Statisztikai Szemle

Közzététel: 2019. július 10.

A tanulmány címe:

Klaszterelemzési eljárások halandósági adatokra

Szerzők:
Ágoston Kolos Csaba, a Budapesti Corvinus Egyetem egyetemi docense, az MTA KRTK Közgazdaság-tudományi Intézetének tudományos munkatársa, E-mail: kolos.agoston@uni-corvinus.hu
Burka Dávid, a Budapesti Corvinus Egyetem egyetemi tanársegédje, az MTA KRTK Közgazdaság-tudományi Intézetének tudományos segédmunkatársa, E-mail: david.burka@uni-corvinus.hu
Kovács Erzsébet, a Budapesti Corvinus Egyetem egyetemi tanára, E-mail: erzsebet.kovacs@uni-corvinus.hu
Vaskövi Ágnes, a Budapesti Corvinus Egyetem egyetemi tanársegédje, E-mail: agnes.vaskovi@uni-corvinus.hu
Vékás Péter, a Budapesti Corvinus Egyetem egyetemi adjunktusa, E-mail: peter.vekas@uni-corvinus.hu

DOI: https://doi.org/10.20311/stat2019.7.hu0629

Klaszterelemzési eljárások halandósági adatokra*

Ágoston Kolos Csaba,
a Budapesti Corvinus Egyetem egyetemi docense,
az MTA KRTK Közgazdaságtudományi Intézetének tudományos munkatársa
E-mail: kolos.agoston@uni-corvinus.hu

Burka Dávid,
a Budapesti Corvinus Egyetem egyetemi tanársegédje,
az MTA KRTK Közgazdaságtudományi Intézetének tudományos segédmunkatársa
E-mail: david.burka@uni-corvinus.hu

Kovács Erzsébet,
a Budapesti Corvinus Egyetem egyetemi tanára
E-mail: erzsebet.kovacs@uni-corvinus.hu

Vasövi Ágnes,
a Budapesti Corvinus Egyetem egyetemi tanársegédje
E-mail: agnes.vaskovi@uni-corvinus.hu

Vékás Péter,
a Budapesti Corvinus Egyetem egyetemi adjunktusa
E-mail: peter.vekas@uni-corvinus.hu

A halandósági adatok elemzése nagy múltra tekint vissza a matematikai statisztikában. Több módszer is ismert, melyekkel elemezni vagy tesztelni lehet két vagy több sokaság túlélési valószínűségének azonosságát. A kockázatközösségek halandósági szempontból való homogén csoportokba sorolására azonban kevesebb figyelem jut. A tanulmány több eljárást is bemutat a halandósági adatok klaszterezésére. A szerzők az eljárásokat valós adatokon is tesztelik és értékelik.

TÁRGYSZÓ:
Biztosítási kockázatközösség.
Túlélési valószínűség.
Klaszterelemzés.

DOI: 10.20311/stat2019.7.hu0629

* Jelen tanulmány elkészítését az Európai Unió, Magyarország és az Európai Szociális Alap társfinanszírozása által biztosított forrásból az EFOP-3.6.2-16-2017-00017 azonosítójú „Fenntartható, intelligens és befogadó regionális és városi modellek” című projekt finanszírozta.

Köszönjük a névtelen lektor szakszerű, a tanulmány értelmezhetőségét javító észrevételeit.

A biztosítók működése során nagy hangsúly helyeződik homogén fogyasztói csoportok, kockázatközösségek létrehozására. A heterogén biztosítotti állomány (viszonylag) homogén csoportokba sorolását kockázati klasszifikációnak (angolul risk classification) hívja a szakirodalom (lásd *Crocker–Snow* [1986], [2000]). A kockázati klasszifikáció talán a leggyakrabban alkalmazott aktuáriusi technika, alkalmazását megfelelően kiválasztott klaszterelemzési eljárások segítik. A biztosítást vásárlók kockázati heterogenitása közismert; azt az antiszelekció[1] jelensége is súlyosbítja. Miközben azonban a biztosító társaságok igyekeznek homogén kockázatközösségeket létrehozni, a díjkalkuláció során nem szerencsés, ha egy-egy kockázatközösség túlságosan elaprózódik.

Életbiztosítások esetén bizonyos heterogenitás már évszázadok óta ismert. Eltérést okoz például a férfi és a női halandósági valószínűségekben tapasztalható különbség. Az utóbbi évtizedekben – a számítástechnika fejlődésével – a nemekről szóló elemzések mellett egyéb témájú írások is napvilágra kerültek. Így ma már köztudott a fehér- és a kékgallérosok halandósági mintázatának eltérése, de mindjobban ismert a halandósági valószínűségek iskolai végzettség vagy dohányzás szerinti különbözősége is. A biztosítótársaságok országhatárokat átívelő működésének köszönhetően pedig nyilvánvalókká váltak az országok és az azok régiói közötti különbségek. Így egyre fontosabb, hogy ezeket az eltéréseket egyidejűleg több szempont szerint vizsgáljuk, és ezért, amennyire lehetséges, indokolt a biztosítottak több változó alapján történő csoportokba sorolása. Az ügyfél-szegmentáció a statisztikai módszerek közül például klaszterelemzéssel végezhető.

Tanulmányunk felépítése a következő. A módszertani szakirodalmi áttekintés után bemutatjuk a halandósági adatok forrásait és az írásunkban szereplő alapfogalmakat. Ezt követi a klaszterelemző eljárások különböző változatainak tárgyalása, majd az eredmények bemutatása és értelmezése.

1. A halandósági mintázatok és modellek szakirodalmi áttekintése

Magyarország régióinak eltérő halandósági mintázatait már *Kovács–Őri* [2009] is elemezte. A szerzők területi és társadalmi különbségeket kerestek a megyék halandó-

[1] Antiszelekció alatt azt a jelenséget értjük, hogy a nagyobb kockázatú egyének hajlamosabbak biztosítást vásárolni, és ezért a biztosítással fedezett kockázatközösség káralakulása összességében kedvezőtlenebb lesz annál, mint amit a teljes népességre vonatkozó statisztikák alapján várna a biztosító.

sági adataiban. Számos további tanulmány között *Klinger* írása [2007] mikrocenzus adatokon vizsgálta a lakóhely és az iskolai végzettség kapcsolatát a megyék halandósági különbözőségeinek vonatkozásában.

A halandósági táblák készítésének módszertana évszázadokra nyúlik vissza (*Halley* [1693], *Gauss* [1900], *Benjamin–Pollard* [1980]). A halandósági valószínűségek csak nagy sokaság esetén adnak pontos képet a tényleges halandóságról. Ráadásul biztosítások esetén a kockázatközösség tagjait nem tudjuk időbeli korlátok nélkül megfigyelni, mert csak az élő biztosítási szerződések esetén vannak a biztosítottakról pontos megfigyeléseink. Ha valakinek lejárt a biztosítási szerződése – vagy visszavásárolta azt –, akkor csak annyit tudunk, hogy a lejárat (vagy a visszavásárlás) pillanatában még élt, de nem ismerjük a halála bekövetkeztének pontos időpontját. Ezt a megfigyelési korlátot a szakirodalom cenzorálásnak hívja. A biztosítotti állomány további sajátossága, hogy jellemzően nem újszülöttek kötnek biztosítást, hanem középkorúak. Ezt csonkolásnak nevezi a szakirodalom. A biztosításban ez a két hatás egyszerre érvényesül. Balról csonkolt, jobbról cenzorált adatok esetén (ezt lásd később) a halálozási valószínűségek becslése nem triviális, rá több módszer is ismert.

A túlélési modell (lásd például *Vékás* [2011]) egy olyan népszerű módszercsalád összefoglaló neve, amely élettartamok valószínűségeloszlásának becslésére alkalmazható. A Kaplan–Meier-modell (*Kaplan–Meier* [1958]) az orvosi statisztikából származó, egyik legszélesebb körben alkalmazott, matematikai szempontból viszonylag egyszerű túlélési modell, melyben az adott élettartam elérésének valószínűségét megadó túlélési függvény egy szakaszonként állandó (lépcsős) függvény. A Kaplan–Meier-féle túlélési függvény értéke a múltban megfigyelt kilépési időpontok (a születéstől a halálozásig ténylegesen eltelt időszakok hosszai) között állandó, a kilépési időpontokban ugyanakkor egy olyan egynél kisebb szorzóval módosul (csökken), amely a kilépési időpontot túlélő és az azt elérő egyedek számának hányadosaként számítható ki az elemi valószínűségszámításból ismert kedvező és összes esetek számának hányadosával analóg módon.

Ha a Kaplan–Meier-féle túlélési függvényt a teljes minta helyett annak almintáira (például férfiakra és nőkre vagy a hét magyarországi NUTS 2 statisztikai régióra) külön-külön akarjuk kiszámítani, akkor gyakorlati szempontból fontos kérdés, hogy miként tesztelhető az élettartamok függetlensége az almintákat meghatározó, csoportosító változótól (például a nemtől vagy a régiótól). Függetlenség esetén az egyes almintákhoz tartozó túlélési függvények megegyeznek, és ezért a csoportosító változó irreleváns például az életbiztosítás díjszámításában. Az alminták túlélési függvényeinek egyezését vizsgáló legnépszerűbb – statisztikai programcsomagokban is elterjedt – eljárás a Mantel–Cox- (más néven log-rank) teszt (*Peto–Peto* [1972]), amelynek tesztstatisztikája vagy *p*-értéke segítségével áttételesen a túlélési függvények egyezőségének vagy különbözőségének mértéke is számszerűsíthető.

Egy másik népszerű – de matematikai szempontból a folytonos idő feltételezése miatt a Kaplan–Meier-módszernél jóval bonyolultabb – túlélési modell a Cox-regresszió (más néven az arányos kockázatok modellje; *Cox* [1972]), amely egyszerre képes tetszőleges számú, folytonos vagy kategorikus prediktor változó (például a nem, a jövedelem és a régió) élettartamra gyakorolt hatását figyelembe venni. A Cox-regresszióban a halandóság pillanatnyi intenzitását az idő függvényében mérő hazárd-függvény alakja független a prediktoroktól, de az egyes pillanatokban felvett értékei a prediktorok lineáris függvényétől függő, időben állandó szorzóval módosulnak.

A klaszterelemzési eljárások rövidebb, néhány évtizedes múltra tekintenek vissza. Klaszterelemzési eljárások esetén analitikus eljárás nem igazán ismert, ezek jellemzően numerikus eljárások, amelyek a számítástechnika fejlődésével virágoztak fel. Több magyar szakkönyv is foglalkozik klaszteranalízissel mint a többváltozós adatelemzési eljárások egyik kiemelt módszertanával. Először *Füstös–Meszéna–Simonné* [1986] tárgyalták a módszercsaládot, később *Hunyadi* [2001]; *Hajdu* [2003]; *Kerékgyártóné et al*. [2008]; *Kovács* [2011], [2014] jelentettek meg összefoglaló szakkönyvet a többváltozós statisztikai módszerek, köztük a klaszterelemzés elméleti hátteréről, kiemelve a módszer közgazdasági és társadalomtudományi alkalmazhatóságát.

Simon [2006] tanulmányában a klaszterelemzés marketingkutatási felhasználhatóságáról ír, amely szoros kapcsolatban van a jelen tanulmány által tárgyalt homogén kockázatközösségek létrehozásával. Számos más tudományterületen is eredményesen alkalmazzák a klaszterelemzést mint szegmentációs módszercsaládot, amely alkalmas különböző megfigyelési egységek csoportosítására és az így létrejött csoportokhoz eltérő szabályrendszerek kialakítására (*Dobos–Michalkó–Nováky* [2017], *Borbély–Vargha* [2017]).

A klaszterelemzési eljárások egyik legelső és máig legnépszerűbb módszere a *MacQue*en [1967] által bemutatott K-középpontú módszer. Ez úgy tudja csoportba sorolni a megfigyeléseket, hogy a csoporton belüli eltérésnégyzetösszeg-növekedés a lehető legkisebb legyen. A K-középpontú módszer korai tanulmányozói *Hartigan*, *Wong* és *Lloyd* voltak (*Hartigan* [1975], *Hartigan–Wong* [1979], *Lloyd* [1982]). A módszer nagyon gyors, de erősen függ a kiinduló klaszterközepek megválasztásától, tehát heurisztikus módszernek tekinthető. Már a klaszterezési elemzések felfutásának időszakában elkezdték tanulmányozni az ún. K-medián (máskor *p*-medián, esetleg K-medoid) problémát. Az ötlet onnan jött, hogy ha nem az eltérésnégyzetösszeget, hanem az abszolút eltéréseket minimalizáljuk, akkor egydimenziós probléma esetén a klaszterközéppontok adatpontok lesznek (vagy legalábbis választhatók adatpontoknak is). Az átlagtól eltérően a többdimenziós térben a medián nem definiált, de az elnevezés a feladathoz ragadt. Többdimenziós K-medián probléma esetén adott egy távolság- vagy hasonlósági mátrix, és a kérdés az, hogy mely adatpontokat válasszuk klaszterközéppontnak, és az adatpontokat melyik klaszterközépponthoz rendeljük hozzá, hogy a klaszterközéppontjuktól számított távolságösszegük minimá-

lis legyen. Fontos hangsúlyozni, hogy a pontok klaszterközéppontjuktól számított távolsága egy mátrixban megadható input paraméterként, tehát a K-medián problémát nemcsak abszolút eltérés esetén lehet használni, hanem tetszőleges távolság vagy hasonlósági mérték esetén is. Ennek a problémának egy speciális vetületét elemzik *García–Labbé–Marín* [2011] tanulmányukban.

A K-medián problémát LP- (lineáris programozási) feladatként *Vinod* [1969] írta le, de a nagyméretű LP-feladatok eleinte a gyakorlatban nem voltak megoldhatók, így a K-medián módszer esetén is ismertek heurisztikus eljárások. Manapság már nagyméretű feladatok esetén is globálisan optimális megoldások találhatók (lásd például *García–Labbé–Marín* [2011]), bár ezek az egzakt megoldások nem érhetők el a szokásos statisztikai szoftverekben.

A K-medián feladatot értelmezhetjük úgy is többdimenziós esetben, hogy a klaszterközéppontokat nem rögzítjük a lehetséges adatpontokhoz, hanem lehetnek közöttük nem megfigyelt pontok is, viszont az adatpontok és klaszterközéppontjaik abszolút eltérését minimalizáljuk (lásd például *Sabo–Scitovski–Vazler* [2013]). Történetesen az abszolút eltérések minimalizálása robusztusabb eredményre vezet, és sok esetben gyorsabb is, mint a K-középpontú algoritmus.

A partícionáló klaszterelemzés mellett gyakran használják az ún. hierarchikus klaszterelemzési eljárásokat is. A hierarchikus klaszterelemzési eljárásoknál szintén a távolság- vagy a hasonlósági mátrix a kiindulópont. Ezek jellemzően agglomeratív eljárásokat alkalmaznak, bár ritkán előfordul, hogy felosztó eljárásokat is (lásd például *Kaufman–Rousseeuw* [1990]). Az agglomeratív eljárások során, melyek algoritmikus szempontból ún. mohó eljárások, a két legközelebbi (leghasonlóbb) klasztert vonjuk össze, majd az összevonást ún. dendrogramon ábrázoljuk. Ehhez különböző agglomeratív elvek léteznek. Csak kevés próbálkozás ismert, ahol megpróbálták meghaladni ezeket a mohó algoritmusokat, és megkíséreltek mutatószámot megfogalmazni az illeszkedés jóságára. Ezek között van *Gilpin–Nijssen–Davidson* [2013] munkája.

Halandósági adatok esetén ismereteink szerint még nem alkalmaztak klaszterelemzési eljárásokat, ezért nem létezik útmutatás arra vonatkozóan, hogy ezen adatkörre a klaszterezés több lehetséges változata közül melyik módszer a leghatékonyabb. A téma szempontjából érdekes *Fu–Simonoff* [2017] tanulmánya, amely bár nem klaszterelemzési eljárást mutat be, de bizonyos tekintetben mégis releváns számunkra. E szerzők a döntési fák módszertanát fejlesztették tovább túlélési modellekre. A döntési fák módszertana a teljes mintát/sokaságot (jellemzően) két részre osztja úgy, hogy a részsokaságok valamilyen eredményváltozó tekintetében – amely esetünkben a kiválást (a halál bekövetkeztét) leíró változó – a lehető legnagyobb mértékben különbözzenek. Nem nehéz egyfajta (felosztó) hierarchikus klaszterelemzést belelátni ebbe a módszerbe, de a kettő között természetesen vannak lényeges különbségek is. A döntési fák esetén a magyarázó változók szerint „vágunk”, tehát a részsokaságok nem előre adottak. Olyan megkötéssel viszont alkalmazható a módszer, hogy egyetlen magyarázó változónk van,

ami egyben a csoportokat leíró nominális változó (például Magyarország megyéi) és a döntési változó is. *Fu–Simonoff* [2017] munkája abból a szempontból kapcsolódik tanulmányunkhoz, hogy meg szeretnénk határozni, mennyire különböznek/hasonlítanak a különböző részsokaságok a halandóság szerint.

2. Halandósági adatok és alapfogalmak

Mivel halandósági adatokat klaszterezünk, először bemutatjuk az adatok lehetséges forrásait. Halandósági adatok jellemzően kétféle helyről gyűjthetők túlélési modellek készítéséhez.

Az egyik forrást az országos statisztikák alkotják, ahol külön-külön megtalálhatók az élőkre, illetve az elhunytakra vonatkozó adatok. Jellemzően aggregált adatok állnak rendelkezésre, amelyek segítségével (nyers) koréves halálozási valószínűségek határozhatók meg. A nyers halálozási adatokat többnyire valamilyen módon simítani szokták. A halálozási valószínűségek elérhetők a legfontosabb bontásokban (nem, iskolai végzettség és régiók szerint), de az egyedi életutak nyomon követése hosszabb távon nehézkes.

A másik forrást a különböző intézetek (például a biztosítótársaságok, a nyugdíjfolyósító, a nyugdíjpénztárak) saját adatai képviselik, melyek jellemzően egyéni szinten érhetők el. Például biztosítótársaságok esetén az 1. táblázatban szereplő formában állnak rendelkezésre.

1. táblázat

Példa a biztosítótársaságok halandósági adataira

ID	Belépési kor	Kilépési kor	Státus	Demográfiai jellemző
P0001	30,6	40	1	férfi; Csongrád megye; …
P0002	32,25	52,25	0	nő; Pest megye; …
P0003	35,167	37,667	0	nő; Pest megye; …

Példánkban az első biztosított 30,6 évesen csatlakozott a kockázatközösséghez. Tegyük fel, hogy e személy 1970. január 1-jén született, és 2000. július 1-jén vásárolt biztosítást. Ez rögtön rávilágít a biztosítói adatok már említett jellegzetességére (ami megkülönbözteti az aktuáriusi elemzéseket a biostatisztikai elemzésektől): az emberek nem újszülött, hanem felnőtt korban vásárolnak biztosítást. Természetesen csak azok tudnak biztosítást vásárolni (és így a kockázatközösséghez csatlakozni), akik életben vannak, tehát szakszóval megfogalmazva, a megfigyelések balról cson-

koltak. A példában a biztosított 40 éves korában, 2010. január 1-jén elhalálozott. Esetében (s_1; e_1; d_1) = (30,6; 40; 1), ahol s_1 a biztosított kezdő belépési korát, e_1 a kilépési korát jelöli; d_1 változó bináris értéke 1, tehát a kilépés oka haláleset (ha 0, a biztosított egyéb okból távozott a kockázatközösségből).

A második biztosított 32,25 évesen vásárolt biztosítást, tegyük fel 20 éves tartammal. A biztosítási díjakat rendre befizette, a biztosítás lejártakor a biztosított életben volt. A biztosítás lejárta után a biztosítónak nincs információja a biztosítottról, halála konkrét időpontját nem ismeri. Csak annyit tud, hogy 52,25 évesen életben volt, amit a szakirodalom (jobboldali) cenzorálásnak hív. A második biztosított esetén (s_2; e_2; d_2) = (32,25; 52,25; 0).

A harmadik biztosított 35,167 évesen biztosítást vásárolt, és a szerződést 37,667 évesen visszavásárolta. Ebben az esetben is a biztosító csak annyit tud, hogy a biztosított 37,667 évesen életben volt. Annyi különbség van az előző példához képest, hogy itt a visszavásárlás időpontja nem ismert előre. A szakirodalom ezt véletlen cenzorálásnak hívja, szemben az előző esettel, amely fix cenzorálás. Esetünkben azonban a fix és a véletlen cenzorálás megkülönböztetésének nincs jelentősége.

Egy biztosítótársaságnál hasonló bontásban jellemzően több ezer személy adata áll rendelkezésre. Amennyiben ezeket Magyarország megyéi szerint szeretnénk csoportosítani, a megfigyeléseket először a megyékre szűrjük, és ezt követően klaszterelemzést végzünk. Könnyen belátható, hogy nyers adatokon nem hajtható végre klaszterelemzés, mivel az adathalmazok dimenziója, mérete nem egyezik meg. Ha az adatokat valamilyen módon mégis közös formára lehetne hozni, akkor is inkább a belépési és a kilépési korok alapján történhetne a klaszterezés, aminek azonban valószínűleg nem sok köze lenne a csoportok halandóságához.

3. A halandóság elemzése klaszterelemzési eljárásokkal

Ha valamilyen klaszterelemzési eljárást szeretnénk alkalmazni, akkor az adatokat előzetesen fel kell dolgozni, hogy azok reprezentálják a csoportok halandóságát. Majd a hasonlósági mérőszám kiválasztását és kiszámítását követően történik a megfelelő klaszterelemző eljárás kiválasztása és alkalmazása.

3.1. A Kaplan–Meier-féle túlélési függvényen alapuló hasonlósági mérték

A csoportok halandóságát talán a legkönnyebben interpretálható módon a túlélési függvény jellemzi. Az $S(t)$ függvény megadja, hogy egy $t = 0$ időpontban a kocká-

zatközösséghez csatlakozott egyed mekkora valószínűséggel lesz életben t (> 0) évesen. Ha feltételezzük, hogy a csoporttagok halandósága homogén, akkor becslést tudunk adni a túlélési függvényre. Legismertebb ilyen becslés a Kaplan–Meier-becslés. A becslés során lényegében olyan kicsi intervallumokra osztjuk az időtengelyt, hogy az intervallumokban ne történjen belépés vagy kilépés. Ezt követően minden intervallumra kiszámítjuk a túlélési valószínűséget: az intervallum végén élők számát osztjuk az intervallum elején megfigyelt személyek számával. Majd a túlélési valószínűségeket összeszorozzuk.

Az élettartamot mint valószínűségi változót T-vel, valamint a növekvő sorrendbe rendezett, különböző kilépési időket (azaz a megfigyelt cenzorálatlan t_i élettartamokat $i = 1, 2, \ldots, n$) a $\tau_1, \tau_2, \ldots, \tau_m$ jelöléssel jelölve, a τ_k időpontban a túlélés becsült (feltételes) valószínűsége:

$$\hat{\pi}_k = \hat{\mathbb{P}}\left(T > \tau_k \middle| T \geq \tau_k\right) = \frac{n_k - \delta_k}{n_k}, \qquad /1/$$

ahol

$$n_k = \#\left\{1 \leq i \leq n : t_i \geq \tau_k\right\} \qquad /2/$$

a τ_k kilépési időt elérő megfigyelések száma,

$$\delta_k = \#\left\{1 \leq i \leq n : t_i = \tau_k, , t_i \text{ cenzorálatlan}\right\} \qquad /3/$$

pedig épp a τ_k időpontban kilépett megfigyelések száma.

Ezekkel a jelölésekkel a túlélési függvény Kaplan–Meier-becslése:

$$\hat{S}(t) = \sum_{\tau_k < t} \hat{\pi}_k. \qquad /4/$$

A fejezet során az eljárás és az eredmények szemléltetésére egy biztosítotti adatállományt használunk. Az adatok a biztosítótársaság számára bizalmas információk, ezért részletes leírásukat nem tudjuk megadni. Az 1. *a*) ábra a biztosítottak egészére vonatkozó túlélési függvény becslését mutatja be, a *b*) pedig külön-külön a férfiak és a nők túlélési görbéit.

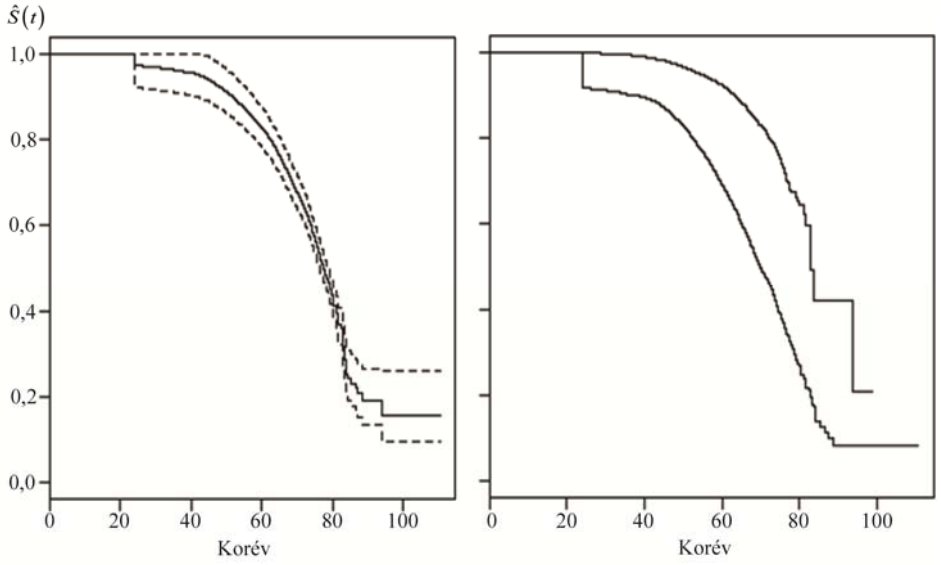
Az 1. *a*) ábrán jól látszik az adatok balról csonkolása. A becsült túlélési függvény értéke 1 egészen a biztosítottak 20-as éveinek közepéig. Ennek az az oka, hogy a legfiatalabb biztosított húszegynéhány éves. Említést érdemel az is, hogy az első

előforduló belépési korok esetén viszonylag nagy a bizonytalanság, hiszen a becslésnél csak pár ember adataira támaszkodhatunk: egy ezrelék körüli halálozási valószínűséget akarunk meghatározni 10-20 biztosított adatai alapján, ezért a konfidenciaintervallum viszonylag széles. Középkorú biztosítottból több van, így a túlélési valószínűség is pontosabban becsülhető, és a konfidenciaintervallum is keskenyebb. Az idős biztosítottak száma ugyancsak alacsony, ráadásul esetükben a halálozási valószínűségek is nagyobbak, ezért a konfidenciaintervallum ismét széles.

Az 1. *b*) ábrán a férfi és a női túlélési függvényeket látjuk. Köztudott, hogy a férfiak túlélési valószínűsége elmarad a nőkétől (ezt a két görbe is mutatja), de a bemutatotthoz képest a valóságban kisebb a különbség a két nem között. Esetünkben a nagy különbséget az okozza, hogy a példánkban szereplő 41 fiatal biztosított közül elhunyt az egyik (körülbelül 24 éves) férfi, ami az összes fiatal biztosított tekintetében 2,5 százalékos halálozási valószínűséget eredményez, az (eredetileg) 12 férfi tekintetében viszont több mint 8 százalékosat, ami nagyon magas érték ebben az életkorban. Tehát a túlélési görbe alakja önmagában nem biztos, hogy jó támpont a becsléshez. Következésképpen a görbe azon részeinek, ahol a becslés számos megfigyelési adaton alapul, nagyobb jelentőséget kell tulajdonítanunk, mint azoknak, ahol csak kevés adaton.

1. ábra. A Kaplan–Meier-féle túlélési függvény becslése a teljes sokaságra

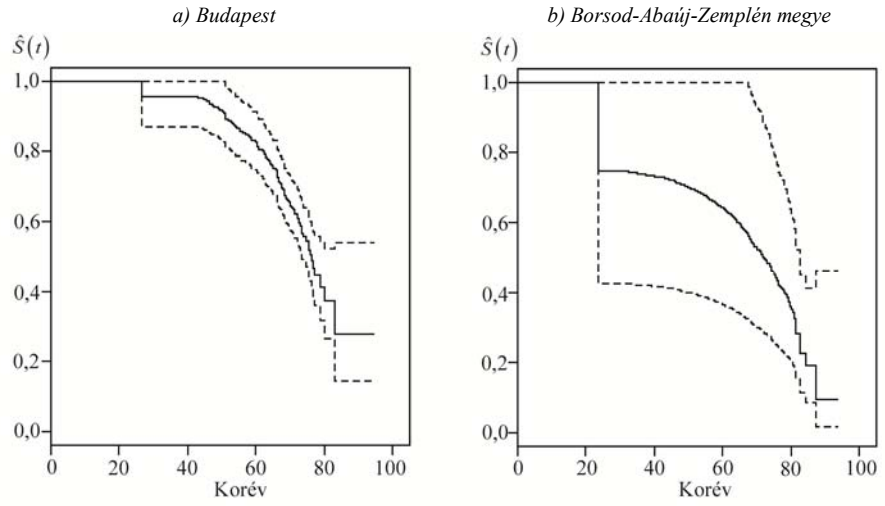
a) az összes biztosítottra *b) férfiakra és nőkre külön-külön*



Megjegyzés. Az 1. *a*) ábrán a két pontozott vonal a *Greenwood* [1926] formulájával becsült, 95 százalékos konfidenciaintervallumot mutatja. Az 1. *b*) ábrán a felső görbe a nőkhöz, az alsó a férfiakhoz tartozik.

Vegyük most példaként Budapest és Borsod-Abaúj-Zemplén megye teljes népességét. (Lásd a 2. ábrát.)

2. ábra. A Kaplan–Meier-féle túlélési függvény becslése Budapest és Borsod-Abaúj-Zemplén megyére



A klaszterelemzéshez először meg kell határozni a két megye távolságát (hasonlóságát). Mivel két függvényről van szó, több lehetőség is adódik:

– Függvényterekben használatos a maximum metrika, amely két függvény abszolút különbségének a maximuma (szuprémuma). A függvényterek tekintetében ez távolságmérték, amely a mi esetünkben könnyen kiszámolható lenne, és alkalmazzák a klaszterelemzés egyéb területein is. Ellene szóló érv azonban, hogy – mint már említettük – fiatal korban kevés a biztosított, és egy „véletlen” halálesetnek nagy a jelentősége.

– A maximum metrika úgy módosítható, hogy ahol több biztosított van, ott a távolságmértéket nagyobb számmal szorozzuk meg, ahol kevesebb, ott kisebbel. Ez ellen viszont az hozható fel, hogy egyáltalán nem biztos, hogy így is távolságmértéket kapunk.

– *Sangalli et al*. [2009], [2010] megadtak egy függvények között értelmezett hasonlósági mértéket, amely alapján *Majstorovic et al*. [2018] növekedési görbéket klaszterezett. A mi esetünkben nehézséget jelent, hogy az említett szerzők folytonos függvényekkel dolgoznak, és továbbra is kezelendő probléma, hogy a Kaplan–Meier-függvények „megbízhatósága” nem minden pontban egyforma. Itt is felmerül a sú-

lyozás kérdése; ezt azonban nem tárgyalják a témával foglalkozó szakirodalmi modellekben.

– A hierarchikus és más klaszterelemzési eljárásokhoz is egy olyan különbözőségi mértéket szükséges definiálni, amely a megfigyelések közötti eltérést – és ezáltal áttételesen azok hasonlóságát is – számszerűsíti. A szakirodalomban ez a különbözőségi mérték gyakran valamely olyan statisztikai próba tesztstatisztikája, amely két egyed egyezését teszteli. Ennek leggyakoribb példája a kategorikus változók terében végzett klaszterelemzés esetén alkalmazott χ^2 tesztstatisztika (lásd például *Elavarasi–Akilandeswari* [2014], *Hussain–Asghar* [2016]). Halandósági adatok elemzése esetén a legismertebb ilyen jellegű statisztika a túlélési függvények egyezését vizsgáló ún. log-rank teszt χ^2 tesztstatisztikája. Ez a két függvény tökéletes egyezése esetén a nulla értéket veszi fel, és minél nagyobb az értéke, annál valószínűbb, hogy a vizsgált túlélési függvények különböznek egymástól a teljes sokaságban. A mi szempontunkból ez azt jelenti, hogy „messzebb" esik, azaz jobban eltér egymástól a két megye halandósága (lényegében ezt a módszert alkalmazza *Fu–Simonoff* [2017] is).

Az előbbiekben bemutatott lehetőségek közül a log-rank tesztstatisztika értékét tekintjük a legalkalmasabb mutatószámnak két megye távolságának (hasonlóságának) meghatározására. A Kaplan–Meier-becslés ismertetésénél már bevezetett jelölésekkel a log-rank tesztstatisztika:

$$\chi^2 = \frac{\sum_{j=1}^{m}\left(\delta_{1j} - \frac{\delta_j}{n_j} n_{1j}\right)^2}{\sum_{j=1}^{m} \frac{1}{n_j - 1} \delta_j \frac{n_{1j}}{n_j}\left(1 - \frac{n_{1j}}{n_j}\right)\left(n_j - \delta_j\right)}, \qquad /5/$$

amely a vizsgált alcsoportokbeli túlélési függvények egyezésére vonatkozó H_0 nullhipotézis fennállása esetén χ^2-eloszlást követ 1 szabadságfokkal. Itt kettős index esetén az „1" első index az első részpopulációra (a mi esetünkben az első összehasonlítandó megyére) vonatkozik, egyszeres index esetén pedig a két összehasonlítandó megye adatainak egyesítésével kapott mutatószámra.

Előnyös tulajdonságai ellenére ennek a különbözőségi mértéknek is vannak korlátai. A tesztstatisztika értéke nem távolság, nem teljesül rá a háromszög-egyenlőtlenség. Vegyünk például három megyét; az első és a második megye esetén

sok biztosított van, és szignifikáns különbséget tapasztalunk közöttük a halandóság tekintetében. A harmadik megyében viszont relatíve kevés a megfigyelés, tehát nem lesz szignifikáns a különbség közötte és az első, illetve a második megye között. Így az első és a második megye „távolsága" nagyobb, mint az első és a harmadik, valamint a második és a harmadik megye „távolságának" összege. Példánkban a log-rank tesztstatisztika értéke Borsod-Abaúj-Zemplén és Komárom megye viszonylatában 5,69, Borsod-Abaúj-Zemplén és Békés megye esetén 0,29, Komárom és Békés megye esetén pedig 0,83. Ha az utóbbiakat összeadjuk, (0,29 + 0,83) 1,12-ot kapunk, ami jelentősen kisebb, mint 5,69. Az, hogy az alkalmazott mérték megsérti a háromszög-egyenlőtlenséget, nem szerencsés, de nem egyedi eset.

A log-rank tesztstatisztika segítségével tehát értelmezhetjük a megyék halandósági adatai közötti hasonlóságot. A következő lépésben klaszterelemzési módszert kell választani.

A leggyakoribb klaszterelemzés a K-középpontú (KMEANS) algoritmus (lásd például *Kovács* [2011], [2014]). A K-középpontú algoritmus esetén a kezdeti magpontokhoz való szétosztás után minden lépésben kiszámítjuk az egy klaszterbe tartozó adatpontok klaszterközepét, majd a pontokat besoroljuk a legközelebbi – számított – klaszterközéphez, és az eljárást iteráljuk. A mi esetünkben az eljárás alkalmazását több tényező nehezíti: az első nehézség az, hogy már magának a klaszterközéppontnak az értelmezése sem triviális. A klaszterközéppont kiszámításához páronként egyesíthetnénk a megfelelő megyék adatait, és az aggregált adatokon meghatározhatnánk a Kaplan–Meier-féle túlélési görbét. De – mint már említettük – a log-rank statisztika nem értelmezhető távolságmértékként, ezért akár végtelen ciklusba is keveredhetünk. Amellett, hogy az eljárás implementálása sem lenne teljesen triviális, K-közép feladat esetén az eltérésnégyzet-összeget minimalizáljuk, aminek a log-rank statisztika esetén nem tudunk értelmet tulajdonítani. Ezért a log-rank statisztika mint hasonlósági mérték kiválasztásakor eltekintünk a K-középpontú algoritmus alkalmazásától.

Az előbbinél sokkal megfelelőbb a K-medián (KMEDIAN vagy *p*-median) módszer. Ebben az esetben úgy akarjuk klaszterbe sorolni a megfigyeléseket, hogy minden klaszter esetén kiválasztunk egy reprezentánst (itt az egyik megyét), és kiszámítjuk a klaszterbe tartozó megfigyelések esetén a reprezentánstól vett távolságok (hasonlóságok) összegét, majd az utóbbiakat aggregáljuk a létrejött klaszterekre. Úgy keressük a csoportba sorolást, hogy a távolságösszeg minimális legyen. Másképp megfogalmazva, n megfigyelést k $\left(k \leq n\right)$ diszjunkt halmazba $\left(S = S_1, S_2, \ldots, S_k\right)$ szétosztunk. Keressük azt az S partíciót és $y_1 \in S_1, y_2 \in S_2, \ldots, y_k \in S_k$ klaszterközepeket, amelyre a

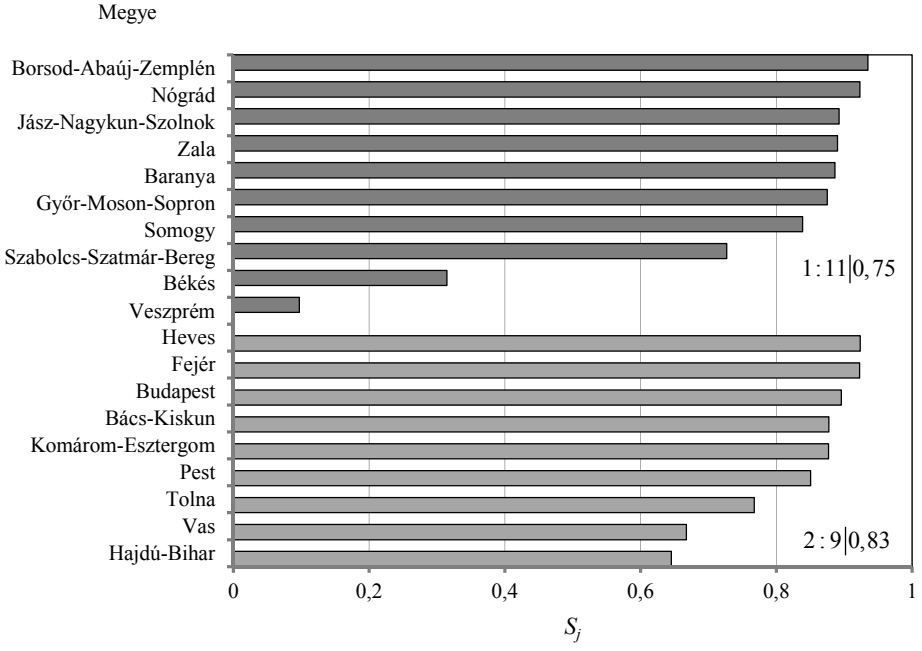
$$\sum_{i=1}^{k} \sum_{x \in S_k} d\left(x, y_k\right) \qquad /6/$$

kifejezés minimális. $d\left(x, y_k\right)$ az x és az y_k megfigyelések távolságát (hasonlóságát) adja meg.

A K-medián feladat esetén a klaszterközepek is adatpontok, ezért csak a megfigyelések (megyék) adatainak távolságára van szükségünk. A log-rank statisztika értékeit mátrixba rendezzük; ez az inputja a K-medián feladatnak.

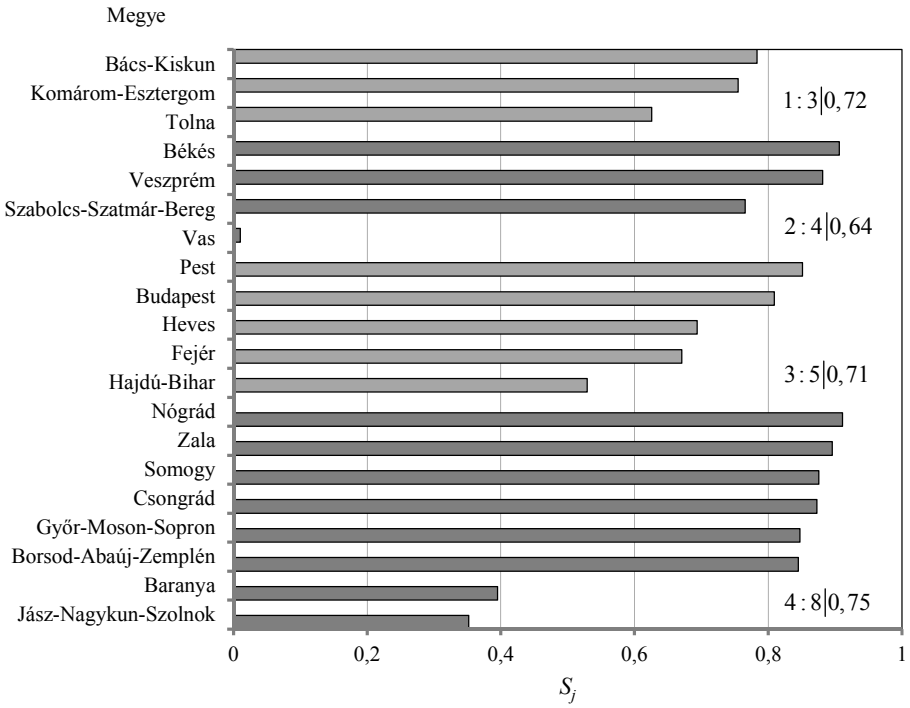
A K-medián feladat felírható vegyes egészértékű LP (lineáris programozás) feladatként (lásd például *Vinod* [1969]). Az LP feladat megoldására a Gurobi programot (8.0 verzió) használtuk. K-medián feladat esetén – akárcsak a K-közép feladatnál – a klaszterek száma az input paraméter az eljárásban, a megfelelő klaszterszám kiválasztása is az elemzés részét képezi. A klaszterszám meghatározásakor a sziluettábrákra hagyatkozunk (*Rousseeuw* [1987]). Halandósági adataink tekintetében „kövér" sziluettek adódtak; a 3. és a 4. ábra alapján a kettő és a négy klaszterszám tűnik jó választásnak, de hasonló következtetést lehet levonni a dendrogram szerint is (lásd az 5. ábrát). $K = 2$ klaszterszám esetén a négyklaszteres felosztáshoz képest a 2-es és a 4-es, valamint az 1-es és a 3-as számú klasztereket aggregáltuk. Egyedül Vas megye a kivétel, amely más megyékkel „társulva" került a 2. klaszterbe.

3. ábra. Sziluettábra K-medián eljárással két klaszterre
($n = 20$)



Megjegyzés. Itt és a következő ábráknál, $j: n_j \left| ave_{i \in C_j} s_j \right.$, ahol n_j a C_j klaszter elemszáma, s_j pedig a sziluettszélesség. Az átlagos sziluettszélesség: 0,79.

4. ábra. Sziluettábra K-medián eljárással négy klaszterre
(n = 20)



Megjegyzés. Az átlagos sziluettszélesség: 0,71.

A K-medián kétklaszteres eljárás eredményét a 2. táblázat, a négyklaszteresét pedig a 3. táblázat foglalja össze.

2. táblázat

A K-medián eljárás eredménye két klaszter esetén

Klaszter száma	Reprezentáns	Klaszterbe tartozó megye
1	Csongrád	Baranya, Békés, Borsod-Abaúj-Zemplén, Csongrád, Győr-Moson-Sopron, Jász-Nagykun-Szolnok, Nógrád, Somogy, Szabolcs-Szatmár-Bereg, Veszprém, Zala
2	Fejér	Bács-Kiskun, Budapest, Fejér, Hajdú-Bihar, Heves, Komárom-Esztergom, Pest, Tolna, Vas

3. táblázat

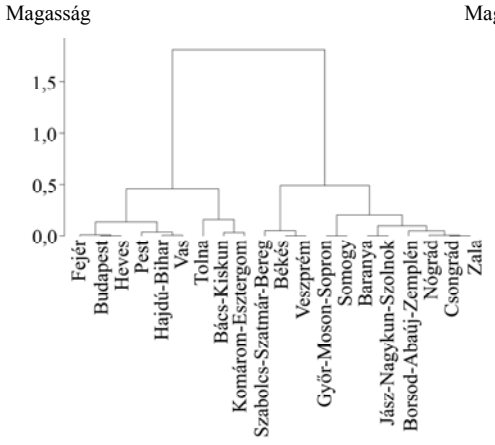
A K-medián eljárás eredménye négy klaszter esetén

Klaszter száma	Reprezentáns	Klaszterbe tartozó megye
1	Bács-Kiskun	Bács-Kiskun, Komárom-Esztergom, Tolna
2	Békés	Békés, Szabolcs-Szatmár-Bereg, Vas, Veszprém
3	Budapest	Budapest, Fejér, Hajdú-Bihar, Heves, Pest
4	Csongrád	Baranya, Borsod-Abaúj-Zemplén, Csongrád, Győr-Moson-Sopron, Jász-Nagykun-Szolnok, Nógrád, Somogy, Zala

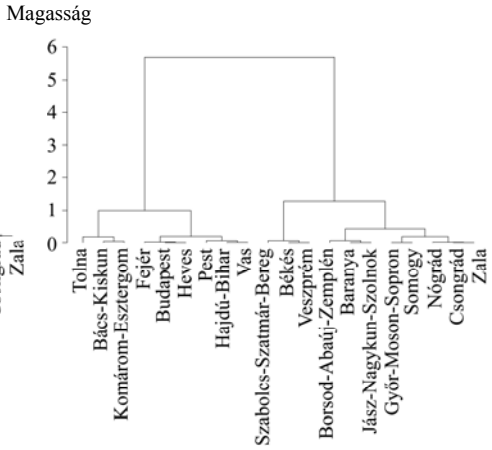
A hasonlósági mátrixnak köszönhetően hierarchikus klaszterelemzési eljárásokban is „gondolkodhatunk”. Az 5. ábra az átlagos lánc és a legtávolabbi szomszéd elvével készült dendrogramokat mutatja be.

5. ábra. Dendrogram

a) az átlagos lánc elve alapján



b) a legtávolabbi szomszéd elve alapján



Az 5. *b*) ábra a legtávolabbi szomszéd elvével készült klaszterelemzést illusztrálja. Az ábrán ráismerhetünk a K-medián eljárás során kapott csoportokra, például a 3. táblázat Tolna, Bács-Kiskun és Komárom-Esztergom megyéből álló, háromelemű 1-es klaszterére. A következő hatelemű klaszter (Fejér megye, Budapest, Heves megye. Pest megye, Hajdú-Bihar megye és Vas megye) lényegében a 3. táblázat 3-as klasztere, amelyből a táblázatban csak Vas megye hiányzik.

Az átlagos lánc elvén alapuló 5. *a*) ábra nem sokban tér el az 5. *b*)-től; csak Borsod-Abaúj-Zemplén megye helyzete lett más: Nógrád, Csongrád és Zala megye hármasához csatlakozik a legtávolabbi szomszéd elve alapján készült dendrogramhoz képest.

Hansen–Jaumard [1997] rámutatnak arra, hogy a legközelebbi és a legtávolabbi szomszéd elve (de a többi agglomerációs elv is) egyfajta mohó eljárás, ahol az adott lépésben mindig csak a lehető legjobb lehetőséget választjuk; azonban előfordulhat, hogy e lépésnek később nagyon rossz következményei lehetnek. *Hansen–Jaumard* [1997] azt is megjegyzik, hogy a klaszteréljárások fejlesztői eddig nem tettek nagy erőfeszítést azért, hogy a legjobb összevonódási sorrendet válasszák meg.

A kevés kivétel egyike *Gilpin–Nijssen–Davidson* [2013] munkája, melyben a szerzők először egy mutatószámot definiálnak az illeszkedés jóságára. Gondolatmenetüket követve mi is kiválasztunk három pontot. Ha az első és a második pontot hamarabb vonjuk össze, mint az első és a harmadik pontot, akkor az első és a harmadik pont távolságából (hasonlóságából) levonjuk az első és második pont távolságát (hasonlóságát). Amennyiben egy (nagy) pozitív számot kapunk, akkor az első és a második pont összevonása megfelelő módon, hamarabb történt meg, mint az elsőé és a harmadiké. Ha viszont negatív, akkor az összevonásnak ellentétes sorrendben kellett volna történnie, ezért csökken az illeszkedés jóságát leíró mutatószám. E számítást az összes számhármasra (a sorrendet is figyelembe véve) megismételtük, és aggregáltuk a távolságkülönbségeket (részleteket lásd *Gilpin–Nijssen–Davidson* [2013]).

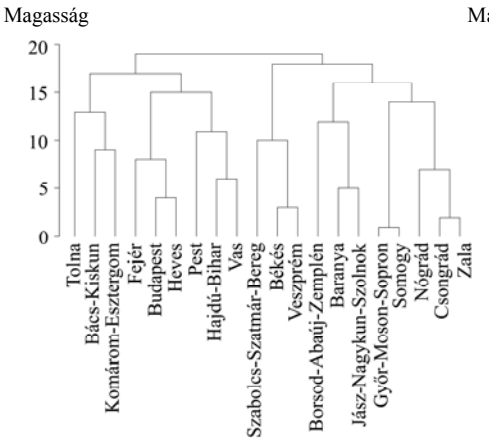
A legközelebbi szomszéd elvének alkalmazásakor a mutatószám értéke 2 083,84, a legtávolabbi szomszéd elvénél 2 875,34, az átlagos lánc elvénél pedig 2 874,61 (a mutató nagyobb értéke jobb illeszkedést mutat). A következőkben azt a dendrogramot szeretnénk megadni, amely esetén a mutatószám értéke a legnagyobb. Ehhez a *Gilpin–Nijssen–Davidson* [2013] által megfogalmazott, egészértékű LP feladatot kell lefuttatnunk; nagy mintánál ez beláthatatlan ideig fut, de 20 megye adatainál csak rövid ideig. A módszer csupán az összevonás optimális sorrendjét adja meg, az összevonás távolságszintjét nem tudjuk értelmezni. Az összevonás sorrendjét a 6. ábra ismerteti. Az illeszkedés jóságát kifejező mutató értéke 2 876,32; ez alig jobb csak, mint a legtávolabbi szomszéd elve esetén kapott érték.

A 6. ábra rávilágít *Gilpin–Nijssen–Davidson* [2013] módszerének arra a sajátosságára (korlátjára), hogy az egy-egy lépésben több pontot vagy klasztert is összevon. Ha például van négy pontunk a számegyenesen (0,1,100,102), akkor a módszer csak annyit tud megállapítani, hogy előbb az első és a második, illetve a harmadik és a negyedik pontot kell összevonni, utána pedig a két kételemű klasztert. Ezáltal a módszer nem tudja érzékelni (meghatározni) azt, hogy vajon érdemesebb-e előbb az első két pontot összevonni és csak utána a második kettőt. Elfogadva ezt a korlátot, azt találjuk, hogy az összevonás sorrendje a legtávolabbi szomszéd elve esetén majdnem optimális, egyedül Borsod-Abaúj-Zemplén megye besorolása változik meg: nem

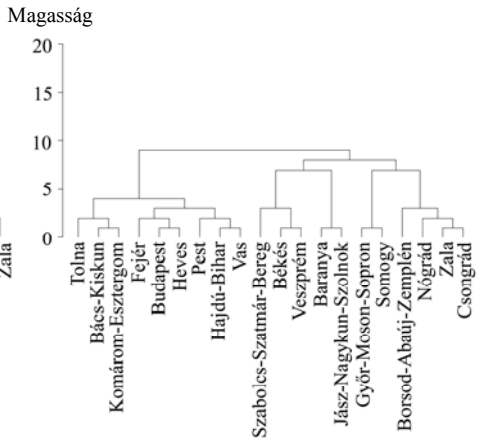
Baranya és Jász-Nagykun-Szolnok megyékhez kell társítani, hanem Nógrád, Zala és Csongrád megyékhez.

6. ábra. Dendrogram

a) a legtávolabbi szomszéd elvének alkalmazása alapján



b) az optimális összevonódási sorrend alapján



3.2. Koréves halálozási valószínűségeken alapuló klaszterezési eljárások

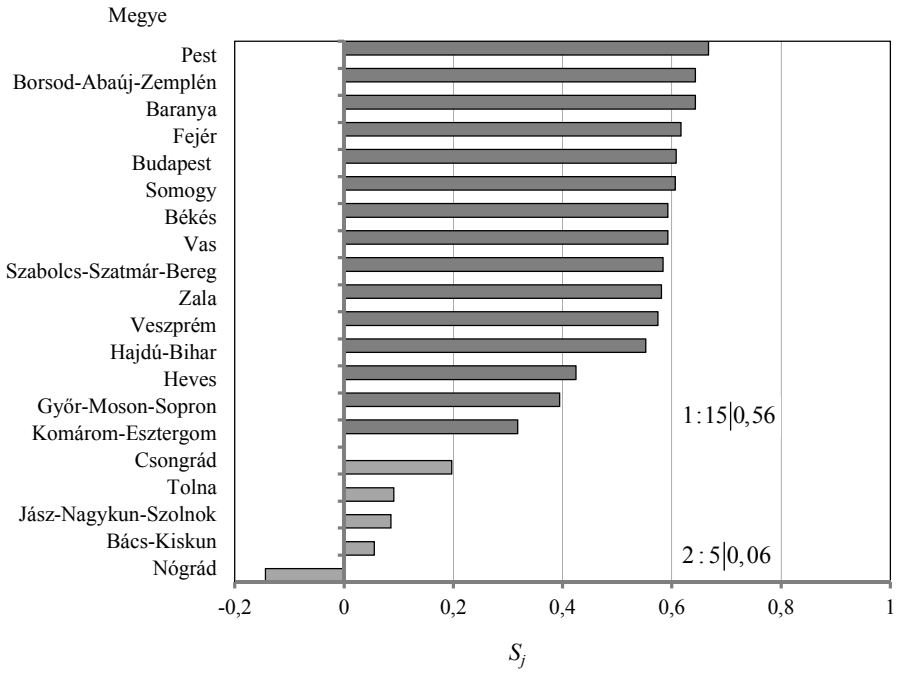
Az előző fejezetben a túlélési függvény Kaplan–Meier-becslését tárgyaltuk, és erre alapozva bemutattuk, hogy miként tudjuk a megyéket csoportosítani. Közösségek halandósági mintáinak jellemzésére a Kaplan–Meier-féle túlélési függvény mellett leggyakrabban a koréves halálozási valószínűségi mutatószámot használják. Országok esetén egyedi adatok szinte sohasem állnak rendelkezésre, csak aggregáltak, amelyből jellemzően koréves halálozási valószínűségeket készítenek. Koréves halálozási valószínűségek becslésére többféle módszer is ismert (lásd például *Benjamin–Pollard* [1980]). Az egyik során először meghatározzuk az ún. nyers koréves halálozási valószínűségeket (erre a Kaplan–Meier-becslés is alkalmazható), majd azokat valamilyen módszerrel simítjuk (egy Magyarországon alkalmazott módszert *Ágoston* [2003] ismertet).

Nekünk a rendelkezésre álló adatok alapján nem volt érdemes megyei bontásban koréves halálozási valószínűségeket számolni, mert ahhoz a megfigyelések száma nem elegendő. Mivel csak a 25–30, a 30–35, ... és a 65–70 korévekre érhető el kellő számú adat, ötéves korcsoportokkal dolgoztunk. A nyers halálozási valószínűségeket nem simítottuk (egyrészt az ötéves korcsoportok esetén kisebb az ingadozás, más-

részt nem ismert, hogy a korcsoportos adatokat milyen módszerrel érdemes simítani. Az óvatosság azért indokolt, mert a kelleténél erőteljesebb simítás eltünteti a létező különbségeket).

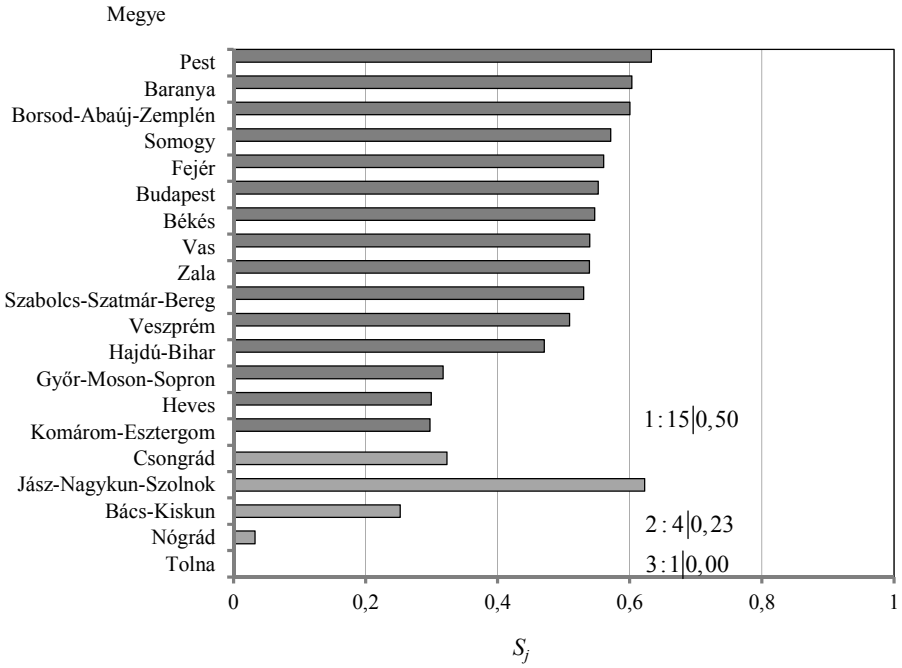
Az ebben a formátumban megadott adatok esetén már alkalmazható a K-középpontú eljárás. Ily módon már csak egyetlen fontos kérdés maradt: standardizáljuk-e az adatokat? Idős korban a halálozási valószínűségek akár nagyságrenddel is nagyobbak lehetnek, mint fiatal korban. Ezért, ha az adatot nem standardizáltuk volna, a csoportok kialakulását lényegében csak az időskori halandóság határozta volna meg. A sziluettábrák alapján a kettő vagy a három klaszterszám a legjobb megoldás (lásd a 7. és a 8. ábrákat), de ezeknél inkább egy nagy elemszámú klaszter keletkezik, a második klaszter a „maradék", három klaszter esetén pedig Tolna megye egyelemű klasztert alkot. (Lásd a 4. táblázatot.)

7. ábra. Sziluettábra K-középpontú eljárással két klaszter esetén



Megjegyzés. Az átlagos sziluettszélesség: 0,43.

8. ábra. Sziluettábra K-középpontú eljárással három klaszter esetén



Megjegyzés. Az átlagos sziluettszélesség: 0,43.

4. táblázat

A K-középpontú eljárás eredménye három klaszter esetén

Klaszter száma	Klaszterbe tartozó megye
1	Baranya, Békés, Borsod-Abaúj-Zemplén, Budapest, Fejér, Győr-Moson-Sopron, Hajdú-Bihar, Heves, Komárom-Esztergom, Pest, Somogy, Szabolcs-Szatmár-Bereg, Vas, Veszprém, Zala
2	Bács-Kiskun, Csongrád, Jász-Nagykun-Szolnok, Nógrád
3	Tolna

A 4. táblázatban bemutatott klaszterek különböznek azoktól, amelyeket a K-medián eljárással kaptunk a log-rank statisztika segítségével. (Lásd a 2. és a 3. táblázatot.)

Az 5. táblázat az ANOVA-teszt *F*-értékeit ismerteti különböző korcsoportokra vonatkozóan. Az *F*-értékek alapján érthetővé válik a különbég a két módszer között:

a mutató az idős korcsoportokban a legnagyobb, tehát éppen azokban, ahol a kevés elemszám miatt a legkevésbé megbízható a becslés. Ha több adatunk lenne, akkor ez a hatás csökkenne, bár továbbra sem szűnne meg. A halálozási valószínűségeken alapuló K-középpontú eljárás vélhetően az idősek (és a fiatalkorúak) adatai alapján fogja csoportokba sorolni a közösségeket, tehát pont azon életkorok alapján, amelyek esetében kevésbé megbízható a becslés. Ezért nem javasoljuk e módszer alkalmazását. Érdemesebb akár koréves halálozási adatokból is log-rank statisztikát előállítani.

5. táblázat

Az ANOVA-teszt F-értékei különböző korcsoportokra

Klaszter száma	Korcsoport (év)								
	25–30	30–35	35–40	40–45	45–50	50–55	55–60	60–65	65–70
2	1,3	0,0	0,0	0,9	2,6	0,1	2,9	6,6	38,1
3	1,2	0,4	0,7	0,6	2,6	3,8	20,5	5,0	18,0

Másik lehetőség egy olyan távolság- vagy hasonlósági mérték megadása, amely figyelembe veszi a korcsoport elemszámát. *Arató et al*. [2009] azt vizsgálták, hogy két halandósági tábla mennyire egyezik meg. Tanulmányukban három hasonlósági mérték is szerepel. Ezek közül kettő az ún. l_x értékeken alapul. Az l_x megmutatja, hogy egy fiktív születési kohorszból hányan érik meg az *x* életkort. Számunkra lényeges szempont, hogy ezek az értékek nem közvetlenül becsültek, hanem becsült halálozási valószínűségeken alapulnak. Tehát a q_x értékek pontatlansága „átragad" az l_x értékekre is. Ezért ezeket a mértékeket a biztosítási adatokra nem használjuk. A szerzők a tanulmányukban szereplő harmadik mértékre QDEV néven hivatkoznak, amelynek képlete:

$$QDEV = \sum_{i=K}^{N} \frac{T_i^a \left(q_i^a - q_i^b \right)^2}{q_i^b} . \qquad /7/$$

A /7/ képlet esetén *K* és *N* a figyelembe vett életkortartomány alsó és felső határa, q_i a halálozási valószínűség, *Ti* pedig a kitettség (kockázatban eltöltött idő). *Arató et al*. [2009] az egyik közösség adatait (a tanulmányban Magyarország [melyre a *b* felső indexű változók vonatkoznak]) „elméleti" értékeknek tekintik, a másik közösségét pedig „mintának" (melyre az *a* felső indexű változók vonatkoznak), és mutatószámuk a két halandósági tábla eltérésének mértékére világít rá. Példánk esetén e hasonlósági mérték aszimmetrikus: két különböző értéket kapunk a Bács-Kiskun és

Baranya megye, illetve a Baranya és Bács-Kiskun megye viszonylatokra. Az aszimmetrikus hasonlóságmértékek használata nem ismeretlen a statisztikai elemzésben, lásd például *Okada* [2000], *Olszewski* [2011].

A QDEV hasonlósági mérték esetén két választási lehetőség kínálkozik: megtarthatjuk a mértéket aszimmetrikusnak, vagy szimmetrikussá tehetjük. A szimmetrikussá tétel két módon történhet: vagy kiszámítjuk mindkét viszonylatra, és átlagoljuk az értékeket, vagy módosítunk a képleten. Mi az utóbbit választottuk.

$$QDEV = \sum_{i=K}^{N} \frac{\min\left(T_i^a,\ T_i^b\right)\left(q_i^a - q_i^b\right)^2}{\frac{q_i^a + q_i^b}{2}} \qquad /8/$$

A /8/ kifejezés magyarázataként egyrészt megemlítendő, hogyha egy kockázatközösség létszáma alacsony, akkor abban a kockázatközösségben megbízhatatlan a becslés, tehát akár a nagy különbségeknek sincs nagy jelentősége; amit a $\min\left(T_i^a,\ T_i^b\right)$ szorzótényező fejez ki. Másrészt azért szerepel a nevezőben a halálozási valószínűségek átlaga, mert például egy 1 százalékos eltérés sokkal lényegesebb, ha a halálozási valószínűség értéke 1 százalék, mint ha 10 százalék. Harmadrészt, sok kis eltérés jobban tolerálható, mint egy nagy, így $\left(q_i^a - q_i^b\right)$ négyzetét vesszük.

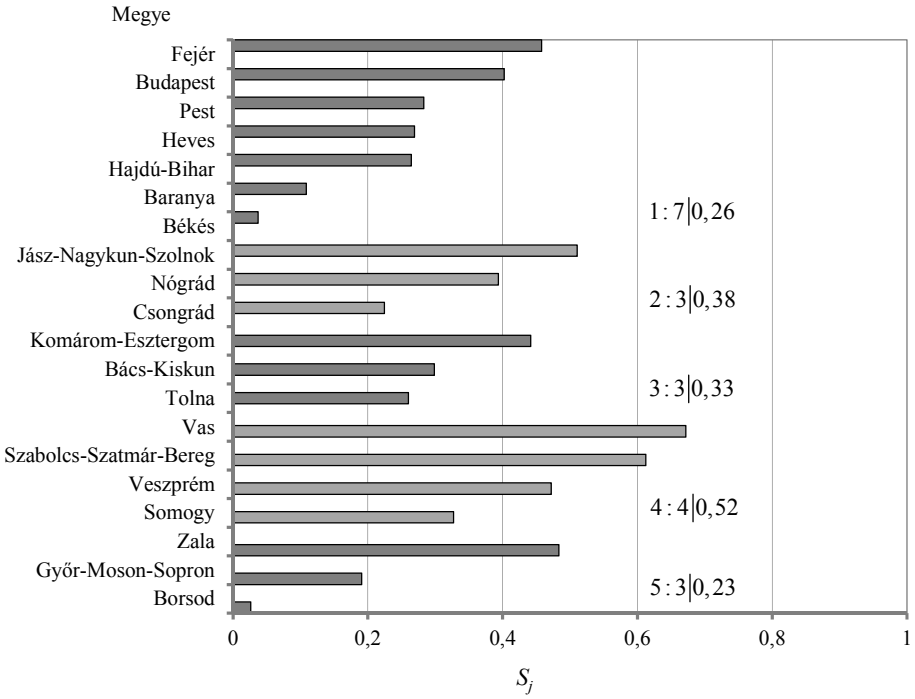
A /8/ képlet használata során a log-rank statisztikánál leírtakhoz hasonló módon járhatunk el. A lehetséges párokra kiszámítjuk a kifejezés értékét, és ezáltal egy hasonlósági mátrixot kapunk. E mátrixra már alkalmazható a K-medián vagy valamilyen hierarchikus módszer. A K-medián módszer alapján példánkban öt klaszter tűnik jó választásnak. (Lásd a 6. táblázatot és a 9. ábrát.)

6. táblázat

A K-medián eljárás eredménye öt klaszter és a QDEV hasonlósági mérték esetén

Klaszter száma	Reprezentáns	Klaszterbe tartozó megye
1	Fejér	Baranya, Békés, Budapest, Fejér, Hajdú-Bihar, Heves, Pest
2	Jász-Nagykun-Szolnok	Csongrád, Jász-Nagykun-Szolnok, Nógrád
3	Tolna	Bács-Kiskun, Komárom-Esztergom, Tolna
4	Vas	Somogy, Szabolcs-Szatmár-Bereg, Vas, Veszprém
5	Zala	Borsod-Abaúj-Zemplén, Győr-Moson-Sopron, Zala

9. ábra. Sziluettábra K-medián eljárással és QDEV hasonlósági mértékkel öt klaszter esetén



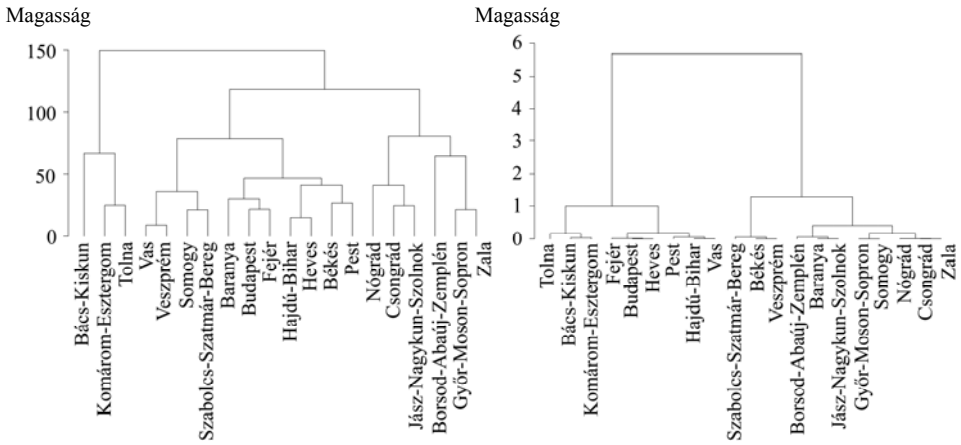
Megjegyzés. Az átlagos sziluettszélesség: 0,34.

A 10. ábra a legtávolabbi szomszéd elve alapján készült. Az összehasonlítás kedvéért a 10. *b*) ábrán szerepeltetjük ugyanezen módszer log-rank statisztikán alapuló eredményét is.

A 6. táblázat és a 10. ábra eredményeit értelmezve megállapítható, hogy észrevehető hasonlóság a kapott klaszterekben a két különböző hasonlósági mérték között. A K-medián módszer esetén például Bács-Kiskun, Komárom-Esztergom és Tolna megye alkotott egy csoportot; ez a csoport a 6. táblázatban is „feltűnik". A hierarchikus módszer esetén a 10. *a*) ábra bal szélén Bács-Kiskun, Komárom-Esztergom és Tolna megye hármasa kapcsolódik össze, a jobb szélén pedig Nógrád, Csongrád, Jász-Nagykun-Szolnok, Borsod-Abaúj-Zemplén, Győr-Moson-Sopron és Zala megye (nagyon hasonló a 3. táblázatban a 4-es klaszter). De vannak különbségek is: például a QDEV hasonlósági mértéken alapuló klaszterelemzés esetén Vas és Veszprém megye „nagyon gyorsan" közös klaszterbe kerül, míg a log-rank statisztikán alapuló klaszterelemzés esetén csak a legutolsó lépésben.

10. ábra. Dendrogram a legtávolabbi szomszéd elvének alkalmazása esetén

a) QDEV hasonlósági mérték alapján b) log-rank hasonlósági mérték alapján



Összességében a három hasonlósági mérték közül a log-rank statisztika adta a leginkább elkülönülő klasztereket, mind a sziluettábrák, mind a dendrogramok szerint. Eddigi elemzéseink alapján ennek a mérőszámnak a használatát javasoljuk életbiztosítási adatok esetében.

4. Összefoglalás

Tanulmányunkban bemutattuk, hogy miként lehet Magyarország megyéit a biztosítást vásárlók halandósági mintázatai alapján homogén csoportokba sorolni. Ismertettünk több lehetséges hasonlósági mértéket, és számba vettük a szóba jöhető klaszterelemzési eljárásokat. Magyarország megyéi jellemzően nem a földrajzi elhelyezkedés szerint sorolódnak csoportokba, sokkal inkább gazdasági fejlettségük szerint, amely meghatározó szerepet tölt be az ott lakó és biztosítást vásárló emberek életében. Ezt az indokolja, hogy többek között a lakosság halandósága is javul, ha jobbak az életkörülményeik.

A leggyakrabban használt K-középpontú eljárásnak korlátozottak a felhasználási lehetőségei ebben a konkrét feladatban, így azt csak összehasonlításul szerepeltettük. A K-medián módszer viszont számunkra kielégítő eredményt adott. A partícionáló eljárások mellett klasszikus és olyan újabbnak számító agglomeratív eljárásokat is kipróbáltunk, amelyekben az összevonás sorrendje jobban igazodik a valós távolságokhoz.

A különböző klaszterelemzési eljárások eredményei vagy hasonló csoportokat eredményeznek, vagy kielégítően magyarázhatók a közöttük tapasztalt különbségek.

A leírt módszerek alkalmazhatók például európai országok csoportosítására is. Ekkor azonban, ha országos adatokat használunk, az adatok nem az itt bemutatott formában érhetők el. Országos adatok esetén ugyanis egyéni életutak nyomon követésére csak ritkán van lehetőség, általában csak aggregált adatok állnak rendelkezésre. Ezen túl a használt hasonlósági mértékeket is újra kell gondolni: módosítani kell a meglévőket, esetleg újakat bevezetve. Országos adatok esetén jellemzően nem reprodukálható a Kaplan–Meier-féle becslés, de az aggregált adatoknál lehetőség van a túlélési függvény másfajta becslésére. A log-rank tesztstatisztikát is elő lehet állítani bizonyos korlátokkal. A halandósági valószínűségek viszont megbízhatóbbak országos adatokra, és ilyenkor a *QDEV* hasonlósági mérték is számítható.

Irodalom

ÁGOSTON K. CS. [2003]: Halálozási valószínűségek, illetve becslésük. In: *Banyár J.* (szerk.): *Életbiztosítás*. Aula. Budapest. 377–390. old.

ARATÓ, M. – BOZSÓ, D. – ELEK, P. – ZEMPLÉNI, A. [2009]: Forecasting and simulating mortality tables. *Mathematical and Computer Modelling*. Vol. 49. Issues 3–4. pp. 805–813. https://doi.org/10.1016/j.mcm.2008.01.012

BENJAMIN, B. – POLLARD, J. H. [1980]: *The Analysis of Mortality and Other Actuarial Statistics*. Heinemann. London.

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

COX, D. R. [1972]: Regression models and life-tables. *Journal of the Royal Statistical Society*. Series B. Vol. 34. No. 2. pp. 187–220.

CROCKER, K. J. – SNOW, A. [1986]: The efficiency effects of categorical discrimination in the insurance industry. *Journal of Political Economy*. Vol. 94. No. 2. pp. 321–344. http://dx.doi.org/10.1086/261376

CROCKER, K. J. – SNOW, A. [2000]: The theory of risk classification. In: *Dionne, G.* (ed.): *Handbook of Insurance*. Kluwer Academic Publishers. Boston, Dordrecht, London.

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

ELAVARASI, S. A. – AKILANDESWARI, J. [2014]: Survey on clustering algorithm and similarity measure for categorical data. *ICTACT Journal on Soft Computing*. Vol. 4. No. 2. pp. 715–722.

FU, W. – SIMONOFF, J. S. [2017]: Survival trees for left-truncated and right-censored data, with application to time-varying covariate data. *Biostatistics*. Vol. 18. No. 2. pp. 352–369. https://10.1093/biostatistics/kxw047

FÜSTÖS L. – KOVÁCS E. – MESZÉNA GY. – SIMONNÉ M. N. [2004]: *Alakfelismerés: Sokváltozós statisztikai módszerek*. Új Mandátum Könyvkiadó. Budapest.

FÜSTÖS L. – MESZÉNA GY. – SIMONNÉ M. N. [1986]: *A sokváltozós adatelemzés statisztikai módszerei*. Akadémiai Kiadó. Budapest.

GARCÍA, S. – LABBÉ, M. – MARÍN, A. [2011]: Solving large *p*-median problems with a radius formulation. *INFORMS Journal on Computing*. Vol. 23. No. 4. pp. 546–556. https://doi.org/10.1287/ijoc.1100.0418

GAUSS, C. F. [1900]: *Eine Ausgleichformel für Mortalitätsdaten*. Werke. Band 8. pp. 161–162. Königliche Gesellschaft der Wissenschaften zu Göttingen. Göttingen.

GILPIN, S. – NIJSSEN, S. – DAVIDSON, I. N. [2013]: *Formalizing Hierarchical Clustering as Integer Linear Programming*. Proceedings of the Twenty-Seventh AAAI (Association for the Advancement of Artificial Intelligence) Conference on Artificial Intelligence. AAAI Press. Palo Alto. pp. 372–378.

GREENWOOD, M. [1926]: The errors of sampling of the survivorship table. *Report on the Public Health and Medical Subjects*. Vol. 33. His Majesty's Stationery Office. London.

HAJDU O. [2003]: *Többváltozós statisztikai számítások*. Központi Statisztikai Hivatal. Budapest.

HALLEY, E. [1693]: An estimate of the degrees of the mortality of mankind, drawn from curious tables of the births and funerals at the city of Breslaw; with an attempt to ascertain the price of annuities upon lives. *Philosophical Transactions*. Vol. 17. pp. 596–610. https://doi.org/10.1098/rstl.1693.0007

HANSEN, P. – JAUMARD, B. [1997]: Cluster analysis and mathematical programming. *Mathematical Programing*. Vol. 79. Issues 1–3. pp. 191–215.

HARTIGAN, J. A. – WONG, M. A. [1979]: Algorithm AS 136: a K-means clustering algorithm. *Journal of the Royal Statistical Society. Series C*. Vol. 28. No. 1. pp. 100–108. https://doi.org/10.2307/2346830

HARTIGAN, J. A. [1975]: *Clustering Algorithms*. Wiley-Interscience. New York.

HUNYADI L. [2001]: *Statisztikai következtetéselmélet közgazdászoknak*. Központi Statisztikai Hivatal. Budapest.

HUSSAIN, T. – ASGHAR, S. [2016]: Chi-square based hierarchical agglomerative clustering for web sessionization. *Journal of the National Science Foundation of Sri Lanka*. Vol. 44. No. 2. pp. 211–222. http://doi.org/10.4038/jnsfsr.v44i2.8002

KAPLAN, E. L. – MEIER, P. [1958]: Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*. Vol. 53. No. 282. pp. 457–481. http://doi.org/10.2307/2281868

KAUFMAN, L. – ROUSSEEUW, P. J. [1990]: *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley Series in Probability and Statistics. New York.

KERÉKGYÁRTÓ GY. L. – BALOGH I. – SUGÁR A. – SZARVAS B. [2008]: *Statisztikai módszerek és alkalmazásuk a gazdasági és társadalmi elemzésekben*. Aula Kiadó. Budapest.

KLINGER A. [2007]: A halandóság társadalmi különbségei Magyarországon a XXI. század elején. *Demográfia*. 50. évf. 2–3. sz. 252–281. old.

KOVÁCS E. [2011]: *Pénzügyi adatok statisztikai elemzése*. IV. bővített kiadás. Tanszék Kft. Budapest.

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

KOVÁCS K. – Őri P. [2009]: Halandósági különbségek. In: *Monostori J. – Őri P. – S. Molnár E. – Spéder Zs*. (szerk.): *Demográfiai Portré 2009*. KSH Népességtudományi Kutatóintézet. Budapest. 53–66. old.

LLOYD, S. P. [1982]: Least squares quantization in PCM. *IEEE Transactions on Information Theory*. Vol. 28. No. 2. pp. 129–137. http://doi.org/10.1109/TIT.1982.1056489

MACQUEEN, J. B. [1967]: Some Methods for Classification and Analysis of Multivariate Observations. In: *Le Cam, L. M. – Neyman, J.* (eds.): *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press. Berkeley. pp. 281–297.

MAJSTOROVIC, S. – SABO, K. – JUNG, J. – KLARIC, M. [2018]: Spectral methods for growth curve clustering. *Central European Journal of Operations Research*. Vol. 26. No. 3. pp. 715–737. http://doi.org/10.1007/s10100-017-0515-6

OKADA, A. [2000]: An asymmetric cluster analysis study of car switching data. In: *Gaul, W. – Opitz, O. – Schader, M*. (eds.): *Data Analysis*. Studies in Classification, Data Analysis, and Knowledge Organization. Springer. Heidelberg. pp. 495–504.

OLSZEWSKI, D. [2011]: Asymmetric *k*-means algorithm. In: *Dobnikar, A. – Lotrič, U. – Šter, B*. (eds.): *Adaptive and Natural Computing Algorithms*. International Conference on Adaptive and Natural Computing Algorithms 2011: Lecture Notes in Computer Science. Vol. 6594. Springer. Berlin, Heidelberg. pp. 1–10.

PETO, R. – PETO, J. [1972]: Asymptotically efficient rank invariant test procedures. *Journal of the Royal Statistical Society*. *Series A*. Vol. 135. No. 2. pp. 185–207.

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

SABO, K. – SCITOVSKI, R. – VAZLER, I. [2013]: One-dimensional center-based l_1–clustering method. *International Journal of Applied Mathematics and Computer Science*. Vol. 24. Issue 1. pp. 151–163. https://doi.org/10.2478/amcs-2014-0012

SANGALLI, L. M. – SECCHI, P. – VANTINI, S. – VENEZIANI, A. [2009]: A case study in exploratory functional data analysis: geometrical features of the internal carotid artery. *Journal of the American Statistical Association*. Vol. 104. Issue 485. pp. 37–48. https://doi.org/10.1198/jasa.2009.0002

SANGALLI, L. M. – SECCHI, P. – VANTINI, S. – VITELLI, V. [2010]: K-mean alignment for curve clustering. *Computational Statistics and Data Analysis*. Vol. 54. Issue 5. pp. 1219–1233. https://doi.org/10.1016/j.csda.2009.12.008

SIMON J. [2006]: A klaszterelemzés alkalmazási lehetőségei a marketingkutatásban. *Statisztikai Szemle*. 84. évf. 7. sz. 627–651. old.

VÉKÁS P. [2011]. Túlélési modellek. In: *Kovács E*. (szerk.): *Pénzügyi adatok statisztikai elemzése*. *IV. bővített kiadás*. Tanszék Kft. Budapest. 173–194. old.

VINOD, H. D. [1969]: Integer programming and the theory of grouping. *Journal of the American Statistical Association*. Vol. 64. Issue 326. pp. 506–519.

Summary

The analysis of mortality data has a long history in the field of mathematical statistics. Multiple methods can be used to analyse or test the equality of survival probabilities in two or more populations. However, less attention has been paid so far to the classification of risk pools into homogeneous groups. The study presents several methods for clustering mortality data and testing them on real datasets.