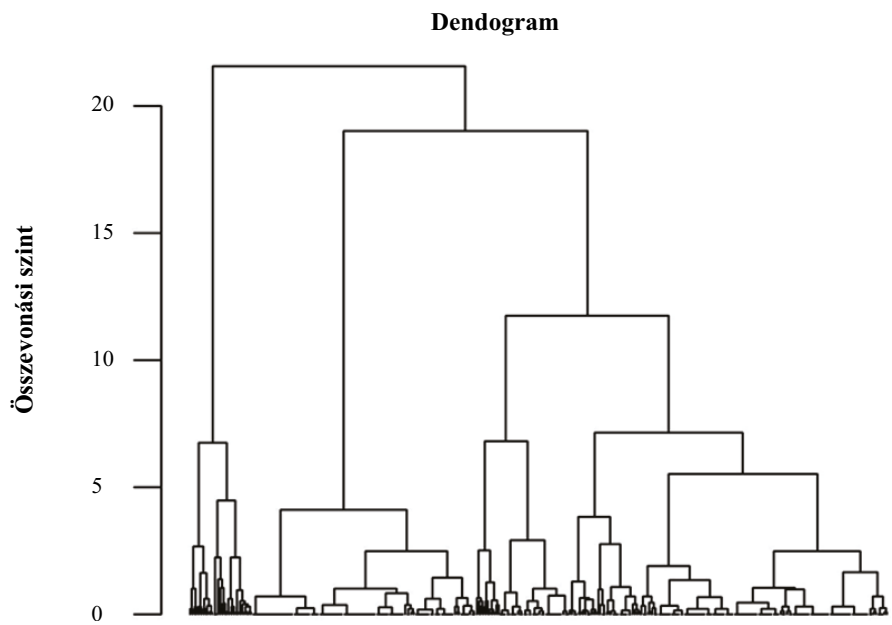
[9] Szokták erre a WPGMA (Weighted Pair Group Method with Arithmetic Mean) elnevezést is használni.

ramokat tartalmazó fájlok nevében megjelenő számok ezen csoportok sorszámát jelzik.

Az anyai kötődéssel kapcsolatban elvégzett AHKA elemzés

Az AnyaElk és az AnyaSzor input változókkal az AHKA-t az alapértelmezés szerinti Ward-módszerrel hajtottuk végre, a változók standardizálásával és a 2-6 klaszterszámokra kérve részletes eredményeket (vö. *1. ábra*). Az ábrák közül dendrogramot és összhiba ábrát kértünk.

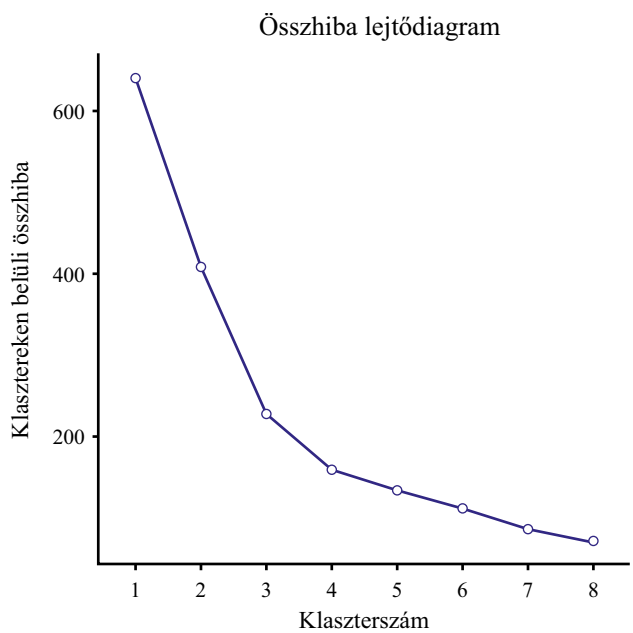
Elsőként a Dendr1.jpg fájlban elhelyezett dendrogramot érdemes szemügyre venni (lásd *2. ábra*). Az ábra vízszintes tengelyén az elemzett minta személyei foglalnak helyet. Egy-egy függőleges vonal alatti hierarchikus struktúra egy-egy klasztert képvisel, melynek magassága az „Összevonási szint" elnevezésű függőleges tengelyen azt méri, hogy a klaszter létrejöttekor annak két alkotó komponense (amelyek összevonásából a klaszter létrejött) milyen távol volt egymástól. Így minél magasabb szinten vonódik össze két klaszter, annál heterogénebb klaszter jön létre. Például a bal oldali első ilyen függőleges vonal alatti klaszter, melynek rövid vízszintes szakasszal jelölt teteje kb. a 7-es összevonási szint magasságában van, az átlagosnál homogénebbnek tűnik. Ugyanezen függőleges vonal felső pontjának magassága azt jelzi, hogy a klaszter milyen messze van attól a klasztertől, amellyel egy későbbi lépésben összevonásra kerül. Ez az említett klaszter esetén érezhetően 20 fölé nyúlik, s ezzel az utolsó lépésben történő összevonással kerül egyetlen nagy klaszterbe a minta összes eleme.



2. ábra. Az AHKA eredményének ábrázolása dendrogram segítségével a ROP-R AHKA moduljában (Ward-módszer, standardizált változók)

Ha a dendrogramot a vízszintes tengellyel párhuzamosan valamilyen szinten egy egyenessel elmetsszük, egy olyan klaszstermegoldást jelölünk ki, amelynél annyi klaszter van, ahány függőleges vonalat átmetszünk. Például egy 20-as szinten meghúzott vonal a 2-, a 15-ös szinten meghúzott a 3-, a 10-es szinten meghúzott pedig a 4-klaszteres megoldást jelöli ki az AHKA klasszifikációs sorozatából. A dendrogram alapján oly módon dönthetünk az optimális klaszterszámról, hogy megnézzük, milyen szinten metszhetjük el vízszintes vonallal a dendrogramot úgy, hogy a metszési szint viszonylag alacsony legyen és a metszéspontok száma se legyen túl nagy. Jelen esetben egy 8-as szint körüli metszés 4-klaszteres megoldása elfogadhatónak tűnik.

Szintén segíthet az optimális klaszterszám meghatározásában a WSSplot1.jpg fájlban elhelyezett összhiba lejtődiagram (lásd *3. ábra*). Ha egy klaszter homogén, akkor az egyedek közel esnek a klaszter centrumához, a centroidhoz. A centroidtól való távolságok összege ezért a klaszter heterogenitását méri. Összegezve ezeket az összes klaszterre, a klasztereken belüli összhibát kapjuk, mely a teljes klaszterstruktúra heterogenitásának egy mérőszáma. Ez olvasható le a *3. ábrá*ról a $k = [1–8]$ klaszterszám tartományra. Az ábra alapján optimálisnak olyan klaszterszám tűnik, amely után az összhiba már kisebb léptékben csökken, mint előtte. A *3. ábrá*n egyaránt nagy esés figyelhető meg $k = 1$, $k = 2$ és $k = 3$ után, de $k = 4$ után a görbe már kissé ellaposodik és innen egyenletes csökkenésbe kezd. Emiatt ez az ábra is a 4-klaszteres megoldást valószínűsíti.



3. ábra. A klasztereken belüli összhiba lejtődiagramja a ROP-R AHKA moduljában (Ward-módszer, standardizált változók)

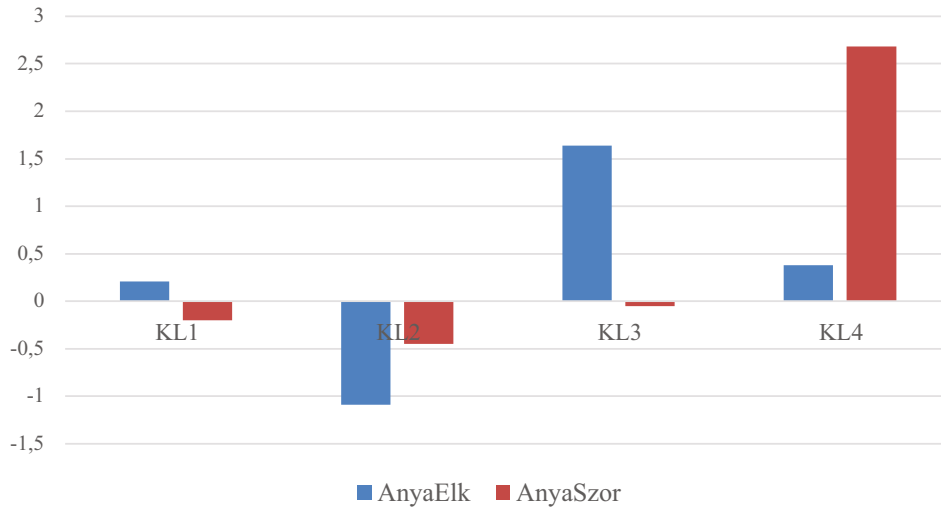
Minthogy a 2-6 klaszterszámokra kértünk részletes eredményeket, ezek alapján áttekinthetjük, hogyan alakulnak e klaszter-megoldásokra a HCátlag, EESS%, XBmod mutatók értékei (lásd *3. táblázat*). A *3. táblázat* adatai szerint a klaszterstruktúra homogenitása az összevonások során 6-tól lefele 4 klaszterig elfogadható (HCátlag nem nagyon nő a 0,50-es szint fölé, EESS% pedig a 75%-os szint alá), $k = 4$ alatt azonban egyértelmű a romlás. A szeparációt mérő XBmod mindenhol a jó struktúrát jelző 0,50-es szint felett van, így a megtekintett ábrákkal összhangban ezek az adatok is a 4-klaszteres megoldást támogatják.

3. táblázat. Különböző klasztermegoldások jellemzői AHKA-ban (Ward-módszer, standardizált változók)

Klaszterszám	HCátlag	EESS%	XBmod
6	0,354	82,74	0,657
5	0,426	79,13	0,585
4	0,505	75,13	0,715
3	0,719	64,38	0,737
2	1,281	36,23	0,856

A több forrás által megerősített 4-klaszteres megoldás szakmai értelmezését az eredménylista $k = 4$ megoldásához tartozó standardizált átlagok táblázata segítségével végezhetjük el, melyből a *4. ábrán* látható diagramot készítettük (Excelben). Az ábrán KL1, KL2, KL3 és KL4 a feltárt négy klasztert jelölik. Tekintve, hogy standardizált átlagokat ábrázolunk, a 0 szint fölé emelkedő oszlopok az átlagosnál nagyobb, míg a 0 alatti oszlopok az átlagosnál kisebb szinteket jeleznek. Ennek alapján KL1 (n = 147, HC = 0,44[10]) egy minden tekintetben átlagos kötődésű típust képvisel. KL2 (n = 103, HC = 0,14) egy átlagosnál jobb, különösen az anyai elkerülés alacsony szintje tekintetében kiemelkedő típus klasztere. KL3 (n = 43, HC = 0,80) ugyanakkor éppen az anyai elkerülés magas szintje tekintetében különbözik a többi klasztertől. Végül a KL4 (n = 29, HC = 1,66) klaszter tűnik a legrosszabb kötődésű csoportnak, az átlagosnál enyhén magasabb anyai elkerüléssel és extrém mértékben magas anyai szorongással. A klaszterek nagyságát és HC-vel mért homogenitását is figyelembe véve azt mondhatjuk, hogy a minta nagy része (78%-a) átlagos vagy átlagosnál jobb kötődésű homogén típusba sorolható, 13%-a erős anyai elkerüléssel jellemezhető átlagosnál rosszabb anyai kötődésű, 9%-a pedig extrém magas anyai szorongással jellemezhető, rossz anyai kötődésű személy.

[10] A klaszterek nagysága és HC-értéke szintén az eredménylista $k = 4$ megoldásához tartozó részén van feltüntetve a klaszterek alapstatisztikái mellett.



4. ábra. A 4-klaszteres AHKA-megoldás standardizált átlagainak oszlopdiagramja (Ward-módszer, standardizált változók)

Összefoglalóul azt mondhatjuk, hogy a kapott eredmény érdekes, de nem egyezik meg Fraley et al. (2011) modelljével, mely ugyan szintén négy típust fogalmaz meg, de azok némileg mások, mint a jelen elemzésben kapottak. A nem problémás kötődésű személyek között markáns típusként jelenik meg a minden tekintetben átlagos kötődésűek típusa, mely teljesen hiányzik a Fraley-féle modellből. Ennek relevanciáját jelzi itt a típust képviselő KL1 klaszter mérete (a teljes minta 46%-a) és homogenitása (HC = 0,44). A másik három klaszter megfeleltethető a Fraley-féle modell egy-egy kissé módosított típusának: KL2 a jó kötődés típusát képviseli, bár a szorongás szintjében lehetne kicsit alacsonyabb; KL3 a magas elkerülés – alacsony szorongás típusát képviseli, de alacsony szorongás nélkül; végül KL4 a magas szorongás – alacsony elkerülés típusát képviseli, de alacsony elkerülés nélkül. Ami a Fraley-féle modellből teljesen hiányzik itt: az az egyaránt magas szorongással és elkerüléssel jellemezhető félelemteli, elkerülő kötődésű személyek típusa.

Érdemes ezért más típusú klaszterelemzéseket is elvégezni, hátha azok – jobban illeszkedve a Fraley-féle modellre – szakmailag megfelelőbb eredményre vezetnek.

Osztódó hierarchikus klaszteranalízis (OHKA)

Ilyen elemzést ROP-R-ben például az osztódó hierarchikus klaszteranalízis (röviden OsztlHierKLA vagy OHKA) modul segítségével is elvégezhetünk. Ez a modul OHKA-t végez a mintabeli eseteken a kijelölt változók felhasználásával. OHKA ugyanazokat az R-package-eket használja fel elemzéseihez, mint az AHKA. Az OHKA menüablak csak annyiban különbözik AHKA-étól,

hogy nincs benne választható módszer, ez ugyanis rögzített, a Kaufman & Rousseeuw (1990, 6. fejezet) nevéhez fűződő DIANA[11] osztódó hierarchikus klaszter-módszer, melynek főbb lépései az alábbiak.

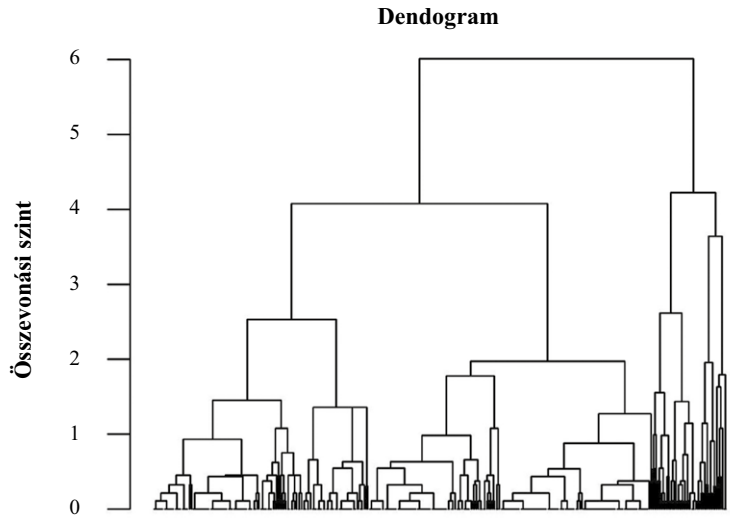
1. Kezdetben a minta összes egyede egy klaszterben van. Ezután minden lépésben megkeressük azt a klasztert, amelyben az egymástól legtávolabbi két egyed a legmesszebb van egymástól, és ezt a következőképpen bontjuk két alklaszterre.
2. Kiválasztjuk azt az egyedet, amelynek a többi egyedtől való átlagos távolsága a legnagyobb, s ezt egy különálló egyelemű klaszternek tekintjük. Ezután addig csatolunk egyenként ehhez az új klaszterhez a régi klaszterből újabb egyedeket, amíg azok közelebb esnek az új klaszterhez, mint a maradék egyedek által alkotott régi klaszterhez.
3. Ezt addig folytatjuk, amíg végül minden egyed egy-egy 1-elemű klaszterbe kerül.

Egy OHKA elemzés végrehajtása után a c:_vargha\ropstat\aktualis mappában megtaláljuk az elemzéshez elkészített ideiglenes adatfájlt (tmpdat.txt), a kért klaszterszámokhoz tartozó, klaszterváltozókkal kiegészített ideiglenes adatfájlt (tmpdat2.txt), a futtatott R-scriptet (DHCA.r), valamint a kért diagramokat jpg vagy pdf fájlban (pl. Dendr_1.jpg vagy WSSplot_1.pdf). Ha feltételes csoportosító változót is kijelölünk, akkor minden feltételes csoport elemzése során elkészülnek a kért diagramok, a diagramokat tartalmazó fájlok nevében megjelenő számok ezen csoportok sorszámát jelzik. Az optimális klaszterszám meghatározásához kérhető összhiba- és Silhouette-ábra 1 és 12 közötti klaszterszámra ábrázol, de nem az OHKA, hanem az ED személytávolságú és centroid módszerű AHKA elemzés szerinti értékeket adja meg[12].

Az OHKA elemzés szemléltetésére az AnyaElk és az AnyaSzor input változóval az elemzést ismét SED személytávolsággal és a változók standardizálásával végeztük el, ugyancsak a 2-6 klaszterszámokra kérve részletes eredményeket. Az ábrák közül most csak dendrogramot kértünk (lásd *5. ábra*).

[11] DIvisive ANAlysis

[12] Ennek oka, hogy az ábrát készítő *fviz_nbclust* R-függvényben az OHKA opció nem állítható be.



5. ábra. Az OHKA eredményének ábrázolása dendrogram segítségével a ROP-R OHKA moduljában (standardizált változók, SED személytávolság)

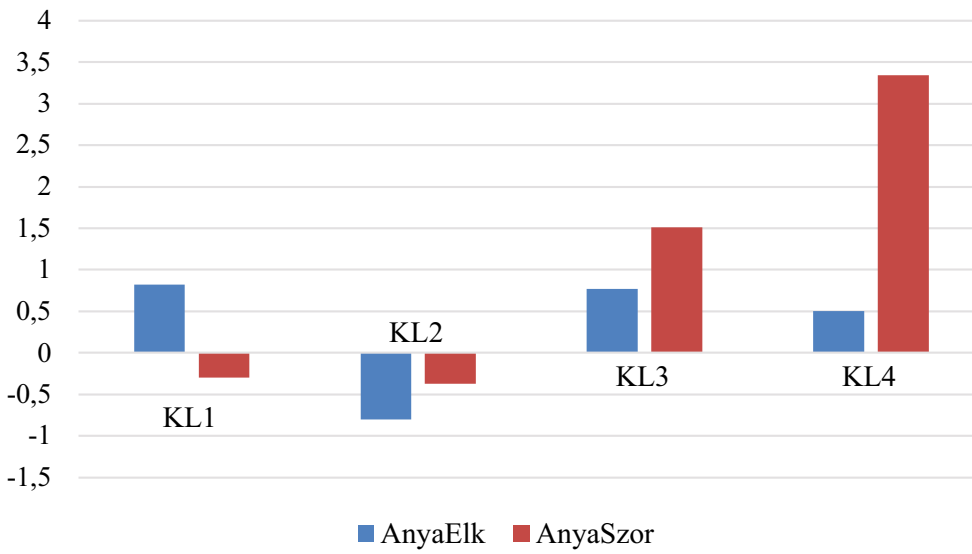
Az *5. ábrá*n látható dendrogram kissé más képet mutat, mint AHKA esetében (lásd *2. ábra*). Itt is oly módon dönthetünk egy optimális klaszterszámról, hogy megnézzük, milyen szinten metszhetjük el vízszintes vonallal a dendrogramot úgy, hogy a metszési szint viszonylag alacsony legyen és a metszéspontok száma se legyen túl nagy. Jelen esetben egy elfogadhatóan alacsony (2-es) szinten történő metszés már 7 klaszteres megoldást mutat.

A 2-6 klaszterszámokra kigyűjtött HCátlag, EESS%, XBmod mutatók értékét a *4. táblázat*ban foglaltuk össze. Ennek alapján a klaszterstruktúra homogenitása az összevonások során ismét $k = 3$ fölött kezd hirtelen változással elfogadható lenni ($k = 4$-nél HCátlag a $k = 3$ esetén látható 1,072-ről 0,508-ra esik le, EESS% értéke pedig a 46,76-ról 74,99-re ugrik). A szeparációt mérő XBmod minden k értéknél kiváló. Tehát az OHKA is arra utal, hogy egy optimális megoldáshoz legalább 4 klaszterre van szükség.

4. táblázat. Különböző klasztermegoldások jellemzői OHKA-ban (SED távolság, standardizált változók)

Klaszterszám	HCátlag	EESS%	XBmod
2	1,179	41,33	0,839
3	1,072	46,76	0,692
4	0,508	74,99	0,811
5	0,465	77,16	0,827
6	0,430	78,90	0,774

A 4-klaszteres megoldás standardizált átlagainak ábráját elkészítve (lásd *6. ábra*) és összevetve azt a 4-klaszteres AHKA megoldás hasonló ábrájával (lásd *4. ábra*) azt látjuk, hogy a KL2 és a KL4 klaszter jó egyezést mutat az AHKA megoldás azonos nevű két klaszterével, KL1 is hasonló mintázatú, mint AHKA-ban KL3, de az OHKA KL3 klasztere, mely Fraley et al. (2011) modelljében a félelemteli elkerülő kötődés típusának felel meg, a 4-klaszteres AHKA megoldásban nem jelent meg. Vajon milyen struktúrát tár majd fel a *k*-középpontú és a modell-alapú klaszteranalízis?



6. ábra. A 4-klaszteres OHKA-megoldás standardizált átlagainak oszlopdiagramja

K-középpontú nemhierarchikus klaszteranalízis (KKA)

A *k*-középpontú vagy *k-centrumú* klaszteranalízis (röviden KKA) ugyanúgy számos klaszteranalízis gyűjtőneve, ahogy HKA. A *k*-középpont jelző arra utal, hogy a KKA-ban mindig egy előre megadott *k* számú klasztert hozunk létre, másrészt arra, hogy a klaszterek középpontjai (centrumai) kiemelt fontosságú többdimenziós pontok, a klaszterbeli személyek reprezentáns értékmintázatai. Centrumként szóba jöhet a többdimenziós átlag (centroid), a geometriai medián, valamint a medoid is (lásd alább). Minden KKA elemzés lényege, hogy először készítünk egy kezdeti felosztást *k* centrummal, majd jön egy többlépéses iterációs folyamat, ahol a relokáció segítségével addig javítjuk a klaszterstruktúrát a személyek egyik klaszterből a másikba átrakásával, amíg el nem érünk egy lehetséges maximumot a klaszterek összhomogenitása tekintetében. KKA-ban a személytávolság rögzített, a *k*-közép módszer esetén egységesen a SED (illetve a vele ekvivalens ASED), a többi esetben pedig ED.

A KKA ROP-R-beli modulja (*k*-centrumú klaszteranalízis, röviden k-centrKLA vagy KKA) többnyire ugyanazokat az R-package-eket (*stats*, *cluster*, *factoextra*, *ggplot2*) használja fel elemzéseihez, mint az AHKA és az OHKA, de ha *k*-medián elemzést futtatunk, akkor még a Gmedian package-et is (Cardot, 2022). Ezzel kapcsolatban megjegyezzük, hogy a *k*-medián elemzés eredménye kismértékű random ingadozást mutat, ezért érdemes többször futtatni, majd az eredmények közül a legjobbat kiválasztani.

Egy KKA elemzés végrehajtása után a c:_vargha\ropstat\aktualis mappában megtaláljuk az elemzéshez elkészített ideiglenes adatfájlt (tmpdat.txt), a kért klaszterszámhoz tartozó, klaszterváltozókkal kiegészített ideiglenes adatfájlt (tmpdat2.txt), a futtatott R-scriptet (KCA.r), valamint a kért diagramot (diagramokat) jpg fájlban.

A KKA modul lehetséges beállításait a menüablak *7. ábr*án látható része mutatja be. Ezek közül csak a „Módszer" és az „Optimális klaszterszámhoz ábrakészítés" panel használata igényel részletesebb magyarázatot.

7. ábra. A *k*-középpontú klaszteranalízis moduljának lehetséges beállításai ROP-R-ben

A KKA módszere

A KKA elemzések elsőként abban különböznek egymástól, hogy mi a választott klaszterecentrum típusa. Attól függően, hogy ez a centrum a centroid, a geometriai medián, illetve a medoid, beszélünk *k*-közép, *k*-medián vagy *k*-medoid módszerről. Egy *p*-dimenziós minta geometriai mediánjának azt a *p*-dimenziós pontot nevezzük, amelyiknek az átlagos euklideszi távolsága a mintabeli elemektől a legkisebb. A medoid pedig a többdimenziós mintának az az eleme, amelyik a legközelebb van a mintabeli többi egyedhez. A medoiddal ellentétben a centroid és a geometriai medián olyan többdi-

menziós pont, amelyik nem biztos, hogy fellép a minta valamelyik egyedénél. A *k*-medián és a *k*-medoid módszert erősen ferde eloszlású, szélsőséges adatokat is tartalmazó minták esetén szokták az alapértelmezett *k*-közép módszer helyett javasolni.

A k-közép módszer

Minden *k*-közép elemzés minden iterációs lépésében minden személyt megpróbálunk úgy átrakni egy másik klaszterbe, hogy a klaszterstruktúra javuljon. Ha az átrakással nem lenne javulás, nem rakjuk át. Az induló klasztercentrumokat úgy határozzuk meg, hogy véletlenszerűen kiválasztunk a mintából *k* számú személyt, s ezeket tesszük meg induló centrumoknak. Alkalmazott algoritmusa alapján három *k*-közép típust szoktak megkülönböztetni: Hartigan-Wong, MacQueen és Lloyd[13]. Ezek részletes ismertetését lásd Vargha (2022, pp. 136–137). Röviden összefoglalva a Hartigan-Wong algoritmus a klasztereken belüli összhibát (vö. *3. ábra*) próbálja minimalizálni (és egyben az EESS% megmagyarázott varianciaarányt maximalizálni), a másik kettő pedig a relokációk iterációs lépéseiben akkor tesz át egy személyt egy másik klaszterbe, ha annak centrumához közelebb van, mint a sajátjához. A Lloyd algoritmus csak annyiban különbözik a MacQueen-félétől, hogy itt a klasztercentroidok újraszámolása nem történik meg minden áthelyezésnél, csak egy-egy teljes iterációs lépéssorozat végén.

Az iterációk száma 1 és 200 között szabadon megválasztható (alapértelmezés: 25) minden *k*-közép elemzésnél. Az elemzés akkor áll le, ha a program elvégezte az optimalizációt a megadott iterációs lépésszámra, vagy ha két egymást követő iterációs lépésben nem javul a klaszterstruktúra. A ROP-R minden *k*-közép elemzéshez 10-szer választ egy véletlen induló klaszterközép együttest, mindegyikre végrehajtja az elemzést és a kapott 10 eredmény közül a legjobbat, vagyis a legkisebb összhibájú (a legnagyobb EESS% értékű) struktúrát tekinti végleges megoldásnak.

A k-medián módszer

A *k*-medián módszer algoritmusa ugyanaz, mint a MacQueen-féle *k*-közép módszeré, annyi különbséggel, hogy klasztercentrumként nem a centroidot, hanem a geometriai mediánt, személytávolságként pedig nem SED-et, hanem a sima ED euklideszi távolságot használjuk.

A k-medoid módszer

A *k*-medoid módszer algoritmusa ugyanaz, mint a Hartigan-Wong-féle *k*-közép módszeré. A különbség mindössze annyi, hogy a centroidok szerepét a medoidok veszik át, és nem a klasztereken belüli összhibát, hanem a saját klasztermedoidtól való személytávolságok összegét próbáljuk minimalizálni. Az alkalmazott személytávolság ROP-R-ben itt is a sima euklideszi.

Ábrák az optimális klaszterszám meghatározásához

A KKA modulban kérhető a végső klaszter-megoldás ábrája, valamint egy ábra az optimális klaszterszám meghatározásához. Ez utóbbi a menüablak jobb alsó paneljén jelölhető ki (lásd *7. ábra*), ahol egy elemzéshez egy ábra választható a következő négy közül: Silhouette, EESS%, átlagos heterogenitás vagy f(K) torzulás ábrája, a maximális klaszterszám 5 és 20 közötti beállításának lehetőségével. ROP-R a választott ábrát a KKA futtatása után a már többször említett „aktua-

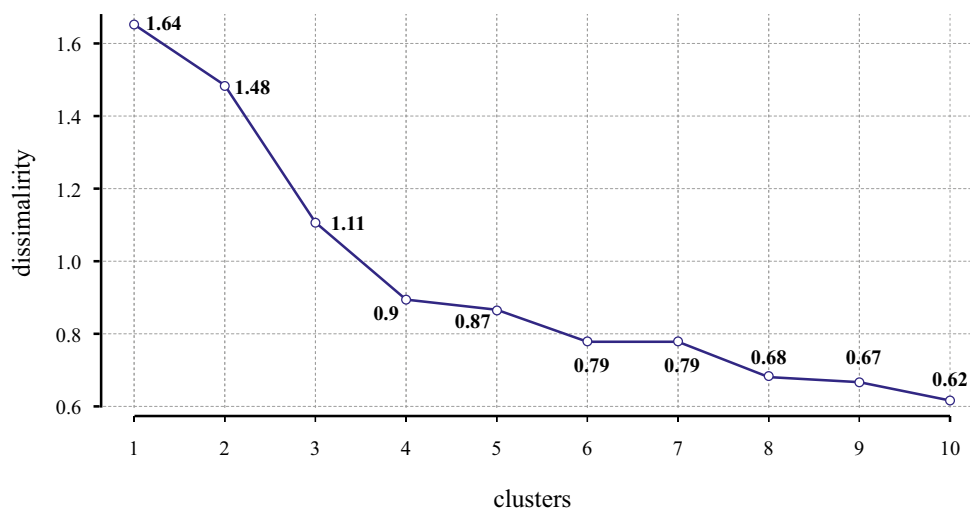
[13] más néven Forgy

lis” almappában, az ábra sorszámának megfelelően optk1_1.jpg, optk2_1.jpg, optk3_1.jpg vagy optk4_1.jpg nevű képfájlba menti.

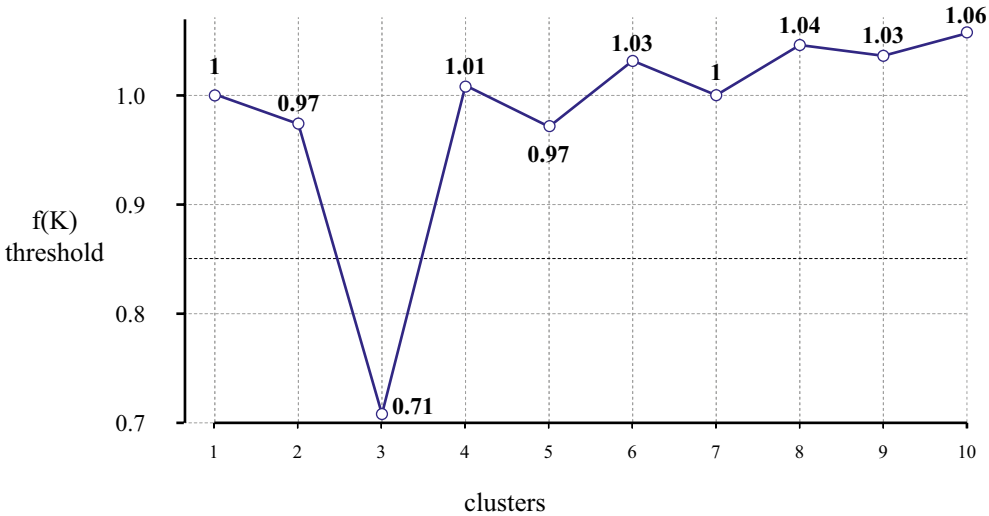
A Silhouette- és az EESS%-ábra értelmezése nem igényel magyarázatot. Az átlagos heterogenitás a klasztereken belüli összes páronkénti személytávolság átlaga euklideszi távolsággal, egyfajta HCátlag mutató, az f(K) mutató pedig egy random egyenletes eloszláson végzett KKA eredményéhez viszonyítja a kapott struktúrát (vö. Pham et al., 2005). Olyan – lehetőleg alacsony – *k* érték tűnik optimálisnak, amelyre a Silhouette-együttható, illetve EESS% értéke nagy, az átlagos heterogenitás és f(K) értéke pedig kicsi, utóbbi lehetőleg 0,85-nél is kisebb. Megjegyezzük, hogy a Silhouette-együttható gyakran torzít a kis *k* értékek felé, pusztán azon okból, hogy kevés klaszter jobban el tud különülni egymástól, mint sok. Ezek az ábrák mind *k*-közép elemzésekkel készülnek, és nem függnek attól, hogy a menüablakban milyen KKA módszert (pl. *k*-medoid vagy *k*-medián) választunk.

Az anyai kötődéssel kapcsolatban elvégzett KKA elemzések

Az AnyaElk és az AnyaSzor input változóval a KKA-t először az alapértelmezés szerinti Hartigan-Wong algoritmusú *k*-közép módszerrel hajtottuk végre, standardizált változókkal, a HKA elemzések tapasztalatait is felhasználva $k = 4$ klaszterszámra, minden optimumkereső ábrát megnézve. A Silhouette-ábra nem meglepő módon $k = 2$-nél érte el maximumát, az EESS%-ábra pedig $k = 3$-nál nagy ugrással (2-ről 3-ra 29 százalékpontot javulva) érte el a 70%-os értéket, majd erősödött tovább $k = 4$-nél 78%-ra. Az átlagos heterogenitás is $k = 4$-nél kezdett ellaposodni (lásd *8. ábra*, amelyen *dissimilarity* az átlagos heterogenitást jelzi). Az f(K)-ábra szerinti optimum egyértelműen a $k = 3$ klaszterszám (lásd *9. ábra*, amelyen a *threshold* felirat a 0,85-ös küszöb szintjét jelzi). Ezek az eredmények 3- vagy 4-klaszteres megoldást valószínűsítenek.



8. ábra. Az átlagos heterogenitás függése a klaszterszámtól a *k*-közép módszer esetén



9. ábra. Az f(K)-torzulás függése a klaszterszámtól a *k*-közép módszer esetén

A KKA ROP-R-beli eredménylistáján (ugyanúgy, mint AHKA és OHKA esetében) igen informatív a standardizált átlagok mintázatának táblázata is (lásd *5. táblázat*).

5. táblázat. A standardizált átlagok mintázata a *k*-közép módszer 4-klaszteres megoldásában (standardizált változók, M=Magas, A=Alacsony, a + jelek száma az extremitás mértékét jelzi)

Klaszter	AnyaElk	AnyaSzor	KLgyak	HC
KL1	M+	.	63	0,37
KL2	A+	.	100	0,18
KL3	.	.	123	0,31
KL4	(M)	M++++	36	1,81

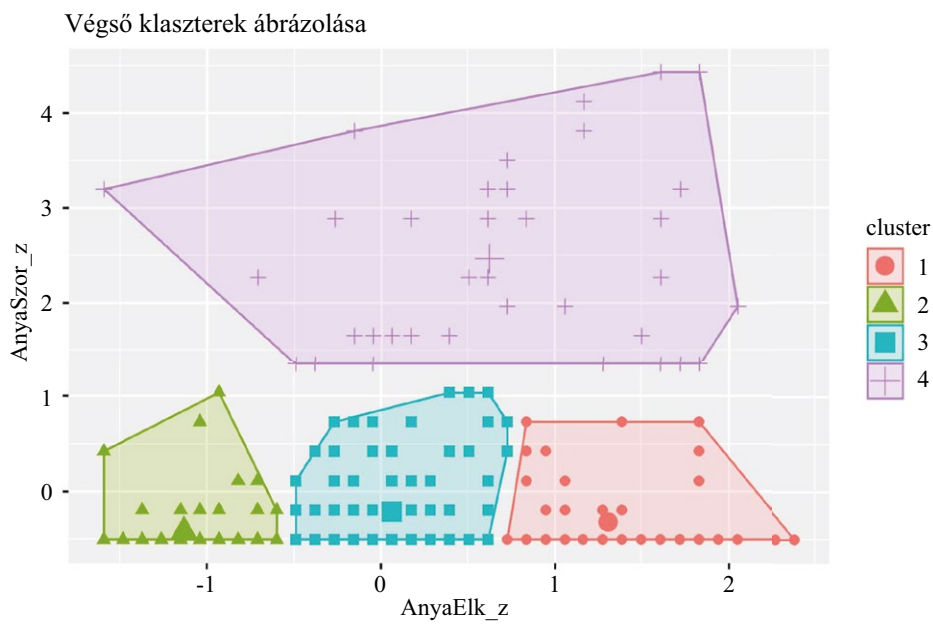
Az *5. táblázat*ból azt olvashatjuk ki, hogy a 4-klaszteres *k*-közép megoldásban az első három klaszter (KL1, KL2 és KL3) erősen homogén (0,40-nél kisebb HC-értékű), átlagos anyai szorongással. Amiben különböznek: az anyai elkerülés szintje. Ez KL1 esetében igen magas (M+), KL2 esetében igen alacsony (A+), KL3 esetében pedig átlagos (.). Ugyanakkor a KL4 klaszter rendkívül heterogén (HC = 1,81), extrém mértékben magas anyai szorongással (M++++).

A Hartigan-Wong helyett a MacQueen- vagy a Lloyd-féle algoritmust választva minimális eltéréssel ugyanerre az eredményre jutunk. Hasonlóképpen a három QC mutató értéke is alig variál: EESS% mindhárom algoritmus esetén 77,87%, HCátlag rendre 0,450, 0,450 és 0,449, XBmod pedig 0,699, 0,699 és 0,682.

3-klaszteres *k*-közép elemzést végezve a megoldás csak abban különbözött a 4-klaszterestől, hogy az első három (KL1,

KL2 és KL3) klaszterből két átlagos anyai szorongású klaszter formálódott: egy magas (M) és egy alacsony (A) anyai elkerülésű klaszter, melyek kis mértékben heterogénebbek, mint a 4-klaszteres megoldásban KL1 és KL3. Mivel itt az EESS% 70,17-re esik vissza, HC-átlag pedig 0,603-ra nő, a 4-klaszteres megoldást érdemes megtartani.

A „Végső klaszterek ábrázolása" opciót bejelölve, a *10. ábrán* látható diagramhoz jutunk, mely a két standardizált változó terében ábrázolja a 4-klaszteres megoldás klasztereit. Erről is leolvashatjuk, hogy az első három klaszter erősen homogén, típusképző, míg KL4 egy erősen heterogén (főleg az anyai elkerülés tekintetében), de az első háromtól mégis markánsan szeparálható klaszter, melynek fő jellemzője a magas anyai szorongás. Ebben a csoportban minden személy anyai szorongás értéke jóval több mint 1 szórással az átlag fölött van. Megjegyezzük, hogy kettőnél több input változó esetén az ábrán az *X*- és az *Y*-tengely az input változókon végzett főkomponens-elemzés első két főkomponense lesz (az ábrán feltüntetve a két főkomponens által megmagyarázott varianciaarányt is).



10. ábra. A végső klasztermegoldás ábrázolása a *k*-közép módszer esetén

Felmerülhet, hogy 4-nél több klaszteres *k*-közép elemzést végezve az igen heterogén KL4 klaszter esetleg felbomlik több homogénebb alklaszterre. Ezt jelen esetben az igen alacsony (n = 36) elemszámú KL4-gyel sajnos, nem tudjuk megtenni, de a *10. ábra* sem jelez ilyen jellegű szeparálódást KL4-en belül.

Figyelembe véve az AnyaSzor változó igen magas ferdeségét felmerülhet, hogy a *k*-medoid vagy a *k*-medián KKA módszer jobban strukturált, homogénebb megoldás-

ra vezet. Elvégezve mindkettőt (*k*-medián esetén három futásból a legjobbat választva), az *5. táblázat*ban láthatóval minden tekintetben megegyező struktúrájú, de kicsit heterogénebb megoldást kaptunk (*k*-medoid és *k*-medián esetén EESS% rendre 77,55, illetve 77,20; HCátlag 0,456 és 0,463; XBmod 0,618 és 0,640 volt), ami ismét a 4-klaszteres *k*-közép megoldás preferálását erősíti. A *k*-medián módszerrel kapcsolatban még megjegyezzük, hogy ebben az esetben ROP-R nem tud olyan tetszetős ábrát készíteni a kapott klaszterekről, mint amilyen a *10. ábr*án is látható. Ez esetben kettőnél több input változó esetén nem készül klaszterábrázolás, két változó esetén pedig egy kétdimenziós pontdiagramot kapunk, ahol a különböző klaszterek egyedei eltérő színekkel vannak jelölve és ki van emelve minden klasztercentrum.

A 4-klaszteres *k*-közép megoldást összevetve az AHKA elemzés 4-klaszteres megoldásával a *4.* és a *10. ábra* alapján, elég jó egyezést találunk A 4. klaszter (KL4) a két megoldásnál teljesen ugyanolyan, a maradék három klaszter pedig a két elemzésnél az anyai szorongás tekintetében átlag körüli, az elkerülés tekintetében pedig rendre az alacsony, az átlagos és a magas szintet képviseli. Ugyanakkor a *k*-közép elemzés megoldása érezhetően homogénebb (EESS% 2,74-gyel magasabb, HCátlag 0,055-tel alacsonyabb).

Az a körülmény, hogy az elemzések a legjobb 4-klaszteres megoldásban csak három megfelelően homogén, típusképző klasztert azonosítottak, arra hívja fel a figyelmet, hogy a feltárt klaszterstruktúra olykor csak részleges eredményre vezet, mert például esetünkben az extrém mértékben heterogén KL4 klaszter aligha tekinthető egy jól körülírt típus képviselőjének. Mindazonáltal szakmailag értékelnünk kell az ilyen részleges eredményt is.

Modell-alapú klaszteranalízis (MKA)

A modell-alapú klaszteranalízis (MKA) modelljében az a kiinduló feltételezés, hogy mintánk adatai egy többdimenziós – többnyire normális – keverékeloszlást követnek, ahol a komponensek mindegyike többdimenziós normális eloszlású, de más centrumokkal, esetleg más varianciákkal és kovarianciákkal. Ebben a keretben MKA célja, hogy azonosítsa az összekevert többdimenziós eloszlások számát (optimális klaszterszám) és megadja az eloszlások jellemzőit (Fraley & Raftery, 2002; Gergely & Vargha, 2021; Vargha, 2022, 7. fejezet). Egy feltárt eloszlásstruktúrában minden eloszlás egy klasztert képvisel. Minden személyt abba a klaszterbe sorolunk, amelyhez tartozó többdimenziós eloszláskomponens a személy értékmintázatát a legnagyobb valószínűséggel produkálja, vagyis amelynél maximális az eloszlás többváltozós sűrűségfüggvényének a személy értékmintázatához tartozó értéke.

Ha a kétdimenziós térben három kétváltozós normális eloszlást összekeverünk (ez felel meg két input változó esetén a háromklaszteres modellnek), a klaszterstruktúra fő típusa 14-féle lehet. A keverékarány által meghatározott relatív klaszternagyság szerint ezek lehetnek egyforma (E) vagy különböző, variábilis (V) méretűek. A főtengelyeik hossza által meghatározott alakjuk szerint szintén egyformák (E) vagy eltérőek, variábilisak (V). Végül az eloszlások két főtengelyének orientációja szerint a kétdimenziós tér *X* és *Y* tengelyével – és ekkor természetesen egymással is – megegyező irányúak (I), a két tengellyel ferde szöget bezáró, de egymással egyező irányultságúak (E), illetve eltérő, változó (V) irányultságú-

ak (lásd részletesebben Vargha, 2022, 7.1. táblázat és 7.2. ábra). Például EEV az egyforma klaszterméret, egyforma eloszlásalak és variábilis orientáció típusa, VII a variábilis klaszterméret és egyforma varianciájú korrelálatlan összetevőkkel jellemezhető eloszlások típusa, VEE a variábilis klaszterméret, egyforma eloszlásalak és azonos orientáció típusa, VVV a minden tekintetben eltérő, illetve variábilis eloszlások típusa stb. Ha kettőnél több input változónk van vagy háromnál több klaszterben gondolkodunk, akkor persze a lehetséges modelltípusok száma 14-nél jóval nagyobb.

A legjobb klaszterstruktúra megtalálása egy statisztikai probléma, amelynek megoldása abból áll, hogy különböző számú és típusú komponensekből álló keverékmodelleket összehasonlítunk és kiértékelünk. A ROP-R ezt a feladatot az *mclust* R-package (vö. Scrucca et al., 2016) segítségével végzi el. Konkrétan az MKA elemzés abból áll, hogy a futtatott program a beállított klaszterszám limitek által meghatározott klaszterszámok és a kijelölt modelltípusok minden kombinációja által meghatározott modellt megpróbálja az adatmintára ráilleszteni és ehhez kiszámít egy bayesi információs kritérium (röviden: BIC) értéket.

Minden modellhez tartozik tehát egy BIC érték, és két modellt összehasonlítva annak a BIC-értéke a nagyobb, amelyik igaz volta esetén nagyobb a vizsgálat alapjául szolgáló minta bekövetkezésének valószínűsége (az elemzett változók tekintetében). Ez okból a program a legnagyobb BIC-értékű modellt fogadja el optimális (legjobb) megoldásként.

Biernacki et al. (2000) szerint a BIC kritérium alkalmas a komponenseloszlások feltárására, de nem optimális egy olyan klaszterbesorolásra, ahol feltételezzük, hogy minden személy egyetlen klaszterbe tartozik. Emiatt, ha a fentebb felvázolt keverékfelbontási módszert valódi klaszteranalízisre akarjuk használni, akkor érdemes a BIC kritériumon kissé módosítani úgy, hogy BIC képletében a klaszterek átfedését egy entrópia tag „büntesse". Ez az entrópia akkor magas, ha több személy esetén is nagy a bizonytalanság afelől, hogy a személy melyik klaszterbe tartozik. Ezt a módosított BIC kritériumot ICL (Integrated Complete-data Likelihood) kritériumnak nevezik. Az ICL-re építő módszer nem javasol más eljárást a komponenseloszlások becslésére, mint a BIC-re építő, csak a végső modellválasztás során BIC helyett az ICL kritérium értéke alapján kell dönteni arról, hogy melyik modell a legjobb.

Biernacki et al. (2000) szerint az ICL jobban megfelel a klaszteranalízis diszkrét szemléletének, mely szerint az MKA-val nem folytonos komponenseloszlásokat akarunk feltárni, hanem diszkrét klasztereket. Egyes vizsgálatok szerint azonban az ICL-alapú döntéssel ritkán jutunk jobb klaszterstruktúrához, mint a BIC-en alapulóval (Gergely & Vargha, 2021).

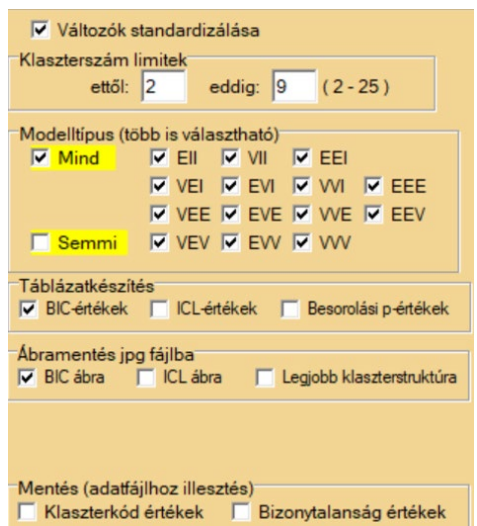
Ha már megvan a program szerint legjobb modell, akkor ennek segítségével kiszámíthatók az ún. klaszterbesorolási (vagy röviden besorolási) valószínűségek is, amelyek a minta minden személye esetén megadják, hogy a személy ebben a modellben milyen valószínűséggel tartozik egyik vagy másik klaszterbe. A program természetesen minden személyt abba a klaszterbe sorol, amelyre vonatkozóan ez a besorolási valószínűség a legnagyobb (p_{max}). Minél közelebb van ez az érték 1-hez, annál egyértelműbb a személy klaszterbesorolása és minél kisebb 1-nél, annál nagyobb a besorolási bizonytalanság, amit az $1 - p_{max}$ mutatóval mérhetünk (Gormley et al., 2023). A besorolási bizonytalanság tehát minden személy esetén az idegen (nem saját) klaszterekre vonatkozó besorolási valószínűségek összege.

Végül megjegyezzük, hogy a klaszstermodellek kiértékelésére itt kiszámított BIC mutató (mely a minta adott modellhez tartozó bekövetkezési valószínűségének egy monoton növő függvénye) hasonló logikájú, mint a CFA-modellek kiértékelése során használt BIC kritérium (lásd 6.1. táblázat), azzal a különbséggel, hogy a CFA esetén ez át van fordítva pozitívra (kényelmi okokból, hogy értéke pozitív legyen). Emiatt a CFA elemzések során mindig a kisebb BIC-érték jelez jobb illeszkedést, míg az MKA elemzések esetén mindig a nagyobb.

MKA elemzés ROP-R-ben a modell-alapú klaszteranalízis (röviden ModellKLA vagy MKA) modullal végezhető. Ehhez a ROP-R az *mclust* (Scrucca et al., 2016), *factoextra* (Kassambara & Mundt, 2020), *ggplot2* (Wickham, 2016) R-package-eket használja fel.

Az MKA modul menüablaka

Ez a menüablak szolgál az elemzésbe bevonandó változók kiválasztására (input változók ablaka), a lehetséges klaszterszámok alsó és felső határának a beállítására, az elemzendő modelltípusoknak a kiválasztására, valamint egy sor opció kijelölésére (lásd *11. ábra*), amelyekhez az alábbi magyarázatokat fűzzük.



11. ábra. A modell-alapú klaszteranalízis moduljának lehetséges beállításai ROP-R-ben

Modelltípus

Ezen a panelen lehet a program számára megadni, hogy milyen modelltípusokon belül keresse a legjobb megoldást. A választáshoz 14 modelltípus áll rendelkezésre. Ha kettőnél több input változónk van vagy háromnál több klaszterben gondolkodunk, akkor persze a lehetséges modelltípusok száma 14-nél jóval nagyobb. A ROP-R MKA modulja ilyenkor sem tud több típust felajánlani, mert az *mclust* R-package csak ennek a 14 modelltípusnak a választását teszi lehetővé. Ez a 14 típus azonban elég szokott lenni egy elfogadható klaszterstruktúra feltárására, ha ilyen egyáltalán létezik.

Táblázatkészítés

ROP-R kiírja az eredménylistára a szerinte legjobb (legnagyobb BIC értékű) modell típusát és komponensszámát (az optimális klaszterszámot), valamint (alapértelmezés szerint) az összes megvizsgált modell BIC-értékének táblázatát. Külön kérésre elkészül az ICL-értékek táblázata is mind a 14 választható modellre, a kijelölt klaszterszámokkal. Ha itt bejelöljük a „Besorolási p-értékek" opciót, akkor ROP-R kiszámítja minden érvényes személyre, hogy az általa azonosított legjobb modell esetén a személy milyen valószínűséggel tartozik az egyes klaszterekbe és ezek táblázatát fájlba menti (pval1.txt néven az „aktualis" almappában). E táblázat alapján például megállapítható, hogy kik azok, akik több klaszterhez is kötődnek. Ha egy személy nem vesz részt az MKA elemzésben, mert valamelyik input változóra nincs érvényes értéke, ebben a táblázatban a személyhez tartozó esetsorszám mellett egy „NA" jelölést találunk.

Ábramentés

Ha nem tiltjuk le, akkor ROP-R elkészíti a BIC-értékeknek nemcsak a táblázatát, hanem az ábráját is, illetve külön kérésre az ICL-értékek ábráját (BIC1.jpg, illetve ICL1.jpg néven az „aktualis" almappában). Ezekről könnyen leolvasható, hogy melyik modell a „legjobb" (a legnagyobb BIC, illetve ICL értékű). Ez az automatikus döntés, azonban néha szuboptimális eredményre vezet (Gergely & Vargha, 2021). A BIC vagy ICL értékek grafikonján a szabálytalan, hirtelen ugrásokkal jellemezhető vagy torzó grafikonok ritkán azonosítják a legjobb megoldást még olyan esetekben is, amikor a maximális BIC értékű modell itt található.

Ezen a rovaton kérhető a legjobb klaszterstruktúra ábrázolása is, amelyet már a KKA modulban is említettünk (lásd pl. *10. ábra*), továbbá 3-nál nem több input változó esetén három ábra a legjobb megoldás értelmezésének megkönnyítésére:

1) Klasszifikációs ábra: ez az adatok pontdiagramja, amelyen a különböző klaszterekbe tartozó személyek eltérő színekkel és formákkal vannak jelölve.

2) Bizonytalanságok ábrája: ez az adatok olyan pontdiagramja, amelyen a különböző klaszterekbe tartozó személyek eltérő színű kitöltött körökkel vannak jelölve úgy, hogy a nagyobb bizonytalanság ($1 - p_{max}$) értékű személyt nagyobb kör képviseli.

3) Sűrűségábra: a többdimenziós adatok klaszterenkénti sűrűségeloszlása.

Megjegyezzük, hogy a ROP-R eredménylistájának a végén olvasható a létrehozott képfájlok neve pontos elérési úttal.

Mentés (adatfájlhoz illesztés)

Külön kérésre a ROP-R elmenti (új változóként az adatállományhoz illeszti) a legjobb megoldás személyenkénti klaszterkódját, illetve besorolási bizonytalanságát.

MKA végrehajtása

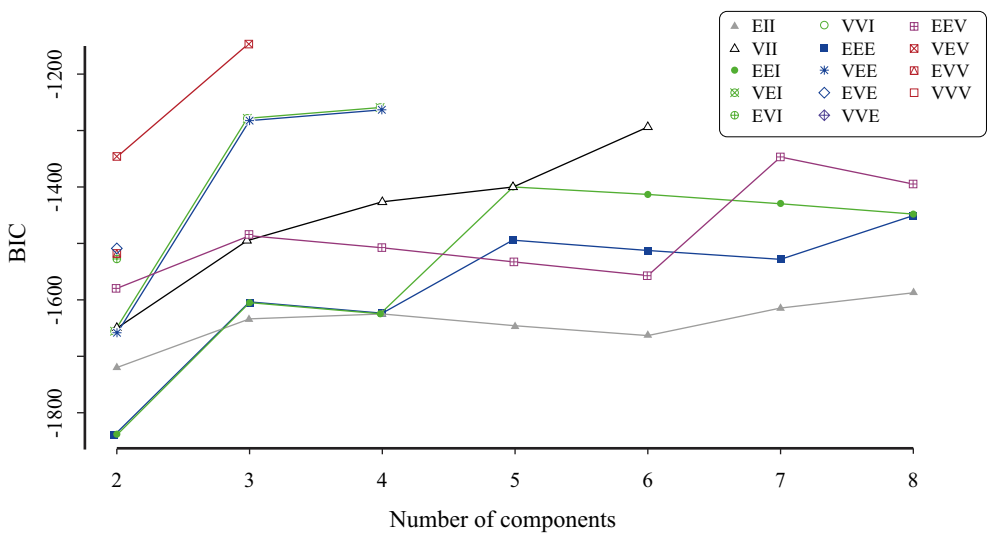
MKA-t több lépésben célszerű végrehajtani. Kezdetben jelöljünk ki minden modelltípust (ez az alapbeállítás) és adjunk meg olyan klaszterszám limiteket, amelyek minden bizonnyal közrefogják az optimális klaszterszámot. Kérjünk BIC- és ICL-ábrát. Az első futtatás után ezen ábrák megszemlélése alapján döntsünk az alaposabb elemzésre érdemes modellekről, amelyeket a klaszterszámok és a modelltípusok alkalmas szűkítésével jelölhetünk ki újabb futtatásokra. E futtatások kiértékelése a KKA esetében már megismert formátumú eredménylista alapján történhet.

Egy MKA elemzés végrehajtása után az „aktualis” mappában találjuk az elemzéshez elkészített ideiglenes adatfájlt (tmpdat.txt), a legjobb megoldás klaszterszámához tartozó klaszterváltozóval kiegészített ideiglenes adatfájlt (tmpdat2.txt), a futtatott R-scriptet (MBCA.r), valamint a kért ábrákat. Ha feltételes csoportosító változót is kijelölünk, akkor minden feltételes csoport elemzése során elkészülnek a kért diagramok (a csoportindexek megjelennek a fájlnevekben).

Az anyai kötődéssel kapcsolatban elvégzett MKA elemzések

Az AnyaElk és az AnyaSzor input változóval MKA-t először a 2–8 klaszterszámokkal az összes modelltípusra, BIC- és ICL-ábrát kérve futtattuk. Az ICL-ábra kis eltérésekkel ugyanúgy nézett ki, mint a BIC-ábra (lásd *12. ábra*). Az ábrán nem látható mind a 14 típus esetén grafikon, illetve a látható grafikonok se húzódnak több esetben végig a teljes 2–8 klaszterszám tartományon. Ilyen eset azért fordulhat elő, mert a modellbecslés során előfordulhat, hogy nem konvergál a megoldás és ilyenkor nincs érvényes BIC-érték sem. Ez a ROP-R eredménylistán a BIC-értékek – itt nem közölt – táblázatából is kiolvasható.

A legnagyobb BIC-értékű, vagyis a legjobb megoldás a 3-klaszteres VEV (röviden VEV3) modell, melynél az eredménylistáról leolvashatóan HCátlag = 0,923, EESS% = 54,16 és XBmod = 0,639. Ez egy meglehetősen heterogén megoldást sejtet, ami kiolvasható a standardizált átlagok mintázatának táblázatából is (lásd *6. táblázat*).



12. ábra. A modell-alapú klaszteranalízis BIC-ábrája ROP-R-ben

6. táblázat. A standardizált átlagok mintázata MKA-ban a 3-klaszteres VEV megoldásban (standardizált változók, M=Magas, A=Alacsony, a + jelek száma az extremitás mértékét jelzi)

Klaszter	AnyaElk	AnyaSzor	KLgyak	HC
KL1	.	(A)	157	0,67
KL2	A+	(A)	79	0,08
KL3	.	M+	86	2,15

A *6. táblázat* szerint a VEV3 modellben egyetlen kellően homogén klasztert találunk (KL2-t, ahol HC = 0,08). Tekintetbe véve, hogy a modell grafikonja k = 3-nál hirtelen megszakad, talán jobb modellt kapunk, ha ezt a típust elhagyjuk a lehetőségek közül. Ily módon újra futtatva az MKA-t, a 4-klaszteres VEI (VEI4) modell lett a legjobb, HCátlag = 0,778, EESS% = 61,40, XBmod = 0,604 QC-értékekkel. Ezek minden tekintetben jobb struktúrára utalnak, mint a VEV3 esetében, amit megerősít a standardizált átlagok mintázatának táblázata is (lásd *7. táblázat*).

7. táblázat. A standardizált átlagok mintázata MKA-ban a 4-klaszteres VEI megoldásban (standardizált változók, M=Magas, A=Alacsony, a + jelek száma az extremitás mértékét jelzi)

Klaszter	AnyaElk	AnyaSzor	KLgyak	HC
KL1	.	(A)	142	0,41
KL2	A+	(A)	76	0,08
KL3	.	M+	86	2,15
KL4	M+++	(A)	18	0,10

A *7. táblázat* szerint a VEI4 modellben két erősen homogén (HC: 0,08 és 0,10) és egy elfogadhatóan homogén (HC = 0,41) klasztert látunk egyetlen igen heterogén klaszter (HC = 2,15) mellett. Ez a struktúra annyiban hasonlít a legjobb k-közép megoldásra (vö. *5. táblázat*), hogy ez is 4-klaszteres, ennél is három típusképző és egy extrém mértékben heterogén klasztert találunk, s ezen felül még a KL2 jelzetű klaszter mintázata is hasonló, igen alacsony (A+) anyai elkerüléssel és átlag körüli (.) anyai szorongással. A fő különbség abból fakad, hogy a k-közép módszer a teljes struktúra lehető legnagyobb homogenitását tűzi ki célul (ennek köszönhetően itt EESS% = 77,87, szemben a VEI4 modell 61,40-es értékével), míg az MKA az egyes komponenseloszlásokra keres jól illeszthető modellt, ami több igen homogén klasztert eredményez (VEI4-ben két klaszter, KL2 és KL4 is homogénebb a legjobb 4-klaszteres k-közép megoldás leghomogénebb KL2 klaszterénél).

Tovább próbálkozva az MKA modellekkel, a modelltípus rovatból elhagytuk a VEI típust is, ekkor azonban a legjobb VEE4 megoldás minden tekintetben megegyezett VEI4-gyel. A VEE típust is elhagyva a legjobb a modell VII6 lett, HCátlag = 0,542, EESS% = 73,20, XBmod = -0,550 QC-értékkel. Ez ugyan homogénebb struktúra, mint VEI4, viszont a negatív XBmod-érték azt jelzi, hogy egyes klaszterek túlzott mértékben hasonlítanak egymásra. A standardizált

átlagok mintázatának táblázata (lásd *8. táblázat*) alapján a legnagyobb hasonlóság KL2 és KL5 között van, amelyek egyaránt a VEI4/KL2 típust replikálják. A VEE típust is elhagyva a legjobb modell EEV7 lett, minden tekintetben rosszabb, mint VEE4 és VII6.

8. táblázat. A standardizált átlagok mintázata MKA-ban a 6-klaszteres VII megoldásban (standardizált változók, M=Magas, A=Alacsony, a + jelek száma az extremitás mértékét jelzi)

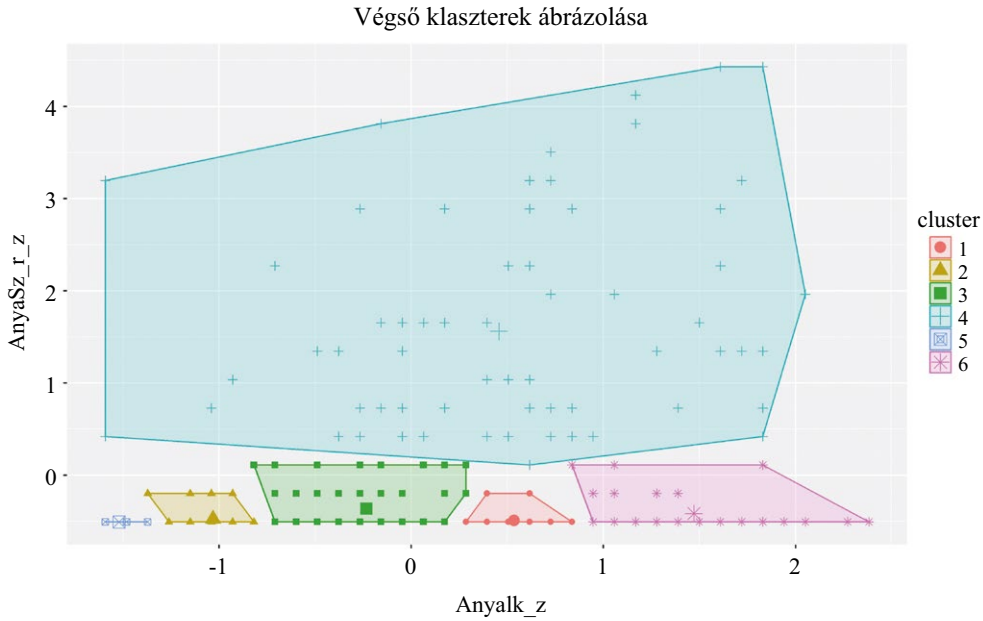
Klaszter	**AnyaElk**	**AnyaSzor**	**KLgyak**	**HC**
KL1	(M)	(A)	41	0,04
KL2	A+	(A)	50	0,03
KL3	.	.	86	0,13
KL4	(M)	M++	70	2,13
KL5	A++	(A)	31	0,01
KL6	M++	.	44	0,25

A 6-klaszteres VII megoldás (VII6) KL2 és KL5 klaszterének nagyfokú hasonlósága a klasztercentroidok páronkénti ASED távolságainak táblázatából is kiolvasható (lásd *9. táblázat*), mely a ROP-R eredménylistáján szintén megtalálható. Eszerint KL2 és KL5 centroidjának ASED távolsága, vagyis a standardizált változók szerinti átlagok átlagos négyzetes eltérése a két klaszterre vonatkozóan) csak 0,12, mely 0,12 átlagos varianciaként értelmezhető. 0,15 alatti *d* távolságértékek esetén igen közeli, 0,50 alatti távolságértékek esetén közeli klasztercentrumokról beszélhetünk. A jelen esetben centrumuk alapján közelinek tekinthető még KL1 és KL3 (d = 0,30), KL1 és KL6 (d = 0,44), valamint KL2 és KL3 (d = 0,33) is.

9. táblázat. A klasztercentroidok páronkénti *d* ASED távolságai a 6-klaszteres VII megoldásban

Klaszter	**KL1**	**KL2**	**KL3**	**KL4**	**KL5**	**KL6**
KL1	0	1,23	0,30*	2,11	2,11	0,44*
KL2	1,23	0	0,33*	3,18	0,12**	3,13
KL3	0,30*	0,33*	0	2,09	0,84	1,45
KL4	2,11	3,18	2,09	0	4,09	2,47
KL5	2,11	0,12**	0,84	4,09	0	4,48
KL6	0,44*	3,13	1,45	2,47	4,48	0

*Jelölés: *: d* < 0,50; **: *d* < 0,15



13. ábra. A végső VII6 klasztermegoldás ábrázolása az MKA módszer esetén

Mindent összevetve MKA-ban a VII6 modellt érdemes elfogadni, mint végső megoldást (lásd *13. ábra*), mert ez szolgáltatja a leghomogénebb klasztereket (a 6-ból 5 klaszter HC-értéke nem haladja meg a 0,25-öt). Persze ilyen sok klaszter stabilitását már nagyobb mintán szükséges ellenőrizni. A megoldás sokban emlékeztet a legjobb *k*-közép megoldásra (vö. *10. ábra*). A *13. ábrá*ról az is leolvasható, hogy a *9. táblázat* alapján közelinek ítélt klaszterek az ábrán mind szomszédos alakzatok.

Az összes klaszterelemzés eredményét figyelembe véve azt láthatjuk, hogy mintánkban van egy nagyon heterogén alcsoport, amelynek anyai elkerülése e skála teljes spektrumán átível, de akik mind hasonlítanak egymásra abban, hogy anyai szorongásuk az átlagosnál nagyobb (lásd *9.* és *13. ábra*). Ezek a személyek a bizonytalanul kötődők. A többi alcsoport anyai szorongásszintje alacsonyabb, ezeket az anyai elkerülés szintje különbözteti meg egymástól. A legjobb kötődésűek a *k*-közép/KL2, illetve a VII6/KL2 és a VII6/KL5 klaszterbe tartozó személyek. Egyébként ezt a két utóbbi klasztert egyesítve még mindig nagyon homogén klasztert kapunk, 0,08-as HC-értékkel[14]. Markáns még a kötődés szempontjából teljesen átlagos személyek klasztere (*k*-közép/KL3, illetve VII6/KL3), amelyhez hasonló a VII6/KL1 klaszter is (vö. *9. táblázat*). E homogén alcsoportok feltárásában az MKA hatékonyabbnak tűnik, ezért ha csak részleges típusfeltárást kapunk eredményül, akkor az MKA-t érdemes lehet a többi módszerrel

[14] Ehhez az elmentett VII6 klaszterváltozóban át kell kódolni az 5 értéket 2-re a „Szerkesztés/Keres, cserél" vagy a „Transzformációk/Átkódolás" menüpont segítségével, majd ezt az immár 5-klaszteres kódváltozót a ROPstat Validálás moduljában kiértékelni.

szemben előnyben részesíteni. Azt azonban megjegyezzük, hogy automatikus kiértékeléssel ritkán kapjuk meg MKA-ban a legjobb megoldást. Több ígéretes modellt is meg kell vizsgálni, és egyes egymáshoz igen hasonló klaszterek összevonásával olykor még ezután is javítható a klaszterstruktúra.

Fraley et al. (2011) modelljével összevetve a kapott eredményeket azt mondhatjuk, hogy nem nyert megerősítést az egyszerű négy típusos (alacsony-alacsony, magas-magas, alacsony-magas, magas-alacsony) kötődésmodell. Mintánk több klasztere hasonlít egyik-másik típusra, de úgy látszik, hogy az anyai kötődés viszonylatában a valóság bonyolultabb, mint az elmélet. Elképzelhető az is, hogy a kapott eredmények részben a minta viszonylag alacsony méretéből és/vagy nem reprezentatív voltából fakadnak és talán egy ebből a szempontból megfelelőbb mintán elvégzett elemzés az elmélettel jobb egyezést mutatna.

Megbeszélés

A klaszteranalízis a személy-orientált pszichológiai kutatások kedvelt módszere. A személy-orientált megközelítés hangsúlyozza, hogy a személyt jellemző adatokat, változóértékeket feldarabolatlan egységként kell tekinteni és kezelni. A személy-orientált többváltozós statisztika olyan eljárásokra fókuszál, amelyek esetében központi szerepet játszanak az egyének közti kvalitatív jellegű különbségek. Ezek hátterében típusmodellek állnak, amelyek jellemzően klasszifikációs módszerekkel tárhatók fel. Cikkünkben áttekintettük a klaszteranalízis alapfogalmait (alkalmazási feltételek, személyek távolsága, klaszterek távolsága), típusait (összevonó, agglomeratív hierarchikus klaszteranalízis: AHKA, osztódó hierarchikus klaszteranalízis: OHKA, nem hierarchikus, *k*-középpontú klaszteranalízis: KKA, modell-alapú klaszteranalízis: MKA).

ROP-R egy ingyenesen elérhető[15], kétnyelvű (magyar és angol), R szoftver alapú, de ROPstat keretben használható többváltozós statisztikai programcsomag, amelynek tíz modulja a többváltozós statisztika három témakörében kínál teljes körű statisztikai elemzéseket: regresszióelemzés, dimenzióredukció (főkomponens- és faktoranalízis), valamint klaszteranalízis (Vargha & Bánsági, 2022; Vargha et al., 2024). Klaszterelemzések végzésére a ROP-R négy modulja áll rendelkezésre (AHKA, OHKA, KKA és MKA), számos olyan lehetőséggel (pl. OHKA, *k*-medoid KKA, *k*-medián KKA, MKA), amelyek más felhasználóbarát menüvezérelt szoftverekben nem találhatók meg. Cikkünkben egy kötődéskutatás (Jantek & Vargha, 2016) adatain szemléltettük, hogy lehet a ROP-R szoftver segítségével különféle klaszterelemzéseket végrehajtani és ezek eredményeit tetszetős ábrák és hasznos táblázatok segítségével kiértékelni.

Ami ROP-R-ből a klaszteranalízissel kapcsolatban hiányzik, az az utóelemzések lehetősége (vö. Vargha, 2022, 8. fejezet). Ezek közé tartoznak az alábbiak:

1. A kapott klaszterstruktúra kiértékelése a ROP-R-ben nem elérhető több más mutató (például PB, CLdelta, MORI-index; vö. Vargha, 2022, 4.4. táblázat, illetve 8.1. alfejezet) segítségével.
2. A kapott klaszterstruktúra kapcsolatának vizsgálata olyan külső változókkal, mint a nem, az életkor, az iskolázottsági szint stb.

[15] lásd www.ropstat.com

3. Különböző klasztermegoldások összevetése, hasonlóságuk mérése.

Mindezen lehetőségek a ROPstatban rendelkezésre állnak és mivel ROP-R egy ROPstat keretben futtatható szoftver, a ROP-R-ben elmentett klaszterváltozókkal kibővített adatfájlok ROPstatban mindenféle konverzió nélkül beolvashatók és ezek az utóelemzések ROPstat segítségével mind elvégezhetők (részletesen lásd Vargha, 2022, 8. fejezet). Ezt az is megkönnyíti, hogy a ROPstat legfrissebb letölthető verziója már úgy működik, hogy ha a ROPstat szoftver észleli, hogy mellette a ROP-R is telepítve van, akkor a ROPstat megjeleníti a ROP-R többváltozós menüjét is. Ami azt jelenti, hogy gyakorlatilag a ROPstaton belül végrehajtható pl. egy MKA elemzés, ennek elmenthető a klaszterváltozója, majd ezt a megoldást a ROPstat „Mintázatfeltáró elemzések/Validálás" menüpontjában a QC mutatók széles választéka és MORI segítségével egyben ki is lehet értékelni.

Azt azért megjegyezzük, hogy a HCátlag, mely a változók saját skáláján (standardizált változók esetén szórásléptékkel) méri a klaszterstruktúra homogenitását (kis értékek nagy homogenitást jeleznek), az EESS% mutató, mely a regressziós elemzések R^2 megmagyarázott varianciaarány mutatójával rokon (itt a magas értékek jeleznek nagy homogenitást), valamint XBmod, mely a klaszterek egymástól való elkülönülését mérő szeparációs mutató, együtt alkalmasnak tűnnek különböző klasztermegoldások összehasonlítására (lásd *3.* és *4. táblázat*).

Fontos kiemelni, hogy a ROP-R moduljai közül a klaszterező modulok vannak a legnagyobb számban képviselve és bár ezek a klaszterelemzési módszerek széles tartományát lefedik, nem merítik ki az összes klaszterezési lehetőséget (vö. angolul Kaufman & Rousseeuw, 1990; Hastie et al., 2009; Moisl, 2015; illetve magyarul Füstös et al., 2004).

A jelen cikk a személyfókuszú kutatások szempontjából fontos módszerekre helyezte a hangsúlyt, s ebből a szempontból – szemléltető pszichológiai példáink által alátámasztva – a ROP-R minden klaszterező modulja tudott értékes információval szolgálni. A kapott eredményeket áttekintve azt mondhatjuk, hogy a hierarchikus klaszteranalízisek (AHKA és OHKA) sikeresen jelezték azt a minimális klaszterszámot (itt $k = 4$), amelyre adott mintánk bontandó. A nem hierarchikus KKA módszerek egységesen homogénebb struktúrát eredményeztek ugyanannyi klaszterrel, mint a hierarchikus elemzések. A három altípus (*k*-közép, *k*-medoid, *k*-medián) eredménye között nem volt érdemi különbség, kis előnnyel a *k*-közép (azon belül is Hartigan-Wong algoritmusú) elemzés produkálta a leghomogénebb struktúrát.

Mindamellett pszichológiai példánkban sem 4, sem több klaszter választása nem tudott csupa homogén klasztert azonosítani, mert volt a mintának egy nagyon heterogén része (az átlagosnál magasabb anyai szorongással jellemezhető személyek), ami minden klasztermegoldásban megjelent. Ilyenkor meg kell elégednünk azzal a részeredménnyel, hogy csak néhány megfelelően homogén klaszterképző típust tudtunk azonosítani.

E homogén alcsoportok feltárásában a modell-alapú klaszteranalízis (MKA) hatékonyabbnak tűnt, ezért részleges típusfeltárás esetén az MKA-t érdemes lehet a többi módszerrel szemben előnyben részesíteni. Azt azonban figyelembe kell venni, hogy a BIC- vagy az ICL-ábrán, illetve az azok táblázatain alapuló automatikus kiértékeléssel ritkán kapjuk meg MKA-ban a legjobb megoldást. Több ígéretes modellt is meg kell vizsgálni, és olykor egyes egymáshoz igen

hasonló klasztereket össze kell vonni a klaszterstruktúra javításának érdekében.

Köszönetnyilvánítás

A jelen cikk elkészítését a Károli Gáspár Református Egyetem Bölcsészet- és Társadalomtudományi Karának *Pszichológiai kutatások módszertani platformja* című kutatói pályázata támogatta (témaszám: 20754B800/2022). Köszönettel tartozom Bánsági Péternek, hogy szerzőtársként nagy lelkesedéssel és hozzáértéssel vett részt a ROP-R klaszterező moduljainak programozásában. Itt szeretnék köszönetet mondani Jakab Zoltánnak és a kéziratot bíráló anonim kollégáknak is, hogy alaposan elmélyedtek a kézirat első változatában és számos hasznos észrevétellel és javaslattal járultak hozzá a cikk színvonalának emeléséhez. Hálás köszönetem végül Jantek Gyöngyvérnek, hogy kötődéskutatási adatait a jelen cikk szemléltető példáiban felhasználhattam.

Summary

Cluster Analysis In Psychological Research Using The ROP-R Software

Background and aims: Cluster analysis is a popular method in person-oriented psychological research. While variable-oriented research is limited to examining indicators (e.g. mean, correlation, etc.) that characterize variables rather than individuals, person-oriented approaches focus on individual differences, and regard variable values characterizing an individual as a unitary representation. Person-oriented multivariate statistics focus on processes in which qualitative differences between individuals play a central role. These are based on type models, which are typically explored using classification methods. In this article, we review the basic concepts of cluster analysis and then use data from real psychological research to show how hierarchical, k-centers non-hierarchical, and model-based cluster analyses can be performed using the free multivariate statistical software ROP-R.

Keywords: person-oriented multivariate statistics, cluster analysis, ROP-R

Irodalom

Aggarwal, C. C. (2017). *Outlier Analysis, 2nd Edition.* Springer.

Bartha, L. (1980) (szerk.). *Pszichológiai alapfogalmak kis enciklopédiája.* Tankönyvkiadó.

Bergman, L. R., & Lundh, L. G. (2015). Introduction: The person-oriented approach: Roots and roads to the future. *Journal for Person-Oriented Research, 1*(1–2), 1–6. https://doi.org/10.17505/jpor.2015.01

Bergman, L. R., Magnusson, D., & El-Khouri, B. M. (2003). *Studying individual development in an interindividual context. A Person-oriented approach.* Lawrence-Erlbaum Associates.

Biernacki, C., Celeux, G., & Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE transactions on pattern analysis and machine intelligence*, *22*(7), 719–725. https://doi.org/10.1109/34.865189

Cardot, H. (2022). *Gmedian: Geometric Median, k-Medians Clustering and Robust Median PCA. R package version 1.2.7.* https://CRAN.R-project.org/package=Gmedian

Fraley, R. C., Heffernan, M. E., Vicary, A. M., & Brumbaugh, C. C. (2011). The Experiences in Close Relationships-Relationship Structures questionnaire: A method for assessing attachment orientations across relationships. *Psyhological Assessment, 23*(3), 615–625. https://doi.org/10.1037/a0022898

Fraley, C., & Raftery, A. E. (2002). Model-based clustering, discriminant analysis and density estimation. *Journal of the American Statistical Association, 97*(458)*,* 611–631. https://doi.org/10.1198/016214502760047131

Füstös, L., Kovács, E., Meszéna, Gy., & Simonné Mosolygó, N. (2004). *Alak-felismerés (Sokváltozós matematikai módszerek)*. Új Mandátum.

Gergely, B., & Vargha, A. (2021). How to Use Model-Based Cluster Analysis Efficiently in Person-Oriented Research. *Journal for Person-Oriented Research*. *7*(1), 22–35. https://doi.org/10.17505/jpor.2021.23449

Gormley, I. C., Murphy, T. B., & Raftery, A. E. (2023). Model-Based Clustering. *Annual Review of Statistics and Its Application*, *10*, 573-595. https://doi.org/10.1146/annurev-statistics-033121-115326

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. *Springer Series in Statistics.* Springer Science & Business Media.

Jantek, G., & Vargha, A. (2016). A felnőtt kötődés korszerű mérési lehetősége: A közvetlen kapcsolatok élményei – kapcsolati struktúrák (ECR-RS) kötődési kérdőív magyar adaptációja párkapcsolatban élő felnőtt személyeknél. *Magyar Pszichológiai Szemle, 71*(3), 447–470. http://dx.doi.org/10.1556/0016.2016.71.3.3

Kassambara, A., & Mundt, F. (2020). *factoextra: Extract and Visualize the Results of Multivariate Data Analyses.* R package version 1.0.7. https://CRAN.R-project.org/package=-factoextra

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding Groups in Data: an Introduction to Cluster Analysis*. John Wiley & Sons.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., & Hornik, K. (2022). *cluster: Cluster Analysis Basics and Extensions.* R package version 2.1.4. https://CRAN.R-project.org/package=cluster

Moisl, H. (2015). *Cluster analysis for corpus linguistics*. Walter de Gruyter GmbH.

Pham, D. T., Dimov, S. S., & Nguyen, C. D. (2005). Selection of K in K-means clustering. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, *219*(1), 103–119. https://doi.org/10.1243/095440605X8298

R Core Team (2021). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20*(1)*,* 53-65. https://doi.org/10.1016/0377-0427(87)90125-7

Scrucca L., Fop, M., Murphy, T. B., & Raftery, A. E. (2016). mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. *The R Journal, 8*(1), 289–317. https://doi.org/10.32614/RJ-2016-021

Takács, Sz., Makrai, B., & Vargha, A. (2015). Klasszifikációs módszerek mutatói. *Psychologia Hungarica Caroliensis*. *3*(1), 67–88. https://doi.org/10.12663/PsyHung.3.2015.1.5

Vargha, A. (2016). A ROPstat statisztikai programcsomag. *Statisztikai Szemle*, *94*(11-12), 1165–1192. https://doi.org/10.20311/stat2016.11-12.hu1165

Vargha, A. (2019). *Többváltozós statisztika dióhéjban: változó-orientált módszerek.* Pólya Kiadó.

Vargha, A. (2020). *Normális vagy? És ha nem? Statisztikai módszerek nem normális eloszlású változókkal pszichológiai kutatásokban.* Pólya Kiadó.

Vargha, A. (2022). *Személy-orientált többváltozós statisztika: Klasszifikációs módszerek.* Pólya Kiadó.

Vargha, A. & Bánsági, P. (2022). ROP-R: a free multivariate statistical software that runs R packages in a ROPstat framework. *Hungarian Statistical Review, 5*(2), 3–29. https://doi.org/10.35618/HSR2022.02.en003. Letölthető: https://www.ksh.hu/hungarian-statistical-review#/year/2022?c=h#02

Vargha, A., Bánsági, P., & Jantek, G. (2024). Statisztikai elemzések a ROP-R szoftver segítségével és szemléltetésük egy kötődéskutatás adataival. *Mentálhigiéné és Pszichoszomatika, 25*(1), 36–55. https://doi.org/10.1556/0406.2024.00028

Vargha, A. & Bergman, L. R. (2019). MORI coefficients as indicators of a "real" cluster structure. *Hungarian Statistical Review, 2*(1), 3–23. https://doi.org/10.35618/hsr 2019.01.en003

Vargha, A., Bergman, L. R., & Takács, S. (2016). Performing cluster analysis within a person-oriented context: Some methods for evaluating the quality of cluster solutions. *Journal for Person-Oriented Research*, *2*(1–2), 78–886. https://doi.org/10.17505/jpor.2016.08

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis.* Springer Verlag.