Tóth Richárd

Klasszifikációs módszerek alkalmazásának lehetőségei az ellenőrzésekben

A tanulmányom célja az Állami Számvevőszék feladatellátásának klasszifikációs módszerekkel történő támogatási lehetőségeinek elemzése, alkalmazhatóságának vizsgálata. Ezek a klasszifikációs módszerek többféle ellenőrzés – beleértve a számvevőszéki ellenőrzéseket és elemzéseket is *– során nyújthatnak támogatást, hozzájárulva az értékteremtéshez.*

Kulcsszavak: klaszterelemzés, diszkriminancia-analízis, klasszifikációs módszerek, rétegzett mintavétel, számvevőszéki ellenőrzés
Jelkód: C38, H83

Kutatás célja

Az ellenőrzésekben a mintavételi eljárásokat egyes klasszifikációs módszerekkel – elsősorban klaszterelemzéssel – lehetne támogatni. Kutatásom során ezen módszereknek kizárólag az ÁSZ feladatellátásában történő alkalmazási lehetőségeit vizsgáltam. A klasszifikációs módszerek alkalmazásával az ÁSZ egyes ellenőrzéseit több szakaszban és elemzéseit is lehetne fejleszteni, ami által javulhatna az ÁSZ elemzéseiben és jelentéseiben megfogalmazott következtetések statisztikai megalapozottsága, amivel az ÁSZ még jobban hozzájárulhatna a „jó kormányzáshoz". Az előbbiekben írtakon túl a témaválasztás aktualitását az ÁSZ helyi önkormányzatokat érintő ellenőrzéseinek magas aránya szolgáltatta.

Az ÁSZ Intézményi Stratégiájában stratégia célként rögzítésre került, hogy a Számvevőszék[1] *új megközelítésű, elemzéssel alátámasztott mintavétellel* és ellenőrzési eljárásokkal törekszik a helyszíni ellenőrzések számának csökkentésére (ÁSZ, 2010).

A kutatás célja a fentiek gyakorlati példán – a gyermekjóléti alapellátások helyi önkormányzatok általi 2015. évre vonatkozó feladatellátásán – keresztüli bemutatása és javaslat megfogalmazása volt az ÁSZ tevékenységei során történő alkalmazására vonatkozóan, ami meggyőződésem szerint hozzáadott értéket teremtene.

Nemzetközi kitekintés

Az INTOSAI[2] által 2016-ban jóváhagyott ISSAI 4000 *Compliance Audit Standard* a teljesítmény-ellenőrzésekhez kapcsolódó követelmények között rögzíti, hogy a megfelelő és elegendő bizonyosság elérése érdekében ahol szükséges, mintavétel alkalmazható. Az irányelv tartalmazza azt is, hogy *a mintavételi eljárások megtervezésekor az ellenőrzés célját és a mintavétel alapjául szolgáló sokaság jellemzőit is figyelembe kell venni*. Az INTOSAI által 2013-ban jóváhagyott ISSAI 1530 *Audit Sampling* irányelv részletesen foglalkozik az ellenőrzésben alkalmazható mintavételi eljárásokkal a pénzügyi ellenőrzésekhez kapcsolódóan. Az irányelv fogalmai között megtalálható a rétegzés is.[3] Magasabb rendű ISSAI standardok a mintavétel témakörét csak általánosan érintik, mint az ellenőrzésben lehetségesen alkalmazandó eszköz. (pl. ISSAI 200, 300, 400 stb.) Az osztályozási módszerek alkalmazásával kapcsolatosan egyik ISSAI standard sem tesz említést (INTOSAI, 2001-2016).

[1] Állami Számvevőszék
[2] Legfőbb Ellenőrzési Intézmények Nemzetközi Szervezete
[3] Az egyes rétegekben szereplő egyedek hasonló jellemzőkkel kell, hogy rendelkezzenek. A rétegzett mintavételi eljárással kapcsolatban részletesen az irányelv 1. sz. melléklete tartalmaz iránymutatást.

Míg az ECA[4] 2015-ben közzétett Teljesítmény-ellenőrzési Kézikönyve kizárólag a lehetséges adatgyűjtési módszerek között említi a mintavételi eljárásokat, addig a 2012-ben publikált Pénzügyi és Szabályszerűségi Ellenőrzési Kézikönyv több fejezetben is résztelesen foglalkozik a mintavétellel. Osztályozási módszerek alkalmazásával kapcsolatban az ECA honlapján nyilvánosan hozzáférhető dokumentumokban nem található utalás (ECA, 2012, 2015).

A NAO[5] 2001-ben jelentette meg honlapján rövid mintavételi útmutatóját, míg a GAO[6] már 1992-ben publikálta a részletes mintavételi eljárások alkalmazására vonatkozó útmutatását (GAO, 1992). Sem az egyesült királyságbeli-, sem az amerikai egyesült államokbeli legfőbb pénzügyi ellenőrző szerv honlapján nem található klasszifikációs módszerek alkalmazásával kapcsolatos nyilvánosan elérhető dokumentum.

A helyi önkormányzatok kötelező feladatait a Mötv.[7] II. fejezetének 13.§ (1) bekezdése rögzíti, mint helyi közügyek, valamint a helyben biztosítható közfeladatok körében ellátandó helyi önkormányzati feladatok, amelybe beletartoznak a gyermekjóléti szolgáltatások is. A gyermekvédelemre vonatkozó alapvető ágazati szabályokat a Gyvt.[8] tartalmazza, aminek a 15. § (2) bekezdése határozza meg a gyermekjóléti alapellátásokat. A települési önkormányzat a személyes gondoskodást nyújtó alapellátások keretében a gyermekjóléti szolgáltatást, a gyermekek napközbeni ellátását és a gyermekek átmeneti gondozását biztosítja, illetve szervezi a máshol igénybe vehető ellátásokhoz való hozzájutást.

Szakirodalmi összefoglaló

Rétegzett mintavétel

A rétegzett mintavételi eljárnál az egyes részsokaságok jellemzőire is szeretnénk következtetéseket megfogalmazni – nem kizárólag a teljes sokaságra vonatkozóan, mivel a módszer megköveteli az alapsokaság valamennyi rétegének mintabeli reprezentációját (Szilágyi, 2011). Fontos előfeltétel a módszer alkalmazásánál, hogy rendelkezzünk előzetes ismeretekkel arra vonatkozóan, hogy a sokaságot – a vizsgált ismérv szempontjából – hogyan lehet homogén – vagy legalábbis kevésbé heterogén – osztályokba sorolni. Ezt a vizsgált ismérvet *rétegképző ismérv*nek is szokás nevezni. A rétegképző ismérvnek a vizsgált jellemzővel szoros kapcsolatban kell lennie (Malhotra, 2007). Megfelelően alkalmazott rétegzéssel és elosztással a reprezentatív megfigyelés hatékonysága növelhető. A rétegzés során szem előtt kell tartani, hogy a vizsgált ismérv szempontjából egynemű elemek azonos rétegekbe kerüljenek (Szilágyi, 2011). Amint azt a későbbiekben látni fogjuk, ebben nyújthat segítséget a klaszterelemzés.

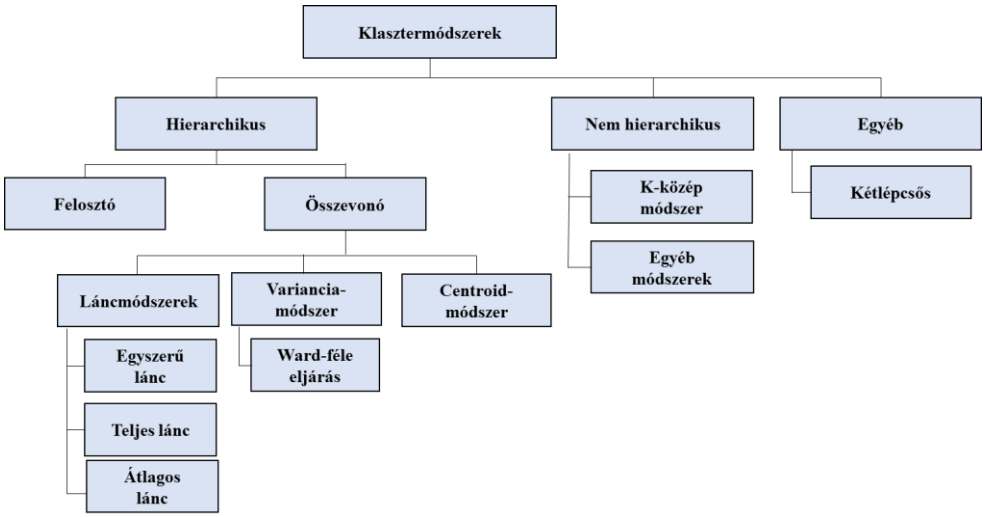
Klasszifikációs módszerek

Az objektumok osztályozása, csoportosítása a többváltozós statisztikai vizsgálatok egy jellegzetes területe (Móri–Székely, 1986). Mind a diszkriminancia-analízis, mind a klaszterelemzés csoportképző-, vagyis osztályozási módszer. A két eljárás között alapvető különbség, hogy míg a diszkriminancia-analízis alkalmazásának a feltétele, hogy rendelkezésre álljanak előzetes ismeretek a csoportokba tartozást illetően, addig a klaszterelemzésnek ez nem feltétele, mivel az eljárás különféle módszerek segítségével előzetes feltételezések nélkül végzi el az adott csoportba sorolást. Ha a csoportokat, osztályokat, klasztereket rétegeknek tekintjük, akkor könnyen észrevehető az osztályozási módszerek hasonlósága a rétegzett mintavétellel.

[4] Európai Számvevőszék
[5] Brit Számvevőszék
[6] Egyesült Államok Számvevőszéke
[7] Magyarország helyi önkormányzatairól szóló 2011. évi CLXXXIX. törvény
[8] A gyermekek védelméről és a gyámügyi igazgatásról szóló 1997. évi XXXI. törvény

Klaszterelemzés

A klaszterelemzés[9] olyan többváltozós módszer, amelyeknek alkalmazásával csoportok (klaszterek) hozhatók létre. Az elemzés végén az analízisbe bevont változók alapján a hasonló megfigyelési egységeket, mint a sokaság egyedeit, homogén csoportokba, klaszterekbe soroljuk be oly módon, hogy minden egyed hozzárendelésre kerül egy klaszterhez és pontosan egyhez. Akkor eredményes az eljárás, ha a folyamat végén minden egyed pontosan egy klaszterbe kerül besorolásra, az egy klaszterbe tartozó egyedek hasonlítanak egymáshoz, azonban különböznek a más csoportba tartozó egyedektől (Malhotra, 2007). A klaszterelemzés alapvető feladata megtalálni azokat a változókat, amelyek a csoportok közötti különbözőséget okozzák (Besenyei et al., 2007). Többféle módszer létezik arra vonatkozóan, hogy az egyes egyedek hogyan vonhatók össze – ezeket a szakirodalom hierarchikus vagy nem hierarchikus összevonó eljárásoknak[10] nevezi – (1. ábra).

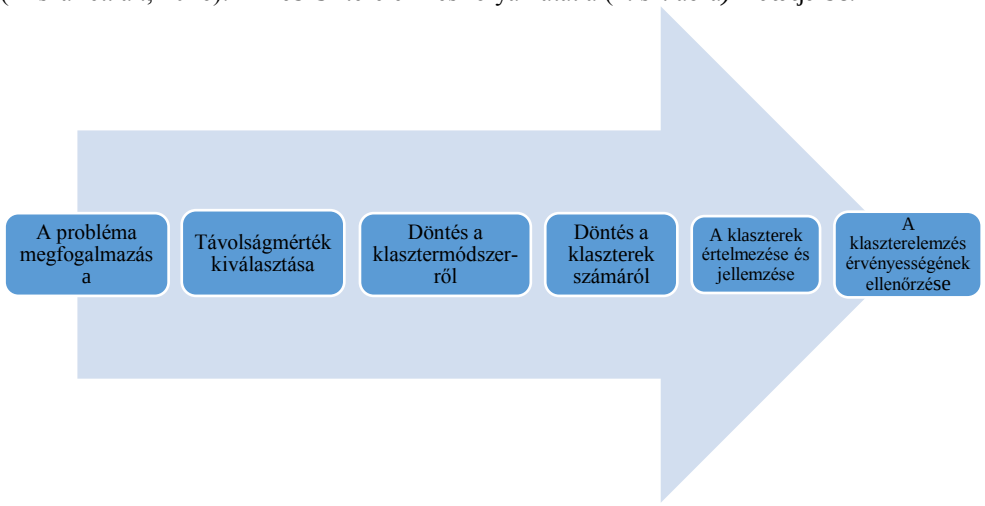


1. ábra: A klasztermódszerek csoportosítása
Grouping cluster analysis methods
Forrás: Malhotra, 2007. 643. oldala alapján saját szerkesztés

Nagyobb esetszámnál[11] bonyolultabb a hierarchikus klaszterezés elvégzése, ezért ilyen esetben javasolt a nem hierarchikus – vagy dinamikus – eljárásokba tartozó K–közép módszert alkalmazni. Ennek a módszernek az alkalmazásakor a kialakítandó klaszterek számát előre meg kell határozni, majd a kiinduló klaszterek fokozatos, lépésenkénti, azaz iterációs módosulásával jön létre a végső K számú klaszter az eredetileg feltételezett üres klaszterekhez tartozó klaszterttközéppontokból (Ketskeméty, 2011). A klaszteren belüli szórások alapján, valamint a Wilks' lambda kiszámításával ítélhető meg a klaszterek homogenitása, míg a klaszterek egymáshoz való viszonya a centroidjaik közötti távolságmátrixok alapján jellemezhető (Hajdú, 2003). A nem hierarchikus klaszterezés[12] kevésbé függ a hasonlóság- és távolságmértékektől, valamint a kiugró értékektől és amellett, hogy gyorsabb, megbízhatóbb eredményt is ad.[13] Míg a hierarchikus el-

[9] Klasszifikációs elemzésnek és numerikus taxonómiának is nevezik.(Malhotra, 2007)
[10] Ismert még a kétlépcsős eljárás, amely automatikusan megállapítja a klaszterek optimális számát (Malhotra, 2007).
[11] Ez 2 000 feletti megfigyelési egységnek felel meg, vagyis nagyobb adathalmaznak.
[12] A nem hierarchikus eljárásokat a szakirodalom partícionáló módszereknek is nevezi.
[13] Részletesebben lásd Sajtos–Mitev (2007), a 325-326. oldalakon.

járás elsősorban feltáró, addig a nem hierarchikus klaszterezés megerősítő elemzésre szolgál (Fliszár et. al., 2016). A klasz9+terelemzés folyamatát a (2. sz. ábra*)* mutatja be.



2. ábra: A klaszterelemzés lépései
Steps of cluster analysis
Forrás: Malhotra, 2007. 642. oldala alapján saját szerkesztés

Olyan változókat célszerű választani a klaszterezés alapjául, amelyek között gyenge a kapcsolat, ellenkező esetben torzításhoz vezethet. A klaszteranalízis előtt az adatokat standardizálni szükséges az összehasonlíthatóság érdekében, amennyiben az egyes változók értékei eltérő skálán szerepelnek az elemzés alapjául szolgáló adatbázisban.

Fontos lépés a kiugró értékek kezelése, amely történhet az elemzés közben, vagy már az elemzéshez felhasználandó adatbázis tisztítása alkalmával is kiszűrhetők az outlierek, mivel a klaszterelemezés érzékeny ezekre. A kiugró adatok kezelésén túlmenően a klaszteranalízis többi alkalmazhatósági feltételeinek vizsgálatára is különös figyelmet kell fordítani az elemzés teljes folyamatában, mint pl. a minta reprezentativitása, a konstans értékű ismérvek kiszűrése, és, az elemzésbe bevont változók korrelációja[14] (Sajtos–Mitev, 2007). Ezt követően meg kell határozni a hasonlóság- és távolságértékeket[15], valamint a klaszterek számát. Itt szeretném felhívni a figyelmet, hogy ugyanúgy, mint a rétegzett mintavételi eljárás esetében a rétegek száma, klaszterelemzés esetében a kialakítandó klaszterek száma hasonlóan – a mintavételt, illetve a klaszteranalízist végzők által – manuálisan történik.

Amennyiben nem az optimális klaszterösszetételt kapjuk az eljárás végén, akkor az elemzést másik módszerekkel is célszerű elvégezni, kombinálva a klaszterek számának változtatásával. A klaszterelemzés végén tudunk a kialakított klaszthereknek nevet adni és elemezni tudjuk az egyes klasztereket. (dr. Varga - Szilágyi, 2011) A klaszterelemzés egyik megoldatlan problémája, hogy a kialakuló klaszterek nem függetlenek az egyedek adatbázisban elfoglalt sorrendjétől. A klaszterelemzés további hátrányaként ismert, hogy minden esetben feltár csoportosítási struktúrákat, még akkor is, ha azok nem valósak. Emiatt az elemzés előkészítésére kiemelt figyelmet kell fordítani. Törekedni kell arra, hogy csak *olyan ismérvek vegyenek részt a csoportosításban, amelyek a vizsgálat tárgya és tartalmi kerete szempontjából relevánsnak tekinthetőek*. Ez azt is

[14] Nagyon erős korrelációról 0,9, vagy e feletti érték esetén beszélhetünk két ismérv esetében.
[15] Lásd részletesen (Sajtos–Mitev, 2007, 285. és 290-294. oldalak.) A Mahalanobis-távolság használata alkalmas a változók közötti korreláció kiszűrésére is (Simon, 2006).

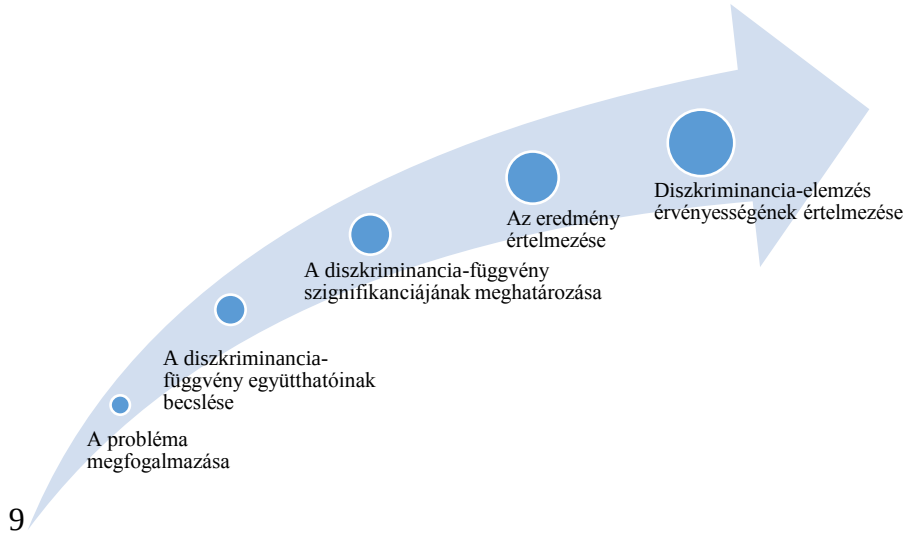
jelenti, hogy a vizsgált jelenség szempontjából jelentéktelennek minősíthető ismérveket a csoportosítási eljárásból ki kell hagyni (Simon, 2006).

Újabb változók bevonása az analízisbe vagy már meglévő változók törlése az adatbázisból jelentősen befolyásolja a klaszterelemzés eredményét. Mivel nem csak egyetlen legjobb megoldás létezhet, ezért a végeredmény attól is függ, hogy az egyes lépésekben milyen módszert választunk. Az eltérő eljárások, valamint a számos hasonlósági- és távolságmértékek különböző felosztásokat eredményeznek (Kovács, 2014). Emiatt célszerű többféle módszerrel is elvégezni a klaszterelemzést egy adott adatbázison. Mivel gyakran használják a klaszterelemzést adatcsökkentéshez, piacszegmentáláshoz, ezért gondoltam, hogy alkalmas lehet a magyarországi települések fontosabb jellemzőit tartalmazó adatbázis alapján az egyes települések klaszterekbe sorolására, elemzésére, amely inputot szolgáltathat a helyi önkormányzati rendszer egészére következtetést megállapító, kivetítő, annak megalapozásául szolgáló rétegzett mintavételi eljáráshoz[16], [17].

Diszkriminancia-analízis

A diszkriminancia-analízis olyan klasszifikációs és adatelemzési módszer, ahol a függő változó többnyire kategorizált, nominális skálán mért – azaz nem metrikus –, a független változó pedig minimum intervallum skálán mért változó [18]. Olyan többváltozós módszer, amely alkalmazásával a független változók egy csoportja alapján prognosztizálható a csoportokba tartozás. A módszer segítségével meghatározhatók azok a független változók, amelyek a függő változók szignifikáns különbségét okozzák (Sajtos–Mitev, 2007). Az elemzés információt szolgáltat arra vonatkozóan, hogy a csoportok jelentősen különböznek-e egymástól, és ha igen, akkor azt melyik változók okozzák.

A diszkriminancia-elemzésnek két típusát különböztetjük meg: A kétcsoportos- és a többcsoportos diszkriminancia-elemzést. Az elemzés lépéseit a (3. sz. ábra) mutatja be.



3. ábra: A diszkriminancia-analízis menete
The course of discriminant analysis
Forrás: Malhotra, 2007. 587. oldala alapján saját szerkesztés

[16] (Csicsman, 1979) ezt a hipotézisemet megerősítette.
[17] Ennek az a feltétele, hogy biztosítani kell, hogy a mintavétellel nyert adatbázisban minden egyes csoportban megfelelő mennyiségű elem legyen ahhoz, hogy az alapsokaságban megfelelően reprezentáltak legyenek az egyes alsokaságok (Simon, 2006).
[18] A független változó jelölésére x-et, míg a függő változó jelölésére y-t használtam.

A minta minden egyes egyedére vonatkozóan rendelkezésre kell, hogy álljanak a független változók. Emellett legalább annyira fontos, hogy *olyan függő és független változókra szükséges alkalmazni, amelyeknek van értelme, célja*. A független változóknak metrikus – számmal kifejezett, azaz mérhető – változóknak kell lennie, azonban elegendő számú mennyiségi változó esetében a diszkriminancia-analízist legalább öt kategóriával rendelkező ordinális skálán mért változókra, vagy dichotóm változókra is lehet alkalmazni. A függő változónak kategorizált változónak kell lennie és legalább kettő ismérvváltozattal kell rendelkeznie. Életszerű, hogy a változók alapból nem kategorizáltak, hanem pl. ordinális skálán mértek, ekkor az egyértelmű, következetes bekategorizálást az elemzés előtt el kell végezni. Ebben az esetben figyelni kell arra, hogy az egyes kategóriák hasonló méretűek legyenek, mert ellenkező esetben az elemzés eredményei torzíthatnak.

A fentieken túl érdemes szem előtt tartani, hogy minél nagyobb a minta elemszáma, annál megbízhatóbb az elemzés. A mintanagyságnak célszerű legalább négyszer nagyobbnak lenni a független változók számánál.[19] Az adatok függetlenségét valamint a csoportok kizárólagosságát is biztosítani kell. Utóbbi azt jelenti, hogy a függő változó valamennyi megfigyelési egységét be kell sorolni valamelyik csoportba úgy, hogy minden megfigyelési egységnek valamelyik csoport tagjának kell lennie. A csoportok nagyságának is hasonlónak kell lenni, továbbá minden csoportnak legalább két esetet kell tartalmaznia.[20]

Az együtthatók becslését tekintve két módszert különböztetünk meg, a közvetlen módszert és a lépcsőzetes diszkriminancia-módszert. A várható értékek és a kovariancia-mátrixok becslését követően a paraméterek értékei meghatározhatóak. A súlyok, illetve az együtthatók becslésénél kulcsfontosságú, hogy a lehető legjobban eltérő csoportokat hozzuk létre a diszkriminancia-függvény értékei alapján. (Hajdu, 2003) Az együtthatók becslése után felírható a diszkriminancia-függvény, – a diszkriminancia-elemzés modellje – ami az x_1, x_2,…, x_p változók egy olyan függvénye, ami az m csoport szétválasztására a legjobban alkalmas. Ez megvalósítható a változók lineáris kombinációval való közelítéssel, az alábbiak szerint:

$$z = b_1x_1 + b_2x_2 + b_px_p$$ [21]

Amennyiben az adatok nem standardizáltak, akkor a fenti képlet jobb oldala kiegészül egy további konstans taggal (Sajtos–Mitev, 2007). A nagyobb diszkriminancia-koefficienssel rendelkező független változók differenciáló képessége jobb. A diszkriminancia-függvény akkor értelmezhető, ha a kritikus szignifikanciaszint 5 % alatti. Több lehetséges lineáris kombináció is létezhet, amiből következik, hogy több diszkriminancia-függvény is felírható. A csoportosítás becsléséhez az egy adott megfigyeléshez tartozó változóértékeket be kell helyettesíteni a függvényekbe. Ez követően az adott megfigyelést abba az osztályba soroljuk, amelyre a legnagyobb függvényértéket kaptuk (dr. Varga–Szilágyi, 2011).

[19] A diszkriminancia-analízis alkalmazásának további feltétele a többváltozós normális eloszlás. Kiemelt feladat a kiugró értékek kezelése és kiszűrése, mivel azok torzíthatják az átlagot, a szórást és a megbízhatóságot. A homoszkedaszticitásnak, a multikollinearitásnak és a független változók linearitásának is teljesülnie kell.

[20] A diszkriminancia-analízis alkalmazásának feltételeit részletesen lásd (Sajtos–Mitev, 2007) a 332-336. oldalakon.

[21] Ahol z a diszkriminancia-változó, b pedig a diszkriminancia-koefficiens.

A kutatás módszere

Adatbázisok összeállítása

Az elemzés adatbázisául egyrészt a KSH[22] honlapjáról a *Kiadványok* között található Excel (.xls) formátumban. letölthető *Helységnévtár*[23] „*Helységek 2016. 01. 01*" elnevezésű munkafüzete szolgált.[24]

Ehhez a 3 179 önkormányzatot tartalmazó adatbázishoz hozzáfűztem a KSH honlapjáról letölthető a „*Tájékoztatási adatbázis*"-on belül a „*Területi statisztikák*" között nyilvánosan hozzáférhető „*Éves településstatisztikai adatok 2015-ös településszerkezetben*" elnevezésű adatbázisban lekérdezhető, az „*Egészségügy, szociális ellátás*" kategórián belül a „*Gyermekjóléti alapellátások*" alkategóriában szereplő alábbi változókat: „*Bölcsődébe beírt gyermekek száma*", „*Bölcsődék száma*" és „*A gyermekjóléti szolgálat által gondozott kiskorúak száma*", valamint a „*Terület, lakónépesség, állandó népesség*" kategórián belül a „*Terület*" alkategóriában található „*A település területe (km^2)"* mutatókat tartalmazó Excelben exportált adatbázist a 2015. évre vonatkozóan. Tekintettel arra, hogy a 23 db fővárosi kerület esetében nem álltak rendelkezésre a fenti adatok, ezért azokat az adatbázisból töröltem. Az adatbázisból a Budapesthez – mint fővároshoz tartozó adatok is törlésre kerültek, mivel a klaszterelemzés során ez minden bizonnyal külön klasztert képezne és ez által kiugró értéknek minősülne. Az így összeállított adatbázis 3 154 önkormányzathoz tartozó település adatait tartalmazza – a fővárosi kerületek és a főváros nélkül.

Az elemzéshez rétegzett mintavételi eljárással a *Helységnévtár*ból (mint alapadatbázisból, illetve sokaságból) levett 70 elemű mintát is felhasználtam kiindulási adatbázisként tesztelési céllal, amelyben a változók törlése és átalakítása az előzőekben felsoroltakhoz hasonlóan történt. Ezt az adatbázist kiegészítettem további két bináris változóval, amely arra vonatkozik, hogy a gyermekjóléti alapellátások közül legalább egyik esetében társulás érintett volt-e a feladat ellátásában a 2015. évben, illetve, hogy a gyermekek napközbeni ellátásában társulás érintett volt-e a 2015. évben. Az adatok forrásai az egyes települések honlapjáról elérhető nyilvános dokumentumok és információk, valamint a MÁK[25] Törzskönyvi nyilvántartásában szereplő dokumentumok voltak. A bináris változókat a klaszterelemzésbe nem vontam be, de a klaszterek jellemzésekor további információval szolgáltak a feladatellátás tekintetében.

A létrehozott 64, illetve 3 154 település adatait tartalmazó adatbázisok alapján végeztem el a klaszterelemzést és a diszkriminancia-analízist a 23. verziószámú SPSS® támogatásával. Mindkét adatbázis esetében a klaszterelemzések során azonosítani tudtam kiugró értékeket, melyek eliminálásával az azokat már nem tartalmazó csökkentett elemszámú, 62, illetve 3 148 települést tartalmazó adatbázisokon hajtottam végre ismételten a különféle klaszterezési eljárásokat. A klaszterelemzéstől azt vártam, hogy segítségével a gyermekjóléti alapellátásokra fókuszálva, a bevont metrikus változók segítségével kialakíthatóak lesznek olyan klaszterek, amelyek alapján az optimális rétegzés megvalósítható lenne.

[22] Központi Statisztikai Hivatal

[23] A fájl neve: hnt_letoltes_2016.xls.

[24] Az adatbázis tartalmazza többek között a fővárossal és a fővárosi kerületekkel együtt összesen 3 179 hazai település, önkormányzat és kerületi önkormányzat jogállását, területét hektárban, lakosságszámát, lakások számát és a településen található helyi nemzetiségi önkormányzatra és a közös önkormányzati hivatalra vonatkozó adatokat.

[25] Magyar Államkincstár

A kutatás eredményei

Klaszterelemzés a 62 települést tartalmazó adatbázis használatával
Elsőként a 62 települést tartalmazó előzetesen összeállított, majd újabb változókkal kiegészített adatbázis alapján végeztem el a klaszterelemzést *A gyermekjóléti szolgálat által gondozott kiskorúak száma (fő)* és a *bölcsődébe beírt gyermekek száma (fő)* metrikus változókkal[26] a 2015. évre vonatkozóan a hierarchikus módszerek közé tartozó Egyszerű láncmódszerrel, Átlagos láncmódszerrel és Ward-eljárással távolságmértéknek négyezetes euklideszi távolságot választva, végül a Centroid-módszerrel.

Az elemzések eredményeként optimális klaszterszámnak kettő adódott, amely meghatározásához az SPSS®-ben kimenetként készült a koefficienseket tartalmazó táblázatokat és az azok alapján készített vonaldiagramokat, a dendogramokat és a jégcsapdiagramokat is felhasználtam. Végeredményként az egymástól heterogén, de egy adott klaszteren hasonló tulajdonsággal rendelkező településeket tartalmazó csoportok jöttek létre a klaszterelemzésben bevont változók alapján.

A két-klaszteres megoldás megbízhatóságáról a K-közepes – nem hierarchikus – klaszterezési eljárás lefuttatásával győződtem meg, a létrehozni kívánt klaszterek számának kettőt adtam meg. A települések egyes klaszterekbe történő besorolása mindegyik módszerrel elvégzett klaszterelemzés esetében megegyezett.[27] Az egyes klaszterek elemszámát tartalmazó táblázat is megerősíti, hogy homogén csoportok jöttek létre, mivel az elemzésbe bevont metrikus változók esetében mindkét klaszternél a szórások és az átlagok jelentősen eltértek. Az 1. klaszterbe tartozó települések mindegyikénél *A gyermekjóléti szolgálat által gondozott kiskorúak száma (fő)*, mind a *Bölcsődékbe beírt gyermekek száma (fő)* átlag alatti, míg a 2. klaszterbe tartozó települések esetében átlag feletti volt a 2015. évben Az 1. klaszter centroidjának (átlagának) koordinátái 77,73 és 31,00, míg a 2. klaszteré 482,67 és 257,33 voltak.

Az SPSS®-ben több beépített funkció is rendelkezésre áll az egyes klaszterek további jellemzésére, mint pl. a kereszttáblás elemzés, az ANOVA-tábla, a jelentéskészítés az esetek összegzéséről, az átlagok összehasonlítása és a diszkriminancia-analízis[28] (IBM, 2015). Nem metrikus változó vizsgálatakor a klaszterek jellemzésére kereszttáblás elemzést hajtottam végre. Pl. a *Helyiségek jogállása* és a *Lakónépesség a Gyvt. 94.§ (e) bekezdése a)-d) pontjai alapján* nominális-, valamint a csoportképző változóra (CLU2_2) végrehajtott a kereszttáblás elemzés[29] eredményeként adódott, hogy az 1. klaszterbe került települések 45,8 %-a község, 45,5 %-a város, az itt szereplő települések 66,1 %-ának lakosságszáma 10 000 fő alatt, 27,1 %-ának pedig 10 000 és 20 000 fő között volt a 2015. évben. A 2. klaszterben található települések 66,7 %-a megyei jogú város, 33,3 %-a város és az itt található települések 66,7 %-ának a lakosságszáma 40 000 fő felett volt 2015-ben. A fentieken túl a 2. klaszterekben található települések mindegyike, míg az 1. klaszterben szereplő települések 52,5 %-a rendelkezett bölcsődével a 2015. évben. Az 1. klaszterbe szereplő települések 15,3 %-a – ez 9 település – látta el társulással kötött megállapodás útján a gyermekek napközbeni ellátását a vizsgált esztendőben. Az esetek összegzéséről készített jelentésben[30] további hasznos információhoz jutottam a kialakult klaszterek jel-

[26] A klaszterelemzésbe A gyermekjóléti szolgálat által gondozott kiskorúak száma (fő) és a Bölcsődébe beírt gyermekek száma (fő) metrikus változókat vontam be, melynek oka, hogy e két változó esetében volt a leggyengébb a korreláció a vizsgált, relevánsnak tekinthető változók között.

[27] Pécs és Szeged településeket, mint azonosított kiugró értékeket a klaszterelemzésbe nem vontam be.

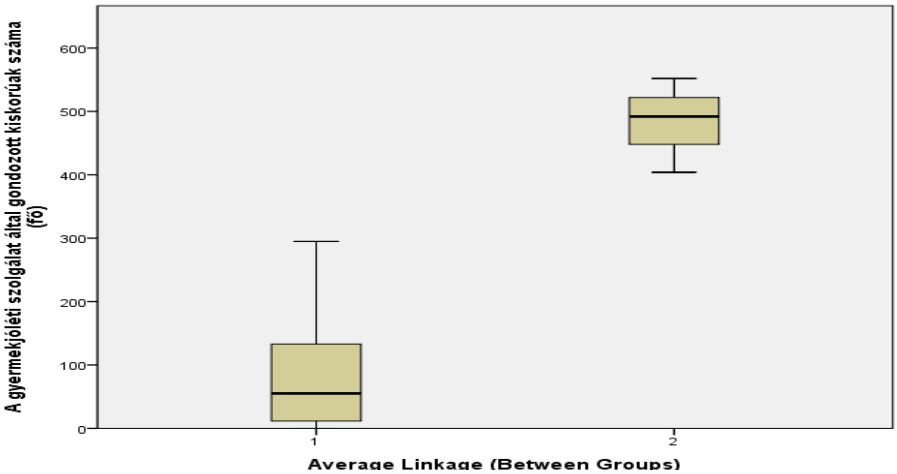
[28] Az IBM® SPSS®–szel közvetlenül nem, – a Kétlépcsős klaszterezés kivételével - viszont a RopStat® statisztikai programcsomag használatával lehetséges mérni a létrejött klaszterstruktúrák jóságát többféle klaszteradekvációs mutató (pl. EESS%, klaszterhomogenitási együtthatók,pontbiszeriális korreláció, Silhouette-mutató, Xie-Beni index) segítségével (Takács et. al., 2015).

[29] IBM® SPSS® Statistics-ben az Analyze->Descriptive Statistics->Crosstabs paranccsal hajtható végre.

[30] IBM® SPSS® Statistics-ben az Analyze->Reports ->Case Summaries paranccsal futtatható le. Az egyes klaszterekbe tartozó települések, zárójelben az egyes klaszterekbe tartozó települések száma:

lemzését illetően. Példaként említhető, hogy az 1. klaszterben szereplő települések lakosainak a száma átlagosan 7 502 fő, míg a 2. klaszterben található települések esetében 46 352 fő volt a 2015. évben.

A létrejött klaszterek vizuálisan megjeleníthetőek az SPSS®-ben beépített különféle diagramok segítségével is. A *Boxplot* (más néven doboz diagram) a leíró statisztikai elemzés szemléletes kiegészítője (Jánosa, 2011). A (4. sz. ábra) jeleníti meg *A gyermekjóléti szolgálat által gondozott kiskorúak száma (fő)* metrikus változó felvett értékeit a csoportképző változó függvényében. Az 1. klaszter esetében a medián – a dobozt megosztó vastag vonal – a magasabb értékek felé, a 2. klaszternél pedig az alacsonyabb értékek felé való elnyúlásra utal. A 2. klaszternél esetében az alsó-, a felső kvartilis és a medián is magasabb értéket képvisel, mind az 1. klaszter esetében. Mindkét csoportnál a „bajszok" (angolul: whiskers) távolsága alapján leolvasható a metrikus változó terjedelme. A *Boxplot* alkalmas az outlierek és az extrém outlierek azonosítására is. Mivel ebben az esetben ilyen nem volt, ezért az ábra is megerősíti, hogy a létrejött klaszterek homogének és optimálisnak tekinthetőek.



4. ábra: Boxplot/vagy doboz ábra
Box-and whisker plot
Forrás: Az IBM® SPSS® Statistics outputja

Klaszterelemzés a 3 148 települést tartalmazó adatbázis alapján

A klaszterelemzést végül Magyarország 3 148 települését tartalmazó – az összes település kivéve a főváros, illetve a fővárosi kerületek, valamint a kiugró értékeknek tekintett települések eliminálásával módosított – adatbázisán hajtottam végre ugyanazokkal a metrikus változókkal a 2015. évre vonatkozóan, amelyekkel előzőekben a kisebb adatbázison, a hierarchikus módszerek közé tartozó Átlagos láncmódszerrel és Ward-eljárással, távolságmértéknek a négyzetes euklideszi távolságot választva, valamint a nem-hierarchikus módszerek közé tartozó K-közepes klaszterezési eljárással, végül a Kétlépcsős klaszterezéssel. Az alkalmazott hierarchikus és nem hierarchikus módszerek közé tartozó eljárások végeredménye nem egyezett meg a klaszterekbe tartozó településeket és a települések számát tekintve. Ebben az esetben is kettő lett az optimális

1. klaszter (59): Abony, Aldebrő, Alsózsolca, Arnót, Balmazújváros, Barcs, Bezi, Börcs, Budakalász, Búcsúszentlászló, Detek, Dombóvár, Dunakeszi, Fábiánsebestyén, Győrújbarát, Inárcs, Kalocsa, Kárász, Katafa, Kesznyéten, Kisoroszi, Kokad, Lajosmizse, Ládbesenyő, Maglód, Makó, Marcali, Mád, Márokpapi, Mátészalka, Mezőcsát, Monorierdő, Mórahalom, Nagybajom, Nagykapornak, Nyírcsászári, Orgovány, Örvényes, Pécs, Pécsudvard, Piliscsaba, Pirtó, Püspökladány, Ráckeve, Rétság, Siófok, Sokorópátka, Solymár, Szarvas, Szentendre, Szentlőrinc, Szentgotthárd, Sződliget, Tápiószecső, Tök, Törökbálint, Újkígyós, Vásárosbéc és Zebegény; 2. klaszter (3): Dunaújváros, Hajdúböszörmény, Sopron.

klaszterszám. A Kétlépcsős klaszterelemzés az előbbiektől is eltérő eredményt hozott az egyes klaszterekbe tartozó településeket és a települések számát illetően, azonban a végeredményül adódott klaszterszám megerősítette a két-klaszteres megoldás elfogadását.

A települések klaszterekbe történő optimális besorolását a korábban bemutatott valamennyi klaszterezési eljárás során kapott csoportképző nominális változók és a klaszterelemzésbe bevont metrikus változók használatával elvégzett átlagok és szórások összehasonlításával határoztam meg Az egyes klaszterekben minden esetben a Ward-módszerrel létrejött két-klaszteres megoldás esetében volt a legalacsonyabb szórás, így ezt a megoldást tekintettem véglegesnek. A 2. klaszter 196, míg az 1. klaszter 2 952 települést tartalmazott.

Az első klaszter esetében újabb szakmailag indokolt és releváns metrikus változók bevonásával elvégzett klaszterelemzéssel homogén alklasztereket lehetne kialakítani, amelyek majd az altétegzést alapozhatják meg. Véleményem szerint az általam gyakorlati példán keresztül bemutatott megoldás finom hangolható lenne a feladatellátás formáit (pl. társulással kötött megállapodás útján, saját intézmény fenntartásával, vagy ellátási szerződés keretében más egyházi vagy állami szerv útján) is figyelembe vevő újabb metrikus változók klaszterelemzésbe történő bevonásával. Az 1. klaszterbe tartozó települések esetében mind a gyermekjóléti szolgálat által gondozott kiskorúak száma, mind a bölcsődékbe beírt gyermekek száma átlag alatti, míg a 2. klaszterbe tartozó települések esetében jóval átlag feletti volt a 2015. évben. Az egyes klaszterek további jellemzése a kisebb adatbázisnál írtakhoz hasonlóan végezhető el.

A fentiekből is látszik, hogy hiába az elemzésbe bevont változók közötti magas korreláció, a klaszterelemzés többféle módszerrel történő végrehajtásakor a hierarchikus és a nemhierarchikus klaszterelemzési eljárások esetében kettő adódott optimális klaszterszámnak. A Kétlépcsős klaszterezési eljárás alkalmazása is ezt eredményezte, amely eljárás sajátossága, hogy automatikusan határozza meg a különböző klaszterezési megoldásokban a modellválasztási kritérium értékeit, amik alapján határozza meg a klaszterek optimális számát (Malhotra, 2008).

Mindezekből következik a klaszterelemzés alkalmazhatósága feltételeinek „lágysága”, valamint, hogy a teljesen automatizált Kétlépcsős klaszterezési eljárás eredményei egyrészt nem feltétlenül jelentik automatikusan a legjobb megoldást, másrészt lehet, hogy épp a hierarchikus és a nem hierarchikus klaszterelemzési eljárások eredményeitől való eltérése – a települések klaszterekbe történő besorolását illetően – jelezheti a klaszterezésbe bevont változók magas korrelációját.

Véleményem szerint mindkét esetben a kialakult klaszterek a rétegzett mintavételi eljárás során az optimális rétegzés támogatásához felhasználhatóak lennének.

Diszkriminancia-analízis

A klaszterelemzések végeredményeként kialakult klaszterek megfelelőségét diszkriminancia-analízissel végeztem el. Függő változóknak minden esetben a megfigyelési egységeket csoportosító változókat, függetlennek pedig a klaszterelemzésbe bevont metrikus változókat választottam. Ellenőriztem, hogy az egyes klaszterekhez való tartozás becsülhető-e, és ha igen, hány százalékban az adott független változókkal. A diszkriminancia-analízis alkalmas a klaszterelemzések során kialakult csoportok további jellemzésére is, függő változónak a csoportképző változót, függetlennek pedig a klaszterelemzésbe be nem vont metrikus változókat választva. Ezt terjedelmi okok miatt gyakorlati példával nem illusztráltam, az adatbázisban szereplő változók közül véleményem szerint a *Lakosságszám (fő)* metrikus változó – mint független változó – bevonásával lenne érdemes elvégezni.

Diszkriminancia-analízis 62 települést tartalmazó adatbázis alapján
Függő változónak az Átlagos láncmódszerrel, távolságmértéknek a négyezetes euklideszi távolságot választva végrehajtott klaszterelemzés során a két-klaszteres megoldásra kapott, a megfigyelési egységeket csoportosító változót, független változóknak pedig a klaszterelemzésbe bevont *A gyermekjóléti szolgálat által gondozott kiskorúak száma (fő)* és a *Bölcsődébe beírt gyermekek száma (fő)* metrikus változókat választva végeztem el a diszkriminancia-analízist a 62 település adatait tartalmazó adatbázison *listwise* módszerrel. Az (1. táblázat) szemlélteti az egyes változók szignifikanciáját.

1. táblázat: ANOVA-táblázat
(ANOVA table)

	Wilk's lambda	F	df1	df2	Sig
A gyermekjóléti szolgálat által gondozott kiskorúak száma (fő)	0,434	78,316	1	60	0,000
Bölcsődébe beírt gyermekek száma (fő)	0,540	51,072	1	60	0,000

Forrás: Az IBM® SPSS® Statistics outputja (az Analyze->Classify->Discriminant parancs alkalmazásával) alapján saját szerkesztés

A táblázatban szereplő Wilks' lambda értékek mutatják, hogy mind *A gyermekjóléti szolgálat által gondozott kiskorúak száma (fő)*, mind a *Bölcsődébe beírt gyermekek száma (fő)* független változók közepes mértékben járultak hozzá diszkriminancia-függvényhez, azonban *A gyermekjóléti szolgálat által gondozott kiskorúak száma (fő)* változónak valamivel nagyobb hatása volt. A táblázat utolsó oszlopa azt mutatja, hogy mindkét változónak szignifikáns hatása volt a csoportokba tartozást illetően. A multikollinearitás tesztelésére vonatkozó korrelációs mátrixot és a variancia-kovariancia mátrixok homogenitásának tesztelésére vonatkozó táblázatokból megállapítható, hogy a diszkriminancia-analízisbe bevont változók közötti korreláció nem volt erős (0,454) és mivel a Box's M-mutatóhoz táblázatban található eredmény nem szignifikáns (0,003), ezért a varianciahomogenitási feltétel is teljesült. A létrejött diszkriminancia-függvényt a (2. táblázat) jeleníti meg, a függvény a varianciák 100 %-át magyarázza. A táblázat utolsó oszlopában szereplő kanonikus korreláció magas értéke (0,776) is azt jelezte, hogy a diszkriminancia-függvény jelentős részt magyaráz a teljes varianciából. A kanonikus korreláció négyzete azt mutatta, hogy a függő változó varianciájának 60,2 %-át magyarázza a független változók egy csoportja a diszkriminancia-függvény esetében.

2. táblázat: A diszkriminancia-függvény saját értékei, a magyarázott variancia és a kanonikus korrelációt
(The eigenvalues of the discriminant function, the explained variance and canonical correlation)

	62 települést tartalmazó adatbázis	3 148 települést tartalmazó adatbázis
A diszkriminancia-függvény saját értéke	1,511	1,190
Magyarázott variancia (%)	100	100
Kanonikus korreláció	0,776	0,737

Forrás: Az IBM® SPSS® Statistics outputja alapján saját szerkesztés

A fentieket a Wilk' lambát és a Khí-négyzetet is tartalmazó táblázat is megerősítette, amely alapján kijelenthető, hogy a diszkriminancia-függvény hatása jelentős és szignifikáns volt. A standardizált diszkriminancia-együttható, a Pearson-féle korrelációs együttható mátrix és az

egyes csoportok középpontértékeit tartalmazó táblázat vizsgálata további elemzésre nyújtott lehetőséget. A standardizált diszkriminancia-együttható, amely parciális, együttható, vizsgálata során megállapítható volt, hogy a diszkriminancia-függvény esetében *A gyermekjóléti szolgálat által gondozott kiskorúak száma (fő)* független változó volt a legfontosabb, mivel ez különböztette meg leginkább a csoportokat, amelyet a Pearson-féle korrelációs együttható mátrix is megerősített. Az elemzés kimenetéül adódott klasszifikációs táblázatban szerepel az eredeti, illetve a magyarázó változók alapján becsült csoportosítás, mind abszolút, mind relatív mértékkel megadva a helyes találati arányt. Azt, hogy hány százalékban kategorizált helyesen a modell, találati aránynak. hívjuk Az elemzés érvényességének vizsgálata a gyakorlatban fontos követelmény. A kapott eredményeket csak azután hasznosíthatjuk sikeresen, ha meggyőződtünk helyességükről. A táblázat alapján megállapítható, hogy a mindkét klaszter esetében a helyesen kategorizált esetek száma 100 % volt. Összességében a program az esetek 98,4 %-át tudta helyesen kategorizálni az adott független változók alapján, amely alapján megállapítható, hogy a klaszterelemzés eredményei megfelelőnek bizonyultak. A diszkriminancia-függvény a független változók alábbi lineáris kombinációjaként írható fel:

$$\boldsymbol{z} = \left(\begin{matrix}0{,}01 * A\ gyermekjóléti\ szolgálat\ által\ gondozott \\ kiskorúak\ száma\end{matrix}\right) + \\ +(0{,}08 * Bölcsődébe\ beírt\ gyermekek\ száma) - 1{,}258$$

Diszkriminancia-analízis végrehajtása a 3 148 települést tartalmazó adatbázison

Függő változónak a Ward-módszerrel végrehajtott klaszterelemzés során a két-klaszteres megoldás során létrejött, a megfigyelési egységeket csoportosító változót, független változóknak pedig az előző szakaszban választott metrikus változókat kiválasztva hajtottam végre a diszkriminancia-analízist a 3 148 település adatait tartalmazó adatbázison *stepwise* módszerrel Mahalanobis-távolságot alkalmazva. Az egyes változók szignifikanciáját is tartalmazó ANOVA-táblázat adatai alapján megállapítható volt, hogy a Wilks' lambda értékeiben annyi a különbség az előző szakaszban írtaktól, hogy a *Bölcsődébe beírt gyermekek száma (fő)* független változó a közepesnél gyengébb mértékben járult hozzá diszkrimimancia-függvényhez. A többi megállapítás hasonló volt az előző szakaszban írtakkal a diszkriminancia-analízisbe bevont változók közötti-, valamint a kanonikus korrelációt és a diszkriminancia-függvény hatására és szignifikáns voltára (2. sz. táblázat), továbbá a legjelentősebb, a klasztereket leginkább megkülönböztető változóra vonatkozóan is.

Eltérés volt tapasztalható viszont a Box's M-mutatót tartalmazó táblázatban, mivel a 0,000 szignifikanciaszint azt jelzi, hogy a varianciahomogenitási feltétel ebben az esetben nem teljesült. A független változók egy csoportja a függő változók varianciájának már csak 54,3 %-át magyarázta az itt létrejött diszkriminancia-függvény esetében. A klasszifikációs táblázat adatai alapján megállapítást nyert, hogy a mindkét klaszter esetében összesen a helyesen kategorizált esetek száma 97,3 %. volt, amely jelezte a klaszterelemzés megfelelőségét. Különbség mutatkozott azonban a 2. klaszter esetében jól kategorizált esetek arányában az előző szakaszban rögzített eredményekhez képest, mivel itt már csak a települések 56,6 %-át sorolta a helyes klaszterbe az eljárás. Érdekesség, hogy a *listwise* módszerrel végrehajtott diszkriminancia-analízis esetében ezek az arányok 98,6 %, illetve 78,1 % voltak, amelyek a két módszer különbözőségével magyarázhatóak**.** A diszkriminancia-függvény a független változók alábbi lineáris kombinációjaként írható fel, amely mind a *listwise***,** mind a *stepwise* módszerek alkalmazásával ugyanaz:

$$z = \left(\begin{matrix}0{,}019 * A\ gyermekjóléti\ szolgálat\ által\ gondozott \\ kiskorúak\ száma\end{matrix}\right) + \\ +(0{,}03 * Bölcsődébe\ beírt\ gyermekek\ száma) - 0{,}653$$

Az előző szakaszban létrejött diszkriminancia-függvénnyel összevetve is jól látható, hogy 3 148 település esetében *A gyermekjóléti szolgálat által gondozott kiskorúak száma (fő)* változó hatása a diszkriminancia-függvényre valamivel jelentősebb volt, mint 62 település tekintetében.

Következtetések, javaslatok

A klaszterelemzés és a diszkriminancia-analízis többféle ellenőrzés – beleértve a számvevőszéki ellenőrzéseket és elemzéseket is – során nyújthat támogatást hozzájárulva az értékteremtéshez, az ÁSZ esetében az Intézményi Stratégiájában megfogalmazott célkitűzésekkel összhangban, melyek – a teljesség igénye nélkül – az alábbiak lehetnek:

⇒ Olyan ellenőrzések során, ahol az ellenőrzöttek számának – mint sokaságnak – a nagysága indokolttá teszi valamely, lehetőleg rétegzett, mintavételi eljárás alkalmazását az ellenőrzöttek kiválasztása érdekében. Ebben az esetben előzetes – célzott – adatgyűjtés után összeállított adatbázis alapján, amely az ellenőrzés tárgya és célja szempontjából releváns, szakmailag indokolt egy vagy több adott időszakra vonatkozó adatokat és metrikus változókat is tartalmazza, az adatbázison végrehajtott *klaszterelemzés alkalmas a sokaság homogén klaszterekre történő felosztására, amelyek alapján történhet meg a rétegzés*. Már megvalósított rétegzés esetében annak helyességét is lehet klaszterelemzéssel ellenőrizni. Reprezentatív rétegzett mintavételi eljárás alkalmazása esetén a számvevőszéki jelentéstervezet, vagy elemzés összeállításakor a mintabeli információk, megállapítások alapján nem csak a sokaságra vonatkozóan lehetnének megfogalmazhatóak következtetések, hanem az egyes rétegekre külön-külön is. Megfelelően végrehajtott rétegzéssel és választott elosztással a reprezentatív megfigyelés hatékonysága tovább növelhető lenne. Lehetséges alkalmazási területek pl.: önkormányzatok egyes feladatellátásának ellenőrzése, a költségvetés végrehajtásának ellenőrzése keretében az egyéb költségvetési szervek ellenőrzése, egy adott közszolgáltatást végző gazdasági társaságok ellenőrzése.

⇒ Abban az esetben, ha egy ellenőrzési területet illetően a sokaság valamennyi egyedének ellenőrzését az ÁSZ már befejezte, az ellenőrzés, illetve az elemzés tárgya és célja szempontjából releváns és szakmailag indokolt egy vagy több adott időszakra vonatkozó adatait és változóit tartalmazó adatbázison lefuttatott *klaszterelemzés segítségével* olyan egymástól heterogén, de az egyes klasztereket külön-külön vizsgálva homogén, csoportok lehetnének kialakíthatóak, amelyek további elemzésével *feltárhatóak az egyes csoportok jellemzői, szerkezeti sajátosságai* is egy összefoglaló elemzés vagy tanulmány keretein belül. Lehetséges alkalmazási területek pl.: megyei hatókörű múzeumok, valamint egy adott közszolgáltatást végző gazdasági társaságok ellenőrzése, elemzése.

⇒ Nem utolsó sorban *a klaszterelemzés* az ellenőrzés előkészítése, az *ellenőrzöttek kockázatelemzéssel ellenőrzésre történő célzott kiválasztásának támogatására is alkalmas lehet* olyan esetekben, amikor nincs szükség, vagy lehetőség – pl. korlátozott kapacitás okán – az ellenőrzöttek mintavétellel történő kiválasztására. Ebben az esetben a megfelelően előkészített – az ellenőrzés, valamint az elemzés tárgya és célja szempontjából relevánsnak és szakmailag indokolt egy vagy több adott időszakra vonatkozó adatait és változóit tartalmazó – adatbázison alkalmazott klaszterelemzés vészcsengető, egyfajta compliance, funkciót is betölthetne, mivel lehetséges lenne az adott ellenőrzési- cél, illetve terület szempontjából „nem megfelelő” ellenőrzött szervezet(ek)re fókuszálni az ellenőrzésre történő kiválasztáskor.

Az egymással erősen korreláló metrikus változók esetében alternatív megoldásként javaslom, hogy faktoranalízis segítségével e változókat egy faktorba tömörítsük. A létrejövő faktorok bevonásával is elvégezhető a klaszterelemzés. Érdemes a klaszterelemzést több módszerrel több távolság-és hasonlóságmértékkel és több releváns metrikus változó-kombináció bevonásával megvalósítani, egyrészt az optimális klaszterszám meghatározása, másrészt a klaszterelemzés eredményei megbízhatóságáról való minden kétséget kizáró meggyőződés érdekében. A klaszterelemzést követően végrehajtott diszkriminancia-analízissel elvégezhető a klaszterelemzés eredményeinek validálása, valamint a beazonosíthatóak azok a változók, amelyek az egyes

klasztereket szignifikánsan megkülönböztetik egymásról. Megfelelő szakmai tudás – mind az ellenőrzés, elemzés tárgyát, mind a klasszifikációs módszereket illetően – hiányában rendkívül bonyolult az eredmények értékelése, ami téves következtetések levonásához vezethet.

Hivatkozott irodalom

ÁSZ (2010) Az Állami Számvevőszék Stratégiája In: ÁSZ honlapja https://www.asz.hu/storage/files/files/ASZ_strategia/asz_strategia.pdf (letöltés: 2017.01.13.)

BESENYEI L. et al. (2007): Klaszteranalízis alkalmazásának lehetősége az innovációs potenciál mérése során, In: Innovációmenedzsment kutatás és gyakorlat, Miskolci Egyetem Innovációmenedzsment Kooperációs Kutatási Központ, Miskolc, 53-64.o.

CSICSMAN J. (1979): A klaszter-elemzés módszerei és alkalmazásának lehetőségei a statisztikában, Statisztikai Szemle, 137-145. o.

ECA (2012): Pénzügyi és Szabályszerűségi Ellenőrzési Kézikönyv, In: ECA honlapja, http://www.eca.europa.eu/Lists/ECADocuments/FCAM_2012/FCAM_2012_HU.pdf (letöltés: 2017.01.23.)

ECA (2015): Teljesítmény-ellenőrzési Kézikönyv, In: ECA honlapja, http://www.eca.europa.eu/Lists/ECADocuments/PERF_AUDIT_MANUAL/PERF_AUDIT_MANUAL_HU.PDF (letöltés: 2017.01.23.)

FLISZÁR V. et. al. (2016): Többváltozós adatelemzési számítások, Budapesti Corvinus Egyetem, In: http://unipub.lib.uni-corvinus.hu/, http://unipub.lib.uni-corvinus.hu/2438/ (letöltés: 2017. február 27.)

GAO (1992): Using statistical sampling, In: GAO honlapja, http://www.gao.gov/assets/80/76112.pdf (letöltés: 2017.01.23.)

HAJDU O. (2003): Többváltozós statisztikai számítások, Központi Statisztikai Hivatal, Budapest

Helységnévtár In: KSH honlapja, http://www.ksh.hu/apps/hntr.egyeb?p_lang=HU&p_sablon=LETOLTES, (letöltés: 2017.01.22.)

HUNYADI L.-VITA L. (2006): Statisztika közgazdászoknak - Statisztikai módszerek a társadalmi és gazdasági elemzésekben, Központi Statisztikai Hivatal, Budapest

IBM® (2015): IBM® SPSS®® Statistics 23 Core System User's Guide, In: IBM honlapja,ftp://public.dhe.ibm.com/software/analytics/SPSS®/documentation/statistics/23.0/en/client/Manuals/IBM_SPSS®_Statistics_Core_System_User_Guide.pdf (letöltés: 2017. március 15.)

INTOSAI (2001-2016): Megfelelőségi ellenőrzésekre vonatkozó ISSAI irányelvek In: INTOSAI honlapja, http://www.issai.org/en_us/site-issai/issai-framework/4-auditing-guidelines.htm (letöltés: 2017.01.22.)

JÁNOSA A. (2011): Adatelemzés SPSS®® használatával, ComputerBooks Kiadói Kft., Budapest

KETSKEMÉTY L. et. al. (2011): Bevezetés az IBM® SPSS®® Statistics programrendszerbe, Artéria Stúdió Kft., Budapest

KOVÁCS E. (2014): Többváltozós adatelemzés, Typotex, In: http://etananyag.ttk.elte.hu/, http://etananyag.ttk.elte.hu/FiLeS/downloads/14_KOVACS_E_Tobbvalt_adatelemzes.pdf (letöltés: 2017. március 6.)

KSH (2017): Éves településstatisztikai adatok 2015-ös településszerkezetben In: KSH honlapja, http://statinfo.ksh.hu/Statinfo/haDetails.jsp?query=kshquery&lang=hu, (letöltés: 2017.03.23.)

MÁK (2017): Törzskönyvi nyilvántartás In: MÁK honlapja, http://www.allamkincstar.gov.hu/hu/ext/ torzskonyv/2/ (letöltés: 2017.03.23.)
MALHOTRA, N. K. (2007): Marketingkutatás, Akadémia Kiadó, Budapest
MÓRI F. T.–SZÉKELY J. G. (1986): Többváltozós statisztikai analízis, Műszaki Könyvkiadó, Budapest
NAO (2001): A practical guide to sampling, In: NAO honlapja, https://www.nao.org.uk/wp-content/uploads/2001/06/SamplingGuide.pdf (letöltés: 2017.01.23.)
SAJTOS L.–MITEV A. (2007): SPSS®® Kutatási és adatelemzési kézikönyv, Alinea Kiadó, Budapest
SIMON J. (2006): A klaszterelemzés alkalmazási lehetőségei a marketingkutatásban, Statisztikai Szemle 84. évfolyam 7. szám, 627-650. o.
SZILÁGYI R. (2011): Mintavételen alapuló becslések hibáinak kezelése, különös tekintettel a nemválaszolás okozta problémákra, Ph.D. értekezés, Miskolci Egyetem, Gazdaságtudományi Kar, Vállalkozáselmélet- és gyakorlat Doktori Iskola, Miskolc
TAKÁCS Sz. et. al. (2015): Klasszifikációs módszerek mutatói, In: Károli Gáspár Református Egyetem honlapja és Psychologia Hungarica, http://www.kre.hu/portal/images/doc/Klasszifikacios_vegleges.pdf, III/1. szám 67-88. o. (letöltés: 2017.04.27.)
VARGA B.–SZILÁGYI R. (2011): Virtuális vállalatok témakör 08. Kvantitatív információképzési technikák c. modul In: http://miskolc.infotec.hu, http://miskolc.infotec.hu/ilias.php?baseClass=ilSAHSPresentationGUI&ref_id=1517 (letöltés: 2017.01.02.)