Statisztikai Szemle

Közzététel: 2019. május 7.

A tanulmány címe:

Klaszterszám-meghatározási módszerek összehasonlítása

Szerző:
Szüle Borbála, a Budapesti Corvinus Egyetem egyetemi docense

E-mail: borbala.szule@uni-corvinus.hu

DOI: 10.20311/stat2019.5.hu0421

Klaszterszám-meghatározási módszerek összehasonlítása*

Szüle Borbála,
a Budapesti Corvinus Egyetem egyetemi docense
E-mail: borbala.szule@uni-corvinus.hu

A matematikai előfeltevések hiánya miatt a klaszterelemzés rugalmasan alkalmazható többféle adatbázis esetén, továbbá a klaszterezés megfelelőségének kérdése is sok szempontból közelíthető meg. Jelen tanulmány az optimális klaszterszám meghatározásával, valamint a különböző módszerek konzisztenciájával foglalkozik. A gyakran alkalmazott „könyök-" és a sziluettmódszert homogén és heterogén adatbázisok esetében hasonlítja össze a szerző. A gyakorlat szempontjából fontos eredmény, hogy mindkét eljárás esetén jelentős különbség van az alkalmazott kétféle adatbázis között: a heterogén adatbázisnál az optimális klaszterszám egyértelműbben megtalálható, és ennek értékét a dimenziószám (a változók száma) csak kismértékben befolyásolja. A tanulmány eredményei alapján arra lehet következtetni, hogy a dimenziószámtól függetlenül mindkét módszer egyformán jól alkalmazható az adatbázis heterogenitásának felderítésére, és a klaszterszám-meghatározás során is csak kis különbségek tapasztalhatók.

TÁRGYSZÓ:
Klaszterelemzés.
Klaszterszám.
Dimenzionalitás

DOI: 10.20311/stat2019.5.hu0421

* A szerző köszönetét fejezi ki a tanulmány lektorának értékes észrevételeiért és javaslataiért.

A kvantitatív kutatások során gyakran felmerül a kérdés, hogy vannak-e az adatbázisban egymástól elkülönülő csoportok (klaszterek). Erre a kérdésre nagy (2-nél több változót tartalmazó) adatbázisoknál általában nem egyszerű a vizualizált adatok megtekintése alapján válaszolni, de a helyes válasz megtalálása jelentős előnyökkel járhat (például, ha sikerül olyan fogyasztói szegmenst találni, amely részére jövedelmezően lehet értékesíteni valamely terméket). A klaszterelemzés mint tanító nélküli (unsupervised) módszer (*Kovács–Legány–Babos* [2006]) pontosan az ilyen típusú kutatási kérdéseknél alkalmazható, a többdimenziós csoportok megtalálásához nyújthat hatékony segítséget. Jelen tanulmány fő kérdése, hogy más-más módszerekkel számolva mennyire konzisztensek az optimális klaszterszámra vonatkozó eredmények különböző feltételek esetén (például eltérő heterogenitású és dimenzionalitású adatbázisoknál).

A klaszterelemzésnek általában nincsenek matematikai előfeltevései (például a változók eloszlására vonatkozóan), ezért az eredmények „jóságának" mérésére sincs egyetlen optimális indikátor. Az eredmények értékelése során általában a kohézió (a klaszteren belüli közelség) és a szeparáció (a klaszterek közötti távolság) szempontjait szokás figyelembe venni (*Mur et al.* [2016]). Az általánosan alkalmazható „jósági" mutatószám kialakítását az is nehezíti, hogy nagyon sokféle klaszterezési módszer van; a gyakran alkalmazott nemhierarchikus és hierarchikus (divizív vagy agglomeratív) klaszterelemzésen belül is számos algoritmus választható az elemzéshez (*Liang et al.* [2012]). Ezenkívül a szakirodalom is folyamatosan bővül, az új módszerek fejlesztésekor az egyik fontos szempont például a számítási gyorsaság (*Zhou–Xu* [2018], *Kolesnikov–Trichina–Kauranne* [2015], *Tîrnăucă et al.* [2018]).

A klaszterezési algoritmusok eredményeinek értékelésekor a klasztervaliditás elnevezést szokás alkalmazni (*Charrad et al.* [2014]), amely a szakirodalom (például *Charrad et al.* [2014]) alapján 3 szempontból vizsgálható: a *külső kritériumok* esetén a klaszterelemzés eredményét egy külsőleg adott csoportosításhoz lehet hasonlítani, a *belső kritériumoknál* a klaszterezési eljárás során létrejött illeszkedési adatok elemezhetők, a *relatív kritériumok* tekintetében pedig egy adott klaszterezési eredményt egyéb (azonos algoritmussal, de más paraméterekkel számolt) klaszterezési eredményekkel lehet összevetni. A relatív kritériumos klasztervaliditási vizsgálatok közé tartozik a klaszterszámváltozás eredményekre gyakorolt hatásának felmérése is.

A klaszterszám kiválasztása a klaszterelemzés egyik központi és nem kizárólag matematikailag megoldható feladata. A szakirodalom (például *Simon* [2006]) alapján a klaszterszámra vonatkozóan lehet előzetes vagy elméleten alapuló feltevés, illetve egyéb megfontolások (például a hierarchikus klaszterelemzésnél a dendrogram) segíthetik a klaszterszámválasztást, például a relatív kritériumos klasztervaliditási in-

dexek (*Charrad et al.* [2014]). A sokféle, elméletileg kiszámolható klasztervaliditási index között nincs olyan, amelyik minden szempontból „jobb" lenne, mint a többi, a gyakorlati elemzésekben is sokféle mutatószám fordul elő. A tanulmány további fejezeteiben a „*klaszterkönyök*-" (például *Deb–Lee* [2018], *Paternina et al.* [2018], *Estiri–Omran–Murphy* [2018], *Natale–Carvalho–Paulrud* [2015], *Lino et al.* [2019]) és a *sziluettmódszerrel* (például *Zhou–Xu* [2018], *Lord et al.* [2017], *Bhargavi–Gowda* [2015], *Fujita–Takahashi–Patriota* [2014]) foglalkozunk, amelyek viszonylag gyakran szerepelnek a klasztervaliditási indexeket összehasonlító vizsgálatokban, ugyanakkor hangsúlyozni kell, hogy nagyon sok egyéb mutatószám is elemezhető lenne. A könyök- és sziluettmódszer fontosságát jelzi, hogy néhány elemzés kiemelten foglalkozik ezek eredményeivel (például *Yahyaoui–Own* [2018], *Masud et al.* [2018]). Vizsgálatunkban e két módszer összehasonlítása – egyetlen mutatószám elemzése helyett – amiatt lehet előnyös, mert a klasztervaliditási indexeken alapuló megoldásoknál nem garantálható, hogy a különböző klaszterezési algoritmusok és adatstruktúrák esetében az eredmény konzisztens lesz (*Masud et al.* [2018]). A sokféle mutatószám együttes áttekintése helyett e két, széleskörűen ismert és alkalmazott módszer összehasonlítása, elsősorban terjedelmi okok miatt lehet előnyös.

A tanulmányban szimulációval előállított adatbázisok alapján számított eredményeket hasonlítunk össze a könyök- és sziluettmódszer segítségével. Az optimális klaszterszám a könyökmódszernél az ábrázolt értékek nagymértékű meredekségváltozása (például *Kovács* [2014] 80.old.), a sziluettmódszernél pedig az átlagos sziluettérték maximuma alapján azonosítható (*Rousseeuw* [1987]). Mivel a klaszterelemzési algoritmusok konkrét eredményei általában nem vezethetők le matematikai módszerekkel, ezért a szimulációval előállított adatbázisok elemzése gyakran szerepel a szakirodalomban (például *Zhou-Xu* [2018], *Estiri–Omran–Murphy* [2018], *Lord et al.* [2017], *Bhargavi–Gowda* [2015], *Kothari–Pitts* [1999], *Fang–Wang* [2012], *Fujita–Takahashi–Patriota* [2014], *Hardy* [1996], *Zhang et al.* [2017], *Yu–Liu–Wang* [2014]).

Elemzésünk hasonlít a korábbi szakirodalomhoz a szimulációs módszertan alkalmazása miatt, valamint a homogén és heterogén adathalmazok összehasonlítása sem ritka (például *Fujita–Takahashi–Patriota* [2014], *Hardy* [1996]). A saját új klaszterező módszereket bemutató tanulmányokhoz képest az egyik különbség, hogy mindössze két, a korábbi szakirodalomból ismert könyök- és a sziluettmódszer tulajdonságait hasonlítjuk össze egy kiválasztott hierarchikus klaszterelemzési eljárással (amely azonban viszonylag hasonló több más klaszterezési algoritmushoz). A nem új módszereket bemutató, hanem elsősorban a meglévők tulajdonságainak leírására fókuszáló szakirodalomhoz (például *Charrad et al.* [2014], *Kovács–Legány–Babos* [2006]) képest különbség, hogy a dimenzionalitás hatása az elemzésünk egyik fő témája. A tanulmányban az optimális klaszterszám lehetséges definícióit az 1. fejezet foglalja össze, a 2. fejezetben találhatók a saját számítások eredményei, a következtetéseket a 3. fejezet összegzi.

1. Az optimális klaszterszám definíciója

A klaszterszám-meghatározás legfontosabb jellemzője, hogy nincs olyan mutatószám, amely minden esetben alkalmazható lenne. A klaszterelemzés során a kutató hozhat döntést, hogy melyik klaszterszámhoz tartozik a legjobbnak tekinthető megoldás (például olyan kutatás, amelyben van előzetes vagy elméleten alapuló feltevés az elérendő klaszterszámról) (*Simon* [2006]). Amikor a klaszterek száma nem előre adott az elemzésben, sok módszer segítheti a klaszterszám kiválasztást. Kisebb adatbázisok hierarchikus klaszterelemzésénél alkalmazható lehet például a dendrogram (*Simon* [2006], *Dobos–Michalkó–Nováky* [2017], *Rencher–Christensen* [2012] 413. old.), és számos egyéb mutatószám is rendelkezésre áll. A gyakorlatban a klasztervaliditás kritériumai közül gyakran választják a relatív mutatószámokat, amikor valamely klasztervaliditási index értékei összehasonlíthatók, és az optimális érték (illetve az index értékhez kapcsolódó egyéb megfontolás és számítás eredménye) utalhat a „legmegfelelőbb" klaszterszámra. Néhány kiválasztott klasztervaliditási index jellemzőit *Charrad et al.* [2014] alapján az 1. táblázat foglalja össze.

Az 1. táblázatban szereplő indexek esetében *Charrad et al.* [2014] tartalmazza a képleteket és az indexeket részletesen leíró tanulmányok bemutatását, a következőkben ezeket az indexeket az optimális klaszterszám meghatározásának módjára fókuszálva hasonlítjuk össze.

1. táblázat

Klasztervaliditási indexek és az optimális klaszterszám meghatározása

Klasztervaliditási index	A megfelelő klaszterszám jellemzője
Calinski–Harabasz-index	Maximális érték
Duda–Hart-index	A legkisebb klaszterszám, amelynél az index nagyobb vagy egyenlő, mint egy meghatározott érték
Pseudo t^2-index	A legkisebb klaszterszám, amelynél az index kisebb vagy egyenlő, mint egy meghatározott érték
C-index	Minimum érték
Gamma-index	Maximum érték
Beale-index	*F* eloszláshoz hasonlítva határozható meg
CCC-index	Maximális érték
Pontbiszeriális korreláció	Maximális érték
Gplus-index	Minimum érték
Davies–Bouldin-index	Minimum érték

(A táblázat folytatása a következő oldalon.)

(Folytatás.)

Klasztervaliditási index	A megfelelő klaszterszám jellemzője
Frey–Van Groenewoud-index	Egy számolt hányados értékének 1 alá csökkenése alapján számolható
Hartigan-index	Hierarchiaszintek közötti maximális távolság alapján számolható
Tau-index	Maximális érték
Ratkowsky–Lance-index	Maximális érték
Scott–Symons-index	Hierarchiaszintek közötti maximális távolság alapján számolható
Marriot-index	Egymást követő szintek közötti maximális távolság alapján számolható
Ball–Hall-index	Szintek közötti maximális távolság alapján számolható
A klasztereken belüli „pooled” kovarianciamátrix nyoma	Szintek közötti maximális távolság alapján számolható
A klasztereken belüli szóródási (dispersion) mátrix nyoma	A második különbségértékek maximuma
Friedman–Rubin- [1967] cikkből az egyik index	Értékek közötti maximális távolság alapján számolható
Friedman–Rubin- [1967] cikkből a másik index	Szintek közötti második távolságok minimuma alapján számolható
McClain–Rao-index	Minimum érték
Krzanowski–Lai-index	Maximális érték
Sziluettindex	Maximális érték
Gap-index	A legkisebb klaszterszám, amelynél a gap-értékek közötti különbségre adott összefüggés teljesül
D-index	A klaszterezési „haszon” minimalizálásával számolható
Dunn-index	Maximális érték
Hubert–Arabie-statisztika	Második különbségértékek alapján számolható
SD-index	Minimum érték
SDbw-index	Minimum érték

Forrás: Saját szerkesztés *Charrad et al.* [2014] alapján.

A klasztervaliditási indexek között fontos különbség lehet, hogy néhányuk csak hierarchikus klaszterelemzésben alkalmazható, például az 1. táblázatban ez jellemző a pseudo t^2- és a Frey–Van Groenewoud-indexre. A klasztervaliditási indexek alkalmazásakor az optimális klaszterszám esetében gyakran a klaszterszám függvényében változó valamely érték minimumát vagy maximumát keressük, erre az 1. táblázatban is számos példa található. A klasztervaliditási indexek összefügghetnek statisztikai hipotézisvizsgálattal is, a Beale-index esetében például az egyetlen klaszter meglétére vonatkozó (statisztikai) nullhipotézis elfogadása vagy elvetése *F*-eloszlású teszt-

statisztikával vizsgálható, és az elemzés során különböző klaszterszámokat lehet figyelembe venni (*Charrad et al.* [2014]). Az 1. táblázatban nem említett mutatószámok köre is széles, a klaszterezés megfelelőségének értékelésénél alkalmazható például a *Vargha–Borbély* [2017] által említett módosított Xie–Beni-index (amely azt mutatja meg, hogy mennyivel kisebb az átlagos távolság a saját klaszter középpontjától, mint az egymáshoz legközelebbi két klaszter távolsága).

A klaszterelemzés „jóságának" két fontos jellemzője a kohézió, valamint a szeparáció (*Mur et al.* [2016]), és mindkettő sokféleképpen mérhető. Az egyik lehetséges megközelítésben a mérésnél a klasztereken belüli és klaszterek közötti „szóródási" (például kovariancia-) mátrixok alkalmazhatók. Az 1. táblázatban szereplő mutatószámok közül több is kapcsolódik ezekhez a mátrixokhoz, például a Calinski–Harabasz-index számításában mindkét szóródási mátrix szerepet játszik, de vannak olyan mutatószámok (például a Duda–Hart-index) amelyeknek a számításához csak a klaszteren belüli szóródási mátrix szükséges. Néhány mutató nem közvetlenül ezekből a mátrixokból számítható, érdekes például a pontbiszeriális korreláció, amely a kezdeti távolságmátrix és egy ennek megfelelő, 0 és 1 értékeket tartalmazó mátrix elemeiből állítható elő (a mátrixban az ugyanolyan klasztertagságra 1 értékkel utalva) (*Charrad et al.* [2014]).

A klasztervaliditási indexek, illetve az optimális klaszterszám meghatározásához alkalmazható mutatószámok köre nagyon széles, az 1. táblázatban szereplő indexeken kívül is sokféle mutatószám szerepel a gyakorlati számításokban. *Kadlecsik* [2013] a kétfázisú klaszterezés során, a változók normális eloszlását feltételező loglikelihood távolságmérték és Schwarz-féle bayesi információs kritérium segítségével a modell által „javasolt" klaszterszámot alkalmazza elemzésében. *Yu–Liu–Wang* [2014] egy új klasztervaliditást értékelő függvényt és egy olyan hierarchikus klaszterezési algoritmust mutatnak be, amely automatikusan megáll a tökéletes klaszterszámnál. *Chakraborty–Das* [2018] olyan algoritmust írnak le, amely szimultán módon foglalkozik a klaszterszám megtalálásával és a különböző változókhoz súlyok rendelésével. *Shen et al.* [2005] dinamikus validitási indexet mutatnak be, *Fang–Wang* [2012] pedig bootstrap módszerrel határozzák meg a klaszterezési instabilitást, és a becsült klaszterezési instabilitást minimalizáló klaszterszámot választják az elemzésben. Különböző klaszterminőségi mutatószámok összehasonlításához kapcsolódóan *Vargha–Bergman–Takács* [2016] arra is felhívják a figyelmet, hogy a klaszterezés eredményeinek megfelelőségénél érdemes azt is vizsgálni, hogy „igazi" klaszterstruktúráról van-e szó. *Vargha–Bergman–Takács* [2016] ezen vizsgálathoz a MORI- (measure of relative improvement – relatív megfelelőség mértéke) mutatószámot javasolják, amely a valódi és szimulációkkal (független változókkal) előállított adatokkal számítható klaszterminőségi mutatószámok összevetésével határozható meg.

Az optimális klaszterszám kiválasztása során a különböző indexek számítása néha elegendő, más esetekben az egyéb megfontolások kiegészítő eredményt jelente-

nek. *Kosztyán–Telcs–Török* [2015] például inhomogenitási index minimalizálásával határozták meg a klaszterek számát, míg *Nezdei–Alpek* [2018] a klaszterszám meghatározásakor a Calinski–Harabasz- és a Duda–Hart-index értékeit is kiszámították, valamint az értelmezhetőséget és a klaszterek elemszámait is figyelembe vették.

A klaszterelemzés folyamatosan fejlődő szakirodalmában a klaszterezés „jóságának" szempontjairól is szó van. *Lee–Olafsson* [2013] a klaszterminőség egy új jellemzőjeként mutatja be a diszkonnektivitást, amely elemzésükben kiegészíti a korábbi szakirodalomban elterjedt kompaktsági szempontokat a klaszterszám meghatározása során. *Lee–Olafsson* [2013] megállapítása szerint több kompaktsági mutatószám (például a „gap") esetében a klaszterszám meghatározás egy kompaktsági ábrán könyökpont keresésével függ össze, olyan módon, hogy a kompaktsági mutatószám értéke monoton csökken a klaszterszám növekedésekor úgy, hogy bizonyos klaszterszám felett már „laposabb" az ábrázolt függvény. Elemzésükben a kompaktsági és diszkonnektivitási szempontok együttes figyelembe vételével *Lee–Olafsson* [2013] is könyökpont keresésével foglalkoznak.

Az 1. táblázatban felsorolt indexek között a „klaszterkönyök-mutatószám" ilyen néven nem szerepel, bár ez a klaszterszám meghatározási módszer is nagyon elterjedt. *Zhang et al.* [2017] megállapítják, hogy a megfelelő klaszterszám keresésekor néhány módszer esetén valamely értékelő függvényt szokás készíteni, amelynél a vízszintes tengelyen a klaszterszám, a függőleges tengelyen pedig az adott függvény értékei szerepelnek, és ezen az ábrán egy „térdet", illetve „könyököt" szokás keresni. *Zhang et al.* [2017] elgondolásuk hasonló: az értékelő függvény a klasztereken belüli variancia (amely monoton csökkenő), a helyes klaszterszámnak pedig a függvény „térdénél" szereplő értéket tekintik. A könyök csökkenő függvényen keresése a klaszterszám meghatározásakor elterjedt a szakirodalomban (például *Lino et al.* [2019], *Deb–Lee* [2018], *Paternina et al.* [2018]). *Lino et al.* [2019] úgy vélik, a könyökkritérium amiatt alkalmazható, mert a klaszterszám függvényében a megmagyarázott varianciát elemzi. A könyökkritérium tehát nem kötődik szorosan egy adott értékelő függvényhez, inkább gyakran egy olyan szemlélet alkalmazását jelenti, amely esetén a klaszterszám akkor megfelelő, ha már eléggé nagy a magyarázott variancia. Azonban ez a meghatározás sem teljesen átfogó, mivel a szakirodalom egy részében (például *Sajtos–Mitev* [2007] 308. oldal) a könyökkritérium alapján úgy határozható meg a klaszterszám, hogy a hierarchikus klaszterelemzésnél a könyök keresése azon az ábrán történik, amelyen a vízszintes tengelyen az összevonási lépések sorszáma, a függőleges tengelyen pedig az összevonásokhoz tartozó együtthatók értékei szerepelnek. Az 1. táblázatban szereplő mutatószámok közül a gap-index kapcsolatban van az elterjedt „klaszterkönyökmódszerrel", *Estiri–Omran–Murphy* [2018] szerint a gap-módszer a könyökmódszer szerinti összehasonlításokat standardizálja.

A könyökmódszert *Estiri–Omran–Murphy* [2018] úgy definiálják, hogy ebben az esetben a teljes klaszteren belüli négyzetösszeg értékét számítják (és ábrázolják)

különböző klaszterszámokra, és az optimális klaszterszám az, amelyiknél ezen az ábrán egy „térd” található. *Yahyaoui–Own* [2018] ehhez hasonlóan írja le a könyökmódszert, azt is kiemelve, hogy a teljes klaszteren belüli négyzetösszeg az összes objektumnak a klasztercentroidjától való távolságai négyzetének összegeként számítható. A szakirodalomban elterjedt könyökmódszer-definíciókban (például *Estiri–Omran–Murphy* [2018], *Yahyaoui–Own* [2018], *Masud et al.* [2018]) a klaszterkönyökábra gyakran csökkenő, azonban van másfajta megközelítés is. *Kovács* ([2014] 62. old.) leírása alapján a klaszterkönyök ábrázolásához először két klaszteres esetben szórásfelbontó (ANOVA [analysis of variance – varianciaanalízis]) táblázatban ellenőrizzük a klaszterelemzésben szereplő változók megkülönböztető erejét, és ha a változókra vonatkozó *F*-statisztika értékei megfelelően alacsonyak, akkor kettőnél több klaszter esetében is összehasonlítjuk az eredményeket. A klaszterkönyök ábrázolásakor a klaszterazonosítóknál és a klaszterelemzésben szereplő változóknál (figyelembe véve, hogy a változók gyakran standardizáltak) ANOVA segítségével összegezzük a változókra számolt külső eltérések négyzetösszegét, valamint a teljes eltérések négyzetösszegét, és e két összeg hányadosait ábrázoljuk a klaszterszám függvényében, a grafikonon pedig egy „könyökértéket” keresünk. A *Kovács* ([2014] 62. old.) által leírtak alapján készíthető klaszterkönyökábra a klaszterszám függvényében monoton növekvő, de a klaszterszám-kiválasztás e módszer esetében hasonló szemléletű, mint a monoton csökkenő könyökábránál (például *Estiri–Omran–Murphy* [2018], *Yahyaoui–Own* [2018]). A továbbiakban a könyökmódszernél a *Kovács* ([2014] 62. old.) által leírtak alapján készült számításokat alkalmazzuk.

A klaszterkönyök meghatározása kapcsán (például *Estiri–Omran–Murphy* [2018]) két fontos probléma adódhat: egyrészt az eljárás „számításintenzív” (ami miatt nagy adatbázisoknál viszonylag lassan számíthatók az eredmények), másrészt a könyökpont megtalálása olykor nehézségekbe ütközik, hiszen az ábrán nem mindig rajzolódik ki a könyök alakzat (*Masud et al.* [2018]).

A szakirodalomban a könyökmódszeren kívül sokféle mutatószám terjedt el a klaszterszám kiválasztásához kapcsolódóan, ezek közül a továbbiakban a (*Rousseeuw* [1987]) által leírt sziluettmódszerrel foglalkozunk. A sziluettérték minden elemre („megfigyelésre”, „objektumra”) számítható (*Rousseeuw* [1987]):

$$s_i = \frac{b_i - a_i}{\max\left(a_i,\ b_i\right)},$$

ahol az a_i érték az i-dik eset („megfigyelés”, „objektum”) saját klaszteren belüli más esetektől való átlagos távolságát mutatja, a b_i érték pedig úgy számítható, hogy először az adott elem más klaszterekbeli elemektől vett átlagos távolságát határozzuk

meg minden más klaszterre külön, aztán ezen értékek közül a minimális b_i értéket vesszük. *Rousseeuw* [1987] a sziluettmódszer definiálásakor hangsúlyozza, hogy a távolság értékek arány mérési szinten mértek (ratio scale), például az euklideszi távolság arány mérési szintűnek tekinthető. Az optimális klaszterszám az az érték, amely esetében az átlagos sziluettérték maximális (*Rousseeuw* [1987]), és ha a sziluettérték 1-hez közeli, akkor a klaszterezés eredménye meglehetősen jónak tekinthető.

A sziluett képletéből jól látható, hogy az adatbázisban minden esetre külön számolt értékről van szó, és a klaszterek „közös" tulajdonságai nincsenek figyelembe véve a számítások során. *Zhou–Xu* [2018] ezt a módszer hiányosságának tekintik, és írásukban egy olyan klasztervaliditási indexet mutatnak be, amely néhány esetben a sziluettmódszer gyorsabb alternatívájának tekinthető.

A továbbiakban a könyök- és a sziluettmódszert alkalmazzuk. Az empirikus számítások során elsősorban annak a kérdésnek a megválaszolására törekszünk, hogy a két módszer eredményei különböző dimenzionalitású adatbázisokban mennyire tekinthetők konzisztensnek.

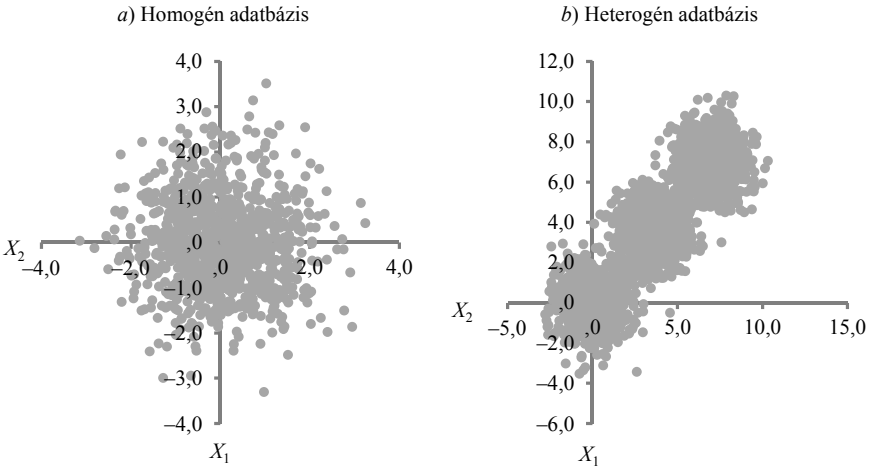
2. Szimulációs adatokkal számolt eredmények

Az optimális klaszterszámra vonatkozó eredmények általában nem vezethetők le matematikai képletekkel, ezért a következőkben a szimulációval előállított adatbázisokat elemezzük, amelyre gyakori példát találunk a szakirodalomban (például *Zhou-Xu* [2018], *Estiri–Omran–Murphy* [2018], *Lord et al.* [2017], *Bhargavi–Gowda* [2015], *Kothari–Pitts* [1999], *Fang–Wang* [2012], *Fujita–Takahashi–Patriota* [2014], *Hardy* [1996], *Zhang et al.* [2017], *Yu–Liu–Wang* [2014]). A tanulmányban viszonylag nagyméretű adatbázisok jellemzőivel foglalkozunk: 1000 eleme van a homogén és 3000 a heterogén (3 csoportot tartalmazó) adatbázisnak. E méretek miatt a számítások időigénye is viszonylag nagy. A klaszterek optimális számának kiválasztása során figyelembe vesszük a szakirodalomban említett egyik hüvelykujjszabályt a klaszterek maximális elemszámáról: n elem esetében a maximális klaszterszám $\sqrt{n/2}$ (*Kovács* [2014] 62. old.). Ennek alapján az 1000 elemnél $\sqrt{1000/2} = 22{,}36$ lehetne a maximális klaszterszám, ezért a klaszterszámokat 2 és 23 között vesszük figyelembe a számítások során. A heterogén adatbázisnál ennél nagyobb klaszterszám adódna a hüvelykujjszabály alapján, de a homogén adatbázissal való összehasonlítás miatt a heterogén adatbázisnál is 23 a legnagyobb

klaszterszám. Ez a maximális klaszterszámválasztás a számítások időigényessége szempontjából is előnyös, ezenkívül az elemzésben a heterogén adatbázisnál 3 a csoportok száma, ezért a 23-nál nagyobb klaszterszámoknál az eredményeknek nincs igazán jelentős információtartalma olyan szempontból, hogy megfigyelhető-e a 3 mint optimális klaszterszám.

A szakirodalom alapján (például *Hajdu* [2003] 122. old.) néhány esetben az outlierek kizárása lenne javasolható, a tanulmányban ezzel a feladattal nem foglalkozunk. A szimulációval előállított adatbázisoknál nem minden esetben nagy az elemszám, vizsgálatunkban ez a feltételezés elsősorban amiatt szerepel, hogy az eredmények lehetőleg ne néhány egyedi adatra vonatkozzanak, illetve csökkenjen a kiugró értékek befolyásoló hatása. E célok teljesülését segítheti elő az is, hogy a szimulációval előállított adatokat p dimenziós „gömbök" alkotják (esetünkben p 2 és 8 közötti, a „gömb" kifejezés pedig arra utal, hogy a változók elméletileg korrelálatlannak tekinthetők). A „gömbstruktúra" azzal függ össze, hogy az adatok szimulációja során a normális eloszlás nagy szerepet kap. A $p = 2$ esetében előállított adatbázisokat az 1. ábra szemlélteti. Érdemes megemlíteni, hogy a gyakorlati adatbázisokban általában nem tapasztalható a változók közötti korrelálatlanság, ez a feltételezés az elemzésben mindössze a minél áttekinthetőbb eredmények számítása érdekében szerepel.

1. ábra. Szimulációval előállított adatok p = 2 esetén



A normális eloszlás feltételezése a szakirodalomban például *Fujita–Takahashi–Patriota* [2014] írásában szerepel, ahol mindegyik szimulációval előállított adathalmaz 100 olyan megfigyelésből áll, amelyeket kétváltozós normális eloszlású sokaságból származónak lehet tekinteni (a feltételezések szerint a kovarianciamátrix az egységmátrix). Jelen tanulmányban a változók korrelálatlanságának feltevése hason-

lít a *Fujita–Takahashi–Patriota* [2014] által alkalmazotthoz, a normális eloszlásra vonatkozóan ugyanakkor a vizsgálatunkban mindössze azt feltételezzük, hogy a p dimenziós gömböknél a változók peremeloszlása normális, a többdimenziós normális eloszlás tesztelésével nem foglalkozunk.

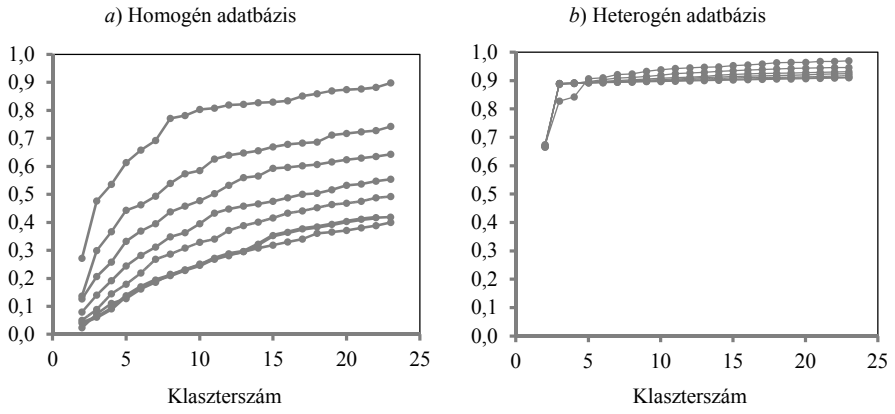
Az elemzésben minden dimenziószámhoz (2 és 8 dimenzió között) egy homogén és egy heterogén adatbázist állítottunk elő. A homogén adatbázisnál minden dimenzió (szimulációval előállított változó) esetében a normális eloszlás várható értéke 0, az elméleti szórása 1 volt (a változók elméleti függetlenséget feltételezve). A heterogén adatbázis a feltételezések szerint 3 csoportból (p dimenziós „gömb") állt, és az adatok előállítására egyenként – a homogén adatbázishoz hasonló módon, de részben más paraméterekkel – került sor: a normális eloszláshoz tartozó elméleti szórás értéke egységesen 1, az elméleti várható érték pedig 0, 3,5 és 7 volt mindegyik dimenzió (szimulációval előállított változó) esetében. E feltevések alapján a heterogén adatbázisban nem volt egymástól tökéletesen elkülönülő a 3 csoport, bár meglétük egyértelmű az 1. ábra alapján. A heterogén adatbázis hasonlít *Zhang et al.* [2017] tanulmányukban szereplő egyik adatbázisra (bár *Zhang et al.* [2017] elemzésében a változók szórása a 2 dimenziós adatbázisban nem egyenlő a 3 csoportban). A szakirodalomban az is előfordul, hogy a szimulációval előállított adatbázisok között homogén és heterogén adatbázisok is vannak, *Fujita–Takahashi–Patriota* [2014] írásában például az egyik (szimulációval előállított) adatbázis egyklaszteres (vagyis homogénnek tekinthető), valamint *Hardy* [1996] tanulmányában is szerepel homogén és heterogén adathalmaz egyaránt. A homogén adathalmazok jellemzőinek vizsgálata többek között amiatt lehet érdekes, mert néhány módszer akkor is jelezheti csoportok jelenlétét, amikor ténylegesen nincsenek az adatbázisban (*Hardy* [1996]).

Elemzésünk hierarchikus klaszterelemzéssel készült. Mivel a szimulációval előállított változók arány mérési szintűnek tekinthetők, ezért az elemek közötti távolság mérésére az euklideszi távolságot alkalmaztuk, az összevonást pedig a legtávolabbi szomszéd módszerrel hajtottuk végre. *Simon* [2006] alapján a legtávolabbi szomszéd módszer a tértágító eljárások közé tartozik. *Kovács* ([2014] 56. old.) szerint a tértágító hatás arra utal, hogy a hierarchikus klaszterezési algoritmus alkalmazásakor valamely lépésben inkább új klaszterek képződnek, és nem a meglevőkhöz kapcsolódnak újabb elemek. *Simon* [2006] megállapítása, hogy a legtávolabbi szomszéd módszernél a klaszterek összevonása a legtávolabbi pontok alapján történik, és az a két klaszter vonható össze, amelyeknek a legkisebb a távolsága. Ezzel a módszerrel nagyjából hasonló nagyságú csoportok képezhetők. A választott klaszterezési algoritmus (a legtávolabbi szomszéd módszer és az euklideszi távolság alkalmazása) jelen elemzésben megfelelőnek tekinthető, hiszen az nem feltételezhető, hogy sok egyelemű klaszter képződne, hanem hasonló méretű csoportok várhatók (hasonlóan a heterogén adatbázishoz, amelyben a tényleges csoportok egyenlő méretűek). A szimulációval létrehozott adatbázisainkban a változók standardizált változata szerepelt, a klaszter-

elemzési és sziluettértékekkel kapcsolatos számítások az SPSS programmal készültek (IBM Corporation Released 2017. IBM SPSS Statistics for Windows, Version 25.0. Armonk, New York). A Függelékben található leírás (2 dimenziós példákon keresztül) az elemzés során alkalmazott adatbázisok struktúráját és a különböző klaszterszám esetén számolható eredményeket illusztrálja. Ezek az adatbázisok (mivel ezekben elméletileg mindegyik szimulációval előállított érték változhat) nem pontosan ugyanazok, amelyek a könyökábra és a sziluettértékek számításánál szerepelnek, ugyanakkor a nagy elemszám miatt feltételezhető a hasonlóságuk.

Elsősorban a dimenzionalitás hatásának elemzésével törekszünk hozzájárulni a korábbi szakirodalomhoz (például *Vargha–Bergman–Takács* [2016] mutattak rá, hogy a változók száma jelentősen befolyásolhatja a klaszterelemzés eredményeit). Az elemzésekben a dimenziószám (vagyis az elemzésekben szereplő változók száma) 2 és 8 között változik, és a klaszterkönyökértéket, valamint a sziluettértékeket is különböző dimenziószám esetében hasonlítjuk össze a homogén és heterogén adatbázisban. A klaszterkönyök-számítással kapcsolatos eredményeket a 2. ábra foglalja össze, az átlagos sziluettértékek (a klaszterszám függvényében) pedig a 3. ábrán találhatók.

2. ábra. A klaszterkönyök-számítás eredményei



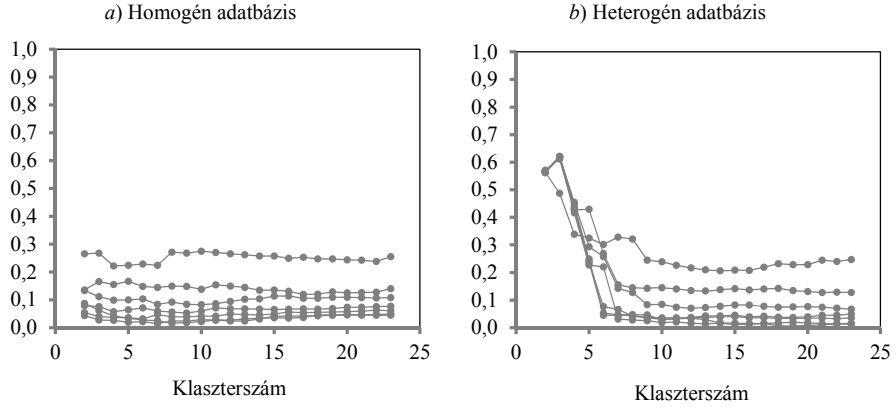
A klaszterkönyök-számítások a *Kovács* ([2014] 62. old.) által leírtak alapján készültek, vagyis az ábrán az értékek monoton növekedők, és az 1-hez közeli értékek utalnak a klaszterezés jó eredményére. A sziluettszámításoknál is az 1-hez közeli érték utal arra, hogy az adott klaszterszám mellett jónak tekinthető a klaszterezés. Ennél pontosabban általában nem lehet meghatározni a jóság feltételét, a szakirodalom mindössze néhány mutatószám esetében utal konkrétabb értékre. *Simon* [2006] a könyökpont-kritériumnál arra utal, hogy a klaszterszám függvényében ábrázolni lehet, mikor csökken jelentősen a belső varianciaösszeg, és ha nincs jelentős ugrás

(vagyis nincs könyök), akkor egy hüvelykujjszabály alapján az 50-50 százalékos belső és külső varianciaarányt érdemes figyelembe venni. *Vargha* [2016] megállapítása alapján sikeres osztályozás esetén a mutatószám (a megmagyarázott varianciaarány, amit úgy lehet definiálni, hogy összevetjük az adott osztályozás összheterogenitását és a lehető legnagyobb összeheterogenitást) eléri a 0,65 értéket.

A tanulmányban a klaszterkönyökkel kapcsolatos számításokban a 2.a) ábrán a legnagyobb értékek a legkisebb dimenziószámhoz ($p = 2$ esethez) tartoznak, és nagyobb dimenziószámnál általában kisebbek a klaszterkönyökértékek. A 2.b) ábra azt mutatja, hogy dimenziószámtól függetlenül nagyjából ugyanolyan a klaszterkönyökérték, ugyanakkor az is látható, hogy nagyobb dimenziószám esetén pontosabban kirajzolódnak a könyökpontok, mint 2 dimenziót (változót) véve, ahol ezek azonosítása nem annyira egyértelmű. Ez az eredmény szemlélteti az *Estiri–Omran–Murphy* [2018] által említett problémát, mivel kizárólag a klaszterkönyökábra értelmezése alapján (a tényleges csoportszámot figyelmen kívül hagyva) akár 5 is lehetne a választott optimális klaszterszám, miközben a tényleges csoportszám az adatbázisban 3.

A homogén és heterogén adatbázis különbségei a klaszterkönyökábrán nagyon jól felismerhetők, azonban ebben az esetben is figyelemre méltó, hogy alacsony dimenziószámnál kizárólag az ábra alapján (amikor az elemzésben szereplő változók száma 2), optimális megoldásként jó választásnak tűnhet a 8 klaszter, ennyi klaszter esetében az ábrázolt értékek is viszonylag magasak. Ez az eredmény a *Hardy* [1996] által leírt problémát szemlélteti, vagyis hogy néhány esetben a klaszterszám-választáshoz alkalmazott módszer akkor is jelezheti csoportok jelenlétét, amikor azok ténylegesen nincsenek az adatbázisban (vagyis homogén az adatbázis).

3. ábra. Átlagos sziluettértékek a klaszterszám függvényében



Az átlagos sziluettértékek esetében az eredmények hasonlók a klaszterkönyök-módszerrel meghatározottakhoz, ezért a tanulmány egyik fő kérdésére az a válasz

adódik, hogy az elemzésben szereplő adatbázisoknál a klaszterkönyök- és a sziluettmódszer eredményei konzisztensnek tekinthetők. A sziluettmódszernél is teljesül az (ami a klaszterkönyökmódszer esetében megfigyelhető volt), hogy a homogén és a heterogén adatbázisoknál számolt eredmények nagymértékben különböznek, és az eredményeket a dimenzionalitás (az elemzésben szereplő változók száma) is befolyásolja, bár ennek hatása nem tekinthető nagymértékűnek. A sziluettmódszer esetében a legkisebb dimenziószámnál (amikor az elemzésben szereplő változók száma 2) a módszer által optimális klaszterszámként azonosítható érték (az elemzésben szereplő adatbázisban 2) nem egyezik meg a tényleges csoportok számával a heterogén adatbázist tekintve (ebben az adatbázisban 3). A homogén adatbázis esetében az átlagos sziluettértékeknél nem található olyan kiemelkedő maximális érték, ami az ábrán egyértelműen optimális klaszterszámként lenne azonosítható, és az is megfigyelhető, hogy a homogén adatbázisnál a legkisebb dimenziószámhoz tartoznak a legnagyobb értékek.

Egészében véve az eredmények arra utalnak, hogy a klaszterkönyökábra és az átlagos sziluettértékeket a klaszterszám függvényében ábrázoló ábra nagy segítséget jelenthet az adatbázisok heterogenitásának, valamint homogenitásának megítélésében, illetve, hogy nagyobb dimenziószámnál (amikor az elemzésben több változó szerepel) kevésbé valószínű, hogy a heterogén adatbázisoknál a tényleges csoportszámtól eltér a módszer eredményeként választott optimális klaszterszám.

3. Következtetések

A klaszterelemzés során az egyik központi kérdés, hogy mennyi klasztert lehet megkülönböztetni az adatbázisban. A klaszterszám lehet szakmai szempontok miatt külsőleg adott, de gyakran különböző klasztervaliditási indexek segítenek az optimális klaszterszám kiválasztásában. Tanulmányunkban a sok lehetőség közül a klaszterkönyök- és sziluettmódszereket hasonlítottuk össze. Ezek meglehetősen gyakoriak a klaszterelemzés dinamikusan fejlődő szakirodalmában, amihez mi a dimenzionalitás hatásának kutatásával törekedtünk hozzájárulni. Az elemzésünk során alkalmazott adatok is hasonlítanak korábbi tanulmányokban szereplő adatbázisokhoz. Egyik fontos eredményünk, hogy a vizsgálatunkban szereplő adatok esetében mindkét módszerrel egyaránt jól felismerhető, hogy homogén vagy heterogén adatbázisról van-e szó. Egy másik, a gyakorlati kutatások számára lényeges eredmény arra utal, hogy az optimális klaszterszám a heterogén adatbázisoknál mindkét módszerrel hasonló értékű (a vizsgált adatbázisokban), és mindössze alacsonyabb dimenziószámnál fordul elő kismértékű eltérés.

A tanulmány eredményeinek értelmezése során természetesen figyelembe kell venni, hogy a klaszterelemzési algoritmusok sajátosságai miatt nem matematikailag levezetett eredményekről van szó, így az alkalmazott adatbázisok tulajdonságai nagymértékben befolyásolják azokat. A témával kapcsolatos további kutatásokban emiatt elsősorban az elemzésben szereplő adatbázisok paramétereinek (méretének, a változók eloszlásának) változtatásával érdemes foglalkozni.

Függelék

Az R program (R CORE TEAM [2018]: *R: A language and environment for statistical computing. R Foundation for Statistical Computing*. Vienna. https://www.r-project.org/) alkalmazásával lehetséges a tanulmányban szereplő szimulált adatokhoz hasonlók előállítása, valamint a legtávolabbi szomszéd és az euklideszi távolság módszerével végzett hierarchikus klaszterelemzés során kialakítható klaszterek elhelyezkedésének szemléltetése.

A homogén adatbázisban kialakítható klaszterekhez tartozó illusztráció:

```
X1=rnorm(1000,mean=0,sd=1)
X2=rnorm(1000,mean=0,sd=1)
data1=data.frame(scale(X1),scale(X2))
clust1=hclust(dist(data1,method="euclidean"),method="complete")
for (i in 1:22) {plot(X1,X2,col=cutree(clust1,k=i+1))}
```

A heterogén adatbázisban kialakítható klaszterekhez tartozó illusztráció:

```
X1a=rnorm(1000,mean=0,sd=1)
X2a=rnorm(1000,mean=0,sd=1)
X1b=rnorm(1000,mean=3.5,sd=1)
X2b=rnorm(1000,mean=3.5,sd=1)
X1c=rnorm(1000,mean=7,sd=1)
X2c=rnorm(1000,mean=7,sd=1)
X1=c(X1a,X1b,X1c)
X2=c(X2a,X2b,X2c)
data1=data.frame(scale(X1),scale(X2))
clust1=hclust(dist(data1,method="euclidean"),method="complete")
for (i in 1:22) {plot(X1,X2,col=cutree(clust1,k=i+1))}
```

Irodalom

ARRIETA PATERNINA, M. R. – ZAMORA-MENDEZ, A. – ORTIZ-BEJAR, J. – CHOW, J. H. – RAMIREZ, J. M. [2018]: Identification of coherent trajectories by modal characteristics and hierarchical agglomerative clustering. *Electric Power Systems Research.* Vol. 158. May. pp. 170–183. http://dx.doi.org/10.1016/j.epsr.2017.12.029

BHARGAVI, M. S. – GOWDA, S. D. [2015]: A novel validity index with dynamic cut-off for determining true clusters. *Pattern Recognition.* Vol. 48. Issue 1. pp. 3673–3687. http://dx.doi.org/10.1016/j.patcog.2015.04.023

CHAKRABORTY, S. – DAS, S. [2018]: Simultaneous variable weighting and determining the number of clusters – A weighted Gaussian means algorithm. *Statistics and Probability Letters.* Vol. 137. June. pp. 148–156. http://dx.doi.org/10.1016/j.spl.2018.01.015

CHARRAD, M. – GHAZZALI, N. – BOITEAU, V. – NIKNAFS, A. [2014]: NbClust: An R package for determining the relevant number of clusters in a data set. *Journal of Statistical Software*. Vol. 61. Issue 6. pp. 1–36. http://dx.doi.org/10.18637/jss.v061.i06

DEB, C. – LEE, S. E. [2018]: Determining key variables influencing energy consumption in office buildings through cluster analysis of pre- and post-retrofit building data. *Energy and Buildings*. Vol. 159. January. pp. 228–245. http://dx.doi.org/10.1016/j.enbuild.2017.11.007

DOBOS I. – MICHALKÓ G. – NOVÁKY E. [2017]: Habilitáltak publikációs adatainak vizsgálata többváltozós statisztikai módszerekkel. *Statisztikai Szemle.* 95. évf. 7. sz. 669–691. old. http://dx.doi.org/10.20311/stat2017.07.hu0669

ESTIRI, H. – OMRAN, A. B. – MURPHY, S. N. [2018]: kluster: An efficient scalable procedure for approximating the number of clusters in unsupervised learning. *Big Data Research.* Vol. 13. September. pp. 38–51. http://dx.doi.org/10.1016/j.bdr.2018.05.003

FANG, Y. – WANG, J. [2012]: Selection of the number of clusters via the bootstrap method. *Computational Statistics and Data Analysis.* Vol. 56. Issue 3. pp. 468–477. http://dx.doi.org/10.1016/j.csda.2011.09.003

FUJITA, A. – TAKAHASHI, D. Y. – PATRIOTA, A. G. [2014]: A non-parametric method to estimate the number of clusters. *Computational Statistics and Data Analysis.* Vol. 73. May. pp. 27–39. http://dx.doi.org/10.1016/j.csda.2013.11.012

HAJDU O. [2003]: Többváltozós statisztikai számítások. In: *Hunyadi L.* (szerk.): *Statisztikai módszerek a társadalmi és gazdasági elemzésekben*. Központi Statisztikai Hivatal. Budapest.

HARDY, A. [1996]: On the number of clusters. *Computational Statistics & Data Analysis*. Vol. 23. Issue 1. pp. 83–96. http://dx.doi.org/10.1016/S0167-9473(96)00022-9

KADLECSIK R. [2013]: A feldolgozóipari vállalkozások statisztikai elemzése jövedelmezőségi és hatékonysági mutatók alapján. *Statisztikai Szemle*. 91. évf. 11. sz. 1072–1091. old.

KOLESNIKOV, A. – TRICHINA, E. – KAURANNE, T. [2015]: Estimating the number of clusters in a numerical data set via quantization error modeling. *Pattern Recognition.* Vol. 48. Issue 3. pp. 941–952. http://dx.doi.org/10.1016/j.patcog.2014.09.017

KOSZTYÁN ZS. T. – TELCS A. – TÖRÖK Á. [2015]: Felsőoktatásba jelentkezők preferenciáinak térbeli és időbeli szerkezete, teljesítményfüggése. *Statisztikai Szemle*. 93. évf. 10. sz. 917–942. old.

KOVÁCS E. [2014]: *Többváltozós adatelemzés*. Typotex. Budapest.

KOVÁCS, F. – LEGÁNY, CS. – BABOS, A. [2006]: *Cluster Validity Measurement Techniques*. Proceedings of the 5th WSEAS International Conference on Artificial Intelligence Knowledge Engineering and Data Bases. 15–17 February. Madrid. https://pdfs.semanticscholar.org/581c/71da74bd3baa06693cc6d0751e7c60f81bb3.pdf

KOTHARI, R. – PITTS, D. [1999]: On finding the number of clusters. *Pattern Recognition Letters.* Vol. 20. Issue 4. pp. 405–416. http://dx.doi.org/10.1016/S0167-8655(99)00008-2

FRIEDMAN, H. P. – RUBIN, J. [1967]: On some invariant criteria for grouping data. *Journal of the American Statistical Association*. Vol. 62. Issue 320. pp. 1159–1178. http://dx.doi.org/10.1080/01621459.1967.10500923

LEE, J.-S. – OLAFSSON, S. [2013]: A meta-learning approach for determining the number of clusters with consideration of nearest neighbors. *Information Sciences.* Vol. 232. May. pp. 208–224. http://dx.doi.org/10.1016/j.ins.2012.12.033

LIANG, J. – ZHAO, X. – LI, D. – CAO, F. – DANG, C. [2012]: Determining the number of clusters using information entropy for mixed data. *Pattern Recognition.* Vol. 45. Issue 6. pp. 2251–2265. http://dx.doi.org/10.1016/j.patcog.2011.12.017

LINO, A. – ROCHA, Á. – MACEDO, L. – SIZO, A. [2019]: Application of clustering-based decision tree approach in SQL query error database. *Future Generation Computer Systems.* Vol. 93. April. pp. 392–406. http://dx.doi.org/10.1016/j.future.2018.10.038

LORD, E. – WILLEMS, M. – LAPOINTE, F-J. – MAKARENKOV, V. [2017]: Using the stability of objects to determine the number of clusters in datasets. *Information Sciences.* Vol. 393. July. pp. 29–46. http://dx.doi.org/10.1016/j.ins.2017.02.010

MASUD, M. A. – HUANG, J. Z. – WEI, C. – WANG, J. – KHAN, I. – ZHONG, M. [2018]: I-nice: a new approach for identifying the number of clusters and initial cluster centres. *Information Sciences.* Vol. 466. October. pp. 129–151. http://dx.doi.org/10.1016/j.ins.2018.07.034

MUR, A. – DORMIDO, R. – DURO, N. – DORMIDO-CANTO, S. – VEGA, J. [2016]: Determination of the optimal number of clusters using a spectral clustering optimization. *Expert Systems with Applications.* Vol. 65. December. pp. 304–314. http://dx.doi.org/10.1016/j.eswa.2016.08.059

NATALE, F. – CARVALHO, N. – PAULRUD, A. [2015]: Defining small-scale fisheries in the EU on the basis of their operational range of activity The Swedish fleet as a case study. *Fisheries Research.* Vol. 164. April. pp. 286–292. http://dx.doi.org/10.1016/j.fishres.2014.12.013

NEZDEI CS. – ALPEK B. L. [2018]: Vásárlói csoportok a Balaton kiemelt üdülőkörzet piachelyeinek példáján. *Tér és Társadalom*. 32. évf. 1. sz. 145–160. old. https://doi.org/10.17649/TET.32.1.2874

RENCHER, A. C. – CHRISTENSEN, W. F. [2012]: *Methods of Multivariate Analysis*. Third Edition. Wiley & Sons Inc. Hoboken. http://dx.doi.org/10.1002/9781118391686

ROUSSEEUW, P. J. [1987]: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*. Vol. 20. November. pp. 53–65. http://dx.doi.org/10.1016/0377-0427(87)90125-7

SAJTOS L. – MITEV A. [2007]: *SPSS kutatási és adatelemzési kézikönyv*. Alinea Kiadó. Budapest

SHEN, J. – CHANG, S. I. – LEE, E. S. – DENG, Y. – BROWN, S. J. [2005]: Determination of cluster number in clustering microarray data. *Applied Mathematics and Computation.* Vol. 169. Issue 2. pp. 1172–1185. http://dx.doi.org/10.1016/j.amc.2004.10.076

SIMON J. [2006]: A klaszterelemzés alkalmazási lehetőségei a marketingkutatásban. *Statisztikai Szemle*. 84. évf. 7. sz. 627–650. old.

TÎRNĂUCĂ, C. – GÓMEZ-PÉREZ, D. – BALCÁZAR, J. L. – MONTAÑA, J. L. [2018]: Global optimality in k-means clustering. *Information Sciences.* Vols. 439–440. May. pp. 79–94. https://doi.org/10.1016/j.ins.2018.02.001

VARGHA A. [2016]: A ROPstat statisztikai programcsomag. *Statisztikai Szemle*. 94. évf. 11–12. sz. 1165–1192. old. http://dx.doi.org/10.20311/stat2016.11-12.hu1165

VARGHA A. – BERGMAN, L. R. – TAKÁCS, SZ. [2016]: Performing cluster analysis within a person-oriented context: some methods for evaluating the quality of cluster solutions, *Journal of Person-Oriented Research*. Vol. 2. Nos. 1–2. pp. 78–86. http://dx.doi.org/10.17505/jpor.2016.08

VARGHA A. – BORBÉLY, A. [2017]: Új klasszifikációs módszerek alkalmazása a kétnyelvűség és az etnikai identitás kutatásában. *Statisztikai Szemle*. 95. évf. 8–9. sz. 805–822. old. https://doi.org/10.20311/stat2017.08-09.hu0805

YAHYAOUI, H. – OWN, H. S. [2018]: Unsupervised clustering of service performance behaviors. *Information Sciences.* Vol. 422. January. pp. 558–571. http://dx.doi.org/10.1016/j.ins.2017.08.065

YU, H. – LIU, Z. – WANG, G. [2014]: An automatic method to determine the number of clusters using decision-theoretic rough set. *International Journal of Approximate Reasoning.* Vol. 55. Issue 1. pp. 101–115. http://dx.doi.org/10.1016/j.ijar.2013.03.018

ZHANG, Y. – MAŃDZIUK, J. – QUEK, C. H. – GOH, B. W. [2017]: Curvature-based method for determining the number of clusters. *Information Sciences.* Vol. 415–416. November. pp. 414–428. http://dx.doi.org/10.1016/j.ins.2017.05.024

ZHOU, S. – XU, Z. [2018]: A novel internal validity index based on the cluster centre and the nearest neighbour cluster. *Applied Soft Computing.* Vol. 71. October. pp. 78–88. http://dx.doi.org/10.1016/j.asoc.2018.06.033

Summary

Clustering is a widespread and very popular data analysis method. Although the lack of mathematical assumptions makes cluster analysis applicable in numerous databases, the „goodness" of clustering can also have several aspects. Among „goodness" related questions, this paper focuses on the determination of the optimal number of clusters, and on the consistency of the methods for cluster number selection. The author uses the frequently applied cluster elbow method and the silhouette method for the analysis of homogeneous and heterogeneous databases. The empirical results suggest that both methods highlight the difference between homogeneous and heterogeneous databases: in the heterogeneous database, the optimal number of clusters can be identified more clearly, on which the number of dimensions (variables) has only limited effect. Overall, the paper indicate that both methods can be applied to explore the heterogeneity of a database, and the optimal numbers of clusters identified by these methods are similar.