DOI: 10.14751/SZIE.2014.035

Doktori (PhD) értekezés

A FOGYASZTÓI MAGATARTÁS VIZSGÁLATÁNAK MÓDSZERTANI TOVÁBBFEJLESZTÉSE

Készítette: Ruff Ferenc

Gödöllő
2014

DOI: 10.14751/SZIE.2014.035

A DOKTORI ISKOLA

MEGNEVEZÉSE: Gazdálkodás és Szervezéstudományok Doktori Iskola

TUDOMÁNYÁGA: gazdálkodás- és szervezéstudomány

VEZETŐJE: Dr. Szűcs István
egyetemi tanár,
az MTA doktora,
SZIE, Gazdaság- és Társadalomtudományi Kar,
Közgazdaságtudományi, Jogi és Módszertani intézet

TÉMAVEZETŐ: Dr. Szelényi László
egyetemi docens,
a mezőgazdasági tudományok kandidátusa,
SZIE, Gazdaság- és Társadalomtudományi Kar,
Közgazdaságtudományi, Jogi és Módszertani intézet

....
Az iskolavezető jóváhagyása A témavezető jóváhagyása

DOI: 10.14751/SZIE.2014.035

Tartalomjegyzék

DOI: 10.14751/SZIE.2014.035

DOI: 10.14751/SZIE.2014.035

DOI: 10.14751/SZIE.2014.035

DOI: 10.14751/SZIE.2014.035

1. fejezet

Bevezetés

1.1. A téma aktualitása

A hatalmas adatvagyonok létrehozása nehéz és költséges feladat. A létrehozott adatbázisok nagyon sok információt tárolnak, melyeknek kinyerése sok éve képezi kutatómunkák alapját. Manapság nagyon sok ilyen eljárás ismert, és évről évre újabbak születnek, melyeknek figyelemmel kísérése szinte lehetetlen feladat. Vannak közöttük olyan eljárások, amelyek meghonosodtak a felsőfokú oktatásban, és több éve képezik a hallgatók kutatómunkáinak módszertani részeit („best practice": pl. faktoranalízis, klaszteranalízis, stb.). Vannak azonban olyanok, amelyek az oktatásban csak néhol találhatók meg, vagy választható tárgyak anyagát képezik („trónkövetelők": pl. döntési fák, logisztikus regresszió, neurális hálók, hasonlóságelemzés, fuzzy logic, genetikus algoritmusok, stb.).

Egy-egy – a probléma megoldása céljából kiválasztott – módszer alkalmazása esetén, a kapott eredmények értékelésekor, figyelembe kell venni, hogy az adott módszer mennyire megbízható ill. az eljárás paramétereinek megváltoztatása mennyiben befolyásolja a kapott eredményeket, a következtetések levonásában található-e különbség. Természetesen nemcsak az alkalmazott módszerekkel lehetnek problémák, hanem már a felhasznált adatbázisokkal is, vajon az összegyűjtött információkból a valóság megismerhető-e, vagy az adatbázisunk (pl. egy minta) nem jó (pl. konzisztencia problémák, a mintavétel nem reprezentatív volta).

A módszerek közötti különbségek feltárása összehasonlító elemzéseket kíván, amelyre találunk példákat a szakirodalomban (pl. a klaszteranalízis területéről [Jay et al., 2012]), de a különböző területeken végzett empirikus vizsgálatok eltérő eredményekhez vezethetnek, az adott téma adatbázisainak szerkezeti, minőségi, mennyiségi eltérései miatt. Ezért az ilyen vizsgálatokból levonható következtetésekkel körültekintően kell bánni, ill. a szerzőnek nyilvánvalóvá kell tennie az érvényesség határait.

DOI: 10.14751/SZIE.2014.035

A felkínált módszerek közül a számunkra „legjobb" módszer kiválasztása újabb problémákat jelent (vö. szakértői rendszerek). Természetesen, sokszor nincs lehetőség arra, hogy egy másik algoritmussal megerősítsük a számításokat, ill. módszer-összehasonlításokat végezzünk a kapott eredmények fényében, majd döntést hozzunk az eredmény elfogadhatóságáról. Ilyenkor azokra az elemzésekre támaszkodhatunk, amelyek ezt már megtették, és azok segítségével tudjuk a megfelelő algoritmust kiválasztani, majd alkalmazni.

Úgy, ahogyan egy probléma megoldása céljából létrehozott modell módosítható, pontosítható, az adatelemzésekben felhasznált módszerek is fejlődnek, módosulnak kutatási munkák eredményeképpen (ld. pl. az osztályozás módszerének fejlődését Everitt et al. [2011] művének bevezetőjében). Ezen új módszerek elfogadása is hosszabb folyamat, főleg az alkalmazói oldal részéről. Itt lehet fontos szerepe a felsőoktatásban oktatóknak, hogy a jövőbeli felhasználókat nyitottságra neveljék, a problémák többirányú megközelítésének lehetőségét megismertessék velük.

1.2. A vizsgálat célja, köre

A kvantitatív elemzési módszerek folyamatos fejlesztésének lehetősége, a problémákban rejlő érdekességek vezettek azon vizsgálatokhoz, amelyeket dolgozatomban bemutatok.

A vizsgálatok körének lehatárolása

A marketingkutatások egyik fontos területe a megfigyelési egységek csoportosítása, szegmentálása, mely probléma megoldására a legszélesebb körben alkalmazott módszer a klaszteranalízis [Malhotra, 2002]. Ezen módszerrel kapcsolatban egy már meglévő tudományos eredmény további vizsgálatát végzem el, *valamint javaslatot teszek annak fejlesztésére.* A vizsgálat lényege, hogy keressük a klaszteranalízis által létrehozott klaszterek számának „optimumát" (vagyis azt a klaszterszámot, amelyik legjobban lefedi az adatbázisban - feltételezésünk szerint meglévő - klasztereket). Erre többféle módszer található a szakirodalomban, melyek közül talán a legismertebb a BIC index [Schwarz, 1978] használata. Vannak azonban olyan eljárások is, melyek a klasztereken belüli és azokon kívüli sűrűségvizsgálatok alapján döntenek bizonyos klaszter-felosztások mellett. A szakirodalomban körüljártam egy ilyen eljárás [Tong és Tan, 2009] előzményeit, jelenlegi állapotát, eddigi eredményeit, majd ezek után javaslatot tettem annak módosítására. Ezután a módosított eljárást összevetettem az eredetivel elméleti és gyakorlati vizsgálatok keretében is.

DOI: 10.14751/SZIE.2014.035

A Tong féle index és előzményeinek áttanulmányozása, valamint tesztelése során olyan hibákat figyeltem meg, melyek kijavítására lehetőséget láttam, továbbá feltételezhető volt az ezáltal kapott eredmények javulása az eredeti index eredményeihez képest. (1. kutatási cél.)

A második vizsgálat a jövőbeli fogyasztói magatartás előrejelzésével kapcsolatos. Ezen a téren is sokféle elemzési technika létezik, melyek közül a legfontosabbak megtalálhatók az irodalom feldolgozásban. Ezek közül választottam ki egyet [VAN OEST és KNOX, 2011], melynek módosítását hajtottam végre azért, mert az általuk létrehozott modell felállításának feltételrendszerét nem tartottam megalapozottnak. Az ő munkájuk is egy modell továbbfejlesztése [FADER, HARDIE és LEE, 2005a]. Én is ezen utóbbi modellhez nyúlok vissza, azonban a fejlesztés iránya más, mint a VAN OEST és KNOX [2011] modellé. Ezen előzmények leírása a szakirodalom feldolgozásában szintén megtalálható.

Az általam végzett módosítás lényege annak keresése, hogy további paraméterek bevonásával pontosabbá tehető-e a módszer. A több paraméter egyrészt az adatgyűjtés kiterjesztését jelenti (a megfigyelési időszakban), ezáltal információtöbbletet eredményez, ugyanakkor az adatokból visszakövetkeztethető valószínűségeloszlások száma (így ezen eloszlások paramétereinek száma) is megnövekszik, ami ezen utóbbi paraméterek becslésének számításigényét, komplexitását növeli meg.

Vajon ezen változások eredője az eredményekre kimutatható hatással lesz-e? Ha kimutatható a különbség, akkor az a modelleredmények pontosságát növeli vagy csökkenti? (2. kutatási cél.)

A módosított modell és a gyakorlatban sokszor alkalmazott ún. heurisztikus modell [WÜBBEN és WANGENHEIM, 2008] *eredményeinek összehasonlításából milyen következtetés lesz levonható az alkalmazás hasznosságának tekintetében, vagyis a valószínűségi modellek alkalmazásához szükséges többletmunka megtérülő befektetésnek tekinthető-e?* (3. kutatási cél.)

A dolgozat szerkezete

A dolgozat első részében a marketingkutatás területén alkalmazott módszerek történeti és szakterület szerinti vizsgálatát végeztem el. Érdekes látni, hogy a mai napig széleskörűen alkalmazott módszerek hová nyúlnak vissza, ill. milyen robbanásszerű fejlődést tett lehetővé a számítógépek mindennapi használatának elterjedése. Itt részletesebben foglalkozom azzal a két területtel, ahol a vizsgálataimat végeztem.

Az anyag és módszer fejezetben az adott területen eddig elért legjobb eredményeket mutatom be, ill. ezek kritikai elemzését végzem el, mert ezek képezik kutatómunkám alapját. Ezután a tesztelésekhez felhasznált adatok származá-

DOI: 10.14751/SZIE.2014.035

sát, kiválasztásuk elméleti hátterét, valamint a kapott eredmények vizsgálatának módszereit összegzem ebben a fejezetben.

Az eredmények fejezetben kerül bemutatásra a meglevő index ill. modell továbbfejlesztése, amik kutatási munkám új eredményeit jelentik. Szintén ebben a fejezetben a létrehozott index valamint modell teszteredményeinek bemutatására és értékelésére is sor kerül. Ezen eredmények alapján lehet megfogalmazni az elvégzett munka tudományos értékét, melyek az 5. fejezetben lettek összefoglalva.

DOI: 10.14751/SZIE.2014.035

2. fejezet

Szakirodalmi feldolgozás

2.1. A marketingkutatás fogalma, célja, eszközei

A marketingkutatás definíciójának megadása előtt célszerű a marketing fogalmának tisztázása is, melyeket az Amerikai Marketing Szövetség (AMA) megfogalmazásában[1] mutatok be.

Marketing
„Szervezésből és eljárásokból álló tevékenység, mely olyan ajánlatok létrehozásával, kommunikációjával, szállításával és cseréjével foglalkozik, mely értékkel rendelkezik a fogyasztók, az ügyfelek, a partnerek, és a társadalom egésze számára."

Marketingkutatás
„A marketingkutatás az a funkció, mely

- összeköti a fogyasztót, a nyilvánosságot –információkon keresztül – a gyártókkal ill. forgalmazókkal, mely információkat marketing problémák és lehetőségek azonosítására, megoldására használnak,
- marketingakciókat hoz létre, fejleszt, és értékel ki,
- figyelemmel kíséri a marketing eseményeit,
- segít megérteni a marketinget, mint folyamatot."

A marketingkutatás megadja a szükséges információkat a fent említett célok eléréséhez, valamint megadja az adatgyűjtés módszereit, irányítja és végrehajtja az adatgyűjtési eljárást, elemzi az adatokat, közzéteszi az eredményeket és javaslatot tesz a felhasználásra.

Azaz, egy igen összetett folyamattal állunk szemben, melynek célja a marketingtevékenység kiszolgálása, információkkal való ellátása. A marketingkutatás egy másik megfogalmazását találhatjuk meg Malhotra [2002] magyar

[1] http://www.marketingpower.com/AboutAMA/Pages/DefinitionofMarketing.aspx

DOI: 10.14751/SZIE.2014.035

nyelven is megjelent, összefoglaló művének 51. oldalán:
„Marketingkutatáson az információk szisztematikus és objektív feltárását, összegyűjtését, közlését, valamint felhasználását értjük, amelynek célja a marketingtevékenység során felmerülő problémák (és lehetőségek) megoldására irányuló vezetői döntések elősegítése."

Ez utóbbi megfogalmazás érthetőbben írja le ennek a marketing-területnek a lényegét: tudományos módszerekkel alátámasztott döntések meghozatalának elősegítése. A tudományos módszerek használata (bár nincs kimondva) az *objektív* jelző definícióba való bevonásából következik, mely természetes igénnyel merül fel minden gazdasági döntés-előkészítés esetén.

Mint a fenti megfogalmazások is mutatják, a probléma felszínre kerülése és megfogalmazása után a megfelelő adatok összegyűjtése történik. Már ezen a ponton szükség van azon ismeretekre, amelyek ezen a területen születtek az utóbbi évek - évtizedek alatt (reprezentativitás, méret, ellentmondás mentesség).

A következő lépés a begyűjtött adatok rendszerezése, előfeldolgozása. Ez a fázis még nem igazán vizsgált terület, azonban Crone, Lessmann és Stahlbock [2006] cikkéből kiderül (mely egy empirikus vizsgálat), hogy különböző adatelemzési módszerek (döntési fa, neurális háló, SVM[2]) esetén szignifikáns pontosság-növekedést tudtak elérni az előrejelzésekben, pusztán az adatbázis előfeldolgozásának segítségével. Vizsgálatukból az is kiderült, hogy nincs univerzális megoldás, az egyes módszerek esetén más és más eljárás mutatkozott célravezetőnek (változók skálázása és kódolása, mintavétel).

Ezt követi az adatok elemzése. Világosan megfogalmazott kérdésekre kereshetjük a választ különböző matematikai – statisztikai eszközökkel. Az elemzés végrehajtójának széles rálátása kell legyen ezen módszerek alkalmazhatósági feltételeire, ugyanazon problémára adható megoldások különbözőségére, az újabb kutatási eredményekre. Egyrészről nem ezt jelzi azonban az a vizsgálat, mely a marketingkutatással kapcsolatosan megjelent tudományos cikkeket vizsgálja. A nemzetközi marketingkutatás területén az 1990-2000-es időszakban a vizsgált tudományos munkák 5%-a foglalkozott módszertani problémával [Nakata és Huang, 2005]. Másrészt azonban tartalmaz olyan összehasonlítást is a cikk, hogy milyen arányban publikáltak kvantitatív ill. kvalitatív munkákat. Az elméleti vizsgálatok esetében ezek aránya 3:5, míg az empirikus vizsgálatok esetében 11:3 és összességében az empirikus kutatások az összes kutatás kb. 70%-át adták az adott időszakban.

Az utolsó fázisban a kapott eredmények bemutatása történik, melyek egy döntés meghozatalának objektív megalapozását szolgálják. A kutató feladata

[2]Support Vector Machine

DOI: 10.14751/SZIE.2014.035

itt véget ér, a labda a döntéshozók térfelére került.

Ezen felsorolás kapcsán nem tértem ki minden egyes pontra, melyeket pl. MALHOTRA [2002] könyvében találunk. Itt csak a dolgozat szempontjából releváns részeket említettem meg a marketingkutatás folyamatából.

2.2. A marketing elméletek történeti fejlődése

A marketingtudomány fejlődését először röviden WILKIE és MOORE [2003] nyomán mutatom be. A szerzők négy szakaszra osztják ezt a folyamatot, melyek legfőbb jellemezőit gyűjtöttem össze.

- *A tudományterület megalapozása* (1900 - 1920)
  Ez a piacok átalakulásának időszaka. Eddig az időszakig azonban a piacokkal az ökonómiában nem foglalkoztak olyan mértékben, mint pl. a termeléssel, a földdel, a tőkével, a munkával. A lokális piacok esetében ez természetesen érthető. Azonban erre az időszakra már egyre jelentősebbé vált a helyi piacok kibővülése. A közgazdászok részéről tehát egyre nagyobb figyelmet kapott ezen területek bevonása a tudományos vizsgálatokba.
  Továbbá az egyetemeken megjelentek marketinggel kapcsolatos kurzusok. A gazdasági folyóiratokban megjelentek ennek a területnek a kutatási eredményei, módszerei, elméletei.

- *A tudományterület formalizálása* (1920 - 1950)
  Ebben az időszakban ugrásszerűen emelkedett a termékek mennyisége, rengeteg új termék került bevezetésre (pl. az elektromos hálózat ugrásszerű kiépítése és a vele párhuzamosan megjelenő elektromos eszközök[3]). Választ kellett keresni a megnövekedett termékmennyiség elosztására, és az ezzel párhuzamosan növekvő vásárlói igények kielégítésére.
  Ezen újonnan megjelent problémák feldolgozását megnehezítette a korszakhoz tartozó világválság, világháború is, melyek újabb és újabb helyzetek megoldása elé állították a terület szakembereit. Ezen környezetben születtek meg a marketing egységesen elfogadott alapelvei. Jelentős fejlődés eredményeként tudományterületté formálódott a marketinggel kapcsolatos elméletek köre.

- *Paradigmaváltás a fő áramlatban* (1950 - 1980)
  A marketingtudomány tudományos infrastruktúrája (BA és MA képzések, tudományos folyóiratokban megjelent cikkek, tudományos társaságok) rohamosan fejlődtek, sokasodtak a korszak során. A korszakban formálódó

[3] A tanulmány az Amerika Egyesült Államokat vizsgálja.

DOI: 10.14751/SZIE.2014.035

új irányra jellemző egyrészt a marketing elméletek tudományos alapokon történő fejlesztése, másrészt a marketing menedzserek szemével való látásmód alkalmazása, és az így szerzett megfigyelések beépítése a menedzseri munkába. Korábban ugyanis a kutatók egy része nem foglalkozott konkrét operatív műveletek elemzésével. Az új irányzatnak az egyetemi oktatásban a gyakorlatorientált szakmai képzés megjelenése lett a gyümölcse. Olyan elméletek megalapozásai születtek meg ebben a korszakban, amik mind a mai napig a vizsgálatok tárgyát képezik: piac szegmentáció, marketing mix, márka arculat, marketing menedzsment. Az időszak szellemi termékei közül kiemelném KOTLER [1967] „Marketing management" c. könyvét, mely nagy hatással volt a fiatal kutatók munkájára és elősegítette a kvantitatív és viselkedés-tudományok bevonását a marketing kutatásokba.

- *A fő áramlat felaprózódása* (1980 -)
  A tudományos műhelyek (tanszékek, folyóiratok) növekvő száma a tudományterület felaprózódását, specializációját eredményezték, természetes módon. A korábbi menedzseri perspektíva – a tudományos munka célja a menedzseri döntések hatékonyságának elősegítése – továbbra is a terület kiemelkedő elve maradt. Ugyanakkor a sorra megjelenő újabb és újabb folyóiratok egyre speciálisabbak, egyre inkább leszűkített területtel foglalkoznak. Pl. BAUMGARTNER és PIETERS [2003] vizsgálata közel ötven tudományos folyóiratra tért ki, melyek elemzése nyomán[4] öt csoportba sorolta a folyóiratokat:
  - közérdeklődésre számító folyóiratok,
  - vásárlói magatartással foglalkozó folyóiratok,
  - menedzsereknek szóló (cég orientált) folyóiratok,
  - marketing alkalmazás orientált folyóiratok,
  - marketing oktatással kapcsolatos folyóiratok.

  WILKIE és MOORE [2003] a cikkében megemlíti, hogy ennek a specializálódásnak az eredménye, hogy a marketing új tudományos eredményeinek áttekintése már meghaladja az egyes ember lehetőségeit, képességeit.
  Igen, ez így van. Azonban ez egy természetes folyamat, egy tudomány fejlődésének természetes állapotváltozása. A marketingkutatás egy fiatal tudomány, szemben például a matematikával, ahol ezzel a „problémával" már régen szembesültek. A tudomány feladata nem a közérthetőség fenntartása, hanem a tudományos műhelyekben megszületett elméletek segítségével

[4] A folyóiratokat egy olyan kétdimenziós térben helyezte el, melyben vertikálisan az elméleti ill. gyakorlati jelleg különbségét, horizontálisan pedig a vásárlókkal ill. a cégekkel való foglalkozás jellegét mérte.

DOI: 10.14751/SZIE.2014.035

a gyakorlati szakemberek többlettudással való ellátása, mely többlettudás alkalmazása, felhasználása a mindennapi életben „haszonnal" jár.

Ezen időszakokon átívelnek azok az elméleti iskolák, amelyek az újabb kutatási területek felismerése után a téma elemzésére, tudományos megalapozására vállalkoztak. SHAWAND és JONES [2005] tíz ilyen iskolát különböztet meg:

- Marketing gyakorlat.
- Áruk.
- Intézmények.
- Régiók közötti szállítás.
- Menedzsment.
- Marketing rendszerek.
- Fogyasztói magatartás.
- Makromarketing.
- Csere.
- Marketing történelem.

Ezek közül kiemelem a fogyasztói magatartással foglalkozó iskolát, mivel dolgozatom témája ebbe a gondolati rendszerbe illeszkedik. A terület az emberi viselkedéssel foglalkozik, ami azt jelenti, hogy az egyik leginkább változó elméleti iskoláról van szó. A terület az 1950-es években indult fejlődésnek, mikor is a vásárlási és fogyasztási szokások vizsgálatába bekapcsolódtak ökonómusok, pszichológusok, szociológusok is, mivel egy összetett területtel álltak szemben. Azonban az első sikerek a 60-as évek végén jelentkeztek, amikor sikerült átfogó, jól megalapozott (a kor ismeretanyagának megfelelő) modelleket létrehozni. Pl. HOWARD és SHETH [1969] modellje azért jelentős, mert rávilágított a vásárlói döntéseket befolyásoló tényezőkre, és ezek rangsorolását vizsgálta az empirikus kutatásaiban. Ezek után a tudományos cikkek és konferenciák egymás után indukáltak újabb kutatási témákat.

1974-ben adták ki első folyóiratukat: Journal of Consumer Research (JCR), mely kiszélesítette a „vásárlás, fogyasztás, használat" kutatási területeket a következő, nem szorosan a témához tartozó területekkel: „családtervezés, foglalkozás választás, mobilitás, termékenységi arányok".

A nem szorosan a marketinggel foglalkozó szakemberek azonban a vásárlói magatartással, mint céllal foglalkoztak, ellentétben azokkal, akik inkább pl. a marketing menedzsment eszközökkel, az eladások felől közelítettek a kérdéshez.

DOI: 10.14751/SZIE.2014.035

A fogyasztói szokások vizsgálata kezdett eltávolodni a marketing tudománytól, mivel látókörébe kerültek nemcsak a vásárlással kapcsolatos, hanem pl. az áruk mozgását befolyásoló egyéb jelenségek (önálló termelés, ajándékozás, jótékonyság, lopás) is. Mára a fogyasztói szokások vizsgálata erősen kapcsolódik a társadalomtudományokhoz, túlmutat azon, hogy csak a marketing egyik elméleti iskolája legyen.

A marketingelméletek egy másféle csoportosítását találjuk meg KOTLER és KELLER [2012, 28-29. old.] művében, akik a vállat piaci orientációjával kapcsolatos marketingelgondolások fejlődésének vizsgálatát tartották szem előtt. Az általuk meghatározott időszakoknak a következő nevet adták:

- *A termelési koncepció*
  Eszerint „a fogyasztók a széles körben hozzáférhető és olcsó termékeket részesítik előnyben". A hangsúly a termelékenységen, az alacsony költségeken, valamint a széles körű elosztáson van. Manapság ez az elgondolás a fejlődő országokban helyénvaló.

- *A termékkoncepció*
  „A fogyasztók a legjobb minőségű, a legjobb teljesítményű és innovatív termékeket fogják előnyben részesíteni." Veszélye a „jobb minőség" csapdája, azaz a jobb termék önmagában nem feltétlenül vonzza majd a vevőket.

- *Az értékesítési koncepció*
  szerint, ha a fogyasztókat magukra hagyjuk, akkor nem vásárolnak eleget a vállalat termékeiből. Ezen koncepciót legerőteljesebben a nem keresett cikkek esetén alkalmazzák. Cél: eladni, amit termelnek.

- *A marketingkoncepció*
  „Találjuk meg a megfelelő terméket a vevő számára." Míg az értékesítés az eladó, addig a marketing a vevő igényeire összpontosít. Ennek eszköze pedig a versenytársakénál vonzóbb vásárlói érték hatékony megteremtése és kommunikációja.

- *A holisztikus marketingkoncepció*
  „olyan marketingprogramok, -folyamatok és -tevékenységek kidolgozására, tervezésére és megvalósítására támaszkodik, amelyek elismerik a feladatok jelentőségét és kölcsönös függőségét". Azaz „felismeri és összehangolja a tevékenységek hatáskörét és bonyolultságát". Négy nagy alkotóelemre bontható:

  – kapcsolati marketing (vevők, partnerek, csatorna),
  – integrált marketing (termékek és szolgáltatások, kommunikáció, csatornák),

DOI: 10.14751/SZIE.2014.035

– belső marketing (felső vezetés, marketing osztály, egyéb osztályok),
– teljesítménymarketing (árbevétel, márka- és vevőérték, etika, környezet, törvényesség, közösség).

Ezen csoportosítás alapján megfigyelhető egyrészt a vevő értékének egyre növekvő felismerése, valamint az értékesítésen túl az egész tervezési, előállítási és értékesítési folyamat egységes rendszerbe foglalása, és ezen rendszer összefüggéseinek vizsgálata. Mára sokféle tudományos módszer áll rendelkezésre ezen összefüggések elemzésére. Dolgozatomban ezen terület irányába haladok tovább.

2.3. A marketingkutatás fejlődése

A marketingkutatás, mint tudományos módszer, hosszú időn keresztül formálódott, alakult ki. Időközben a kvantitatív elemzések módszertana is egyre bővült. Ennek rövid áttekintését Maex [2009] cikke nyomán teszem meg.

Az 1800-as évek végén elindultak az első csomagküldő szolgálatok (ld. pl. Aaron Montgomery Ward), melyek jelentős hatást gyakoroltak több millió család mindennapi életére. Ward munkája során (utazó ügynök) az igényeket és lehetőségeket személyesen tapasztalta meg, és ezen tapasztalatok kiértékelésén keresztül sikerült az új rendszerét kialakítani, megvalósítani.

Ezt követően az első tudományos igényű művek az 1900-as évek első harmadában jelentek meg. Megemlíthetjük itt Claude Hopkins - Scientific Advertising című, 1923-ban megjelent könyvét [Hopkins, 2010], mely könyv általános érvényű gondolatokat tartalmaz az adott területen (pl. mintavétel, visszafizetési garancia, piac tesztelés, kockázatmentes kipróbálás). A mű fő értéke, hogy a tudományos megismerés kritériumainak megfelelően tárgyalja az érintett területeket. A mű elején leszögezi, hogy a reklámozás eddigi tudáshalmaza, bizonyos pontokon, elérte a tudományos státuszt, mely az elvek lefektetésének és az alkalmazott módszerek/módszertanok kidolgozásának köszönhető.

Ezen időszak jellemző kutatási területe a reklám volt, melynek megalapozásában fontos szerepet játszott a máig ajánlott irodalomként számon tartott Tested Advertising Methods [Caples, 1932]. A szerző a reklámozási technikák tesztelésével ill. ezek méréseken alapuló összehasonlításával foglalkozott. A tudományterület meghatározó művei közé való bekerülésének főbb szempontjai:

- a következtetéseket csak mérések alapján fogadta el,
- az új projektek megalapozását e kísérletek tesztelése jelenti,

DOI: 10.14751/SZIE.2014.035

- az egyes projekteket nem tekinti lezártnak, a jobb eredményeket hozó módszerek adaptálását javasolja.

A vizsgálatok ezen korai szakaszában a kutatás fő területei a direkt válaszos tevékenységek voltak, mint pl. a fent említett csomagküldő szolgálatok. A Második Világháborút követően kezdték el a kutatók az újabb matematikai eredmények felhasználását a marketing területén. Így pl. a nagy fejlődésnek indult operációkutatás is a figyelem előterébe került [Magee, 1960]. Magee ebben a cikkében azonban tágabb értelemben használja a fogalmat, mint ahogy az a matematikai terminológiában megszokott volt. A szerző szerint három fontos ponton járulhat hozzá az operációkutatás a marketinggel kapcsolatos kutatásokhoz: a) rendszerek felépítése, b) a kísérletek hangsúlyozása, c) fogyasztói magatartás modellezése.

A marketing fejlődésének következő szakaszában [Wilkie és Moore, 2003] a reklámszakemberek figyelme egyre inkább a tömegmédiák felé fordult, mely a marketing hatékonyságát jelző direkt adatok mennyiségének kicsiny volta miatt a matematikai módszerek használatának határait is kijelölte. A változást ekkor az ökonometriai vizsgálatok marketing területre történő bevonása jelentette, mellyel az összegyűjtött adatokból (beruházások és eladások nyomon követése, kérdőívek, fórumok) a következő változókra kerestek megfelelő becsléseket, előrejelzéseket: márkahűség, eladható mennyiség, profit. További lehetőségei ezeknek a vizsgálatoknak, hogy bepillantást nyerhettek az alkalmazott marketing mix parciális hatásainak, térbeli eloszlásának hatékonyságába, vizsgálható volt az erőforrások hatékonyabb felhasználásainak lehetősége.

A 90-es évek során a vásárlói kapcsolatok előtérbe kerülése, valamint az egyre több rendelkezésre álló adat a marketingkutatás matematika módszereiben is megújulást igényelt. Ilyen terület volt pl. a vásárlói lojalitás vizsgálata [Reichheld és Teal, 2001], ahol kimutathatóak voltak a befektetett erőforrások megtérülései. Egy másik fontos terület a meglevő vásárlók/ügyfelek osztályozása, csoportosítása. Hallberg és Ogilvy [1995] megállapítja, hogy bármely termék vagy szolgáltatás fogyasztói bázisában elkülöníthető egy csoport, amely az adott termék profitjának nagy részét adja. Majd a fogyasztók adatainak személyre szabott vizsgálatából előrejelzések is levezethetők voltak, melyekhez a sok, elektronikusan keletkező adat szolgáltatta az információt. Ezek kinyerése céljából az addig már jól ismert módszerek (regressziós modellek, diszkriminancia analízis) mellett újabb eljárások születtek: neurális hálók, döntési fák.

Az ezt követő információs robbanás nagyságrendekkel megnövelte a rendelkezésre álló bináris adatok mennyiségét és magával hozta az online kiértékelés lehetőségét is. Ennek alapját a digitális kommunikáció rohamos növekedése

DOI: 10.14751/SZIE.2014.035

teremtette meg. Pl. a marketing-kutatás egyik adatforrása a weblapokat látogatók tevékenysége, szokása, mely adatok „gyors" kiértékelése után válaszolni lehet a feltárt problémákra ill. lehetőségekre.

2.4. A marketingkutatás matematikai módszerei az utóbbi 20 évben

2.4.1. Időrendi áttekintés

Egy 1995-ben végzett kutatás [Hussey és Hooley, 1995] azt vizsgálta, hogy a marketingkutatás egyes szereplői (kutatók, oktatók, elemzők) milyen kvantitatív módszereket használnak, tartanak felhasználhatónak ezen a területen. A leginkább használt módszerek a következők voltak:

- szignifikancia teszt,
- adatok grafikus ábrázolása, leíró statisztikák, gyakorisági táblázatok,
- faktoranalízis, klaszteranalízis, többdimenziós skálázás, diszkriminancia analízis, AID (Automatic Interaction Detector), log-lineáris analízis,
- kétváltozós lineáris regresszió, többszörös lineáris regresszió, exponenciális simítás.

A matematikai módszerek alkalmazásának feltételei egyre javulnak, hiszen a szoftverfejlesztő cégek egyre inkább felhasználó-barát termékekkel jelennek meg a piacokon, melyeket a cégek által alkalmazott, vagy külső megbízással rendelkező szakemberek felhasználnak elemzéseikhez.
Ugyanakkor érdekes, hogy a kutató cégek véleménye szerint jelentős passzivitás mutatkozik az elemzők részéről az újabb módszerek alkalmazása iránt [Hussey és Hooley, 1995]. Igaz, nagy különbségek vannak a cégek között a tekintetben, hogy mennyi idő telik el egy-egy újabb tudományos eredmény megjelenése és gyakorlati bevezetése között, mennyire fogékonyak újabb módszerek kipróbálására. Természetesen, ezen módszerek „diffúziójának" ideje – a módszer paraméterein kívül – függhet magától a területtől is, ahol alkalmazható.

A nemzetközi összehasonlításokat is tartalmazó cikk szerint, a cégek által alkalmazott adatelemzési technikák között a legnépszerűbbek a grafikus megjelenítések, a táblázatok, és a leíró statisztikák. Az alkalmazott matematikai módszerek rangsora a következő: regresszió elemzés, klaszteranalízis, faktoranalízis. Igaz, a paletta ennél sokkal szélesebb, de ezeket a módszereket a megkérdezett cégek kb. 15-20%-a alkalmazta (míg a fent említett adatelemzési technikák esetében ez 80-90% volt).

DOI: 10.14751/SZIE.2014.035

Az 1994-es vizsgálat óta eltelt idő nagyon sok változást hozott a számítás-teljesítményekben, adatbázis-méretekben, továbbá a valós idejű elemzések terén. Kérdés, hogy mennyiben változtatta meg ez a módszertant, vannak-e az eddig alkalmazott kvantitatív módszerek mellett újabbak?

Az Amerikai Marketing Társaság (American Marketing Association - AMA) 2010-ben meghirdetett kurzusának (American Marketing Association Applied Research Methods) - mely a praktikus alkalmazások bemutatására fókuszált - több előadása is foglalkozott a kvantitatív elemzési technikákkal. Ezek közül az egyik előadásában a többváltozós módszerek közül a következőket mutatta be: faktoranalízis, klaszteranalízis, regresszió elemzés, diszkriminancia analízis, conjoint analízis [Chakrapani, 2010]. További hagyományos kvantitatív módszerek is szerepeltek a kurzus tematikájában, pl. a piaci szegmentáció témakörében a klaszteranalízis és a döntési fák [Mulhern, 2010].

Összehasonlítva a korábbi cikk elemzésével, látható, hogy ezek a módszerek képezik az elemzések alapjait a kutatásokban. A születésüktől kezdve azonban hosszú utat jártak be, állandóan finomították, a felmerülő problémákhoz igazították őket (vö. klaszterek képzésének módszerei). Így tehát ugyanaz a gyűjtőnév már más, újabb algoritmusokat is takarhat, mint korábban.

2.4.2. A módszerek csoportosítása

Egy módszer megválasztása körültekintést igényel a kutatótól. Számos kérdésre kell a választás előtt válaszolni. Mi a cél? Milyen adatok állnak rendelkezésre? Mik a feltételei az egyes módszerek alkalmazásának?

Az 1-es táblázat a marketingkutatásban használt egyváltozós módszerek, a 2. és a 3. táblázatok a többváltozós módszerek csoportosításának lehetőségét mutatják. Ezek azok a módszerek, amelyek ma a marketingképzésben módszertani oldalról megtalálhatók, ezeknek az elsajátítását várják el a kikerülő marketing szakemberektől. Természetesen a spektrum ennél sokkal szélesebb, ill. az egyes módszerek is már sokszor módszerek sokaságát jelentik (ld. klaszteranalízis, döntési fák, diszkriminancia analízis stb.). Egy oktatási célra szánt tananyag nem tartalmazhatja az összes létező fejlesztést, melyek az utóbbi időben születtek a módszertanban, de képességet kell kifejleszteni a felhasználókban ezen újabb eredmények befogadására.

Moutinho és Meidan [2003] által írt könyvfejezet szintén kvantitatív elemzési módszerek ismertetésével foglalkozik. Az általuk felvázolt paletta már jóval szélesebb (1. ábra, 22. old.), mint az előbb említett táblázatokban. Ennek oka, hogy sokkal szélesebb kutatási területet vizsgálnak, melyekhez már egyéb módszerek is szükségesek (pl. operációkutatás). Továbbá tartalmaznak olyan módszereket is, melyek még újdonságnak számítanak ezen tudományterületen

DOI: 10.14751/SZIE.2014.035

1. táblázat. Egyváltozós statisztikai módszerek osztályozása.
Forrás: MALHOTRA [2002].

Adatok típusa	Minták száma	Minták kapcsolata	Módszer
Metrikus	Egy	–	t-póba
Metrikus	Egy	–	z-póba
Metrikus	Kettő vagy több	Független	Kétmintás t-póba
Metrikus	Kettő vagy több	Független	z-próba
Metrikus	Kettő vagy több	Független	Egy szempontos ANOVA
Metrikus	Kettő vagy több	Összefüggő	Páros t-próba
Nem metrikus	Egy	–	Gyakoriság
Nem metrikus	Egy	–	χ^2-próba
Nem metrikus	Egy	–	Kolm. - Szmirnov próba
Nem metrikus	Egy	–	Sorozatpróba
Nem metrikus	Egy	–	Binomiális próba
Nem metrikus	Kettő vagy több	Független	χ^2-próba
Nem metrikus	Kettő vagy több	Független	Mann - Whitney próba
Nem metrikus	Kettő vagy több	Független	Medián próba
Nem metrikus	Kettő vagy több	Független	Kolm. - Szmirnov próba
Nem metrikus	Kettő vagy több	Független	Kruskal - Wallis féle egysz. ANOVA
Nem metrikus	Kettő vagy több	Összefüggő	Előjelpróba
Nem metrikus	Kettő vagy több	Összefüggő	Wilcoxon próba
Nem metrikus	Kettő vagy több	Összefüggő	McNemar próba
Nem metrikus	Kettő vagy több	Összefüggő	Mann - χ^2-próba

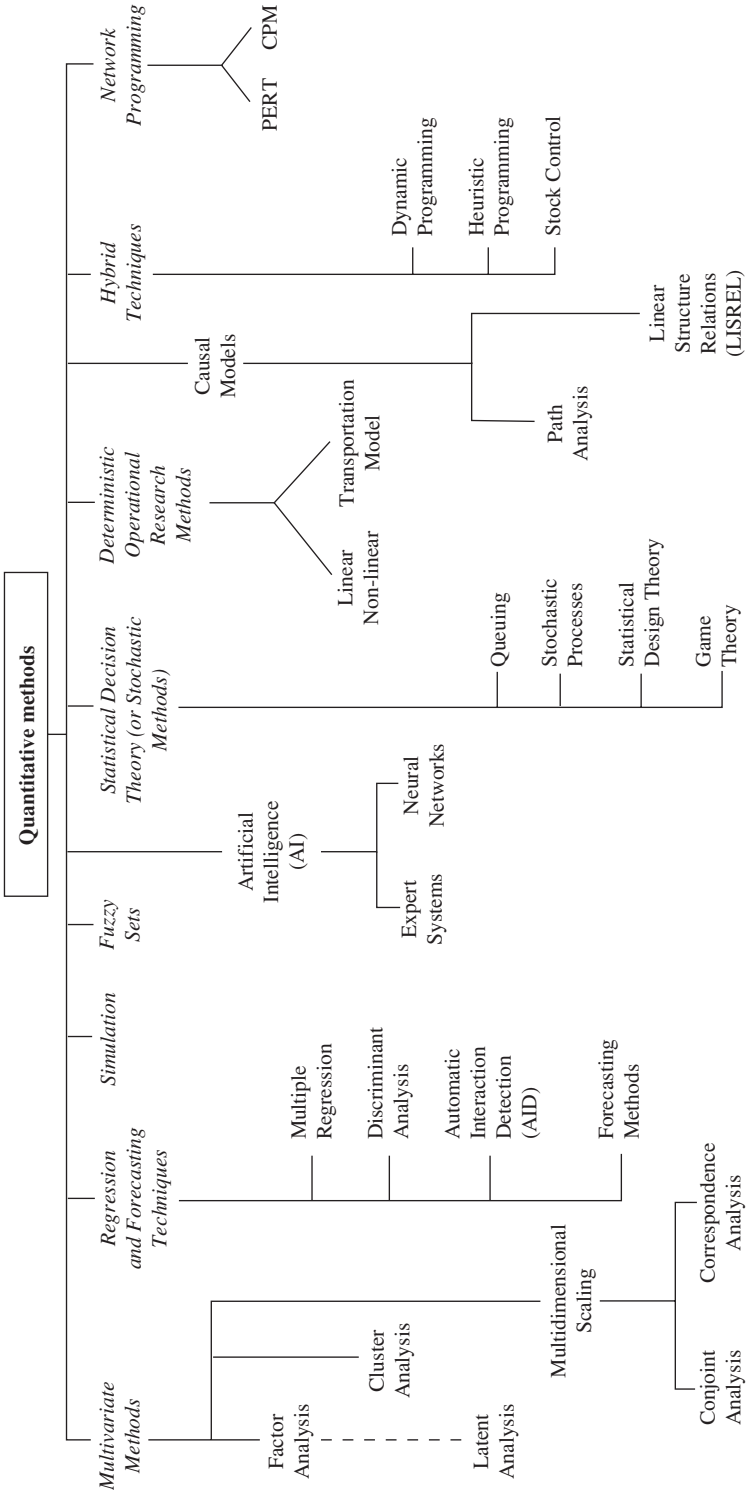
2. táblázat. Függőségen alapuló többváltozós statisztikai módszerek.
Forrás: MALHOTRA [2002].

Függő változók száma	Módszer
Egy	Kereszttábla
Egy	Variancia- és kovariancia elemzés
Egy	Regresszióelemzés
Egy	Kétcsoportos diszkriminancia elemzés
Egy	Conjoint elemzés
Egynél több	Többváltozós variancia- és kovar. elemzés
Egynél több	Kanonikus korreláció elemzés
Egynél több	Többszörös diszkriminancia elemzés

(pl. mesterséges intelligencia).

A módszer megválasztásának első két lépése: az adatbázis milyen mérési szintű adatokat tartalmaz (pl. 1. táblázat), ill. a változók milyen kapcsolatban vannak egymással (ok-okozati vagy kölcsönösen összefüggő). A marketingkutatások esetében általában kölcsönösen összefüggő változókkal találkozunk, mint például: termék, ár, elosztás, reklámozás [MOUTINHO és MEIDAN, 2003]. A változók visszahatnak egymásra. Ezért használatosak a többváltozós módszerek, hiszen általában nem egy-egy változó hatását szeretnénk kideríteni, hanem fontos az egyes változók kapcsolatának hatása is.

DOI: 10.14751/SZIE.2014.035



1. ábra. Kvantitatív módszerek csoportosítása.
Forrás: Moutinho és Meidan [2003], 199. old.

DOI: 10.14751/SZIE.2014.035

3. táblázat. Kölcsönös összefüggésen alapuló többváltozós statisztikai módszerek. Forrás: MALHOTRA [2002].

A vizsgálat célja	Módszer
Változók kölcsönös összefüggésének vizsgálata	Faktorelemzés
Egyedek hasonlóságnak elemzése	Klaszterelemzés
	Többdimenziós skálázás

2.5. A fogyasztói magatartás kvantitatív vizsgálatának eszközei

Az általános ismertetés után leszűkítem azokat a területeket, melyekre dolgozatomban koncentrálok. Először általánosan foglalkozom a fogyasztói magatartás vizsgálatának fontosságával és eszközeivel, majd a következő fejezetben ennek is egy részletét vizsgálom csak tovább.

A témát NGAI, XIU és CHAU [2009] cikke alapján közelítem meg, akik egy irodalom feldolgozást végeztek – a 2000 és 2006 közötti időszakra – az ügyfélkapcsolat kezelésben (angolul Costumer Relationship Management - CRM) alkalmazott matematikai-statisztikai eszközök területén. A CRM egy széles körben ismert és alkalmazott rendszer, melynek több definíciója közül én a következőt választottam: *a vásárlói magatartás megértésének és befolyásolásának vállalati megközelítése, melynek feladata, hogy javítani tudjanak az új ügyfelek megszerzésén, az ügyfélmegtartáson, az ügyfelek hűségén, és az ügyfél-jövedelmezőségen* [SWIFT, 2000, 12. old.]. A piacok telítődése nyomán a 90-es évektől kezdték ezeket a rendszereket kidolgozni, melyeknek a középpontjában a vevő, a vevővel történő egyre személyesebb kommunikáció áll.

Dolgozatomban ennek az analitikai oldala kerül előtérbe a műveleti oldalával szemben. Az említett cikk bemutatja, hogy a vizsgálatok nagy része a CRM rendszeren belül a vevők megtartásával kapcsolatos (a módszertannal foglalkozó cikkek 62.1%-a ebből a témakörből került ki). Látszik tehát, hogy egy súlyponti kérdésről van szó, melynek módszertani háttere a tudományos vizsgálatok előterében van. Ismert tény, hogy egy vevő megtartásának költsége sokkal kisebb, mint egy új vevő megszerzésének költsége - ld. pl. SEO, RANGANATHAN és BABAD [2008] cikkét a telekommunikáció területén végzett vizsgálatáról.

Az alkalmazott módszereket vizsgálva, mivel az osztályozás szerepelt hangsúlyos területként a vizsgált cikkekben (a vásárlói szokások előrejelzésének fontos eszköze), a neurális hálók (24%), a döntési fák (18%), az asszociációs szabályok (14%) és a regresszió (8%) voltak a legnépszerűbbek.

Látható, hogy a „hagyományos” statisztikai módszerek kisebb részét alkotják az alkalmazott módszereknek. TSIPTSIS és CHORIANOPOULOS [2010] könyvében bőséges leírását és alkalmazási lehetőségeit találhatjuk ezeknek a

DOI: 10.14751/SZIE.2014.035

módszereknek a vásárlói élettartam vizsgálatának (ld. később a 2.6.2. alfejezetben) területén. Ők az előbbi megállapításra, vagyis, hogy az adatbányászati módszerek alkalmazásának súlya egyre nagyobb, azt mondják, hogy a nagyon nagy mennyiségű adatokkal való munka esetében a korábbi módszerek nem elég számítás-hatékonyak (az adatbányászati eszközöket pedig éppen ezekre az esetekre fejlesztették ki). Vagyis a gépi erőforrások nagyon nagy részét lefoglalják, ha ki nem merítik, a futási idők magasak.

Saját vizsgálatomban egy harmadik utat, az ún. valószínűségi modelleket fogom vizsgálni, melynek részleteit a következő alfejezetben mutatom be. Lényege, hogy a vásárlói szokások megfigyelése után valószínűségi változók paramétereinek meghatározására, majd azok alapján jövőbeli viselkedések előrejelzésére nyílik lehetőség.

2.6. A dolgozatban vizsgált problémák irodalmának áttekintése

Dolgozatomban két publikált és alkalmazott módszertani eljárást veszek górcső alá. Az egyik a klaszterelemzés esetében az optimális klaszterszám meghatározásával, míg a másik, a vasárlások jövőbeli számának egyéni szintű előrejelzésével kapcsolatos. Az első egy általánosabb, de a marketingkutatásban gyakran alkalmazott eszköz kiegészítése, a második viszont szorosan a marketingkutatáshoz kapcsolódik, azon belül pedig a vásárlói élettartam vizsgálatának fontos részét képezi.

2.6.1. A klaszteranalízis vizsgálata

Ebben az alfejezetben a sok területen alkalmazott klaszteranalízissel foglalkozom. Először körbejárom a fontosabb fogalmakat, típusokat, majd rátérek a konkrét probléma, az optimális klaszterszám meghatározásának lehetőségeire. Ebből is kiválasztok egy módszert, és annak továbbfejlesztésével foglalkozom az eredmények fejezetben.

A klaszteranalízis fogalma

A klaszterezés egy minta *nem felügyelt csoportosítását* jelenti [Jain, Murty és Flynn, 1999], azaz egy feltáró adatelemzési technika.

Más megfogalmazásban: „A klaszterelemzés az alakfelismerés tanító nélküli tanuló algoritmusa. Egyszerűen úgy definiáljuk, hogy a klaszterelemzés megfigyelések egyedeit bontja viszonylag homogén csoportokba p változó értékeinek hasonlósága alapján. A klaszterelemzés az egyedek olyan csoportosítását keresi, amelyekre igaz, hogy egy egyed egy és csakis egy csoporthoz tartozik, és azokhoz az egyedekhez lesz hasonló, amelyekkel egy klaszterbe került, míg a

DOI: 10.14751/SZIE.2014.035

többi klaszterbe tartozó egyedektől különbözik."[Füstös, Kovács, Meszéna és Simonné, 2004, 160. old.]

Klasszifikáló elemzésnek valamint numerikus taxonómiának is nevezik, mely az 1950-es években indult fejlődének [Sneath, 2005]. Azóta nagyon sokféle módszert dolgoztak ki a fenti célok megvalósításának érdekében. Az irodalmak között megjelentek összefoglaló jellegű ill. egy-egy szakterület számára íródott művek [Everitt, Landau, Leese és Stahl, 2011; Kaufman és Rousseeuw, 2005; Theodoridis és Koutroumbas, 2003]. Magyar szerzők tollából származó könyvek is találhatók ezek között [Füstös és Kovács, 1989; Füstös, Kovács, Meszéna és Simonné, 2004; Kovács, Füstös és Meszéna, 2007; Hajdu, 2003; Simon, 2006].

A módszer lényege, hogy az egyedek olyan csoportosítását hozzuk létre, melyben az egyes csoportokba tartozó elemek nagyobb hasonlóságot mutatnak, mint a különböző csoportba esők. Ezáltal az eredeti (nagyon sok adatból álló) adatbázisunkat tömörítjük (veszteségesen), melynek eredményként kapott információk átláthatók, kezelhetők a szakértők számára. Napjainkban leginkább a személyre szabott szolgáltatások terén használják és fejlesztik ezt a módszert. A csoportba sorolás azonban kétféle módon is értelmezhetjük.

Legyen adott egy minta $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$, $N \in \mathbb{N}$,
ahol $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{in})^T \in \mathbb{R}^n$ *az egyes megfigyelési egységeket (objektumokat) jelenti.*[5]
A minta egy csoportosításán értjük a $C = \{C_1, C_2, \ldots, C_k\}$ $(k \leq N)$ *halmazt, ha*

$$C_i \neq \emptyset \quad , \quad i = 1, 2, \ldots, k \tag{2.1}$$

$$\cup_{i=1}^{k} C_i = \boldsymbol{X} \tag{2.2}$$

$$C_i \cap C_j = \emptyset \quad , \quad i, j = 1, 2, \ldots, k \quad ; \quad i \neq j \tag{2.3}$$

Ez a felfogás a szigorú értelemben vett csoportosítás, melyben minden egyed egyetlen csoporthoz tartozik [Xu és Wunsch, 2008]. Beszélhetünk azonban olyan osztályozásról is, ahol minden egyed több csoportba is tartozhat valamilyen valószínűséggel (a valószínűségek összege 1, minden egyed esetében). Ez az ún. fuzzy megközelítés [Yang, 1993].

Jain és Dubes [1988] művében a klaszterezés folyamata a következő lépésekből áll:

- A mintázat reprezentációja, a minta előfeldolgozása.
- Közelségi (hasonlósági) mérték definiálása.
- Csoportosítás.

[5] A minta N db megfigyelési egységet, valamint n db attribútumot $(A_1, A_2, \ldots, A_n)$ tartalmaz.

DOI: 10.14751/SZIE.2014.035

- Adatok absztrakciója, csoportok jellemzőinek meghatározása.

- Eredmények kiértékelése.

A klaszteranalízisnek mára sok fajtája vált használatossá és használatuk módja megtalálható a mindenkori legfrissebb irodalomban (ld. fentebb). A problémát általában a módszer megválasztása jelenti, mivel nincs univerzális eljárás, amely minden ilyen jellegű feladat megoldására egyformán jó lenne [Jain, Murty és Flynn, 1999]. Kérdéses, hogy a sok lehetőség közül melyik lesz az adott probléma szempontjából megfelelő? Sharma és Kumar [2006] a fenti kérdés megválaszolását a hierarchikus és nem hierarchikus módszerek különbözősége alapján próbálja megválaszolni, és arra a következtetésre jut, hogy érdemes lehet a módszereket egy vizsgálaton belül ötvözni. Hiányossága a könyv ezen fejezetének, hogy a nem hierarchikus módszerekkel csak általánosan foglalkozik, holott egy szerteágazó területről van szó, így nem tud érdemben segítséget nyújtani a kutatónak. Ellenben Everitt et al. [2011] egy részletes ismertetését mutatja a legújabb módszereknek mind a hierarchikus, mind a nem hierarchikus algoritmusok esetében. Magyar nyelvű összefoglaló mű pl. Simon [2006] tollából származik, aki kifejezetten a marketingkutatás területén való alkalmazhatóságot tartja szem előtt, és ad a gyakorlati felhasználás szempontjából is hasznos elemzést.

Természetesen vannak egyszerűen eldönthető kérdések a módszer kiválasztásának folyamatában, pl. a változók mérési szintje, a megfigyelési egységek száma, a változók eloszlása. Ezek bizonyos behatárolást jelentenek, de ezen belül még mindig több lehetőség közül lehet választani. Vannak olyan módszerek, melyek már régóta használatosak, beváltak, melyek mindenki számára ismerősek, könnyen elérhetők. Ezek leginkább a hierarchikus módszerek, ill. a particionáló módszerek közül a legelterjedtebb K-közép módszer. Ezen módszereket előszeretettel alkalmazzák a marketingkutatás több területén is: piacszegmentálás, fogyasztói magatartás megértése, új termékek piaci lehetőségeinek feltárása, tesztpiacok kiválasztása, adatcsökkentés [Malhotra, 2002]. Ha pl. a fogyasztói magatartás vizsgálatát figyeljük, látható, hogy a változók egy része (esetleg az összes) ordinális skálán mérhető mennyiségek, melyek esetében a hasonlóság mérésére általánosan alkalmazott euklideszi távolság nem alkalmazható. Ezért áttekintem a hasonlóság/különbözőség mérés alternatív lehetőségeit. Attól függően, hogy adataink milyen skálán mérhetők, más és más hasonlósági mértéket használunk.

DOI: 10.14751/SZIE.2014.035

Hasonlóság, különbözőség mérése

Bináris adatok esetére CHOI, CHA és TAPPERT [2010] foglalta össze az eddig alkalmazott mérőszámokat. Munkájukban 76 hasonlósági ill. távolság definíciót említenek meg, melyek az 1884 - 2005-ig terjedő időszakban születtek. Ebben az esetben két olyan vektor összehasonlítása történik, melyeknek minden komponense 0, vagy 1 lehet. Vagyis az eredmények egy olyan 2×2-es kontingencia táblába összesíthetők, melyekben az egyes típusok összege n, azaz a vektorok dimenziója[6] ($a+b+c+d = n$). A legismertebb ilyen hasonlósági mértékek a következők:

$$S = \frac{a+d}{a+b+c+d} \quad \text{(Sokal és Michener index)} \tag{2.4}$$

$$S = \frac{ad-bc}{ad+bc} \quad \text{(Yule (Q) index)} \tag{2.5}$$

$$S = \frac{n(ad-bc)^2}{(a+b)(a+c)(c+d)(b+d)} \quad \text{(Pearson (1) index)} \tag{2.6}$$

További fontos eredménye ennek a munkának, hogy összehasonlították[7] ezeket a mérőszámokat egymással, hogy melyek vezetnek közel azonos eredményre, így egy használható összefoglalást adtak a gyakorlati felhasználók kezébe.

Kategorikus adatok esetében BORIAH, CHANDOLA és KUMAR [2008] cikkében találhatunk egy összefoglaló elemzést a hasonlósági mértékekről. A cikk fókuszában a kiugró adatok (outliers) kezelése áll. Véleményük szerint nagyon kevés mű foglalkozik ezzel a problémával (kevesebb, mint a folytonos adatok esetében), azok is javarészt a bináris adatokra történő áttérésre (transzformációra) tesznek javaslatot. Holott, ezen adatok is nagy súllyal vannak jelen a tudományos vizsgálatokban (ld. marketingkutatás). A hasonlóság mérését a következő általános képlet adja meg:

$$S(\boldsymbol{x}_j, \boldsymbol{x}_k) = \sum_{i=1}^{n} w_i S_i(x_{ji}, x_{ki}) \tag{2.7}$$

ahol S_i az attribútumonkénti hasonlóságot adja meg, w_i pedig a hozzá tartozó súly. Az attribútumonkénti hasonlóságok meghatározása többféleképpen történhet, BORIAH, CHANDOLA és KUMAR [2008] 14 ilyen definíciót mutat be. A legegyszerűbb, ha egyező komponensek esetében egyes értéket adunk, ellenkező esetben 0-át:

$$S_i(x_{ji}, x_{ki}) = \begin{cases} 1 & \text{ha } x_{ji} = x_{ki} \\ 0 & \text{egyébként} \end{cases}, \quad w_i = \frac{1}{n} \tag{2.8}$$

[6] a és d az egyezések (0-0, 1-1), b és c a különbözők (1-0, 0-1) számát jelöli.
[7] Hierarchikus klaszterezés segítségével.

DOI: 10.14751/SZIE.2014.035

Ez az irodalomban „overlap” (átfedés) hasonlósági mérőszám néven ismert. Az ilyen típusú indexek mellett vannak valószínűségen alapuló, valamint információelméleti mérőszámok is. Valószínűségen alapuló pl. az ún. *Goodall* mérőszám[8]:

$$S_i(x_{ji}, x_{ki}) = \begin{cases} 1 - \sum_{q \in Q} p_i^2(q) & \text{ha } x_{ji} = x_{ki} \\ 0 & \text{egyébként} \end{cases} \quad , \quad w_i = \frac{1}{n} \tag{2.9}$$

Összehasonlítva a két definíciót, az látszik, hogy az első esetében minden egyes egyezés a megfigyelési egység koordinátái között azonos értékkel szerepel az összegben, míg a második esetben azok az egyezések, melyek véletlenszerű előfordulásának valószínűsége kisebb, nagyobb értékkel szerepelnek az összegben. Eredményeik (empirikus összehasonlító elemzés) fényében arra a következtetésre jutnak, hogy nem lehet a hasonlósági mértékek között legjobbat találni, a különböző jellegű attribútumokon különböző eredményeket értek el.

Intervallum skálán mért adatok esetében hasonló a helyzet az előbb bemutatott változókhoz. Nagyon sokféle mértéket dolgoztak ki itt is. Az egyedek (megfigyelési egységek) hasonlóságának ill. különbözőségének mérését legegyszerűbben a távolságfogalom bevezetésével lehet megoldani.

Legyen X egy nem üres halmaz. Ekkor a $d : X \times X \longrightarrow \mathbb{R}$ függvényt, melyre

$$d(x, y) \geq 0 \quad , \quad d(x, y) = 0 \Longleftrightarrow x = y \quad , \quad \forall x, y \in X \tag{2.10}$$

$$d(x, y) = d(y, x) \quad , \quad \forall x, y \in X \tag{2.11}$$

$$d(x, y) \leq d(x, z) + d(z, y) \quad , \quad \forall x, y, z \in X \tag{2.12}$$

teljesül, az X-en értelmezett metrikának nevezzük.

Az egyedek közötti távolságok egy $n \times n$-es mátrixba rendezhetők (n az egyedek számát jelenti), mely bármely két egyed távolságát tartalmazza. Számításigénye $\binom{n}{2}$, ami azonban nagyon sok megfigyelési egység esetén nagyon nagy szám lesz (hiszen n^2-tel arányos), és ez az adatmennyiség egyidejűleg kezelhetetlen lehet [Bodon, 2010].

Cha [2007] cikkében részletes katalogizálását[9] találjuk ezen mértékeknek. A vizsgálat célja az volt, hogy kimutassa a hasonlóságot a vizsgálatba bevont – összesen 56 db – távolság ill. hasonlóság mérték között.

[8] Ahol $Q \subseteq \text{range}(A_i)$, melyre $\forall q \in Q$ esetén $p_i(q) \leq p_i(x_{ji})$, továbbá $p_i(q)$ az A_i attribútum esetén a q érték előfordulásának relatív gyakorisága.

[9] Minkowski típusú, abszolút eltérés típusú, skalárszorzat típusú, geometriai közép típusú, χ^2 típusú, entrópia típusú.

DOI: 10.14751/SZIE.2014.035

Ezek között a legáltalánosabban használt távolság a *Minkowski* távolság[10]:

$$d(\boldsymbol{x}_i, \boldsymbol{x}_j) = \left(\sum_{i=1}^{n} |x_{ji} - x_{ki}|^p \right)^{\frac{1}{p}} \quad , \quad j, k \in \{1, 2, \ldots, N\} \quad , \quad j \neq k \quad (2.13)$$

Ennek speciális esetei a Manhattan ($p = 1$), valamint az Euklideszi ($p = 2$) távolság. Ezen távolságokkal kapcsolatban Boriah, Chandola és Kumar [2008] azon kritikát fogalmazza meg, hogy nem veszi figyelembe a minta további egyedeinek elhelyezkedését, csak a két vizsgált egyedét. További leszűkítését jelenti az alkalmazhatóságnak, hogy az euklideszi távolsággal jól szeparált, kompakt halmazok azonosíthatók jól [Mao és Jain, 1996].

További, gyakran használt távolság mérték az ún. Mahalanobis [1936] féle távolság, mely a fenti kritikák közül figyelembe veszi az összes pont kapcsolatát a korrelációs mátrixon keresztül. A két megfigyelési egység távolságának definíciója:

$$d(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i - \boldsymbol{x}_j)^T \cdot \boldsymbol{\Sigma}^{-1} \cdot (\boldsymbol{x}_i - \boldsymbol{x}_j) \quad (2.14)$$

ahol $\boldsymbol{\Sigma}^{-1}$ az adatok kovariancia mátrixának az inverze.

Kritikai észrevételek

Egyazon változótípus esetén is sokféle mérték alkalmazható, melyek kimenetele nem feltétlenül lesz azonos. Így tehát a kutatónak kell döntést hoznia arról, hogy melyiket használja, melyik eredményt fogadja el és melyiket nem. Emiatt a módszert sok kritika éri.

Hair et al. [2009] szerint a módszer kevéssé alkalmas következtetés levonására, inkább csak leíró jellegűnek tekinthető az érzékenysége valamint esetlegessége miatt. Kellő óvatossággal való alkalmazását azonban hasznosnak tartják, ugyanis ezen csoportosításokkal kapott mintázatok felderítésére más módszerek nem alkalmasak.

Ugyanazon adatbázis esetében azonban egészen eltérő eredményeket kaphatunk még az alkalmazási feltételek betartása mellett is. Ilyen észrevételek is megfogalmazásra kerültek a klaszterezési eljárásokkal kapcsolatban, mint pl. az input adatok „kicsi” megváltoztatása nagy eltéréseket eredményezhet a végeredményben (nem eléggé robusztus), vagy, hogy a sokféle lehetséges megoldás közül (pl. ugyanazon algoritmus esetében kapott különböző klaszterszámok esetében) ki kell választani valamilyen módon a legjobbat [Hoek, Gendall és Esslemont, 1996].

A sok empirikus vizsgálat mellett azonban elméleti megközelítéseket is találhatunk a klaszterezés vizsgálatára. A klaszterezéssel szemben támasztott

[10] Jelöléseket ld. 25. old.

DOI: 10.14751/SZIE.2014.035

igényekkel elméletben is foglalkozott pl. Kleinberg [2003], aki levezette, hogy egy távolság alapú klaszterező algoritmus nem rendelkezhet egyidejűleg a skála invariancia[11], a gazdagság[12] és a konzisztencia[13] tulajdonságokkal. Vagyis az újabb és újabb eljárások elvi akadályokba is ütköznek a pontosságot, „jóságot" illetően.

2.6.2. A vásárlói élettartam kvantitatív vizsgálata

Ebben az alfejezetben a kutatómunkám második részének irodalmi hátterét dolgozom fel, leszűkítve a vizsgálatot egy bizonyos módszerre. Ezen módszer ismertetését, fontosságát először a tágabb marketingterületbe ágyazva mutatom be, annak előzményeivel, alternatíváival együtt. Ezután térek rá a konkrét modellre és annak egy meglévő továbbfejlesztésére, és megfogalmazom kritikáimat, melyek a modell egy módosított irányú továbbfejlesztéséhez vezettek.

A vásárlói élettartam vizsgálat tartalma, célja

A 80-as évektől kezdve a szakemberek figyelme a vásárlói kapcsolatok marketingje felé irányult [Berger és Nasr, 1998]. A cél: hosszútávú kapcsolat kiépítése a vásárlókkal/fogyasztókkal. Ez természetesen költséges tevékenység, mely esetében különbséget kell tenni vásárló és vásárló között [Blattberg, Getz és Thomas, 2001]. Ahogy Kotler és Armstrong [2010, 26. old.] fogalmaz: „a marketing a jövedelmező vásárlókkal való kapcsolatok irányítása" . A cégek jelentős forrásokat fektettek (fektetnek) be vevőkapcsolati rendszerek felállításába (Customer Relationship Management - CRM). Ezen rendszerek megvalósíthatóságának alapja a vásárlókról rendelkezésre álló adatok egyre növekvő mennyisége. Már nemcsak a szerződéses kapcsolatban álló ügyfelekről állnak rendelkezésre adatok, hanem az elektronikus fizetés, megrendelés használatának elterjedésével, a szerződésben nem álló vásárlókról is egyre több információ beszerezhető. Egy vásárlói tranzakciós adatbázis felépítése segítségével lehetővé válik

- annak előrejelzése, hogy a jövőben melyik vásárló marad aktív,
- a jövőbeni tranzakciók szintjének előrejelzése (egyéni és kollektív szinten).

Ezek az információk az alapjai a CLV (Customer Lifetime Value), vagyis a fogyasztói élettartam érték meghatározásának. Ennek definíciója azonban

[11] Ha minden elempár távolsága helyett annak λ szorosát ($\lambda > 0$) vesszük, akkor a klaszterező eljárás eredménye változatlan marad.

[12] Tetszőleges, előre megadott csoportosításhoz meg lehet adni távolságot úgy, hogy az eljárás az adott módon csoportosítson.

[13] Ha az egy csoportba került elemek távolságát csökkentem, valamint a külön csoportba került elemek távolságát növelem, akkor a klaszterezés eredménye ugyanaz lesz mint az előbb.

DOI: 10.14751/SZIE.2014.035

nem teljesen egységes a szakirodalomban, továbbá számítására is különböző modellek születtek. Az egyik megközelítés szerint a *vásárló által előállított profit jelenértékét jelenti arra a jövőbeni időtartamra, ameddig a vásárló kapcsolatban áll a céggel* [Gupta et al., 2006]. Más megközelítés szerint inkább a vásárlóval kapcsolatba hozható jövőbeli cashflow jelenértékét jelenti [Pfeifer, Haskins és Conroy, 2005][14]. Ez utóbbi esetben a szerzők különbséget tesznek a két fogalom között, ami a felmerülő költségek elszámolási különbségei miatt tehető meg. Gupta et al. [2006] nyomán a számítás a következőképpen történik:

$$CLV = \sum_{t=0}^{T} \frac{(p_t - c_t) r_t}{(1+i)^t} - AC \tag{2.15}$$

ahol
p_t: a vevő által fizetett ár (t időpontban),
c_t: a vevő kiszolgálásának direkt költsége (t időpontban),
r_t: annak a valószínűsége, hogy a vásárló még aktív (t időpontban),
i: diszkont ráta,
AC: a vevő megszerzésének költsége,
T: a vizsgált időtartam.

A fogalom lényege, hogy a vevőket egyenként lehet értékelni, jövőbeli aktivitásukat meghatározni. Ezen adatok aggregálásából kapjuk a CE (Customer equity), a vásárlói tőke fogalmát: $CE = \sum_{vásárlók} CLV$. Ezen értékek pontossága (többek között) a fent említett két mennyiség előrejelzésének pontosságán múlik. Ezek meghatározására sokféle modell létezik.

A CLV számítás módszerei

1. Elsőként meg kell említeni a régóta használatos (a mindennapi gyakorlatban népszerű) *RFM modellt* (Recency, Frequency, Monetary). Ehhez a következő adatok összegyűjtése szükséges vásárlói szinten: utolsó vásárlás időpontja, vásárlások gyakorisága, a vásárlások alkalmával elköltött pénz mennyisége.

   A modell lényege, hogy a jövőt a múlt függvényeként írja le, vagyis a múltbeli értékekre számított függvényértékek adják a jövőbeli értékeket, azaz $jövő = f(múlt)$. Ez a modell lényegében egy regressziós modell, mely Malthouse és Blattberg [2005] cikkében a következő alakban található: $g(CLV_i) = f(\boldsymbol{x}_i) + e_i$, ahol $\boldsymbol{x}_i$ tartalmazza az i-edik vásárlóra vonatkozó múltbeli vásárlásokkal kapcsolatos információkat, e_i pedig

[14] A szerzők hangsúlyozzák a fogalmak pontos definiálásának fontosságát (természetesen nemcsak ezen fogalmak esetében).

DOI: 10.14751/SZIE.2014.035

a hibatag. Itt a g függvénynek variancia-stabilizáló szerepe van. A módszer alkalmazásával kapcsolatban azonban vannak ellenvetések. Például a megfigyelt változók (Recency, Frequency, Monetary) csak indikátorai a háttérben meghúzódó összefüggéseknek [Fader, Hardie és Lee, 2005b], vagyis nem ők az okok, melyekből az okozat következik. Márpedig a fent vázolt függvénykapcsolat egy ok-okozati kapcsolatot jelent.

2. Az RFM analízis gyakorlatban alkalmazott legáltalánosabb módja azonban az ún. pontozásos módszer[15] [McCarty és Hastak, 2007], melyekben a változók (R, F, M) értékeit pontokká transzformálják, majd a változókat szakértői vélemény alapján súlyozzák, és ezek után minden megfigyelési egységhez (vásárló) hozzárendelik az ő pontszámát [Miglautsch, 2000]. Fader, Hardie és Lee [2005b] ezzel kapcsolatban is problémákat fogalmaz meg, pl. az így elkészült modell statikus voltát, azaz nem lehet 2-3 periódussal előbbre tekinteni, csak 1 periódusra ad választ (utána megint el kellene készíteni a pontozást). Maguk a változók olyan információkat hordoznak magukban, melyek kiaknázása lehetséges, csak a most említett pontozásos módszer erre nem alkalmas, másként fogalmazva, pontosabb előrejelzés is adható ugyanezen adatok felhasználásával.

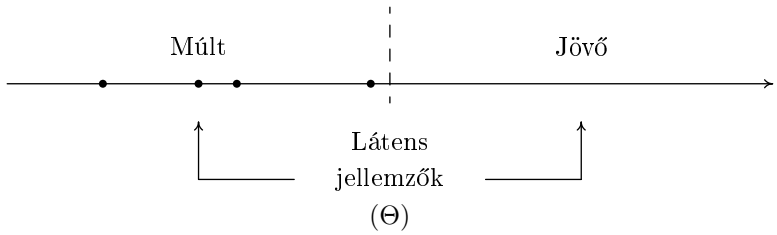
3. A vásárlói élettartam vizsgálat egy másik lehetősége egy úgynevezett *valószínűségi modell* felállítása. A valószínűségi modellben a megfigyelhető viselkedésmódokat úgy tekintjük, mint az egyénre jellemző látens tulajdonságok (melyek egyénenként változnak) által vezérelt sztochasztikus folyamatok megvalósulásai [Gupta et al., 2006]. Ebben az esetben azt mondjuk, hogy a múlt és a jövő között egy nem látható kapcsolat létezik, de nem a korábban említett – a jövő a múltbeli adatok függvénye (ld. RFM modell, 31. old.) – függvénykapcsolat, hanem egy áttételes kapcsolat, melyet megpróbálunk modellezni. Jelölje az adott egyénre jellemző látens tulajdonságokat Θ.

   Ekkor a megfigyelt múlt (vagyis a megfigyelt vásárlói magatartás) ezen jellemzők függvénye, továbbá a megvalósuló jövő is ezen információk függvénye, azaz $múlt = f(\Theta)$, $jövő = f(\Theta)$, amint azt a 2. ábra szimbolizálja [Fader és Hardie, 2009]. Mivel a vásárlói (látens) tulajdonságokat nem ismerjük, az előrejelzést két lépésben tudjuk megtenni. Az első lépésben a megfigyelt viselkedésmódokat írjuk le valamilyen valószínűségi modellel, és a kapott modellből határozzuk meg a látens tulajdonságokat, mint okokat. Ez utóbbi számítására a Bayes tétel[16] nyújt lehetőséget. Az első lépés-

[15] Pl. az SPSS programcsomag 17-es változata is tartalmaz ilyen modult SPSS EZ RFM néven, `http://www.docs.is.ed.ac.uk/skills/documents/3663/SPSSEZRFM17.0.pdf`

[16] „A Bayes-tétel tulajdonképpen egy olyan formula, amely lehetővé teszi azt, hogy valamely A esemény

DOI: 10.14751/SZIE.2014.035



2. ábra. A tranzakciós eredmények valószínűségi modelljének alapja
Forrás: FADER és HARDIE [2009]

ben, tehát az egyén esetében meghatározzuk, hogy milyen eloszlást követ a megfigyelt mintázat[17], majd pedig figyelembe vesszük, hogy a látens karakterisztikák vásárlónként változnak (az eloszlás paraméterei vásárlóról vásárlóra változnak). A vásárlókra jellemző paraméterek meghatározására is valószínűségi eloszlásokat alkalmazunk[18], és egy véletlenszerűen kiválasztott vásárlót a két eloszlás alapján tudunk jellemezni. Vagyis az adatokra illesztett modellből következtetünk az ezen eredményeket okozó látens jellemzőkre (egyénenként megkapjuk az eloszlások paramétereit). Ennek ismeretében, a második lépésben, a megfigyelt változók jövőbeli értékének előrejelzése válik lehetségessé az előbb felállított modell segítségével, az egyének (most már) ismert látens jellemzőinek (az eloszlások paramétereinek) ismeretében (2. ábra).

FADER, HARDIE és LEE [2006] szerint több érv is szól ezen valószínűségi modellek mellett, összehasonlítva őket a regressziós modellekkel. Így például nem kell a meglévő adatbázist kettéosztani (függő és független változókra), hanem az összes meglévő adatból (mint független változóból) építhető fel a modell. Másrészt, a becsülni kívánt időszak hosszát nem kell korlátozni (mint pl. a pontozásos modellnél).
FADER és HARDIE [2009] cikkében két csoportba sorolja ezeket az eljárásokat, olyanokra, melyek szerződéses kapcsolatok leírására, valamint olyanokra, amelyek nem szerződéses kapcsolatok leírására alkalmasak. Lényeges különbség a kettő között, hogy a szerződéses kapcsolat esetében a cég (szolgáltató) visszajelzést kap arról, hogy egy ügyfél elpártolt tőle,

bekövetkezéséből következtessünk a lehetséges $B_1, B_2, \ldots, B_n$ események, »hipotézisek«, »okok« valószínűségére." (Dobó Andor: A „következtetési tétel" következményei, `http://www.freeweb.hu/doboandor/pdfs/a_kovetkeztetesi_tetel_kovetkezmenyei.pdf`).
Bayes tétel: Ha a nem nulla valószínűségű $B_1, B_2, \ldots, B_n \in \mathcal{A}$ események teljes eseményrendszert alkotnak, akkor egy tetszőleges $A \in \mathcal{A}$, $P(A) > 0$ valószínűségű esemény esetén: $P(B_k|A) = \frac{P(A|B_k)P(B_k)}{\sum_{i=1}^{n} P(A|B_i)P(B_i)}$

[17]Pl. a vásárlások időpontjai közötti időtartamot exponenciális eloszlással írjuk le.
[18]Pl. a fenti esetben gamma eloszlást

DOI: 10.14751/SZIE.2014.035

míg nem szerződéses kapcsolat esetén erre vonatkozóan csak valószínűségi kijelentések tehetők. Mindkét kategórián belül megkülönböztetnek még két csoportot: a két tranzakció között eltelt idő folytonos vagy diszkrét valószínűségi változóval írható le.

(a) Tekintsük először a *nem szerződéses kapcsolat* leírását. A modellkészítés alapjai a 60-as évekre nyúlnak vissza, mikor is megjelent Ehrenberg [1959] modellje, mely negatív binomiális eloszlásra (NBD) épül. Célja az volt, hogy a nem tartós fogyasztási cikkek vásárlásának leírását megadja. Az NBD alapú eljárások lényege, hogy

- egy adott vásárló esetén egy időegységre jutó vásárlások számát Poisson eloszlással közelíti, melynek várható értéke λ,
- a vásárlók közötti eltéréseket, vagyis a λ - mint valószínűségi változó - eloszlását, gamma eloszlással közelíti.

E két eloszlás kompozíciója adja a keresett eloszlást.

Ezen módszer továbbfejlesztése született meg Schmittlein, Morrison és Colombo [1987] által, akik bevezették, hogy a vásárló „élettartama" nem végtelen[19]. Amíg a vásárló aktív, addig a vásárlások leírására az előbbi modellt (NBD) alkalmazták. A vásárló élettartamát pedig exponenciális eloszlással, ennek paraméterét – mely egyénenként változik, tehát egy valószínűségi változónak tekinthetjük – gamma eloszlással írták le, melyek együttesen az ún. Pareto eloszlást alkotják. Az így keletkezett modell a Pareto/NBD modell nevet kapta. A modell bemenő adatai, melyekből a négy eloszlás paramétereit meghatározzák, a vásárlások száma (x), valamint az utolsó vásárlás időpontja (t_x) egy vizsgált T időtartam alatt. Vagyis, még az egyes vásárlások időpontja sem kell (mint később látni fogjuk, ezek kiesnek a számítások során), jóllehet ezek az összeállított adatbázisban rendelkezésre állnának. A tesztelések során jó eredményeket mutatott [Schmittlein és Peterson, 1994], mégsem terjedt el a gyakorlati alkalmazások terén. Ennek okát abban látják, hogy a paraméterbecslés nehéz matematikai eljárásokból áll, vagyis a gyakorlati felhasználók képzettsége nem elégséges hozzá. Emiatt születtek egyéb megoldások a modell módosítására.

Az egyik ilyen lehetőség, hogy a paraméterbecslést az MCMC (Markov Chain Monte Carlo) módszer segítségével végezték el, mely az

[19] Ők a „buy till you die" fogalmat használták, én azonban ennek magyar tükörfordítását nem alkalmazom dolgozatomban, helyette a „kapcsolat megszakítása", lemorzsolódása" ill. „inaktívvá válás" kifejezéseket használom szinonimaként. A dolgozat nem foglalkozik azzal a problémával, hogy egy vásárló véglegesen, vagy csak időlegesen pártol el. Ezért a továbbiakban az inaktívvá válás megegyezik a lemorzsolódással.

DOI: 10.14751/SZIE.2014.035

elvi alapjait nem változtatta meg, ám a számításokat (az eloszlások paramétereinek meghatározását) leegyszerűsítette (lényegében kísérletek eredményének átlagát számolja várható érték helyett) [Ma és Liu, 2007].
A másik irány az volt, hogy feltételezték, hogy a vásárló nem szakíthatja meg bármikor a kapcsolatát, hanem minden vásárlás után p valószínűséggel pártol el [Fader, Hardie és Lee, 2005a]. Így ennek leírására már nem az exponenciális eloszlást, hanem geometriai eloszlást használtak, míg a p paraméternek a vásárlók közötti változékonyságát béta eloszlással írták le. A vásárlások közötti időtartamok jellemzése úgy történik, mint a Pareto/NBD modell esetén. Az így kapott modell előnye, hogy a paraméterbecslés sokkal egyszerűbb, így gyakorlati felhasználásra alkalmasabb. Vizsgálataimat ezen modellek továbbfejlesztése terén végeztem, melyek az eredmények fejezetben találhatóak.

(b) A *szerződéses kapcsolatok* leírására is több modellt fejlesztettek ki. Itt azonban a megválaszolandó kérdések egy kicsit mások. Melyik vásárló esetén legnagyobb a kockázata annak, hogy a következő periódusban elpártol a cégtől? Milyen időtartamra (hány periódusra) tervezhetjük, hogy a vásárló kapcsolatban marad a céggel? (Mivel dolgozatomnak nem ez az iránya, ezzel kevésbé részletesen foglalkozom, mint az előző kapcsolattípussal.)
Az első kérdés megválaszolására alkalmas módszerek: logisztikus regresszió, döntési fák, neurális hálók (melyek azonban már átnyúlnak felsorolásunk 4. pontjába) [Berry és Linoff, 2004, 116 - 120. old]. A választ egy kétértékű változó adja, mely értékének meghatározását végezhetjük el a fenti eljárásokkal. Lényegében a korábbi időszakokban a vásárlókról összegyűjtött adatok ismeretében kell csoportosítani a vásárlókat, kik maradnak meg, és kik bontják majd fel a kapcsolatot.
A második kérdés nehezebb, mivel több periódussal előre szeretnénk látni. Ennek egyik lehetséges megoldása az ún. túlélés analízis [Berry és Linoff, 2004], mely statisztikai eljárás az orvosi és műszaki tudományok területéről indult. Egy adott időpontban annak a valószínűsége, hogy a vásárló aktív lesz a következő időpontban is (vö. szerződéses kapcsolat) képezi az alapját annak, hogy bármely későbbi időpontra is valamilyen valószínűséggel meghatározhassuk a vásárló „túlélését". A valószínűségszámítás szorzási szabályának értelmében, annak a valószínűsége, hogy egy ügyfél a következő k periódus mindegyikét túléli, az egyes feltételes valószínűségek szorzataként állítható

DOI: 10.14751/SZIE.2014.035

elő (mely k növekedtével monoton csökken).

4. A számítási kapacitás ugrásszerű megnövekedése folytán folyamatosan születtek meg a különböző adatbányászati eljárások, melyek között vannak olyanok, amelyeknek előrejelző funkciója is van. Ezen algoritmusok segítségével megalkotott modelleket nevezi Gupta et al. [2006] „*informatikai modellek*"-nek. Ilyenek pl.: neurális hálók, döntési fák, általánosított additív modellek, support vector machine. Ezek a módszerek nagy mennyiségű változó (sokdimenziós tér) esetében nyújtanak jól használható megoldásokat.
   Lemmens és Croux [2006] az ún. bagging és boosting eljárásokat alkalmazza a fogyasztók elpártolásának előrejelzésére. A bagging [Breiman, 1996] egy olyan eljárás, mely az előrejelző modellek többszöri lefuttatása (tanulási minták egy sorozatán lefuttatva) esetén kapott eredményeket határozza meg első lépésként. Ezek után meghatározza ezen eredmények várható értékét (gyakorlatban az átlagát). Az így kapott becslés lesz az előrejelzés eredménye. Lemmens és Croux [2006] vizsgálata azt mutatja, hogy a bagging módszerrel elért eredmények szisztematikusan jobbak lettek, mint az összehasonlításra használt döntési fa eredményei. Gupta et al. [2006] szerint ezen módszerek alkalmazása inkább a tudományos kutatásokban találhatók meg, kevéssé ismertek még a marketing gyakorlatban, ám a jövőre nézve nagyobb figyelmet fognak kapni. Ezzel jelen dolgozat szerzője is egyetért, hiszen a napi szinten keletkező adatokból képződő adatbázisok információtartalmának kinyerése a módszertanban is kutatásokat generál, melyek azután megjelennek a gyakorlati alkalmazásokban is.

5. Az eddigi vizsgálatok módszere az volt, hogy az egyének jövőbeni adatainak előrejelzése után, azok összegzésével lehetett megkapni a vásárlói bázissal kapcsolatos információkat. Egy másik lehetséges módja az előrejelzésnek, hogy a múltbeli adatainkat nem egyéni szinten, hanem aggregált formában használjuk fel. Ezen számításokat a „*növekedési modell*"-ek segítségével lehet elvégezni. Pl. a jövőben belépő új vásárlók számának becslése [Gupta, Lehmann és Stuart, 2004] a CLV érték meghatározására, vagy az elpártoló vásárlók hatásának elemzése [Hogan, Lemon és Libai, 2003], ahol az eladások számának meghatározásánál használnak ilyen modellt.

Látható, hogy a tudományos cikkek különböző módszertani megoldásokat tartalmaznak, fejlesztenek ugyanazon probléma, jelen esetben a vásárlók jövőben várható aktivitásának mérésére. Ez a sokféle megközelítés természetes,

DOI: 10.14751/SZIE.2014.035

hiszen egy előrejelzésről van szó, melynek jelentős anyagi vonzata lehet a cégek számára. A számításoknak azonban egyéb „buktatói" is vannak az ügyfél jövőbeni vásárlásszámának és az egyes vásárlások értékének meghatározásán túl. Az egyik fő probléma a különböző költségek vásárlókhoz rendelésének nehézsége (a nem aggregált adatokkal dolgozó módszerek esetében). Másrészről a konkurenciához való átpártolás ill. más termékre váltás motivációjáról nem tudunk a fent említett módon információhoz jutni, csak mintavétel segítségével, holott az eddigi számítások alapja egy olyan adatbázis volt, melyben szereplő adatok minden vásárlóra egyaránt megvoltak. További kérdés, hogy az adatbányászati technikák fölül fogják-e múlni a „hagyományos" (matematikai-statisztikai) módszereket, vagy esetleg egymást kiegészítve adnak majd pontosabb jövőképet.

A BG/NBD modell

Kutatómunkám második részében az előző alszakasz 3. pontjában említett nem szerződéses kapcsolatok modellezését vizsgáltam a vásárlói piacon. Ebben az alfejezetben bemutatom azt a modellt, melynek módosítását, továbbfejlesztését tűztem ki célul. Az ún. BG/NBD modell megalkotása Fader, Hardie és Lee [2005a] nevéhez fűződik[20].

A modell az alábbi feltételezéseken alapszik:

1. A vásárlások között eltelt idő exponenciális eloszlást követ λ paraméterrel. Sűrűségfüggvénye:

$$f(t_j|t_{j-1};\lambda) = \lambda\,\mathrm{e}^{-\lambda(t_j-t_{j-1})} \tag{2.16}$$

   ahol t_j a j-edik vásárlás időpontja, és λ a két vásárlás időpontja között eltelt idő várható értéke.

2. λ (vagyis az egyes vásárlókra jellemző paraméter) változékonyságának értéke gamma eloszlást követ. Sűrűségfüggvénye:

$$f(\lambda|r,\alpha) = \frac{\alpha^r \lambda^{r-1}\,\mathrm{e}^{-\lambda\alpha}}{\Gamma(r)} \tag{2.17}$$

   ahol r („shape") és α („inverse scale") az eloszlás két paramétere[21] , Γ pedig a gamma függvény[22].

[20] Ezen összefoglaló a megadott cikk alapján készült.

[21] A gamma eloszlású valószínűségi változó várható értéke $\frac{r}{\alpha}$, varianciája pedig $\frac{r}{\alpha^2}$.

[22] $\Gamma(r) = \int_0^\infty t^{r-1}\,\mathrm{e}^{-t}\,\mathrm{d}t$

DOI: 10.14751/SZIE.2014.035

3. Minden vásárlás után a vásárló p valószínűséggel inaktívvá válik. Ennek leírására a geometriai eloszlás alkalmas:

   $$\mathrm{P}(\text{inaktívvá válik a } j\text{-edik vásárlás után}) = p(1-p)^{j-1} \quad , \quad j = 1, 2, 3, \ldots$$

4. A p paraméter változékonysága béta eloszlást követ, melynek sűrűségfüggvénye

   $$f(p|a,b) = \frac{p^{a-1}(1-p)^{b-1}}{B(a,b)} \tag{2.18}$$

   ahol a és b („shape") az eloszlás két paramétere[23], B pedig a béta függvény.[24]

5. λ és p vásárlónkénti értékei egymástól függetlenül változnak.

Tegyük fel, hogy az adatgyűjtés a $[0;T]$ időintervallumban készült, és ez idő alatt egy kiválasztott vásárló x db vásárlást hajtott végre, melyek időpontjai: $t_1, t_2, \ldots, t_x$.

Első lépésként a vásárlói szintű Likelihood függvény[25] kerül előállításra.

- Annak, hogy az első vásárlás t_1-kor következik be a likelihood komponense (exponenciális eloszlást feltételezve): $\lambda\,\mathrm{e}^{-\lambda t_1}$.
- Annak, hogy a második vásárlás t_2-kor következik be (aktív marad t_1 után és az időtartam exp. eloszlású) a likelihood komponense:
  $(1-p)\lambda\,\mathrm{e}^{-\lambda(t_2-t_1)}$.
  $\vdots$
- Annak, hogy az x-edik vásárlás t_x-kor következik be (aktív marad t_{x-1} után és az időtartam exp. eloszlású) likelihood komponense:
  $(1-p)\lambda\,\mathrm{e}^{-\lambda(t_x-t_{x-1})}$.
- Annak, hogy nem vásárol a $[t_x;T]$ időintervallumban (az x-edik vásárlás után elpártol, vagy a következő vásárlása T időpont után következik be) a likelihood komponense:
  $p + (1-p)\lambda\,\mathrm{e}^{-\lambda(T-t_x)}$.

[23] A béta eloszlású valószínűségi változó várható értéke $\frac{a}{a+b}$, varianciája pedig $\frac{ab}{(a+b)^2(a+b+1)}$.

[24] $\mathrm{B}(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}\,\mathrm{d}x$

[25] A Likelihood függvény a valószínűségi modell paramétereinek függvénye. Legyen $X_1, X_2, \ldots, X_n$ egy-egy valószínűségi változó, melyeknek értékeit mérjük (az $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^T$ jelentsen egy lehetséges mintát, $\boldsymbol{x}^* = (x_1^*, x_2^*, \ldots, x_n^*)^T$ pedig a mérési eredmények vektorát, azaz a minta realizációját). Ezen valószínűségi változók együttes sűrűségfüggvényének paraméterét (paramétereit) jelöljük Θ-val. Ezt a sűrűségfüggvényt a következőképpen jelölhetjük: $f(\boldsymbol{x}|\Theta)$. Ha behelyettesítjük a mért értékeket (x_i^*-okat) a függvénybe, az ismeretlen Θ lesz a függvény változója. Ezt a függvényt nevezzük Likelihood függvénynek és a következőképpen jelölhetjük: $L(\Theta|\boldsymbol{x}^*)$.

DOI: 10.14751/SZIE.2014.035

Ezen komponensek szorzata adja az egyén-szintű Likelihood függvényt:

$$\begin{aligned} L(\lambda, p|t_1, t_2 \ldots, t_x, T) = \\ &= \lambda \, \mathrm{e}^{-\lambda t_1}(1-p)\lambda \, \mathrm{e}^{-\lambda(t_2-t_1)} \ldots (1-p)\lambda \, \mathrm{e}^{-\lambda(t_x-t_{x-1})} \cdot \\ &\quad \cdot \left(p + (1-p)\lambda \, \mathrm{e}^{-\lambda(T-t_x)}\right) = \\ &= p(1-p)^{x-1}\lambda^x \, \mathrm{e}^{-\lambda t_x} + (1-p)^x \lambda^x \, \mathrm{e}^{-\lambda T} \end{aligned} \tag{2.19}$$

Jelölje $X(t)$ a t időpontig megtörtént vásárlások számát egy adott vásárló esetén. Ekkor annak a valószínűsége, hogy ennek értéke éppen x lesz a következő formulával adható meg:

$$\begin{aligned} \mathrm{P}(X(t) = x|\lambda, p) = &(1-p)^x \frac{(\lambda t)^x \, \mathrm{e}^{-\lambda t}}{x!} + \\ &+ \delta_{x>0} \; p(1-p)^{x-1} \left[1 - \mathrm{e}^{-\lambda t} \sum_{j=0}^{x-1} \frac{(\lambda t)^j}{j!}\right] \end{aligned} \tag{2.20}$$

ahol $\delta_{x>0} = \begin{cases} 1, & \text{ha } x > 0 \\ 0, & \text{ha } x = 0 \end{cases}$

Ezen adatok ismeretében arra keresik a választ, hogy mennyi a t idő alatt bekövetkező tranzakciók számának várható értéke, vagyis $\mathrm{E}\left(X(t)\right)$. Ha a vásárló a t időpontban még aktív, akkor ez λt. Azonban előfordulhat, hogy egy $\tau < t$ időpontban inaktívvá válik, így ennek valószínűségét is számításba kell venni. Ezért

$$\begin{aligned} \mathrm{E}\left(X(t)|\lambda, p\right) &= \lambda t \cdot \mathrm{P}(\tau > t) + \int_0^t \lambda\tau \cdot g(\tau|\lambda, p) \, \mathrm{d}\tau = \\ &= \frac{1}{p} - \frac{1}{p} \, \mathrm{e}^{-\lambda p t} \end{aligned} \tag{2.21}$$

ahol $g(\tau|\lambda, p) = \lambda p \, \mathrm{e}^{-\lambda p \tau}$, az inaktívvá válás időpontjának sűrűségfüggvénye.

Az eddigi megállapítások λ és p ismeretét feltételezték (egy adott vásárlóra érvényesek), azonban ezeket a mintából kellene meghatározni, vagyis ezek is valószínűségi változók (ld. 2.17. és 2.18. egyenletek).

Ezután tehát egy véletlenszerűen kiválasztott vásárló esetében kell megadni ugyanezeket (Likelihood függvény, várható érték). Az előbbi eredményekből látszik, hogy három adatot használtak egy-egy vásárló adatai közül: a vizsgált időtartam (T), a vásárlások száma a vizsgált időtartam alatt (x), és az utolsó vásárlás időpontja (t_x), hiszen a többi vásárlás időpontja kiesett a modellből (ld. 2.19. egyenlet). A populációra számított Likelihood függvény az egyén-

DOI: 10.14751/SZIE.2014.035

szintű Likelihood függvényből származtatható:

$$L_i(r,\alpha,a,b|x_i,t_{x_i},T_i) = \\ = \int_0^1 \int_0^\infty L(\lambda,p|x_i,t_{x_i},T_i)\cdot f(\lambda|r,\alpha)f(p|a,b)\,\mathrm{d}\lambda\,\mathrm{d}p \tag{2.22}$$

Az integrálás elvégzése után a következő kifejezés áll elő:

$$L_i(r,\alpha,a,b|x_i,t_{x_i},T_i) = \frac{\mathrm{B}(a,b+x_i)}{\mathrm{B}(a,b)}\frac{\Gamma(r+x_i)\alpha^r}{\Gamma(r)(\alpha+T_i)^{r+x_i}} + \\ + \delta_{x>0}\frac{\mathrm{B}(a+1,b+x_i-1)}{\mathrm{B}(a,b)}\frac{\Gamma(r+x_i)\alpha^r}{\Gamma(r)(\alpha+t_{x_i})^{r+x_i}} \tag{2.23}$$

A négy paraméter meghatározását a maximum likelihood módszer segítségével végezték. A lényege, hogy olyan paramétereket keresnek, amelyek esetében az adott minta előfordulásának valószínűsége maximális, vagyis a Likelihood függvény maximumát kell meghatározni. Ebben az esetben célszerű a függvény logaritmusának (LL) maximumát keresni. Mivel az összes vásárlóra számított Likelihood függvényt a 2.23. egyenletből így kapjuk:

$$L(r,\alpha,a,b) = \prod_{i=1}^{N} L_i(r,\alpha,a,b|x_i,t_{x_i},T_i) \tag{2.24}$$

Ezért ennek logaritmusa:

$$LL(r,\alpha,a,b) = \ln\left[L(r,\alpha,a,b)\right] = \sum_{i=1}^{N}\ln\left[L_i(r,\alpha,a,b|x_i,t_{x_i},T_i)\right] \tag{2.25}$$

ahol N a vásárlók száma a megfigyelt időszakban, x_i, t_{x_i}, T_i az i-edik vásárló adatai.
Ennek a függvénynek a maximuma numerikus optimalizáló eljárással meghatározható.

Hasonló módon határozható meg a $\mathrm{P}(X(t)|r,\alpha,a,b)$ valószínűség – mely egy véletlenszerűen kiválasztott vásárló esetén adja meg az $X(t)=x$ esemény valószínűségét – a $\mathrm{P}(X(t)|\lambda,p)$ (ld. 2.20. egyenlet) valószínűségből (mely egy adott vásárlóra vonatkozik).

$$\mathrm{P}(X(t)|r,\alpha,a,b) = \int_0^1\int_0^\infty \mathrm{P}(X(t)|\lambda,p)\cdot f(\lambda|r,\alpha)f(p|a,b)\,\mathrm{d}\lambda\,\mathrm{d}p = \\ = \frac{\mathrm{B}(a,b+x)}{\mathrm{B}(a,b)}\frac{\Gamma(r+x)}{\Gamma(r)x!}\left(\frac{\alpha}{\alpha+t}\right)^r\left(\frac{t}{\alpha+t}\right)^x + \\ + \delta_{x>0}\frac{\mathrm{B}(a-1,b+x-1)}{\mathrm{B}(a,b)}\left[1-\left(\frac{\alpha}{\alpha+t}\right)^r\left\{\sum_{j=0}^{x-1}\frac{\Gamma(r+j)}{\Gamma(r)j!}\left(\frac{t}{\alpha+t}\right)^j\right\}\right] \tag{2.26}$$

DOI: 10.14751/SZIE.2014.035

Ezek után az $X(t)$ várható értékét is meghatározták egy véletlenszerűen kiválasztott vásárló esetén, mely a 2.21. egyenletből számolható. Vagyis t idő alatt, átlagosan, ennyi tranzakciót bonyolított egy vásárló.

$$\begin{aligned} \mathrm{E}\left(X(t)|r,\alpha,a,b\right) &= \int_0^1 \int_0^\infty \mathrm{E}\left(X(t)|\lambda,p\right)\cdot f(\lambda|r,\alpha) f(p|a,b)\,\mathrm{d}\lambda\,\mathrm{d}p = \\ &= \frac{a+b-1}{a-1}\left[1-\left(\frac{\alpha}{\alpha+t}\right)^r \cdot {}_2F_1\left(r,b;a+b-1;\frac{t}{\alpha+t}\right)\right] \end{aligned} \tag{2.27}$$

ahol ${}_2F_1$ a Gauss-féle hipergeometriai függvény.

Ebből kiindulva juthatunk el az eredeti kérdés megoldásához, vagyis, hogy egy konkrét vásárló (az előzmények ismeretében) t idő alatt hányszor vásárol a cégtől. A feladat tehát $\mathrm{E}\left(Y(t)|x,t_x,T,r,\alpha,a,b\right)$ kiszámítása[26]. $Y(t)$ jelöli a jövőbeli vásárlások számát $[T;T+t]$ időtartam alatt.

$$\begin{aligned} \mathrm{E}\left(Y(t)|x,t_x,T,r,\alpha,a,b\right) &= \frac{a+b+x-1}{a-1}\cdot \\ &\cdot \frac{1-\left(\frac{\alpha+T}{\alpha+T+t}\right)^{r+x}\cdot {}_2F_1\left(r+x,b+x;a+b+x-1;\frac{t}{\alpha+T+t}\right)}{1+\delta_{x>0}\frac{a}{b+x-1}\left(\frac{\alpha+T}{\alpha+t_x}\right)^{r+x}} \end{aligned} \tag{2.28}$$

Heurisztikus modell

Míg tudományos kutatások egész sora hoz létre újabb és újabb előrejelző modelleket, mint pl. az előbb említett modellek, addig felhasználói oldalon nem igazán mutatkozik széleskörű igény ezek felhasználására [Wübben és Wangenheim, 2008]. A felhasználók sokkal inkább alkalmaznak egyszerű, gyors, információszegény heurisztikákat [Huang, 2012]. Ezekkel kapcsolatban további tudományos vizsgálatok is napvilágot láttak, melyekben összehasonlításra kerültek ezek a heurisztikus modellek más, tudományos (pl. sztochasztikus) modellekkel. Goldstein és Gigerenzer [2009] pl. több területről (sport, üzlet, bűnözés) is felkutatott összehasonlításokat az „egyszerű" és „komplex" modellek között. A vásárlói szokások előrejelzésének területén Wübben és Wangenheim [2008] eredményeit publikálja, melyek szerint a vásárlói aktivitás előrejelzésében a Pareto/NBD ill. a BG/NBD modellek eredményei nem jobbak, mint a menedzseri gyakorlatból ismert heurisztikus modell eredményei.

Az általuk használt heurisztikus modellt „hiatus heuristic"-nek nevezték, melynek lényege, hogy egy adott üzleti területen a vásárlási adatokból úgy határozzák meg, hogy egy adott vásárló még aktívnak tekinthető-e, hogy az utolsó

[26] A levezetés megtalálható a fent hivatkozott cikkben.

DOI: 10.14751/SZIE.2014.035

vásárlásának időpontját egy előre (az eddigi tapasztalatok alapján optimálisnak tűnő) időponthoz („hiatus") viszonyítják. Ha ennél a kritikus időpontnál régebben vásárolt, akkor őt inaktívnak tekintik.

Természetesen a kritikus időpont értéke az üzleti területeken más és más [Wübben és Wangenheim, 2008]. Ennek meghatározásához szükséges az előző időszakok adatai mellett a szakértői intuíció is, aminek segítségével kialakított modell – mint látható a fenti példából – a kisebb eszközrendszere ellenére is tud jó eredményt elérni. Vizsgálataimba én is bevontam ezt a modellt, ezzel tovább bővítve ezen összehasonlító elemzések számát is.

DOI: 10.14751/SZIE.2014.035

3. fejezet

Anyag és módszer

Ebben a fejezetben kutatási munkám előzményeinek bemutatására, kritikai észrevételek megtételére, ill. a továbblépés indoklására kerül sor. A fejlesztések, módosítások már az Eredmények fejezetben kerülnek bemutatásra, hiszen ezek már a kutatás fő részét képezik.

Továbbá ebben a fejezetben térek ki a felhasznált adatbázisok bemutatására, valamint a teszteredmények vizsgálatának módszertani hátterére.

A modellszámításokat és azok tesztelését az R nyelv és környezet [R Core Team, 2013] segítségével végeztem. A szoftver statisztikai számítások végzésére és ábrák előállítására fejlesztett szabad szoftver[1].

3.1. A klaszterelemzés eredményének vizsgálata: a megfelelő klaszterszám kiválasztásának lehetséges megoldása

Az irodalomfeldolgozás során érintett területek alapján látható, hogy az eredmények helyességének ellenőrzése fontos része a vizsgálatnak. Szeretnénk olyan eljárásokkal dolgozni, amelyek valamilyen (objektív) számszerűsíthető eredmények bevonásával a kutató segítséget kapna a döntés meghozatalában (itt pl. az optimális klaszterszám meghatározása esetén).

Dolgozatom középpontjában az a probléma áll, hogy ha az elemzőnek kell megadnia a keresett klaszterek számát (az algoritmus inputjaként), akkor a különböző klaszterszám-beállítások esetén kapott eredmények közül milyen módon választhatja ki a „legjobbat". A probléma akkor is fennáll, ha nem előre kell megadni a klaszterek számát, de utólag a sok lehetséges megoldás közül kell egyet kiválasztani. Ezt a problémát az angol nyelvű irodalomban „cluster validation" néven találhatjuk meg, amely alatt olyan kvantitatív elemzést értenek, mely a klaszteranalízis eredményeként létrejött csoportokat vizsgálja [Theodoridis és Koutroumbas, 2003, 591. oldal]. Ennek megoldására sok eljárás született, melyeket Theodoridis és Koutroumbas [2003] három

[1] GNU General Public License

DOI: 10.14751/SZIE.2014.035

típusba sorol: külső kritérium alapú, belső kritérium alapú valamint relatív kritérium alapú. Egy kicsit más csoportosítást alkalmaz FÜSTÖS, KOVÁCS, MESZÉNA és SIMONNÉ [2004, 205. old.]: „A klaszterek érvényessége (validitása) négy kritérium alapján vizsgálható. Külső követelményként értelmezhető az, ha ismert csoportokba tartozó egyedekből veszünk mintát, és arra végezzük el a klaszterezést. Belső követelménynek tekinthetők azok a mutatók, amelyekkel az eredeti és a származtatott távolságok illeszkedését mérjük. Harmadik megközelítést jelent a megismételhetőség kritériuma, amelynek lényege a kettéosztott megfigyelések klaszterezése és a felosztások összevetése. A klaszterek érvényességének relatív kritériuma az adatmátrix több eljárás szerinti klaszterezését és a felosztások közötti egyezés mérését fogalmazza meg."

LIU ET AL. [2010] munkájában a klaszterszám meghatározása céljából végrehajtott vizsgálatának célja az volt, hogy megfigyeljék, hogy a vizsgált indexek pontosságára (11 ilyen indexet teszteltek) – amelyek külső információt nem tartalmaztak – milyen hatással van az adatok szerkezete (zajos adatok, sűrűség különbségek, alcsoportok, aszimmetrikus eloszlás). Ezek közül az alcsoportok felismerése okozta a legtöbb problémát az ellenőrzés során, ezen esetben a legtöbb index nem adott helyes eredményt. Egy olyan index – az ún. S_Dbw index – volt a 11 között, mely mindegyik esetben helyes döntést hozott. Az eljárást HALKIDI és VAZIRGIANNIS [2001] dolgozta ki, mely a klaszterek közötti sűrűségkülönbségen alapszik. Ezt fejlesztette tovább KIM és LEE [2003] valamint TONG és TAN [2009] abba az irányba, hogy robusztusabb[2] legyen, valamint ne csak gömbszimmetrikus klasztereket ismerjen fel. Ennek fontosságára korábban felhívta a figyelmet LEGÁNY, JUHÁSZ és BABOS [2006] is, akik megfigyelték, hogy az általuk vizsgált indexek (pl. az S_Dbw is) csak jól szeparált, gömbszimmetrikus klaszterek esetén nyújtottak megfelelő segítséget a klaszterek validálásához. Dolgozatomban ezen módszerek vizsgálatával és továbbfejlesztésével foglalkozom.

3.1.1. Az S_Dbw_{new} index

Vizsgálatom kiindulópontja a HALKIDI és VAZIRGIANNIS [2001] által kidolgozott, majd KIM és LEE [2003] valamint TONG és TAN [2009] által továbbfejlesztett módszer – alapja az *S_Dbw* (Scatter and Density between clusters) index – mely a sűrűségkülönbségek és a szórások alapján rendel hozzá egy adott csoportosításhoz egy valós számot. A különböző csoportosításokhoz tartozó értékek alapján lehet a legjobban illeszkedő megoldást kiválasztani.
Itt most csak a legutolsó változattal foglalkozom, mert ez jobb eredményeket

[2] A kiugró adatokra kevésbé érzékenyen határozza meg a klaszterek számát.

DOI: 10.14751/SZIE.2014.035

ért el a tesztelések során, mint az első két változat.

A módszer alapja, hogy a klaszterek közötti hasonlóságot ill. a klaszterek közötti különbséget bizonyos pontok körül kialakított tartományokon belül található megfigyelési egységek számának (mint sűrűségnek) összehasonlítása alapján határozták meg [TONG és TAN, 2009]. Ők az indexet S_Dbw_{new}-nak nevezték (megkülönböztetésül az előzményektől), és ezt a jelölést itt is megtartom.

Legyen adott egy adatbázis, amely N számú egyed, mint megfigyelési egység adatát tartalmazza. Az egyedek tulajdonságait k db változóval írjuk le. Ezen adatok egy $N \times k$ méretű mátrixba rendezhetők. Ezen adatbázison futtassunk le egy klaszterező módszert, így kapjuk a megfigyelési egységeink egy csoportosítását (c db klasztert). Ezen csoportosításhoz fogunk hozzárendelni egy számot, amely az S_Dbw_{new} index egy lehetséges értéke. Az eljárás az említett cikk alapján röviden a következőképpen írható le.[3]

Az indexnek két összetevője van: $Dens_{bw}(c)$ – klaszteren belüli sűrűség, valamint $Scat(c)$ – klaszterek közötti variancia.

$$Dens_{bw}(c) = \frac{1}{c(c-1)} \sum_{i=1}^{c} \left[\sum_{\substack{j=1 \\ j \neq i}}^{c} \frac{\text{density}^*(\boldsymbol{m}_{ij})}{\max\{\text{density}^*(\boldsymbol{v}_i), \text{density}^*(\boldsymbol{v}_j)\}} \right] \tag{3.1}$$

ahol
c: a kialakított klaszterek száma,
$\boldsymbol{v}_i$: az i-edik klaszter középpontja.

$$\text{density}^*(\boldsymbol{m}) = \sum_{i=1}^{n_m} f^*(\boldsymbol{x}_i, \boldsymbol{m}) \tag{3.2}$$

$\boldsymbol{x}_i$: az i-edik megfigyelési egység,
$\boldsymbol{m}$: egy tetszőleges megfigyelési egység,
n_m: a figyelembe vett megfigyelési egységek száma.

$$f^*(\boldsymbol{x}_i, \boldsymbol{m}) = \begin{cases} 1 & , \text{ ha } CI_-^{(p)} \leq d(x_i^{(p)}, m^{(p)}) \leq CI_+^{(p)} \ , \ \forall p \in \{1, 2, 3, \ldots, k\} \\ 0 & , \text{ egyébként} \end{cases} \tag{3.3}$$

k: a megfigyelési változók száma,
$x_i^{(p)}$: az i-edik megfigyelési egység p-edik változójának értéke,
$m^{(p)}$: egy tetszőlegesen kiválasztott megfigyelési egység p-edik változójának értéke,

[3] Ahol az eredeti cikk jelölésrendszere nem volt egészen világos, ott ennek módosítására került sor.

DOI: 10.14751/SZIE.2014.035

továbbá

$$CI_{\pm}^{(p)} = v_l^{(p)} \pm \left(1{,}96 \cdot \frac{\sigma_l^{(p)}}{\sqrt{n_l}}\right) \tag{3.4}$$

$v_l^p(p)$, $\sigma_l^{(p)}$, n_l: a figyelembe vett klaszter elemei p-edik változójának átlaga ill. szórása, valamint a klaszter elemszáma.

Legyen továbbá $\boldsymbol{m}_{ij}$ az i-edik és j-edik klaszter középpontját összekötő szakasz olyan osztópontja, mely a két klasztert „elválasztja", és melynek p-edik komponense:

$$m_{ij}^p = 0.7 \cdot \left(\frac{n_j \cdot v_i^{(p)} + n_i \cdot v_j^{(p)}}{n_i + n_j}\right) + 0.3 \cdot \left(\frac{\text{density}^*(\boldsymbol{v}_i) \cdot v_i^{(p)} + \text{density}^*(\boldsymbol{v}_j) \cdot v_j^{(p)}}{\text{density}^*(\boldsymbol{v}_i) + \text{density}^*(\boldsymbol{v}_j)}\right) \tag{3.5}$$

n_i: az i-edik klaszter elemszáma.

Az algoritmus az $\boldsymbol{m}_{ij}$ számításakor figyelembe veszi a két klaszter elemszámait, valamint a két klaszter középpontja körüli sűrűséget, és a kettő kombinációja[4] adja az osztópontot.

E részindex (3.1. egyenlet) számításának elve tehát, hogy összehasonlítja a klaszterek középpontja körüli, valamint a klaszterközéppontok között kiválasztott pont ($\boldsymbol{m}_{ij}$) körül elhelyezkedő egyedek számát.

A második részindex számításának módja:

$$Scat(c) = \frac{1}{c-1} \sum_{i=1}^{c} \frac{n - n_i}{n} \cdot \frac{\|\boldsymbol{\sigma}^2(\boldsymbol{v}_i)\|}{\|\boldsymbol{\sigma}^2(\boldsymbol{S})\|} \tag{3.6}$$

ahol

$\boldsymbol{\sigma}^2(\boldsymbol{v}_i)$: a $\boldsymbol{v}_i$ középpontú klaszter variancia vektora[5],

$\boldsymbol{\sigma}^2(\boldsymbol{S})$: az adatbázis variancia vektora,

$\|.\|$: vektor euklideszi normája.

Ezekből a részindexekből a következő módon adódik az index:

$$S\_Dbw_{new}(c) = Dens_{bw}(c) + Scat(c) \tag{3.7}$$

Legyen $\boldsymbol{S}$ olyan adatbázis, mely konvex klasztereket tartalmaz. Futtassunk le ezen, különböző klaszterszám beállításával, egy klaszterező eljárást többször. Belátható, hogy az index akkor vesz fel minimális értéket, ha a klaszterező eljárás a tényleges klasztereket találta meg [Halkidi és Vazirgiannis, 2001].

Természetesen nem garantált, hogy a klaszterek képzése során a tényleges

[4] A súlyok (0.7 - 0.3) meghatározása empirikus vizsgálatok tapasztalatai alapján történt [Tong és Tan, 2009].

[5] A koordinátatengelyek irányába számolt varianciákból képzett vektor.

DOI: 10.14751/SZIE.2014.035

klaszterek (ha léteznek) valóban előállnak megoldásként. Ekkor is az index minimumát fogadjuk el megoldásként, mivel ez jelenti a legjobb szeparációt [Halkidi és Vazirgiannis, 2001].

3.1.2. Az S_Dbw_{new} index kritikája

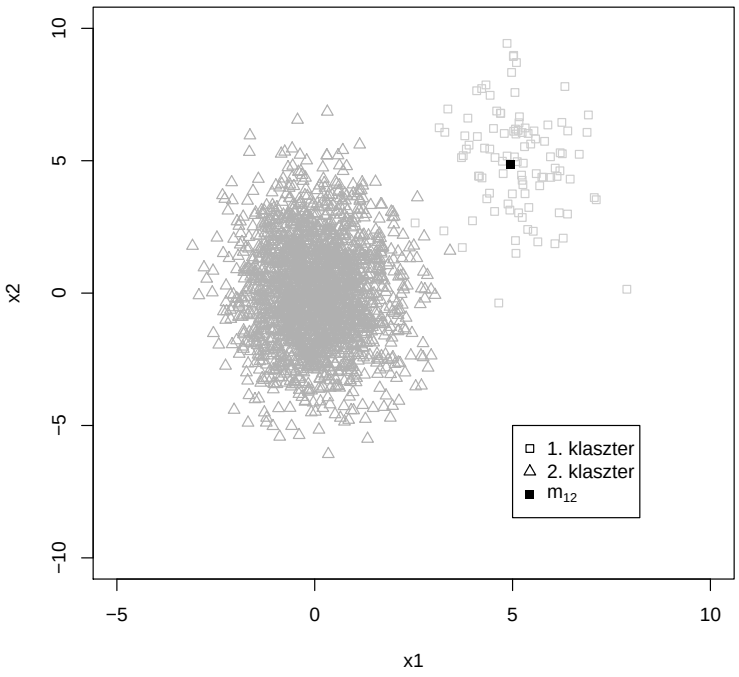
A 3.3. egyenlet megadja, hogy a „sűrűség" számításánál mely egyedeket kell figyelembe venni, és melyeket nem. Az adott pont környezetének definiálása határozza meg ezt a számot. Látható, hogy a CI hossza a klaszter számának (n) növekedésével csökken[6] (3.4. egyenlet). Ez pedig azt jelenti, hogy a lecsökkenő területen (még a nagy egyedszám mellett is) kevés egyed található, vagy egyáltalán nem is találunk egyedet. Ezzel pedig a mérés válik lehetetlenné, hiszen nem lesz alkalmas a sűrűbb és ritkább tartományok elkülönítésére.

Következő észrevételem, hogy az $\boldsymbol{m}_{ij}$ osztópont (3.5. egyenlet) számításánál két szempontot vettek figyelembe. Az első fele a két klaszter-középpontot összekötő szakaszt a klaszterek elemszámának arányában osztja, méghozzá úgy, hogy amelyik klaszternek nagyobb az elemszáma, attól távolabb lesz az osztópont. A második rész pedig a klaszter-középpontok körüli sűrűségek arányában osztja a szakaszt úgy, hogy amelyik klaszter esetében a sűrűség nagyobb volt, ahhoz kerül közelebb az osztópont. Ezen két hatás konvex lineáris kombinációjából állították elő $\boldsymbol{m}_{ij}$-t, méghozzá kísérletekből, tapasztalati úton állították be az együtthatókat (0,7 - 0,3).
Kísérleteim szerint az így kialakított osztópont a két klaszter eltérő elemszáma esetén jelentősen eltolódhat a kevesebb elemet tartalmazó klaszter közelébe. Szélsőséges esetben a nagy elemszámú klaszter „beletolja" az osztópontot a kevesebb elemet tartalmazó klaszterbe. Ez látható a 3. ábrán, amely esetében az elemszámok aránya 1:20. Az osztópont eltolódásának egyik oka, hogy a nagy elemszám miatt a CI értéke olyan kicsi lett, hogy abba nem került elem, így a sűrűség értéke 0, ami azt jelenti, hogy a második része a képletnek (3.5. egyenlet) nem kompenzálja az első rész hatását (hiszen ezek itt éppen egymás ellen hatnának). Sőt, mint ahogyan a 3. ábrán látható eset hátterében is megfigyelhető volt, ha a kisebb sűrűségű klaszter esetében az adott tartományba véletlenül belekerül egy pont, míg a nagyobb sűrűségű esetében nem, akkor az még jobban növeli a torzító hatást (ld. 3.5. egyenlet).

Ezen észrevételeket támasztják alá a következő gondolatmenetek.
A 3. ábra 1. klaszterének adatai: $\boldsymbol{v}_1 = (5\,,\,5)^T$, $\boldsymbol{\sigma} = (1\,,\,2)^T$, $n_1 = 100$, valamint 2. klaszterének adatai: $\boldsymbol{v}_2 = (0\,,\,0)^T$, $\boldsymbol{\sigma} = (1\,,\,2)^T$, $n_2 = 2000$.
Keresem annak a valószínűségét, hogy egy megfigyelési egység a klaszterkö-

[6] A CI 0-hoz tart, ha $n \to \infty$.

DOI: 10.14751/SZIE.2014.035



3. ábra. Klaszterek középpontja közötti TONG és TAN [2009] féle osztópont eltolódása 2 változó esetén. Forrás: saját szerkesztés

zéppont megfelelő (ld. 3.3. és 3.4. egyenlet) környezetébe esik. Jelentse ξ_{1x} ill. ξ_{1y} az első klaszterbe tartozó pont x ill. y koordinátáját (normál eloszlású valószínűségi változók). Továbbá ξ_{2x} ill. ξ_{2y} a második klaszterbe tartozó pont x ill. y koordinátáját.

Az 1. klaszter esetében (x és y irányban):

$$\mathrm{P}\left(5 - 1{,}96 \cdot \frac{1}{\sqrt{100}} < \xi_{1x} < 5 + 1{,}96 \cdot \frac{1}{\sqrt{100}}\right) = 2\Phi\left(\frac{1{,}96}{\sqrt{100}}\right) - 1 = 0{,}155$$

és

$$\mathrm{P}\left(5 - 1{,}96 \cdot \frac{2}{\sqrt{100}} < \xi_{1y} < 5 + 1{,}96 \cdot \frac{2}{\sqrt{100}}\right) = 2\Phi\left(\frac{1{,}96}{\sqrt{100}}\right) - 1 = 0{,}155$$

A keresett valószínűség tehát $p_1 = 0{,}155^2 = 0{,}0241$.

Legyen η_1 egy diszkrét valószínűségi változó, és jelentse a középpont megadott környezetében található egyedek számát. Ennek várható értéke (binomiális eloszlás esetén): $\mathrm{M}(\eta_1) = n_1 \cdot p_1 = 100 \cdot 0{,}0241 = 2{,}41$

A 2. klaszter esetében (x és y irányban):

$$\mathrm{P}\left(0 - 1{,}96 \cdot \frac{1}{\sqrt{2000}} < \xi_{1x} < 0 + 1{,}96 \cdot \frac{1}{\sqrt{2000}}\right) = 2\Phi\left(\frac{1{,}96}{\sqrt{2000}}\right) - 1 = 0{,}034$$

és

$$\mathrm{P}\left(0 - 1{,}96 \cdot \frac{2}{\sqrt{2000}} < \xi_{1y} < 0 + 1{,}96 \cdot \frac{2}{\sqrt{2000}}\right) = 2\Phi\left(\frac{1{,}96}{\sqrt{2000}}\right) - 1 = 0{,}034$$

A keresett valószínűség tehát $p_2 = 0{,}034^2 = 0{,}0012$.

DOI: 10.14751/SZIE.2014.035

Legyen η_2 egy diszkrét valószínűségi változó, és jelentse a középpont megadott környezetében található egyedek számát. Ennek várható értéke (binomiális eloszlás esetén): $\mathrm{M}(\eta_2) = n_2 \cdot p_2 = 2000 \cdot 0{,}0012 = 2{,}44$
Vagyis a 2. klaszter 20-szor annyi elemet tartalmaz, mégis, a középpont megadott környezetében található elemek száma közelítőleg annyi, mint az 1. klaszter esetében, átlagosan 2.44. Annak valószínűsége pedig, hogy egy elem sem esik a megadott környezetbe: $\mathrm{P}(\eta_2 = 0) = (1 - 0{,}0012)^{2000} = 0{,}0906$. A 3. ábrához tartozó eset bekövetkezési valószínűsége pedig (a nagy klaszterben 0, a kicsiben nem 0 a középpont meghatározott környezetében található elemek száma) $0{,}0906 \cdot \left(1 - (1 - 0{,}0241)^{100}\right) = 0{,}088$, azaz 8,8%, ami nem elhanyagolható, tehát bekövetkezésével számolni kell.

Érdekes következmény továbbá, hogy a fent számított két valószínűség (p_1 és p_2) közel azonos. Megvizsgáltam tehát a $\mathrm{M}(\eta)$ értékét az n (klaszterelemek száma) függvényében. A grafikon (4. ábra) egy monoton növekvő függvény képét mutatta. Kiszámítottam a kapott függvény határértékét a végtelenben ((3.8). egyenlet). Vagyis növelve a klaszterek elemszámát, a középpont adott környezetében található elemek számának várható értéke lényegében konstansnak tekinthető. Ennek oka a korábban már említett terület csökkenése, mely terület a klaszter elemszámával fordítottan arányos.

$$\lim_{k\to\infty}\left(2\Phi\left(\frac{1{,}96}{\sqrt{k}}\right)-1\right)^2\cdot k=\lim_{k\to\infty}\frac{\left(2\Phi\left(\frac{1{,}96}{\sqrt{k}}\right)-1\right)^2}{k^{-1}}=$$

$$=\lim_{k\to\infty}\frac{2\left(2\Phi\left(\frac{1{,}96}{\sqrt{k}}\right)-1\right)\cdot 2\phi\left(\frac{1{,}96}{\sqrt{k}}\right)\cdot 1{,}96\cdot\left(-\frac{1}{2}k^{-\frac{3}{2}}\right)}{-k^{-2}}=$$

$$=\lim_{k\to\infty}\frac{4\phi\left(\frac{1{,}96}{\sqrt{k}}\right)\cdot 1{,}96\cdot\left(-\frac{1}{2}k^{-\frac{3}{2}}\right)\cdot 1{,}96\cdot\phi\left(\frac{1{,}96}{\sqrt{k}}\right)}{-\frac{1}{2}k^{-\frac{3}{2}}}+$$

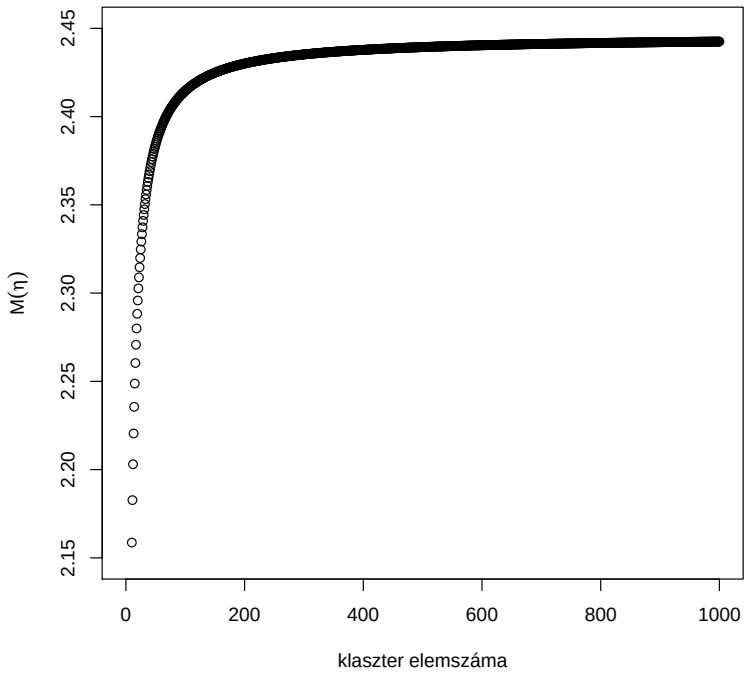
$$+\lim_{k\to\infty}\frac{2\cdot\left(2\Phi\left(\frac{1{,}96}{\sqrt{k}}\right)-1\right)\cdot\phi'\left(\frac{1{,}96}{\sqrt{k}}\right)\cdot 1{,}96^2\cdot\left(-\frac{1}{2}k^{-\frac{3}{2}}\right)}{-\frac{1}{2}k^{-\frac{3}{2}}}=$$

$$=\lim_{k\to\infty}\left[4\phi\left(\frac{1{,}96}{\sqrt{k}}\right)\cdot 1{,}96^2\cdot\phi\left(\frac{1{,}96}{\sqrt{k}}\right)+2\cdot\left(2\Phi\left(\frac{1{,}96}{\sqrt{k}}\right)-1\right)\cdot\right.$$

$$\left.\cdot\phi'\left(\frac{1{,}96}{\sqrt{k}}\right)\cdot 1{,}96^2\right]=4\cdot 1{,}96^2\cdot\left(\frac{1}{\sqrt{2\pi}}\right)^2+0\cdot 0=\frac{2\cdot 1{,}96^2}{\pi}\approx 2{,}4456 \quad (3.8)$$

Végül meghatároztam az η_2 valószínűségi változó eloszlását, és annak egy részletét tartalmazza a 4. táblázat (a várható érték környezete). Ebből is látszik, hogy az 1-3 objektum előfordulásának legnagyobb a valószínűsége, a ma-

DOI: 10.14751/SZIE.2014.035



4. ábra. M(η) a klaszter elemszámának függvényében (a Tong-féle környezet esetén).
Forrás: saját szerkesztés

ximuma 2-nél van. Ezáltal a fejezet elején tett megállapításaimat igazoltam.

3.1.3. Az indexek teszteléséhez használt adatbázisok és az összehasonlítások módszere

Ahhoz, hogy az indexek eredményei összehasonlíthatók legyenek, olyan adatbázisokra van szükség, amelyek esetében ismertek a klaszterek elemei (tehát léteznek csoportok, és minden megfigyelési egység hovatartozása ismert). Ezeket az adatbázisokat véletlenszerű mintavétellel állítottam elő normál eloszlású valószínűségi változók segítségével. Mivel a dolgozatomban kétváltozós eset-

4. táblázat. Az η_2 valószínűségi változó eloszlásának részlete. Forrás: saját számítás.

y_i	$\mathrm{P}(\eta_2 = y_i)$
0	0,0905
1	0,2177
2	0,2613
3	0,2092
4	0,1255
5	0,0602
⋮	⋮

y_i az η_2 valószínűségi változó által felvehető értékeket jelenti.
Mivel binomiális eloszlású valószínűségi változóról van szó, ezért $y_i \in \{0, 1, 2, \ldots, 2000\}$

DOI: 10.14751/SZIE.2014.035

tel foglalkozom, ezért minden megfigyelési egység esetében képezni kellett két értéket: az első és a második változó értékét. Mindkét érték normál eloszlású valószínűségi változó egy-egy lehetséges értéke (véletlen mintavétellel). A különböző klaszterek létrehozását pedig az eloszlás paramétereinek (várható érték, szórás) változtatásával lehetett elérni.

8 db adatbázison tesztelem az indexeket. Ezen adatbázisok előállításának szempontjai a következők voltak:

- legyen kisebb és nagyobb elemszámú klasztereket is tartalmazó adatbázis,
- legyen sűrűbb és ritkább klasztereket is tartalmazó adatbázis,
- legyen jól szeparált, és kevésbé jól szeparált klasztereket is tartalmazó adatbázis.

Az adatbázisok azért két változót tartalmaztak, hogy az eredmények kiértékelésekor legyen lehetőség annak szemléltetésére is, így lehetőséget teremtve annak jobb megértésére. Természetesen a továbbiakban semmi akadálya annak, hogy többváltozós adatbázisok esetén is teszteljük/alkalmazzuk az indexet, azonban ekkor a szemléltetés nehezebb, vagy nem megoldható.

Az 5. táblázat mutatja a létrehozott adatbázisok paramétereit (klaszterek középpontja, szórása, elemszáma). Az első négy esetben 4 klasztert állítottam elő, és az első esetben olyan távol helyeztem el őket, hogy teljesen szeparáltak legyenek. A további 3 esetben közelebb helyeztem őket ill. változtattam az elemszámukat (egyrészt úgy, hogy egyszerre mindegyik kevesebb elemet tartalmazzon, másrészt úgy, hogy különböző legyen az elemszámuk). Az 5. adatbázis egy háromklaszteres elrendezés, melyben K1 és K2 között átfedés van, míg K3 egy távolabb levő klaszter, sűrűségük pedig különböző. A 6. adatbázis tartalmaz ellipszis alakú klasztereket is, ezenfelül a K1 kivételével a többi átfedéseket is tartalmaz. A 7. adatbázis a 6.-ból keletkezett úgy, hogy az első klaszter szórása (x és y irányban is) nagyobb lett, így ez a klaszter is mutat átfedést a többivel. A 8. adatbázis esetében a K3 elkülönül a többitől, a többi három pedig jobban átfedi egymást, mint az eddigi példákban generált klaszterek esetében. Az előállított adatbázisok szemléltetése az A.3 – A.10 mellékletekben megtalálható.

Ezeken az adatbázisokon klaszterező eljárásokat futtatok le különböző paraméterbeállítások mellett, és a kapott klasztereken tesztelem a két indexet. Ezt az eljárást követték mindhárom cikkben, amelyek ennek az indexnek kidolgozásával foglalkoztak. Halkidi és Vazirgiannis [2001] valamint Tong és Tan [2009] elemzésében, többek között, az ún. DBSCAN [Ester, Kriegel, Sander és Xu., 1996] algoritmust alkalmazták. Ez a módszer a sűrűségek vizsgálatán alapszik, és nagyon hatékony nem konvex, de jól szeparált

DOI: 10.14751/SZIE.2014.035

5. táblázat. Az indexek összahasonlításához használt adatbázisok paraméterei. Forrás: saját összeállítás.

	K1			K2			K3			K4		
	$\boldsymbol{v}_1$	$\boldsymbol{\sigma}_1$	N_1	$\boldsymbol{v}_2$	$\boldsymbol{\sigma}_2$	N_2	$\boldsymbol{v}_3$	$\boldsymbol{\sigma}_3$	N_3	$\boldsymbol{v}_4$	$\boldsymbol{\sigma}_4$	N_4
1	(0,0)	(1,1)	500	(7,0)	(1,1)	500	(0,-7)	(1,1)	500	(2,7)	(1,1)	500
2	(0,0)	(1,1)	500	(4,0)	(1,1)	500	(0,-7)	(1,1)	500	(2,5)	(1,1)	500
3	(0,0)	(1,1)	100	(4,0)	(1,1)	100	(0,-7)	(1,1)	100	(2,5)	(1,1)	100
4	(0,0)	(1,1)	500	(4,0)	(1,1)	100	(0,-7)	(1,1)	500	(2,5)	(1,1)	250
5	(2,2)	(1,1)	750	(6,0)	(2,2)	500	(2,-7)	(0.5,0.5)	500			
6	(-4,0)	(1,1)	500	(4,0)	(2,2)	1000	(0,-7)	(3,2)	500	(2,5)	(2,1)	500
7	(-4,0)	(2,2)	500	(4,0)	(2,2)	1000	(0,-7)	(3,2)	500	(2,5)	(2,1)	500
8	(0,0)	(1,1)	500	(4,0)	(1,1)	500	(0,-7)	(1,1)	500	(2,2)	(1,1)	500

K1, K2, K3, K4: Klaszterazonosító
$\boldsymbol{v}_i$: az i-edik klaszter középpontja
$\boldsymbol{\sigma}_i$: az i-edik klaszter elemeinek x és y irányú szórása
N_i: az i-edik klaszter elemszáma

klaszterek elkülönítésére. Ezen vizsgálat fókuszában azonban a konvex és nem feltétlenül teljesen elkülönülő csoportok felismerése áll, ezért ezt az algoritmust a szimulációkban nem használtam. Mindhárom cikkben alkalmazták a K-means klaszterezési eljárást. Ezt az eljárást a marketing kutatásokban is gyakran alkalmazzák, így ennek ismertetésére dolgozatomban nem térek ki, megjegyzem ugyanakkor, hogy az alkalmazott szoftver az ún. Hartigan-Wong algoritmust alkalmazta [Hartigan és Wong, 1979].
A másik alkalmazott módszer a hierarchikus klaszterező eljárások közé tartozó Ward módszer [Ward, 1963], mely szintén gyakran alkalmazott módszer a marketingkutatás területén. Ez a módszer leginkább kompakt és gömbszimmetrikus klaszterek azonosítására alkalmas. Kérdéses, hogy az adatbázisok között található nem ilyen tulajdonságú klaszterek felismerésére mennyire lesz alkalmas.

Természetesen a szimulációval nem lehet minden lehetséges helyzetet ellenőrizni. Itt a cél annak vizsgálata volt, hogy az egymáshoz közelebb levő klaszterek esetében kimutatható különbség van-e a két index eredményei között. Ennek bemutatására került definiálásra a 8-féle adatbázis.

Az összehasonlításhoz azonban minden egyes adatbázist 10-szer állítottam elő az adott paraméterbeállítások (ld. 5. táblázat) mellett, és ezek mindegyikén teszteltem az indexeket. Ezeket az eredményeket értékeltem ki találati pontosság tekintetében: mely index esetében lesz a találatok száma több az egyes klaszterelhelyezkedések esetében, illetve általánosan jobbnak tekinthető-e valamelyik index.

DOI: 10.14751/SZIE.2014.035

3.2. A fogyasztói magatartás előrejelzése: a BG/NBD modell módosítása

3.2.1. A BG/NBD modell bővítése (1)

Az irodalomfeldolgozásban bemutatott modell kibővítését készítette el VAN OEST és KNOX [2011], melynek tömör bemutatására kerül sor ebben az alfejezetben. Azért került a dolgozat ezen részébe, mert az általam elkészített módosításnak ez adta az alapját, tehát a modellfejlesztés „anyagának" tekinthető.

A BG/NBD modell csak a tranzakciók számát, és az utolsó tranzakció időpontját használja fel jövőbeli értékek előrejelzésére. Itt azonban felmerül a kérdés, ha a CRM rendszereken keresztül az egyes vásárlókról sokkal több adat áll rendelkezésre, miért ne használjuk fel azokat is az előrejelzésben. Így született az ún. egyszerű modell most bemutatásra kerülő kibővítése.

A felállított modell a tranzakciós adatokon kívül inputként tartalmazza a vásárlással kapcsolatban felmerülő panasz „történetét" is. Feltételezték, hogy ezek olyan információkat tartalmaznak, melyek figyelembevételével a modell pontosabb eredményre vezet az előrejelzésben.

A modell a következő feltételezéseken alapszik:

1. Amíg a vásárló aktív, addig a vásárlások száma Poisson eloszlást követ, melynek paramétere λ_p, amely egy bizonyos időtartam alatt bekövetkező vásárlások számának várható értéke.
2. λ_p változékonysága gamma eloszlást követ r és α paraméterekkel.[7]
3. Panaszmentes vásárlás esetén a vásárló q_p valószínűséggel válik inaktívvá.
4. q_p változékonysága béta eloszlást követ u_p és v_p paraméterekkel:

$$f(q_p|u_p, v_p) = \frac{q_p^{u_p-1}(1-q_p)^{vp-1}}{\mathrm{B}(u_p, v_p)} \tag{3.9}$$

5. q_p és λ_p vásárlónként egymástól függetlenül változnak.
6. A vásárlás napján bekövetkező panasz μ valószínűséggel következik be.
7. μ változékonysága béta eloszlást követ a és b paraméterekkel.
8. Amíg a vásárló aktív, a nem aznapi (nem a vásárlás napján történő) panaszok száma Poisson eloszlást követ λ_c paraméterrel.
9. λ_c változékonysága gamma eloszlást követ s és β paraméterekkel. λ_c a panaszok számának várható értéke.

[7] Ld. előző (BG/NBD) modell.

DOI: 10.14751/SZIE.2014.035

10. Egy panasz után (aznapi vagy nem aznapi) után a vásárló q_c valószínűséggel inaktívvá válik.

11. q_c változékonysága béta eloszlást követ u_c és v_c paraméterekkel.

12. q_c, λ_c és μ vásárlónként egymástól függetlenül változnak.

13. A vásárlásokkal kapcsolatos paraméterek és a panaszokkal kapcsolatos paraméterek egymástól függetlenül változnak.

Ennek a modellnek a leírásához a következő adatokra volt szükségük:

T a megfigyelési időtartam,
x_p a vásárlások száma,
$x_{c|p}$ az aznapi panaszok száma,
x_c a késleltetett panaszok száma,
t_x az utolsó vásárlás időpontja,
z_c az utolsó vásárlás által generált panaszok száma ($z_c \in \{0, 1\}$).

Először itt is a Likelihood függvényt határozták meg, mely a következő alakot kapta az átalakítások után:

$$\begin{aligned} &L(r, \alpha, u_p, v_p, a, b, s, \beta, u_c, v_c | x_p, x_{c|p}, x_c, t_x, T)) = \\ &= \frac{\Gamma(r + x_p)\alpha^r}{\Gamma(r)(\alpha + t_x)^{r+x_p}} \frac{\Gamma(s + x_c)\beta^s}{\Gamma(s)(\beta + t_x)^{s+x_c}} \frac{\mathrm{B}(a + x_{c|p}, b + x_p - x_{c|p} + 1)}{\mathrm{B}(a, b)} \\ &\cdot \frac{\mathrm{B}(u_p + 1 - z_c, v_p + x_p - x_{c|p} + z_c)}{\mathrm{B}(u_p, v_p)} \frac{\mathrm{B}(u_c + z_c, v_c + x_{c|p} + x_c - z_c)}{\mathrm{B}(u_c, v_c)} \\ &+ \frac{\Gamma(r + x_p)\alpha^r}{\Gamma(r)(\alpha + T)^{r+x_p}} \frac{\Gamma(s + x_c)\beta^s}{\Gamma(s)(\beta + T)^{s+x_c}} \frac{\mathrm{B}(a + x_{c|p}, b + x_p - x_{c|p} + 1)}{\mathrm{B}(a, b)} \\ &\cdot \frac{\mathrm{B}(u_p, v_p + x_p - x_{c|p} + 1)}{\mathrm{B}(u_p, v_p)} \frac{\mathrm{B}(u_c, v_c + x_{c|p} + x_c)}{\mathrm{B}(u_c, v_c)} \end{aligned} \tag{3.10}$$

Jelölje Φ a paraméterek halmazát, *History* pedig az input adatok halmazát. A fenti függvény ekkor így írható föl:

$$L(\Phi | History) = L_{\text{inaktív}}(\Phi | History) + L_{\text{aktív}}(\Phi | History) \tag{3.11}$$

ahol az inaktív ill. aktív megkülönböztetés arra utal, hogy a vásárló aktív marad-e az utolsó vásárlás (t_x) után is a megfigyelési időszakban. Az inaktív esetében az utolsó vásárlás után inaktívvá válik, míg a másik esetben aktív marad, csak a megfigyelési időszakban már nem kezdeményez vásárlást.

A cél szintén az volt, hogy előrejelzést adjanak a későbbi vásárlások számára, vagyis $Y(t)$ várható értékét keresték (ld. 2.28. egyenlet). Az alábbi formulát

DOI: 10.14751/SZIE.2014.035

kapták:

$$\begin{aligned}&\mathrm{E}\left(Y(t)|History,\Phi\right)=\\&=\frac{\frac{1}{M}\sum_{m=1}^{M}\frac{\lambda_p^{(m)}\left(1-\mathrm{e}^{-(\lambda_p^{(m)}(1-\mu^{(m)})q_p^{(m)}+(\lambda_p^{(m)}\mu^{(m)}+\lambda_c^{(m)})q_c^{(m)}+\delta)t}\right)}{\lambda_p^{(m)}(1-\mu^{(m)})q_p^{(m)}+(\lambda_p^{(m)}\mu^{(m)}+\lambda_c^{(m)})q_c^{(m)}+\delta}}{1+\left(\frac{\alpha+T}{\alpha+t_x}\right)^{r+x_p}\left(\frac{\beta+T}{\beta+t_x}\right)^{s+x_c}\left(\frac{u_p}{v_p+x_p-x_{c|p}}\right)^{1-z_c}\left(\frac{u_c}{v_c+x_{c|p}+x_c-1}\right)^{z_c}}\end{aligned}\tag{3.12}$$

ahol
$\lambda_p^{(m)} \sim \Gamma(r+x_p, \alpha+T)$
$q_p^{(m)} \sim \mathrm{B}(u_p, v_p+x_p-x_{c|p}+1)$
$\mu^{(m)} \sim \mathrm{B}(a+x_{c|p}, b+x_p-x_{c|p}+1)$
$\lambda_c^{(m)} \sim \Gamma(s+x_c, \beta+T)$
$q_c^{(m)} \sim \mathrm{B}(u_c, v_c+x_{c|p}+x_c$

Mint látható, a 3.12. egyenlet számlálójában egy átlag található, melynek segítségével becsüli meg a tényleges értéket (Monte Carlo Method). Az ehhez szükséges adatokat (M db minden egyes paraméter esetén) a fent említett eloszlások segítségével véletlenszerűen generálja. Összehasonlítva a 2.28. egyenletben kapott összefüggéssel, szembetűnő, hogy ott már nem találjuk az egyéni szintű paramétereket (λ, p), helyette az α, r, a, b paraméterek (az egész mintára jellemző paraméterek) függvényeként állt elő $Y(t)$ várható értéke. Igazából itt sincs ez másként, csak először az egyéni szintű paraméterek előállítása történik meg, melyek átlagaként kapott (pl. $\lambda_p^{(m)}$) értékek függvényeként áll elő $\mathrm{E}(Y(t))$.

VAN OEST és KNOX [2011] vizsgálatai szerint az általuk létrehozott modell jobb előrejelzéseket ad, mint az, melyből született, azonban ők is jelzik a továbbgondolási lehetőségeket.

Ez a modell valóban többletinformációkat is felhasznál az előzőhöz képest, de nem látszik világosan a kétféle panasz (aznapi ill. késleltetett) közötti különbség. A megvásárolt áru esetében általában hosszabb idő áll a vásárló rendelkezésére, hogy panaszát érvényesíthesse. Továbbá a panasz időpontja függhet a vásárló lakásának az üzlettől mért távolságától is. Így, az eredmények ellenére, nem meggyőző a modell. Ennek egy lehetséges módosítását készítettem el az Eredmények fejezet 2. részében (4.2.1. alfejezet).

3.2.2. A modell teszteléséhez használt adatbázisok

A már meglévő és a megalkotott modellt mesterségesen előállított adatbázisokon teszteltem. A tesztelés lényege, hogy sok adatbázison mérjem az egyes modellek eredményét. Az adatbázisokat a modellek alapjául szolgáló, a tapasz-

DOI: 10.14751/SZIE.2014.035

talati tényekkel leginkább összhangot mutató eloszlások alapján állítottam elő, úgy, hogy az eloszlások bizonyos paramétereit változtattam. Vizsgálataimban 3 ilyen paraméter értékét, valamint az előrejelzési időszak (t) hosszát módosítottam. Mindegyik 3 különböző értéket vehetett fel, így összesen $3^4 = 81$ adatbázison teszteltem a modelleket. Ezeken belül minden adatbázis 1000 vásárló adatait tartalmazza, melyeket a különböző vásárlói tulajdonságok (mint paraméterek) változtatásával generáltam. Az adatbázisok létrehozásakor az alapvető eloszlások az exponenciális és a binomiális eloszlások voltak. Az exponenciális eloszlással az egymás után következő vásárlások között eltelt időt adtam meg, míg a binomiális eloszlás segítségével a lemorzsolódást modelleztem minden vásárlás után (ezen eloszlások paraméterei személyenként változnak, ahogy az a modellel kapcsolatos feltételezések definiálása során már említésre került, ld. 53. old.). Természetesen ebben szerepe van még az általam a modellbe bevont egyéb hatásoknak, nevezetesen, hogy a vásárlás pozitívan elbírált panasszal ill. negatívan elbírált panasszal[8] történt-e. Ezek esetében ugyanis – feltételezésem szerint – különbözik a lemorzsolódás valószínűsége.

A kapott adatbázisok esetében rendelkezésünkre áll, hogy a vizsgálati idő (T) alatt hány vásárlás történt személyenként, mikor volt ebben az időszakban az utolsó vásárlás, mennyi panasz volt. Ezen adatok alapján a modellek meghatározzák, hogy milyen eloszlások (pontosabban azok milyen paraméterei) esetében jöttek ki ezek az eredmények (ld. maximum likelihood módszer), és ezen becsült paraméterek segítségével ad előrejelzést a modell a T időszakot követő t időszakra.

Az adatbázisok előállítására szolgáló kód az A.13. mellékletben az „adatmátrix kezd” és az „adatmátrix vég” sorok között található. A hozzájuk tartozó paraméterek értékeit ezen sorok előtt definiáltam.

3.2.3. A modelleredmények értékelésének módszerei

A modellek által kapott előrejelzések pontosságát vizsgálom az Eredmények fejezetben több szempont szerint. Ezekhez bizonyos mutatószámokat határozok meg, melyek azonosságát ill. különbözőségét mérem statisztikai módszerekkel.

Ezen mutatószámok egyike a Cohen féle kappa mutató, melyet két nominális (jelen esetben kétértékű) változó egyezőségének vizsgálatára fejlesztettek ki [Cohen, 1960]. Ennek értékét a következő képlettel számolhatjuk:

$$\kappa = \frac{p_0 - p_e}{1 - p_e} \tag{3.13}$$

ahol

[8] Részletesebben lásd 4.2.1. alfejezet, 69.old.

DOI: 10.14751/SZIE.2014.035

p_0 az egyezések aránya,
p_e az egyezések aránya függetlenséget feltételezve.
Az index AGRESTI [2010, 250. old.] szerint „nominális skálán a legnépszerűbb egyetértési mutató". Értéke 0 és 1 között lehet, minél nagyobb, annál szorosabb az egyezőség a két változó között. Ennek segítségével mértem a tényleges és az előrejelzett értékek közötti eltéréseket a vásárlói lemorzsolódás esetében.

Az egyes vásárlókra számolt mutatószámok az egyes modellek esetében különböznek egymástól, ezek összehasonlításához a következő módszereket használtam:

- Az eredményeket Boxplot ábrán szemléltettem, mely szemléletesen bemutatja a kapott értékeket, és egyszerűbb összehasonlításokra alkalmas.
- Az eredmények normalitásvizsgálatára a Shapiro-Wilk tesztet tartottam legalkalmasabbnak RAZALI és WAH [2011] eredményei alapján.
- Az egyes modellek esetében kapott eredmények szórásának összehasonlítását, hagyományosan, F-próbával végeztem.
- A modellátlagok összehasonlítására a párosított t-próbát, ha azonban a szükséges feltételek nem teljesültek, akkor a Wilcoxon párosított (nemparaméteres) próbát alkalmaztam. A két mintát azért kell párosított próbával összehasonlítani, hiszen az egy-egy vásárlóhoz tartozó értékek összehasonlítása a cél.

DOI: 10.14751/SZIE.2014.035

DOI: 10.14751/SZIE.2014.035

4. fejezet

Eredmények

4.1. A klaszterezés eredményének ellenőrzése

4.1.1. Az S_Dbw_{new} index módosítása

Az eredmények fejezet első részében a klaszterszámok optimális meghatározásának vizsgálatában elért eredményeimet mutatom be. A korábbi módszerekben felfedezett hibák (ld. 3.1.2. alfejezet) kijavításával egy új módszert mutatok be, melynek teszteredményei meggyőzőek a tekintetben, hogy a módosítás eredményes volt. Az anyag és módszer fejezetben megfogalmazott hibák miatt a tartomány[1] megválasztásának módosítását javaslom. Az eredeti javaslat - 3.3. egyenlet - helyett a következőképpen definiálom az f^* függvényt, amelyet megkülönböztetésül f^{**}-nak nevezek:

$$f^{**}(\boldsymbol{x}_i, \boldsymbol{m}) = \begin{cases} 1 \quad , & \text{ha } m^{(p)} - \alpha \cdot D^{(p)} \leq x_i^{(p)} \leq m^{(p)} + \alpha \cdot D^{(p)} \ , \\ & \forall p \in \{1, 2, 3, \ldots, k\} \\ 0 \quad , & \text{egyébként} \end{cases} \tag{4.1}$$

ahol
$\boldsymbol{m}$ egy tetszőleges egyed
$m^{(p)}$ a tetszőleges egyed p-edik változójának értéke,
$D^{(p)} = \min_i(\sigma_i^{(p)}) \quad , \quad i \in \{1, 2, \ldots, c\}$, a klaszterelemek p-edik változójának szórásai közül a minimális,
α: egy alkalmasan megválasztott konstans.

A módosítás lényege, hogy az az intervallum, amelyen belül a megfigyelési egységeket keresem, már független az n-től (a klaszterelemek számától), így egy adott intervallumba eső megfigyelési egységek száma (az adott térrészben) arányos lesz a klaszterek elemszámával. Másrészt, az $\boldsymbol{m}_{ij}$ osztópontok esetében, a korábban említett torzító hatás is megszűnik.

[1] A két klaszter középpontja ill. a klaszterközéppontokat elválasztó pont körül kijelölt tartomány, amelyben található elemek száma alapján választható szét a két klaszter.

DOI: 10.14751/SZIE.2014.035

Ezt a módosított függvényt használva a $Dens_{bw}$ részindex (3.1. egyenlet) helyett kapjuk a $Dens_{bw}^{**}$ részindexet, melyből a teljes index

$$S\_Dbw^{**}(c) = Dens_{bw}^{**}(c) + Scat(c) \tag{4.2}$$

a 3.7. egyenlet alapján adódik.

4.1.2. A klaszterek közötti mérőszám ($Dens_{bw}^{**}$ részindex) elemzése

A 3.1. egyenlet adja meg a klaszterek közötti sűrűségkülönbség alapján, hogy mely klasztereket tekintünk majd különbözőnek, és melyeket nem tudunk megkülönböztetni. A következő elemzésben két klaszter egymáshoz viszonyított helyének függvényében vizsgálom az index értékét, két változó bevonása mellett.

Legyen adott két klaszter (C_1 és C_2). Középpontjaik: $\boldsymbol{v}_1 = (0\,,\,0)^T$ és $\boldsymbol{v}_2 = (a\,,\,0)^T$. Mindkettő legyen kör alakú, azonos átmérővel (mindkét irányú szórásuk legyen 1-1). Legyen $\alpha = 0{,}5$ (4.1. egyenlet). Az elméleti megközelítés esetében nem konkrét elemekkel megadott klasztereket vizsgálok, hanem a két klasztert két-két normál eloszlású valószínűségi változóval jellemzem ($\xi_{1x}, \xi_{1y}, \xi_{2x}, \xi_{2y}$). Ilyen feltételek mellett vizsgálom az alábbi három valószínűséget:

$p_1 = \mathrm{P}\left((0 - 0{,}5 \cdot 1 < \xi_{1x} < 0 + 0.5 \cdot 1) \wedge (0 - 0{,}5 \cdot 1 < \xi_{1y} < 0 + 0{,}5 \cdot 1)\right)$,
mely arányos a C_1 klaszter középpontja körüli $\alpha \cdot D^{(1)}$, azaz $0{,}5 \cdot 1$ sugarú tartományba, valamint (y irányban) a C_1 klaszter középpontja körüli $\alpha \cdot D^{(2)}$, azaz $0.5 \cdot 1$ sugarú tartományba eső pontok számával (mely tartomány egy téglalap).

$p_2 = \mathrm{P}\left((a - 0{,}5 \cdot 1 < \xi_{2x} < a + 0.5 \cdot 1) \wedge (0 - 0{,}5 \cdot 1 < \xi_{2y} < 0 + 0.5 \cdot 1)\right)$,
mely ugyanaz, mint az előbb, csak a C_2 klaszterre vonatkoztatva.

$p_k = 2{\cdot}\mathrm{P}\left(\left(\frac{a}{2} - 0{,}5 \cdot 1 < \xi_{1x} < \frac{a}{2} + 0{,}5 \cdot 1\right) \wedge (0 - 0{,}5 \cdot 1 < \xi_{1y} < 0 + 0{,}5 \cdot 1)\right)$,
mely jelentése azonos az előzőekkel, csak a két középpontot összekötő szakasz felezőpontjára vonatkoztatva. A 2-vel való szorzás a két eloszlás azonossága miatt alkalmazható.

Ezen mennyiségek segítségével definiálom a következő indexet:

$$ind := \frac{p_k}{\max(p_1, p_2)} \tag{4.3}$$

mely index arányos a (3.1) egyenletben megadott $Dens_{bw}$ indexszel. Értékkészlete a $[0, 2]$, hiszen maximális értéket akkor vesz fel, ha a két klaszter középpontja egybeesik.

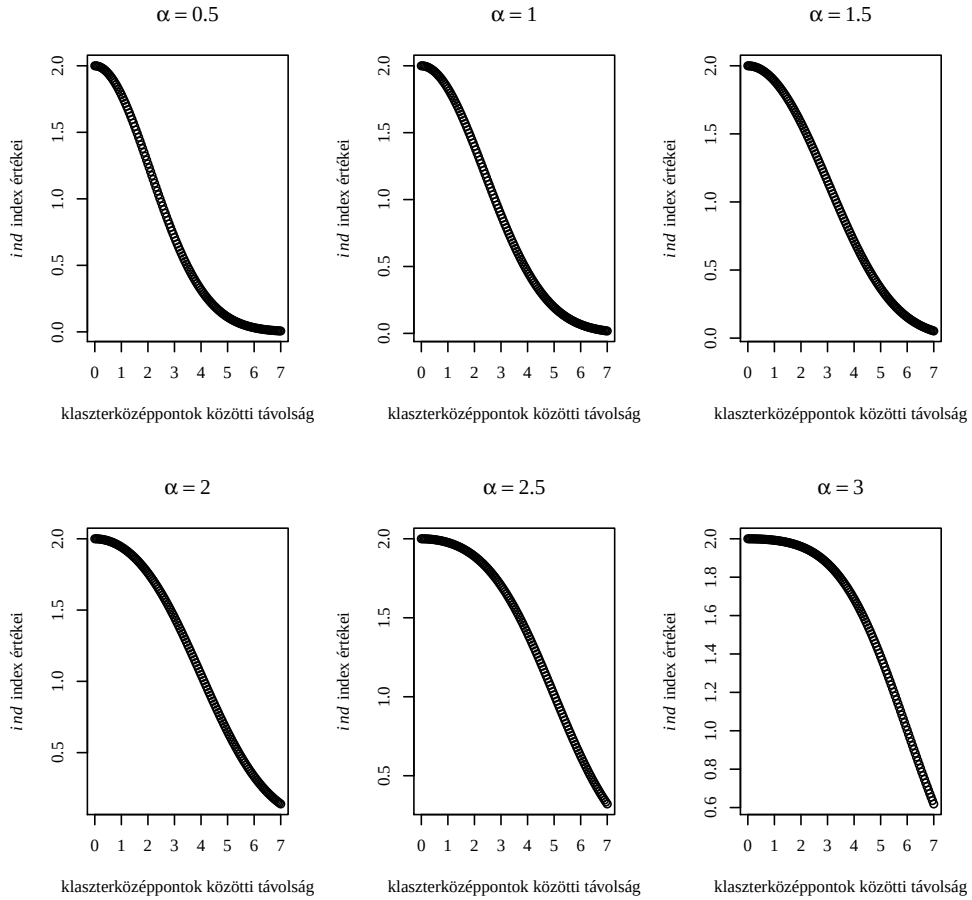
A definiált index vizsgálata során azt figyeltem, hogy miként változik az index értéke a középpontok távolságának függvényében. A távolság értékét 0-

DOI: 10.14751/SZIE.2014.035

tól 7-ig változtattam (a szórás értéke 1), azaz $a \in [0, 7]$. A kapott $ind(távolság)$ függvényt az 5. ábra első grafikonja ($\alpha = 0{,}5$) mutatja.

A függvény az 1-et, mint függvényértéket az $x = 2{,}4$ helyen veszi fel, ami azt jelenti, hogy a két középpont ezen távolsága esetén a két középpont adott környezetében (ld. α) ugyanannyi megfigyelési egység található, mint a két középpontot elválasztó osztópont (jelen esetben felezőpont) ugyanazon környezetében. A három pont tehát a sűrűség szempontjából egymástól nem megkülönböztethető. A távolság további növelésével az index értéke (csökkenő ütemben) tovább csökken.

A fenti kísérletben az α értékét 0,5-nek választottam. Hogy hogyan változnak a függvény értékei más α paraméterértékek esetén, az 5. ábra további grafikonjai mutatják. Látható, hogy az α értéket növelve az $ind = 1$ egyenlet megoldásai – vagyis azon klaszterközéppont távolságok, melyekre az ind index értéke 1 lesz – is egyre növekednek.

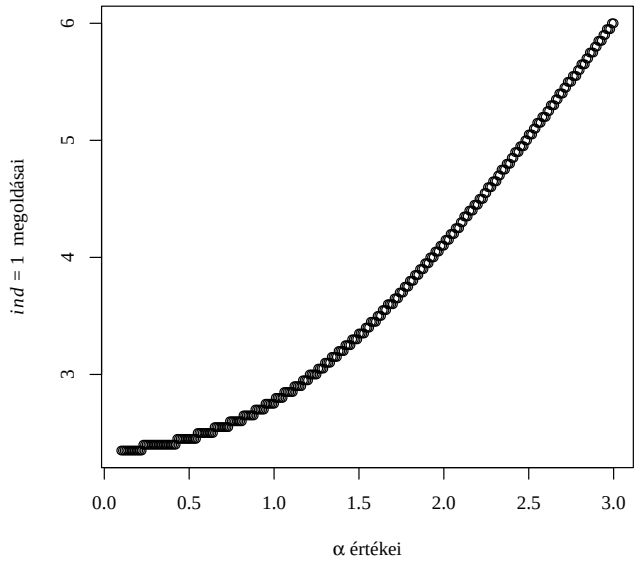


5. ábra. Az „*ind*” index a klaszterközéppontok közötti távolság függvényében különböző *alfa* paraméterek esetén. Forrás: saját szerkesztés.

Vagyis célszerű minél kisebb α értéket választani. Azonban a normál eloszlás esetében az elméleti valószínűségek számolhatók akkor is, ha α nagyon

DOI: 10.14751/SZIE.2014.035

kicsi, addig a konkrét adatbázis esetén ezek a kicsi intervallumok nem tartalmaznak majd megfigyelési egységeket, vagyis nem lesznek alkalmasak az összehasonlításra.
Ennek további vizsgálata céljából megrajzoltam egy függvényt, mely azon klaszterközéppontok távolságát adja meg az α függvényében, melyek esetében az *ind* index értéke 1 lett (6. ábra).



6. ábra. Az $ind = 1$ eredményt adó klasztertávolságok az *alfa* paraméter függvényében.
Forrás: saját szerkesztés.

A grafikon azt mutatja, hogy a változó (α) növekedésével egyre növekvő mértékben növekszik a függvényérték. Más szóval, minél nagyobb α értéket választunk, annál távolabbinak kell lennie a két klaszternek, hogy meg tudjuk különböztetni őket[2]. Mivel a függvény monoton növekvő, ezért az optimális értéket az értelmezési tartomány bal oldalán veszi fel. Itt bizonyos tartományonként konstans értéket vesz fel (ezen a részén a növekedés lassúbb), vagyis ezen a részen kell egy alkalmas α értéket kiválasztani. A továbbiakban legyen ez az érték $\alpha = 0{,}4$, figyelembe véve a korábbi érveket.

További kérdés azonban, hogy a fenti kísérlet eredménye 2 gömbszimmetrikus klaszter esetében lett kiértékelve. Van-e ennek hatása a 6. ábra grafikonjának jellegére? A kísérletet különböző szórás paraméterek mellett megismételve ugyanazt a jellegű görbét adta.

[2]Itt még nem került sor annak vizsgálatára, hogy az index milyen értéke mellet különböztethető meg két klaszter. Erre később kerül sor.

DOI: 10.14751/SZIE.2014.035

4.1.3. A módosított *S_Dbw*** index szerkezetének vizsgálata

A vizsgálat egyik célja, hogy a teljes index értékét a két részindex változásának függvényében figyelhessük meg.
Ennek modellezésére egy három klaszterből álló adatbázist készítettem, amelyben két klaszter helyét nem változtattam, a harmadikat pedig kiindulásként az egyik fix klaszterre helyeztem, majd távolítottam tőle az x tengely mentén (miközben a másik klaszterhez sem közelítettem). A két egymást átfedő klaszter egyszer egynek, majd két különböző klaszternek tekintettem, és vizsgáltam az indexek értékét mindkét változat esetében. A harmadik klaszterra azért volt szükség, hogy minden esetben legyen legalább két klaszter, amire az index számolható.

Először mindhárom (C_1, C_2, C_3) klaszter következő paramétereit azonosra állítom: $\sigma_{1x} = \sigma_{2x} = \sigma_{3x} = \sigma_{1y} = \sigma_{2y} = \sigma_{3y} = 1$, melyek az egyes klaszterek x és y irányú szórását jelentik. A $\boldsymbol{v_1} = (0,0)^T$, $\boldsymbol{v_2} = (d,0)^T$, ahol $d \in [0,7]$, továbbá $\boldsymbol{v_3} = (0,-7)^T$ pedig az egyes klaszterek középpontjait határozzák meg. Mindhárom klaszter 1000 megfigyelési egységet tartalmazott. Először a C_1 és a C_2 klasztert összevontam egy klaszterré, majd pedig külön klaszternek tekintettem őket, és mindkét esetben vizsgáltam az indexek értékét, miközben az d értékét 0-tól 7-ig változtattam bizonyos lépésközönként. Az eredmények a 6. táblázatban láthatók. Az egyes részindexeket, valamint a teljes indexet is párba állítottam a két klaszteres ill. a három klaszteres megoldások esetében. A két utolsó oszlop összehasonlításából látható, hogy az indexek nagyságában kb. 3,5-4 egység távolság ($3{,}5 < d < 4$) esetén váltás történik. Innentől kezdve tehát a három klasztert tartalmazó megoldást fogadjuk el a másikkal szemben, mivel az index minimális értéke esetén kapjuk a legjobb csoportosítást [Halkidi és Vazirgiannis, 2001]. Vagyis, ha a két klaszter szórása 1-1 egység, akkor középpontjuk kb. 4 egység távolságra kell, hogy legyen, hogy két különböző klaszterként értékelje őket az index. Vagyis nem szükséges teljesen átfedés mentesnek lenniük („jól szeparált"), bizonyos átfedés esetén is felismerhető a kettő különbözősége.

A 6. táblázat alapján vizsgálhatjuk a két részindexet is, melyek összegeként áll elő az előbb vizsgált index. A *Scat* részindex méri a klaszthereken belüli szórás értékét. Látható, hogy a két klaszteres számításnál növekszik az értéke, ha növeljük a C_1 és a C_2 klaszterek távolságát (ezt a két klasztert ugyanis egynek tekintjük ekkor). A három klaszteres változat esetében ez a részindex egyre csökken. Magyarázata: míg a három klaszter szórása külön-külön változatlan, addig az összes megfigyelési egység által alkotott „nagy" klaszter szórása növekszik. A (3.6) egyenlet értelmében a hányadosuk csökken.

Ugyancsak a 6. táblázat alapján vizsgálhatjuk a másik, a $Dens_{bw}^{**}$ részinde-

DOI: 10.14751/SZIE.2014.035

6. táblázat. A részindexek és a teljes index értékei a távolság függvényében 2 és 3 klaszter képzése esetén. Forrás: saját számítás.

Távolság d	$Dens\_bw^{**}$ $nc = 2$	$Dens\_bw^{**}$ $nc = 3$	$Scat$ $nc = 2$	$Scat$ $nc = 3$	S_Dbw^{**} $nc = 2$	S_Dbw^{**} $nc = 3$
0,0	0,0053	0,3281	0,0592	0,0776	0,0644	0,4057
0,5	0,0000	0,3076	0,0593	0,0790	0,0593	0,3866
1,0	0,0000	0,2266	0,0608	0,0770	0,0608	0,3036
1,5	0,0093	0,2336	0,0671	0,0792	0,0764	0,3128
2,0	0,0156	0,1911	0,0715	0,0782	0,0872	0,2693
2,5	0,0147	0,1774	0,0779	0,0792	0,0926	0,2566
3,0	0,0294	0,1188	0,0871	0,0776	0,1165	0,1964
3,5	0,0777	0,1004	0,0927	0,0744	**0,1704**	**0,1748**
4,0	0,0437	0,0408	0,1046	0,0723	**0,1483**	**0,1131**
4,5	0,0463	0,0383	0,1140	0,0725	0,1603	0,1108
5,0	0,0756	0,0146	0,1248	0,0693	0,2004	0,0838
5,5	0,1067	0,0099	0,1330	0,0660	0,2397	0,0759
6,0	0,0895	0,0045	0,1444	0,0618	0,2338	0,0662
6,5	0,0806	0,0036	0,1519	0,0600	0,2325	0,0637
7,0	0,1190	0,0056	0,1613	0,0569	0,2803	0,0625

nc : klaszterek száma

xet. A három klaszteres változat eredményeit (3. oszlop) figyelve megállapítható a csökkenő tendencia. Oka: a két távolodó klaszter között egyre kevesebb megfigyelési egység található, ezért a részindex számlálója (ld. (3.1) egyenlet) csökken, míg nevezője változatlan marad. A két klaszteres változat (2. oszlop) esetében, mivel C_1 és a C_2 klaszter alkot egy klasztert, a két klaszter távolodásakor a részindex nevezője csökken, vagyis a tört értéke növekszik.

A két részindex értéke 3 klaszter figyelembevételével csökken (tehát összegük is csökken), 2 klaszter esetében pedig növekszik (tehát összegük is növekszik). Ezen hatások eredményként egy bizonyos távolságban a két index (utolsó két oszlop) nagyságának viszonya megfordul. Innentől a három klaszteres megoldást választjuk a két klaszteres megoldás helyett.

A szimulációt többféleképpen is elvégeztem. Először a klaszterek minden számítás (d érték) esetén ugyanazok voltak, és csak az egyik klaszter (C_2) elemeinek első változóját növeltem a megadott d értékkel („A” változat). A második esetben minden egyes távolság esetén új klasztereket állítottam elő a megfelelő pataméterek alapján („B” változat). Mindkét esetben különböző szórás-beállítások mellett is elvégeztem a szimulációt (σ_{1x}-et és σ_{2x}-et változtattam, a többi értékét konstansnak vettem), amint a 7. táblázatban látható. A szórások növekedése miatt a klaszterközéppontok távolságának is nagyobb tartományt kellett megadni, ez 0–11 egységig terjedt. A két index értékei ismét a fent leírtak szerint változtak (a két klaszteres változat esetében növekedett, a háromklaszteres változat esetében csökkent az index értéke d növekedése esetén), természetesen a szórások értékének változása miatt más-más távolság

DOI: 10.14751/SZIE.2014.035

esetén következett be a váltás.

7. táblázat. A szimulációk száma a három klaszter felismeréséhez szükséges középpontok közötti távolság legkisebb értéke szerint, különböző szórású klaszterek esetén. Forrás: saját számítás.

Kísérlet típusa	σ_{1x}	σ_{2x}	Szimulációk száma az adott távolságeredményekkel															
			3,5	4	4,5	5	5,5	6	6,5	7	7,5	8	8,5	9	9,5	10	10,5	11
A	1	1	2	8														
A	1	2			4	6												
A	1	3				2	5	2	1									
A	2	2						1	2	6	1							
A	2	3								2	2	3	3					
A	3	3												1	2	3	3	1
B	1	1	3	7														
B	1	2			2	7	1											
B	1	3					1	7	2									
B	2	2						1	3	4	2							
B	2	3									3	6	1					
B	3	3											1	2	3	3	1	

σ : szórás

Minden egyes paraméterbeállítás mellett 10-10 futtatást végeztem, és vizsgáltam egyrészt az index növekedését ill. csökkenését a távolság függvényében, másrészt azt a távolságot kerestem, ahol a kétklaszteres eredmény helyett a háromklaszteres eredmény kerül elfogadásra. A 7. táblázat adatai azt mutatják, hogy 10 kísérlet esetén melyik távolság esetén ismerte föl az index a három klaszter jelenlétét.

A táblázat adataiból megállapítható, hogy a három klaszter felismerésének nem feltétele, hogy a klaszterek teljesen szeparáltak legyenek. Az is látható azonban, hogy a szórások növekedése esetén a bizonytalanság is egyre növekszik, tehát a felismerési távolság szórása is nagyobb.[3]

A vizsgálatban használt C_3 klaszter szerepe annyi volt, hogy a C_1 és C_2 összevonása esetén is legyen két klaszterünk, amelyre az index számolható. Ezért ezt a C_1-től és C_2-től szeparáltan helyeztem el, a cél ugyanis a C_1 és C_2 közötti átfedés vizsgálata volt.

4.1.4. Az S_Dbw_{new} és a S_Dbw^{**} index összehasonlítása.

Ebben az alfejezetben az Anyag és módszer fejezetben bemutatott 8 féle adatbázison (ld. A.3 - A.10 mellékletek) tesztelem a két indexet. Minden egyes adatbázist mindkét klaszterező algoritmus (K-means, Ward) segítségével csoportokra bontottam, és a csoportok számát 2-től 7-ig változtattam. Ezután összehasonlítottam a kapott klasztereket a tényleges klaszterekkel úgy, hogy a tényleges klaszterekkel (mivel ismertek) a legtöbb egyezést mutató csoporto-

[3] A vizsgálatok során a klaszterek elemszáma nem változott.

DOI: 10.14751/SZIE.2014.035

sítást választottam legjobbnak. A kapott eredményeket rendeztem az 8. táblázatba. Ez tartalmazza az egyes klaszterező eljárások által előállított legjobb csoportosítás klaszterszámát, illetve az egyes indexek által legjobbnak ítélt csoportosítások klaszterszámát.

8. táblázat. Az indexek összehasonlításának eredményei. Forrás: saját számítás.

	Szimuláció	Klaszterek száma					
	sorszáma	KM	T-KM	S-KM	W	T-W	S-W
1. adatbázis	1	4	4	4	4	4	4
	2	4	4	4	4	4	4
	3	4	5	4	4	5	4
	4	4	4	4	4	4	4
	5	4	5	4	4	6	4
	6	4	4	4	4	4	4
	7	4	4	4	4	4	4
	8	4	5	4	4	4	4
	9	4	4	4	4	4	4
	10	4	6	4	4	5	4
2. adatbázis	1	4	4	4	4	4	4
	2	4	4	4	4	5	4
	3	4	6	4	4	4	5
	4	3	5	6	4	5	4
	5	4	4	4	4	7	4
	6	4	5	4	4	4	4
	7	3	7	5	4	7	4
	8	3	5	5	4	4	4
	9	4	4	4	4	5	4
	10	4	4	4	4	7	4
3. adatbázis	1	4	5	4	4	4	4
	2	5	5	3	4	7	7
	3	4	4	4	4	7	4
	4	4	4	4	4	4	4
	5	4	4	4	4	4	4
	6	4	4	4	4	4	4
	7	5	6	7	4	4	4
	8	4	5	4	4	5	4
	9	6	5	3	4	7	4
	10	5	5	2	4	5	4
4. adatbázis	1	5	5	5	4	6	4
	2	4	4	4	4	4	4
	3	3	7	3	4	5	4
	4	4	4	4	4	4	4
	5	4	7	4	4	7	4
	6	4	4	4	4	3	4
	7	3	6	3	4	4	4
	8	5	7	5	4	4	3
	9	4	5	4	4	3	4
	10	4	4	4	4	7	4

(*a táblázat folytatódik*)

DOI: 10.14751/SZIE.2014.035

8. táblázat. Az indexek összehasonlításának eredményei (folytatás).

	Szimuláció	Klaszterek száma					
	sorszáma	KM	T-KM	S-KM	W	T-W	S-W
5. adatbázis	1	3	4	3	4	5	2
	2	3	6	3	3	5	2
	3	3	3	3	4	2	2
	4	3	7	2	3	2	2
	5	3	3	3	2	3	2
	6	3	5	3	3	2	2
	7	3	5	3	4	2	2
	8	3	4	3	4	4	2
	9	3	3	3	4	2	2
	10	3	5	3	3	5	2
6. adatbázis	1	4	7	6	4	7	5
	2	4	6	7	4	7	7
	3	4	7	4	4	5	4
	4	4	7	7	4	7	4
	5	4	7	7	4	5	4
	6	4	7	7	4	4	6
	7	4	6	7	4	4	4
	8	4	7	7	4	5	6
	9	4	7	7	4	4	4
	10	4	6	5	4	7	7
7. adatbázis	1	4	7	6	4	7	7
	2	4	7	6	4	4	7
	3	4	6	7	4	6	6
	4	4	6	7	4	7	7
	5	4	7	5	4	7	6
	6	4	7	7	4	7	5
	7	4	6	6	4	7	5
	8	4	6	7	4	7	6
	9	4	7	7	4	7	7
	10	4	7	7	4	6	7
8. adatbázis	1	4	4	2	4	3	2
	2	5	5	2	4	3	2
	3	4	4	2	4	3	2
	4	5	7	2	4	7	2
	5	4	7	2	4	7	2
	6	4	3	2	4	5	2
	7	4	6	2	4	3	2
	8	4	4	2	4	7	2
	9	4	7	2	4	4	2
	10	5	7	2	4	6	2

KM: Legjobb csoportosítás klaszterszáma (K-means)
W: Legjobb csoportosítás klaszterszáma (Ward)
T-KM: Tong index eredménye (K-means)
T-W: Tong index eredménye (Ward)
S-KM: saját index eredménye (K-means)
S-W: saját index eredménye (Ward)

A kapott eredmények olyan szempont szerint értékeltem, hogy az egyes indexek eltalálták-e az adott algoritmus által előállított megoldások közül a

DOI: 10.14751/SZIE.2014.035

ténylegeshez legközelebb álló megoldást. Az 1. adatbázis tartalmazott jól szeparált klasztereket, mindkét index ebben jó eredményt ért el.

A 2., 3. és 4. adatbázisok esetében az 1. adatbázis klaszterei közelebb kerültek egymáshoz, ill. az elemszámaik is változtak. Ezekben az esetekben megfigyelhető, hogy a lecsökkentett elemszám (3. adatbázis), valamint az egyenlőtlen elemszám esetén (4. adatbázis) a saját index teljesítménye is romlott. A Tong index viszont ezen klaszterelrendezések esetén már sokkal rosszabb eredményt adott, főként a 4. adatbázis esetében. Az általam módosított index a legjobb csoportosításnak megfelelő klaszterszámokat többször találta el, mint a Tong index. A találatok különbsége jelentős.

Az 5. adatbázis esetében lényeges különbség van az egyes klaszterek sűrűsége között, továbbá a K3 klaszter elkülönül a másik kettőtől. Az eredmények tanulmányozásából az derül ki, hogy a K-means algoritmus esetében a háromklaszteres elrendezés bizonyult a legjobbnak mind a tíz szimuláció esetén, míg a Ward algoritmus mindössze 4 esetben adott az eredetihez hasonló megoldást. Az indexeket vizsgálva, a K-means által előállított klaszterek esetében a saját index jobb eredményt ért el (a tíz szimuláció összesítéseként), mint a Tong féle. Ugyanakkor a Ward módszer által előállított klasztereken végzett szimulációk esetében a saját index mindig a kétklaszteres megoldást részesítette előnyben, és csak egyszer találta el a legjobb csoportosítást. Megfigyelhető még, hogy ezen adatbázis esetén a Ward algoritmus által előállított klaszterek száma változékony volt, 2, 3 és 4 klaszteres megoldás is előállt.

A 6. adatbázis előállításakor a szórások változtatásával olyan klasztereket is képeztem, amelyek nem kör alakúak. Továbbá elemszámban és sűrűségben is van közöttük különbség. A négy klaszter nem teljesen szeparált egymástól. Mind a K-means, mind pedig a Ward legjobb besorolása a négyklaszteres megoldás volt (az eredeti adatbázis is ennyi klasztert tartalmazott). Ennek ellenére mindkét index lényegében rossz besorolást határozott meg. A megoldások véletlenszerűnek tűnnek. Vagyis a módosított index alkalmazhatósága ezen adatbázis esetében már szintén megkérdőjelezhető.

A 7. adatbázis a hatodikból keletkezett úgy, hogy a K1 klaszter szórását mindkét irányban megdupláztam, ezáltal kevésbé szeparálódik el a másik háromtól, mint a 6. adatbázis esetében. Hasonlóan az előzőhöz kísérlethez, mindkét esetben a négyklaszteres elrendezés adta a legtöbb egyezést az eredeti klaszterekkel, de a két index egyike sem tudott konzekvens megoldást találni a 10 szimuláció során. Az eredmények nem értékelhetők.

A 8. adatbázis esetén három klaszter nagyon közel került egymáshoz, míg a negyedik (K3) tőlük jól szeparálva helyezkedik el. Mindkét klaszterező algoritmus 4 klaszteres elrendezés esetén adta a legpontosabb besorolást (igaz, a

DOI: 10.14751/SZIE.2014.035

Ward módszer ebben jobban teljesített), de a Tong-féle index ismét nem tudott segítséget adni a legjobb besorolás kiválasztásához. A saját index azonban végig a kétklaszteres megoldást részesítette előnyben. Az A.10 mellékletben szereplő ábráról látható, hogy a három közeli klaszter esetében a klaszterek közötti sűrűség nagy, így nem várható, hogy a módosított index ezeket a csoportokat meg tudja egymástól különböztetni. Tehát az elvárásainknak megfelelő eredményt kaptunk ebben az esetben.

Összefoglalva az eredményeket, az jelenthető ki, hogy a Tong index semelyik szimulációs kísérletben sem múlta fölül az általam létrehozott index eredményeit, viszont több esetben is jóval gyengébb eredményt adott. Természetesen vannak olyan pontelhelyezkedések, ahol egyik index sem tudott támogatást nyújtani egy megfelelő döntés meghozatalában. Tehát ezen korlátokat is figyelembe véve kimondható, hogy a saját index szélesebb körben alkalmazható, a módosítás tehát az alkalmazhatóságot tovább növelte.

4.2. Az előrejelzési modell bővítése, és a tesztelések eredményei

4.2.1. A BG/NBD modell bővítése (2)

A modell bővítésének iránya, és annak indoklása

A vásárlásszámot a panaszok bevonásával előrejelző modell kritikai észrevételei nyomán merült fel a kérdés, hogy miként lehetne kibővíteni az eredeti modellt más módon. A panaszok bevonását a számításokba jónak tartom, és ezen a vonalon készítettem el saját módosításaimat. Azonban nem a panasz időpontjára koncentráltam, hanem arra, hogy az egyes panaszokra milyen megoldást talált a cég: kezelték a problémát vagy nem[4]. Figyelembe fogok venni panaszmentes, és nem panaszmentes vásárlást, továbbá ez utóbbi kategóriát is két csoportra osztom az előzőek értelmében. Így olyan információkat építek be a modellbe, melyek érdemben befolyásol(hat)ják az eredményt.

Feltételezésem szerint a nem kezelt panaszt nagyobb valószínűséggel követi a lemorzsolódás, még akkor is, ha a panasz nem volt jogos. Ezt a feltételezést a paraméterek beállításánál veszem figyelembe.

A modell megalkotásának feltételei

1. Amíg a vásárló aktív, addig az egységnyi idő alatt bekövetkező vásárlások száma Poisson eloszlást követ, melynek paramétere λ.

[4]Kezelt panasz esetén a továbbiakban azt értem, hogy a vásárló panaszát orvosolták, a panaszt pozitívan bírálták el.

DOI: 10.14751/SZIE.2014.035

2. λ változékonysága gamma eloszlást követ r és α paraméterekkel (ld. 2.17. egyenlet, 37. old.)
3. Panaszmentes vásárlás esetén a vásárló q_p valószínűséggel morzsolódik le.
4. q_p változékonysága béta eloszlást követ u_p és v_p paraméterekkel (ld. 3.9. egyenlet, 53. old.).
5. Panasz μ valószínűséggel következik be egy vásárlás után.
6. μ változékonysága béta eloszlást követ a és b paraméterekkel.
7. Egy panaszt ϵ valószínűséggel jogosnak találnak és kezelnek.
8. ϵ változékonysága béta eloszlást követ e és f paraméterekkel.
9. Kezelt panasz után a vásárló q_{c1} valószínűséggel morzsolódik le.
10. q_{c1} változékonysága béta eloszlást követ u_{c1} és v_{c1} paraméterekkel.
11. Nem kezelt panasz után a vásárló q_{c2} valószínűséggel morzsolódik le.
12. q_{c2} változékonysága béta eloszlást követ u_{c2} és v_{c2} paraméterekkel.
13. Az egyes vásárlókra vonatkozó paraméterek egymástól függetlenül változnak.
14. $\lambda > 0$, továbbá $0 < q_p, q_{c1}, q_{c2}, \mu, \epsilon < 1$.

Bemenő adatok

T	a megfigyelési időtartam,
x	a vásárlások száma T idő alatt,
x_{c1}	a kezelt panaszok száma,
x_{c2}	a nem kezelt panaszok száma,
t_x	az utolsó vásárlás időpontja,
z	az utolsó vásárlás panaszmentes (igen: $z = 1$, nem: $z = 0$),
z_1	az utolsó vásárlást kezelt panasz követett (igen: $z_1 = 1$, nem: $z_1 = 0$),
z_2	az utolsó vásárlást nem kezelt panasz követett (igen: $z_2 = 1$, nem: $z_2 = 0$).

z, z_1, z_2 közül pontosan az egyik 1-es, a többi 0.

A Likelihood függvény előállítása

A lehetőségeket 3 esetre bontom. Mindegyik esetében meghatározom a Likelihood függvény összetevőjét (az első vásárlásra).

DOI: 10.14751/SZIE.2014.035

1. eset (panaszmentes vásárlás): $L_1 = \lambda \, \mathrm{e}^{-\lambda t_1}(1-\mu)$.
2. eset (vásárlás kezelt panasszal): $L_2 = \lambda \, \mathrm{e}^{-\lambda t_1} \, \mu\epsilon$.
3. eset (vásárlás nem kezelt panasszal): $L_3 = \lambda \, \mathrm{e}^{-\lambda t_1} \, \mu(1-\epsilon)$.

Az egyes esetek bekövetkezésének száma ismert (bemenő adatok), így az összes eseményre felírható, mindhárom esetet tartalmazó Likelihood függvény a következő lett:

$$\begin{aligned} L_A = & \lambda^x \, \mathrm{e}^{-\lambda t_x} \, \mu^{x_{c1}+x_{c2}}(1-\mu)^{x-1-x_{c1}-x_{c2}}\epsilon^{x_{c1}}(1-\epsilon)^{x_{c2}} \cdot \\ & \cdot (1-q_p)^{x-1-x_{c1}-x_{c2}+1}(1-q_{c1})^{x_{c1}}(1-q_{c2})^{x_{c2}} \end{aligned} \tag{4.4}$$

További kérdés még, hogy a $[t_x, T]$ intervallumban (utolsó vásárlás után, de a megfigyelési időszakon belül) aktív marad-e, vagy inaktívvá válik. Az ezen lehetőségeket figyelembe vevő Likelihood függvény összetevője:

$$L_B = q_p^z q_{c1}^{z_1} q_{c2}^{z_2} + \mathrm{e}^{-\lambda(T-t_x)}(1-q_p)^z(1-q_{c1})^{z_1}(1-q_{c2})^{z_2} \tag{4.5}$$

Az L_A és az L_B függvények szorzataként kapjuk az egyéni szintű Likelihood függvényt:

$$\begin{aligned} & L(\lambda, q_p, q_{c1}, q_{c2}, \mu, \epsilon | T, t_x, x, x_{c1}, x_{c2}, z, z_1, z_2) = \\ & = \lambda^x \, \mathrm{e}^{-\lambda T} \, \mu^{x_{c1}+x_{c2}}(1-\mu)^{x-1-x_{c1}-x_{c2}}\epsilon^{x_{c1}}(1-\epsilon)^{x_{c2}} \cdot \\ & \cdot (1-q_p)^{x-1-x_{c1}-x_{c2}+z}(1-q_{c1})^{x_{c1}+z_1}(1-q_{c2})^{x_{c2}+z_2} + \\ & + \lambda^x \, \mathrm{e}^{-\lambda t_x} \, \mu^{x_{c1}+x_{c2}}(1-\mu)^{x-1-x_{c1}-x_{c2}}\epsilon^{x_{c1}}(1-\epsilon)^{x_{c2}} \cdot \\ & \cdot q_p^z q_{c1}^{z_1} q_{c2}^{z_2}(1-q_p)^{x-1-x_{c1}-x_{c2}}(1-q_{c1})^{x_{c1}}(1-q_{c2})^{x_{c2}} \end{aligned} \tag{4.6}$$

Jelöljük a $\lambda, q_p, q_{c1}, q_{c2}, \mu, \epsilon$ paraméterek halmazát Φ-vel, és tekintsük a bemenő adatok halmazát $(T, t_x, x, x_{c1}, x_{c2}, z, z_1, z_2)$ *input*-nak. Ekkor a 4.6. egyenletben szereplő függvényt a következőképpen jelölhetjük: $L(\Phi|input)$. A függvény két tényező összegére bontható. Az első tényező azzal kapcsolatos, hogy az utolsó vásárlás után aktív marad a vásárló, a második az ellentettje. Ezt a következő jelöléssel fejezem ki: $L(\Phi|input) = L_{\text{aktív}}(\Phi|input) + L_{\text{inaktív}}(\Phi|input)$

Ahhoz, hogy a sokaságra vonatkozó Likelihood függvényt megkapjuk, szükséges az egyes egyéni paraméterek, mint valószínűségi változók sűrűségfüggvényének[5] ismerete. Feltételeink szerint ezek gamma ill. béta eloszlást követnek,

[5] Ezek az ún. a priori sűrűségfüggvények.

DOI: 10.14751/SZIE.2014.035

tehát:

$$\begin{aligned}
f(\lambda|r,\alpha) &= \frac{\alpha^r \lambda^{r-1}\,\mathrm{e}^{-\lambda\alpha}}{\Gamma(r)} \\
f(q_p|u_p,v_p) &= \frac{q_p^{u_p-1}(1-q_p)^{v_p-1}}{\mathrm{B}(u_p,v_p)} \\
f(\mu|a,b) &= \frac{\mu^{a-1}(1-\mu)^{b-1}}{\mathrm{B}(a,b)} \\
f(\epsilon|e,f) &= \frac{\epsilon^{e-1}(1-\epsilon)^{f-1}}{\mathrm{B}(e,f)} \\
f(q_{c1}|u_{c1},v_{c1}) &= \frac{q_{c1}^{u_{c1}-1}(1-q_{c1})^{v_{c1}-1}}{\mathrm{B}(u_{c1},v_{c1})} \\
f(q_{c2}|u_{c2},v_{c2}) &= \frac{q_{c2}^{u_{c2}-1}(1-q_{c2})^{v_{c2}-1}}{\mathrm{B}(u_{c2},v_{c2})}
\end{aligned} \tag{4.7}$$

A sokaságra számított Likelihood függvényt úgy kaphatjuk meg az egyéniből, hogy kiszámítjuk annak várható értékét (vö. folytonos valószínűségi változó várható értéke). A feladatot azonban két részre osztom: az aktív és inaktív komponenseket külön számolom (megtehető, hiszen összegről van szó).

$$\begin{aligned}
&L_{\text{aktív}}(r,\alpha,u_p,v_p,a,b,e,f,u_{c1},v_{c1},u_{c2},v_{c2}|input) = \\
&\quad = \int_0^\infty\int_0^1\int_0^1\int_0^1\int_0^1\int_0^1 \lambda^x\,\mathrm{e}^{-\lambda T}\,\mu^{x_{c1}+x_{c2}}(1-\mu)^{x-1-x_{c1}-x_{c2}}\epsilon^{x_{c1}}(1-\epsilon)^{x_{c2}}\cdot \\
&\qquad \cdot (1-q_p)^{x-1-x_{c1}-x_{c2}+z}(1-q_{c1})^{x_{c1}+z_1}(1-q_{c2})^{x_{c2}+z_2}\cdot \\
&\quad \cdot \frac{\alpha^r\lambda^{r-1}\,\mathrm{e}^{-\lambda\alpha}}{\Gamma(r)}\frac{q_p^{u_p-1}(1-q_p)^{v_p-1}}{\mathrm{B}(u_p,v_p)}\frac{\mu^{a-1}(1-\mu)^{b-1}}{\mathrm{B}(a,b)}\frac{\epsilon^{e-1}(1-\epsilon)^{f-1}}{\mathrm{B}(e,f)}\cdot \\
&\quad \cdot \frac{q_{c1}^{u_{c1}-1}(1-q_{c1})^{v_{c1}-1}}{\mathrm{B}(u_{c1},v_{c1})}\frac{q_{c2}^{u_{c2}-1}(1-q_{c2})^{v_{c2}-1}}{\mathrm{B}(u_{c2},v_{c2})}\,\mathrm{d}q_p\,\mathrm{d}\mu\,\mathrm{d}\epsilon\,\mathrm{d}q_{c1}\,\mathrm{d}q_{c2}\,\mathrm{d}\lambda = \\
&= \int_0^\infty\int_0^1\int_0^1\int_0^1\int_0^1\int_0^1 \lambda^x\,\mathrm{e}^{-\lambda T}\cdot\frac{\alpha^r\lambda^{r-1}\,\mathrm{e}^{-\lambda\alpha}}{\Gamma(r)}\cdot(1-q_p)^{x-1-x_{c1}-x_{c2}+z}\frac{q_p^{u_p-1}(1-q_p)^{v_p-1}}{\mathrm{B}(u_p,v_p)}\cdot \\
&\quad \cdot \mu^{x_{c1}+x_{c2}}(1-\mu)^{x-1-x_{c1}-x_{c2}}\frac{\mu^{a-1}(1-\mu)^{b-1}}{\mathrm{B}(a,b)}\cdot\epsilon^{x_{c1}}(1-\epsilon)^{x_{c2}}\frac{\epsilon^{e-1}(1-\epsilon)^{f-1}}{\mathrm{B}(e,f)}\cdot \\
&\qquad \cdot (1-q_{c1})^{x_{c1}+z_1}\frac{q_{c1}^{u_{c1}-1}(1-q_{c1})^{v_{c1}-1}}{\mathrm{B}(u_{c1},v_{c1})}\cdot \\
&\qquad \cdot (1-q_{c2})^{x_{c2}+z_2}\frac{q_{c2}^{u_{c2}-1}(1-q_{c2})^{v_{c2}-1}}{\mathrm{B}(u_{c2},v_{c2})}\,\mathrm{d}q_p\,\mathrm{d}\mu\,\mathrm{d}\epsilon\,\mathrm{d}q_{c1}\,\mathrm{d}q_{c2}\,\mathrm{d}\lambda
\end{aligned} \tag{4.8}$$

DOI: 10.14751/SZIE.2014.035

Ez az integrál hat integrál szorzatára bontható, melyeket $A_1, A_2, \ldots, A_6$ szimbólumokkal jelölök.

$$A_1 = \int_0^\infty \lambda^x \, \mathrm{e}^{-\lambda T} \cdot \frac{\alpha^r \lambda^{r-1} \, \mathrm{e}^{-\lambda\alpha}}{\Gamma(r)} \, \mathrm{d}\lambda = \frac{\alpha^r}{\Gamma(r)} \int_0^\infty \lambda^{x+r-1} \, \mathrm{e}^{-\lambda(T+\alpha)} \, \mathrm{d}\lambda =$$

(helyettesítéses integrálás $c := \lambda(T+\alpha)$ helyettesítéssel)

$$= \frac{\alpha^r}{\Gamma(r)} \int_0^\infty \left(\frac{c}{\alpha+T}\right)^{x+r-1} \mathrm{e}^{-c} \, \frac{1}{\alpha+T} \, \mathrm{d}c =$$

$$= \frac{\alpha^r}{\Gamma(r)} \frac{1}{(\alpha+T)^{x+r}} \int_0^\infty c^{x+r-1} \, \mathrm{e}^{-c} \, \mathrm{d}c = \frac{\alpha^r}{\Gamma(r)} \frac{1}{(\alpha+T)^{x+r}} \Gamma(x+r) \quad (4.9)$$

$$A_2 = \int_0^1 (1-q_p)^{x-1-x_{c1}-x_{c2}+z} \frac{q_p^{u_p-1}(1-q_p)^{v_p-1}}{\mathrm{B}(u_p, v_p)} \, \mathrm{d}q_p =$$

$$= \int_0^1 \frac{q_p^{u_p-1}(1-q_p)^{x-1-x_{c1}-x_{c2}+z+v_p-1}}{\mathrm{B}(u_p, v_p)} \, \mathrm{d}q_p =$$

$$= \frac{\mathrm{B}(u_p, x-1-x_{c1}-x_{c2}+z+v_p)}{\mathrm{B}(u_p, v_p)} \quad (4.10)$$

$$A_3 = \int_0^1 \mu^{x_{c1}+x_{c2}} (1-\mu)^{x-1-x_{c1}-x_{c2}} \frac{\mu^{a-1}(1-\mu)^{b-1}}{\mathrm{B}(a,b)} \, \mathrm{d}\mu =$$

$$= \int_0^1 \frac{\mu^{x_{c1}+x_{c2}+a-1}(1-\mu)^{x-1-x_{c1}-x_{c2}+b-1}}{\mathrm{B}(a,b)} \, \mathrm{d}\mu =$$

$$= \frac{\mathrm{B}(x_{c1}+x_{c2}+a, x-1-x_{c1}-x_{c2}+b)}{\mathrm{B}(a,b)} \quad (4.11)$$

$$A_4 = \int_0^1 \epsilon^{x_{c1}} (1-\epsilon)^{x_{c2}} \frac{\epsilon^{e-1}(1-\epsilon)^{f-1}}{\mathrm{B}(e,f)} \, \mathrm{d}\epsilon =$$

$$= \int_0^1 \frac{\epsilon^{x_{c1}+e-1}(1-\epsilon)^{x_{c2}+f-1}}{\mathrm{B}(e,f)} \, \mathrm{d}\epsilon = \frac{\mathrm{B}(x_{c1}+e, x_{c2}+f)}{\mathrm{B}(e,f)} \quad (4.12)$$

DOI: 10.14751/SZIE.2014.035

$$A_5 = \int_0^1 (1-q_{c1})^{x_{c1}+z_1} \frac{q_{c1}^{u_{c1}-1}(1-q_{c1})^{v_{c1}-1}}{\mathrm{B}(u_{c1},v_{c1})}\,\mathrm{d}q_{c1} =$$

$$= \int_0^1 \frac{q_{c1}^{u_{c1}-1}(1-q_{c1})^{x_{c1}+z_1+v_{c1}-1}}{\mathrm{B}(u_{c1},v_{c1})}\,\mathrm{d}q_{c1} = \frac{\mathrm{B}(u_{c1},x_{c1}+z_1+v_{c1})}{\mathrm{B}(u_{c1},v_{c1})} \quad (4.13)$$

$$A_6 = \int_0^1 (1-q_{c2})^{x_{c2}+z_2} \frac{q_{c2}^{u_{c2}-1}(1-q_{c2})^{v_{c2}-1}}{\mathrm{B}(u_{c2},v_{c2})}\,\mathrm{d}q_{c2} =$$

$$= \int_0^1 \frac{q_{c2}^{u_{c2}-1}(1-q_{c2})^{x_{c2}+z_2+v_{c2}-1}}{\mathrm{B}(u_{c2},v_{c2})}\,\mathrm{d}q_{c2} = \frac{\mathrm{B}(u_{c2},x_{c2}+z_2+v_{c2})}{\mathrm{B}(u_{c2},v_{c2})} \quad (4.14)$$

Tehát

$$L_{\text{aktív}}(r,\alpha,u_p,v_p,a,b,e,f,u_{c1},v_{c1},u_{c2},v_{c2}|input) = \prod_{i=1}^{6} A_i \quad (4.15)$$

Ezután az inaktív komponensét számolom ki.

$$L_{\text{inaktív}}(r,\alpha,u_p,v_p,a,b,e,f,u_{c1},v_{c1},u_{c2},v_{c2}|input) =$$

$$= \int_0^\infty\int_0^1\int_0^1\int_0^1\int_0^1\int_0^1 \lambda^x\,\mathrm{e}^{-\lambda t_x}\,\mu^{x_{c1}+x_{c2}}(1-\mu)^{x-1-x_{c1}-x_{c2}}\epsilon^{x_{c1}}(1-\epsilon)^{x_{c2}}\cdot$$

$$\cdot\, q_p^z q_{c1}^{z_1} q_{c2}^{z_2}(1-q_p)^{x-1-x_{c1}-x_{c2}}(1-q_{c1})^{x_{c1}}(1-q_{c2})^{x_{c2}}\cdot$$

$$\cdot \frac{\alpha^r\lambda^{r-1}\,\mathrm{e}^{-\lambda\alpha}}{\Gamma(r)} \frac{q_p^{u_p-1}(1-q_p)^{v_p-1}}{\mathrm{B}(u_p,v_p)} \frac{\mu^{a-1}(1-\mu)^{b-1}}{\mathrm{B}(a,b)} \frac{\epsilon^{e-1}(1-\epsilon)^{f-1}}{\mathrm{B}(e,f)}\cdot$$

$$\cdot \frac{q_{c1}^{u_{c1}-1}(1-q_{c1})^{v_{c1}-1}}{\mathrm{B}(u_{c1},v_{c1})} \frac{q_{c2}^{u_{c2}-1}(1-q_{c2})^{v_{c2}-1}}{\mathrm{B}(u_{c2},v_{c2})}\,\mathrm{d}q_p\,\mathrm{d}\mu\,\mathrm{d}\epsilon\,\mathrm{d}q_{c1}\,\mathrm{d}q_{c2}\,\mathrm{d}\lambda =$$

$$= \int_0^\infty\int_0^1\int_0^1\int_0^1\int_0^1\int_0^1 \lambda^x\,\mathrm{e}^{-\lambda t_x}\cdot\frac{\alpha^r\lambda^{r-1}\,\mathrm{e}^{-\lambda\alpha}}{\Gamma(r)}\cdot q_p^z(1-q_p)^{x-1-x_{c1}-x_{c2}}\frac{q_p^{u_p-1}(1-q_p)^{v_p-1}}{\mathrm{B}(u_p,v_p)}\cdot$$

$$\cdot\,\mu^{x_{c1}+x_{c2}}(1-\mu)^{x-1-x_{c1}-x_{c2}}\frac{\mu^{a-1}(1-\mu)^{b-1}}{\mathrm{B}(a,b)}\cdot\epsilon^{x_{c1}}(1-\epsilon)^{x_{c2}}\frac{\epsilon^{e-1}(1-\epsilon)^{f-1}}{\mathrm{B}(e,f)}\cdot$$

$$\cdot\, q_{c1}^{z_1}(1-q_{c1})^{x_{c1}}\frac{q_{c1}^{u_{c1}-1}(1-q_{c1})^{v_{c1}-1}}{\mathrm{B}(u_{c1},v_{c1})}\cdot$$

$$\cdot\, q_{c2}^{z_2}(1-q_{c2})^{x_{c2}+1-z_2}\frac{q_{c2}^{u_{c2}-1}(1-q_{c2})^{v_{c2}-1}}{\mathrm{B}(u_{c2},v_{c2})}\,\mathrm{d}q_p\,\mathrm{d}\mu\,\mathrm{d}\epsilon\,\mathrm{d}q_{c1}\,\mathrm{d}q_{c2}\,\mathrm{d}\lambda \quad (4.16)$$

DOI: 10.14751/SZIE.2014.035

Ez az integrál is hat integrál szorzatára bontható, melyeket $B_1, B_2, \ldots, B_6$ szimbólumokkal jelölök.

$$B_1 = \int_0^\infty \lambda^x \, \mathrm{e}^{-\lambda t_x} \cdot \frac{\alpha^r \lambda^{r-1} \, \mathrm{e}^{-\lambda\alpha}}{\Gamma(r)} \, \mathrm{d}\lambda = \frac{\alpha^r}{\Gamma(r)} \frac{1}{(\alpha + t_x)^{x+r}} \Gamma(x + r) \quad (4.17)$$

(ld. A_1)

$$B_2 = \int_0^1 q_p^z (1 - q_p)^{x-1-x_{c1}-x_{c2}} \frac{q_p^{u_p-1}(1 - q_p)^{v_p-1}}{\mathrm{B}(u_p, v_p)} \, \mathrm{d}q_p =$$

$$= \int_0^1 \frac{q_p^{u_p+z-1}(1 - q_p)^{x-1-x_{c1}-x_{c2}+v_p-1}}{\mathrm{B}(u_p, v_p)} \, \mathrm{d}q_p =$$

$$= \frac{\mathrm{B}(u_p + z, x - 1 - x_{c1} - x_{c2} + v_p)}{\mathrm{B}(u_p, v_p)} \quad (4.18)$$

$$B_3 = \int_0^1 \mu^{x_{c1}+x_{c2}} (1 - \mu)^{x-1-x_{c1}-x_{c2}} \frac{\mu^{a-1}(1 - \mu)^{b-1}}{\mathrm{B}(a, b)} \, \mathrm{d}\mu =$$

$$= \int_0^1 \frac{\mu^{x_{c1}+x_{c2}+a-1}(1 - \mu)^{x-1-x_{c1}-x_{c2}+b-1}}{\mathrm{B}(a, b)} \, \mathrm{d}\mu =$$

$$= \frac{\mathrm{B}(x_{c1} + x_{c2} + a, x - 1 - x_{c1} - x_{c2} + b)}{\mathrm{B}(a, b)} \quad (4.19)$$

$$B_4 = \int_0^1 \epsilon^{x_{c1}} (1 - \epsilon)^{x_{c2}} \cdot \frac{\epsilon^{e-1}(1 - \epsilon)^{f-1}}{\mathrm{B}(e, f)} \, \mathrm{d}\epsilon =$$

$$= \int_0^1 \frac{\epsilon^{x_{c1}+e-1}(1 - \epsilon)^{x_{c2}+f-1}}{\mathrm{B}(e, f)} \, \mathrm{d}\epsilon = \frac{\mathrm{B}(x_{c1} + e, x_{c2} + f)}{\mathrm{B}(e, f)} \quad (4.20)$$

$$B_5 = \int_0^1 q_{c1}^{z_1} (1 - q_{c1})^{x_{c1}} \frac{q_{c1}^{u_{c1}-1}(1 - q_{c1})^{v_{c1}-1}}{\mathrm{B}(u_{c1}, v_{c1})} \, \mathrm{d}q_{c1} =$$

$$= \int_0^1 \frac{q_{c1}^{u_{c1}+z_1-1}(1 - q_{c1})^{x_{c1}+v_{c1}-1}}{\mathrm{B}(u_{c1}, v_{c1})} \, \mathrm{d}q_{c1} = \frac{\mathrm{B}(u_{c1} + z_1, x_{c1} + v_{c1})}{\mathrm{B}(u_{c1}, v_{c1})} \quad (4.21)$$

DOI: 10.14751/SZIE.2014.035

$$B_6 = \int_0^1 q_{c2}^{z_2}(1-q_{c2})^{x_{c2}+1-z_2}\frac{q_{c2}^{u_{c2}-1}(1-q_{c2})^{v_{c2}-1}}{\mathrm{B}(u_{c2},v_{c2})}\,\mathrm{d}q_{c2} =$$

$$= \int_0^1 \frac{q_{c2}^{u_{c2}+z_2-1}(1-q_{c2})^{x_{c2}+v_{c2}-1}}{\mathrm{B}(u_{c2},v_{c2})}\,\mathrm{d}q_{c2} = \frac{\mathrm{B}(u_{c2}+z_2,x_{c2}+v_{c2})}{\mathrm{B}(u_{c2},v_{c2})} \quad (4.22)$$

Tehát

$$L_{\text{inaktív}}(r,\alpha,u_p,v_p,a,b,e,f,u_{c1},v_{c1},u_{c2},v_{c2}|input) = \prod_{i=1}^{6} B_i \quad (4.23)$$

Vagyis a keresett Likelihood függvény a következő:

$$L(r,\alpha,u_p,v_p,a,b,e,f,u_{c1},v_{c1},u_{c2},v_{c2}|input) = \prod_{i=1}^{6} A_i + \prod_{i=1}^{6} B_i \quad (4.24)$$

Vezessük még be a Θ jelölést az $r,\alpha,u_p,v_p,a,b,e,f,u_{c1},v_{c1},u_{c2},v_{c2}$ paraméterek halmazára. Ekkor a függvény a következőképpen írható föl: $L(\Theta|input) = \prod_{i=1}^{6} A_i + \prod_{i=1}^{6} B_i$

A vásárlásszám várható értékének meghatározása

Egy adott vásárló esetén egy tetszőleges t időpontig lezajlott vásárlások számát $X(t)$-vel jelölve, keressük ennek várható értékét, vagyis $\mathrm{E}(X(t))$-t. Ez lesz az alapja annak, hogy a későbbiekben előrejelzést tudjunk adni a T-n túli időszakra.

A problémát két esetre kell bontani: 1. a t időpontnál később válik inaktívvá, 2. a t időpontnál korábban válik inaktívvá a vásárló.

1. Jelöljük az inaktívvá válás időpontját τ-val. Ebben az esetben tehát $\tau > t$. Ekkor a vásárlások számának várható értéke λt. Meg kell még határozni annak a valószínűségét, hogy ez az eset áll elő, vagyis keressük a $\mathrm{P}(\tau > t)$ valószínűséget, másként fogalmazva, annak a valószínűségét, hogy egy $t \in [0;T]$ időpontban a vásárló még aktív.

   A t idő alatt bekövetkező vásárlások száma legyen j. Ebből k esetben következik be panasz ($j-k$ esetben nem), továbbá l db panaszt kezelnek, miközben nem következik be lemorzsolódás. Ennek a valószínűségét keressük, mely a következő alakban írható föl: $\frac{(\lambda t)^j}{j!}\mathrm{e}^{-\lambda t}\cdot\binom{j}{k}\mu^k(1-\mu)^{j-k}\cdot\binom{k}{l}\epsilon^l(1-\epsilon)^{k-l}\cdot q_p^0(1-q_p)^{j-k}\cdot q_{c1}^0(1-q_{c1})^l\cdot q_{c2}^0(1-q_{c2})^{k-l}$
   Az l értéke 0 és k között mozoghat, k értéke 0 és j között, míg j (elvileg) bármilyen nemnegatív egész értéket felvehet (vö. Poisson eloszlás). Ezért

DOI: 10.14751/SZIE.2014.035

a keresett valószínűséget ezen valószínűségek összegeként kapjuk:

$$\mathrm{P}(\tau > t) = \sum_{j=0}^{\infty}\sum_{k=0}^{j}\sum_{l=0}^{k} \frac{(\lambda t)^j}{j!}\,\mathrm{e}^{-\lambda t}\binom{j}{k}\mu^k(1-\mu)^{j-k}\cdot$$
$$\cdot\binom{k}{l}\epsilon^l(1-\epsilon)^{k-l}(1-q_p)^{j-k}(1-q_{c1})^l(1-q_{c2})^{k-l} \quad (4.25)$$

Az összeg meghatározását három részre osztom: C_1, C_2, C_3 (belülről haladva kifelé).

$$C_1 = \sum_{l=0}^{k}\binom{k}{l}\epsilon^l(1-\epsilon)^{-l}(1-q_{c1})^l(1-q_{c2})^{-l} =$$
$$= \sum_{l=0}^{k}\binom{k}{l}\left(\frac{\epsilon(1-q_{c1})}{(1-\epsilon)(1-q_{c2})}\right)^l = \left(1+\frac{\epsilon(1-q_{c1})}{(1-\epsilon)(1-q_{c2})}\right)^k \quad (4.26)$$

C_1 számításához felhasználtam, hogy $\sum_{k=0}^{n}\binom{n}{k}x^k = (1+x)^n$

$$C_2 = \sum_{k=0}^{j}\binom{j}{k}\mu^k(1-\mu)^{-k}(1-\epsilon)^k(1-q_p)^{-k}(1-q_{c2})^k\cdot$$
$$\cdot\left(1+\frac{\epsilon(1-q_{c1})}{(1-\epsilon)(1-q_{c2})}\right)^k =$$
$$= \sum_{k=0}^{j}\binom{j}{k}\left(\frac{\mu(1-\epsilon)(1-q_{c2})}{(1-\mu)(1-q_p)} + \frac{\mu\epsilon(1-q_{c1})}{(1-\mu)(1-q_p)}\right)^k =$$
$$= \left(1+\frac{\mu(1-\epsilon)(1-q_{c2})}{(1-\mu)(1-q_p)} + \frac{\mu\epsilon(1-q_{c1})}{(1-\mu)(1-q_p)}\right)^j =$$
$$= \left(\frac{(1-\mu)(1-q_p)+\mu(1-\epsilon)(1-q_{c2})+\mu\epsilon(1-q_{c1})}{(1-\mu)(1-q_p)}\right)^j \quad (4.27)$$

$$C_3 = \sum_{j=0}^{\infty}\frac{(\lambda t)^j}{j!}\,\mathrm{e}^{-\lambda t}(1-\mu)^j(1-q_p)^j\cdot$$
$$\cdot\left(\frac{(1-\mu)(1-q_p)+\mu(1-\epsilon)(1-q_{c2})+\mu\epsilon(1-q_{c1})}{(1-\mu)(1-q_p)}\right)^j =$$
$$= \mathrm{e}^{-\lambda t}\sum_{j=0}^{\infty}\frac{(\lambda t\,[(1-\mu)(1-q_p)+\mu(1-\epsilon)(1-q_{c2})+\mu\epsilon(1-q_{c1})])^j}{j!} =$$
$$= \mathrm{e}^{-\lambda t}\cdot\mathrm{e}^{\lambda t[(1-\mu)(1-q_p)+\mu(1-\epsilon)(1-q_{c2})+\mu\epsilon(1-q_{c1})]} =$$
$$= \mathrm{e}^{-\lambda t[1-(1-\mu)(1-q_p)-\mu(1-\epsilon)(1-q_{c2})-\mu\epsilon(1-q_{c1})]} \quad (4.28)$$

DOI: 10.14751/SZIE.2014.035

C_3 számításához felhasználtam, hogy $\sum_{k=0}^{\infty} \frac{x^k}{k!} = \mathrm{e}^x$
Mivel $C_3 = \mathrm{P}(\tau > t)$, ezért

$$\mathrm{P}(\tau > t) = \mathrm{e}^{-\lambda t[1-(1-\mu)(1-q_p)-\mu(1-\epsilon)(1-q_{c2})-\mu\epsilon(1-q_{c1})]} \tag{4.29}$$

2. A második eset, hogy $\tau < t$, vagyis a kiválasztott t időpont előtt válik inaktívvá a vásárló. Ebben az esetben a vásárlások számának várható értéke $\lambda\tau$. Mivel τ-t (idő) folytonos valószínűségi változónak tekintjük, és ennek a segítségével határozzuk meg a vásárlások számának várható értékét, elő kell még állítani a τ sűrűségfüggvényét (ld. folytonos valószínűségi változó várható értéke).

   A 4.29. egyenlet azt mutatja, hogy τ egy exponenciális eloszlású valószínűségi változónak tekinthető. Egyszerűsítsük a kifejezést a $c := 1-(1-\mu)(1-q_p)-\mu(1-\epsilon)(1-q_{c2})-\mu\epsilon(1-q_{c1})$ bevezetésével: $\mathrm{P}(\tau > t) = \mathrm{e}^{-\lambda ct}$. Ennek a valószínűségi változónak a sűrűségfüggvénye (változóját x-szel jelölve):

$$f(x) = \lambda c\, \mathrm{e}^{-\lambda cx} \tag{4.30}$$

A vásárlások számának várható értéke ezen információk birtokában meghatározható. A fenti szétválasztásnak megfelelően két tényezőből adódik össze:

$$\mathrm{E}(X(t)|\lambda, q_p, q_{c1}, q_{c2}, \mu, \epsilon) = \lambda t \cdot \mathrm{P}(\tau > t) + \int_0^t \lambda x \cdot f(x)\,\mathrm{d}x =$$

$$= \lambda t \cdot \mathrm{e}^{-\lambda ct} + \int_0^t \lambda x \cdot \lambda c\, \mathrm{e}^{-\lambda cx}\,\mathrm{d}x =$$

$$= \frac{1 - \mathrm{e}^{-\lambda t[1-(1-\mu)(1-q_p)-\mu(1-\epsilon)(1-q_{c2})-\mu\epsilon(1-q_{c1})]}}{1-(1-\mu)(1-q_p)-\mu(1-\epsilon)(1-q_{c2})-\mu\epsilon(1-q_{c1})} \tag{4.31}$$

Az eredmény levezetését az A.12. melléklet (131. old.) tartalmazza.

A vásárlásszám előrejelzése

A modellépítés céljához érkeztünk, vagyis annak meghatározásához, hogy a vizsgált időtartamon túl, t idő alatt várhatóan hány vásárlást bonyolít le egy-egy vásárló ($Y(t)$), ennek segítségével pedig „személyre szabott" marketing eszközöket alkalmazhatunk közöttük.

A cél tehát egyéni szinten $\mathrm{E}(Y(t)|\lambda, q_p, q_{c1}, q_{c2}, \mu, \epsilon, input)$ meghatározása, ill. $\mathrm{E}(Y(t)|r, \alpha, u_p, v_p, a, b, e, f, u_{c1}, v_{c1}, u_{c2}, v_{c2}, input)$ meghatározása a populáció szintjén. Először megint egy konkrét vásárló esetében adjuk meg (vagyis

DOI: 10.14751/SZIE.2014.035

ismertnek tételezzük fel a $\lambda, q_p, q_{c1}, q_{c2}, \mu, \epsilon$ paramétereket, melyeket korábban Φ-vel jelöltem).

Az előrejelzés esetén szükségszerű azonban, hogy a vásárló a T időpontban (vagyis a megfigyelési időszak végén) még aktív legyen. A megoldást azon feltétel mellett keressük, hogy ismerjük a vásárlási „szokásait" az elmúlt időszakban. Keressük tehát $\mathrm{P}(\tau > T|\Phi, input)$-ot. A 4.5. egyenlet (71. old.) 2. része tartalmazza annak valószínűségét, hogy az utolsó vásárlás (t_x) után aktív marad, az egész egyenlet pedig annak a valószínűsége, hogy aktív marad (csak a T időpontig már nem vásárol), vagy inaktívvá válik ebben az időszakban. A kettő hányadosa adja a keresett valószínűséget:

$$\mathrm{P}(\tau > T|\Phi, input) = \\ = \frac{\mathrm{e}^{-\lambda(T-t_x)}(1-q_p)^z(1-q_{c1})^{z_1}(1-q_{c2})^{z_2}}{q_p^z q_{c1}^{z_1} q_{c2}^{z_2} + \mathrm{e}^{-\lambda(T-t_x)}(1-q_p)^z(1-q_{c1})^{z_1}(1-q_{c2})^{z_2}} \quad (4.32)$$

Ha a tört számlálóját és nevezőjét is megszorozzuk L_A-val (4.4. egyenlet), akkor a számlálóban $L_{\text{aktív}}(\Phi|input)$ értékét (4.6. egyenlet), a nevezőben $L(\Phi|input)$-ot kapjuk. Azaz

$$\mathrm{P}(\tau > T|\Phi, input) = \frac{L_{\text{aktív}}(\Phi|input)}{L(\Phi|input)} \quad (4.33)$$

Ezen valószínűség, valamint $\mathrm{E}(X(t)|\Phi)$ (4.31. egyenlet) ismeretében megadhatjuk a vásárlások számának várható értékét a T időpont utáni t időtartamra, vagyis a $[T, T+t]$ időintervallumra. Egy adott vásárló esetén ez:

$$\mathrm{E}(Y(t)|\Phi, input) = \mathrm{E}(X(t)|\Phi) \cdot \mathrm{P}(\tau > T|\Phi, input) = \\ = \frac{1-\mathrm{e}^{-\lambda t[1-(1-\mu)(1-q_p)-\mu(1-\epsilon)(1-q_{c2})-\mu\epsilon(1-q_{c1})]}}{1-(1-\mu)(1-q_p)-\mu(1-\epsilon)(1-q_{c2})-\mu\epsilon(1-q_{c1})} \frac{L_{\text{aktív}}(\Phi|input)}{L(\Phi|input)} \quad (4.34)$$

Mivel az egyes vásárlókra vonatkozó paraméterek (Φ) nem ismertek, hanem az eloszlásukra vonatkozó feltételeket ismerjük (ld. a modell megalkotásának feltételeit), ezen eloszlások segítségével kell az előrejelzést elkészíteni. Ez pedig $\mathrm{E}(Y(t)|\Phi, input)$ várható értéke, a $\lambda, q_p, q_{c1}, q_{c2}, \mu, \epsilon$ paraméterek eloszlásának figyelembevételével, azon feltétel mellett, hogy a vásárlásokról információk állnak rendelkezésünkre ($input$).

$$\int_0^\infty \int_0^1 \int_0^1 \int_0^1 \int_0^1 \int_0^1 \mathrm{E}(Y(t)|\Phi, input) f(\Phi|input)\, \mathrm{d}q_p\, \mathrm{d}\mu\, \mathrm{d}\epsilon\, \mathrm{d}q_{c1}\, \mathrm{d}q_{c2}\, \mathrm{d}\lambda \quad (4.35)$$

DOI: 10.14751/SZIE.2014.035

Itt $f(\Phi|input)$ az ún. a posteriori sűrűségfüggvény, mely a Bayes-tétel segítségével a következő alakra hozható:

$$f(\Phi|input) = \frac{f(input|\Phi)f(\Phi)}{\int f(input|\Phi)f(\Phi)\,\mathrm{d}\Phi} = \frac{L(\Phi|input)f(\Phi)}{L(\Theta|input)} = \\ = \frac{L(\Phi|input)f(q_p)f(\mu)f(\epsilon)f(q_{c1})f(q_{c2})f(\lambda)}{L(\Theta|input)} \quad (4.36)$$

A számítás során felhasználtam, hogy a paraméterek személyről-személyre egymástól függetlenül változnak. Összegezve (4.34. és 4.36. egyenlet behelyettesítése 4.35. egyenletbe):

$$\mathrm{E}(Y(t)|\Theta, input) = \int_0^\infty\int_0^1\int_0^1\int_0^1\int_0^1\int_0^1 \frac{1-\mathrm{e}^{-\lambda tc}}{c}\frac{L_{\text{aktív}}(\Phi|input)}{L(\Phi|input)}\cdot \\ \cdot\frac{L(\Phi|input)f(q_p)f(\mu)f(\epsilon)f(q_{c1})f(q_{c2})f(\lambda)}{L(\Theta|input)}\,\mathrm{d}q_p\,\mathrm{d}\mu\,\mathrm{d}\epsilon\,\mathrm{d}q_{c1}\,\mathrm{d}q_{c2}\,\mathrm{d}\lambda = \\ = \frac{1}{L(\Theta|input)}\int_0^\infty\int_0^1\int_0^1\int_0^1\int_0^1\int_0^1 \frac{1-\mathrm{e}^{-\lambda tc}}{c}L_{\text{aktív}}(\Phi|input)\cdot \\ \cdot f(q_p)f(\mu)f(\epsilon)f(q_{c1})f(q_{c2})f(\lambda)\,\mathrm{d}q_p\,\mathrm{d}\mu\,\mathrm{d}\epsilon\,\mathrm{d}q_{c1}\,\mathrm{d}q_{c2}\,\mathrm{d}\lambda \quad (4.37)$$

Az integrál nem hozható zárt alakra, azonban ez nem más, mint $\frac{1-\mathrm{e}^{-\lambda tc}}{c}\cdot L_{\text{aktív}}(\Phi|input)$ várható értéke, hiszen a feltételekben megkövetelt függetlenség miatt $f(q_p)\cdot f(\mu)\cdot f(\epsilon)\cdot f(q_{c1})\cdot f(q_{c2})\cdot f(\lambda) = f(q_p,\mu,\epsilon,q_{c1},q_{c2},\lambda)$. Ennek közelítő értékét (az átlagot) fogom meghatározni, olyan módon, hogy az egyéni szintű paraméterek (Φ), valamint az $f(q_p)$, $f(\mu)$, $f(\epsilon)$, $f(q_{c1})$, $f(q_{c2})$, $f(\lambda)$ eloszlásfüggvények (4.7. egyenlet) segítségével előállítok véletlen mintákat, és azok átlagaként kapom a várható érték közelítő értékét.

Ezek alapján a következő alakban adható meg a közelítő érték:

$$\mathrm{E}(Y(t)|\Theta, input) \approx \frac{1}{L(\Theta|input)}\frac{1}{N}\sum_{i=1}^{N}\left[\frac{1-\mathrm{e}^{-\lambda_i tc_i}}{c_i}\cdot L_{\text{aktív}}(\Phi_i|input)\right] \quad (4.38)$$

ahol

N a véletlen minta elemszáma,

$c_i = 1-(1-\mu_i)(1-q_{p_i})-\mu_i(1-\epsilon_i)(1-q_{c2_i})-\mu_i\epsilon_i(1-q_{c1_i}),\quad 1\leqslant i\leqslant N,$

továbbá

λ_i a $\Gamma(r,\alpha)$,

q_{p_i} a $\mathrm{B}(u_p,v_p)$,

μ_i a $\mathrm{B}(a,b)$,

DOI: 10.14751/SZIE.2014.035

ϵ_i a $\mathrm{B}(e, f)$,
q_{c1i} a $\mathrm{B}(u_{c1}, v_{c1})$,
q_{c2i} a $\mathrm{B}(u_{c2}, v_{c2})$ eloszlású valószínűségi változó i-edik véletlenszerűen kiválasztott értéke .

4.2.2. A vizsgálatba bevont modellek

Az Anyag és módszer fejezetben bemutatott adatbázisokon (3.2.2. alfejezet) három modellt teszteltem: az eredeti BG/NBD modellt (2.6.2. alfejezet), ennek általam történt módosítását (4.2.1. alfejezet), valamint egy ún. heurisztikus modellt (2.6.2. alfejezet).

Az első két modell részletes leírása megtörtént, ezért itt most csak az alkalmazott heurisztikus módszerrel foglalkozom.

A heurisztikus modell esetében a megfigyelési időszakot minden esetben 2 részre kellett bontani: egy tanulási és egy teszt időszakra. Kísérleteimben a T megfigyelési időszakot két egyenlő $(T/2, T/2)$ részre osztottam. Vagyis megvizsgáltam, hogy mennyi az utolsó vásárlások időpontjának átlaga azon vásárlók esetében, akik inaktívvá váltak az első $T/2$ időszakban[6], és az előrejelzéshez ezen utolsó vásárlások időpontjának átlagát választottam kritikus időpontnak. Természetesen az egész T időpontra számolt „hiatus" érték az előbb számolt kritikus érték duplája. Aki ennél régebben vásárolt (a megfigyelési, azaz a T időszakban), azt inaktívnak tekintettem az előrejelzési (t) időszakra, akinek viszont ennél későbbi az utolsó vásárlásának időpontja, annak a vásárlásszámát a megfigyelési időszak vásárlásszámához mérten számítottam ki (egyenes arányosságot feltételezve a vásárlásszám és az eltelt idő között).

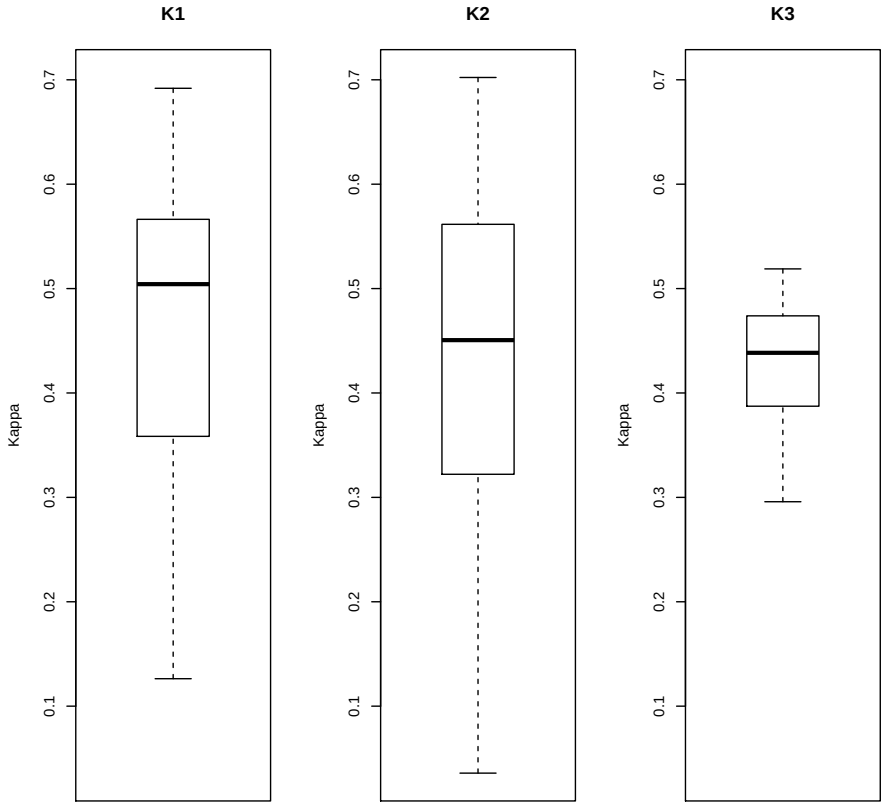
4.2.3. Az előrejelzési időszakban még aktív vásárlók előrejelzésének tesztelése

Ez a vizsgálat arra irányul, hogy az egyes modellek mennyire képesek előrejelezni egy adott vásárló inaktívvá válását a megfigyelési időszak adataiból. A technikai megvalósítás során először előállítottam az adatbázist, majd ezen lefuttattam mindhárom modellt (A.13. melléklet). Minden egyes paraméterbeállítás esetén 10-10 modelleredményt átlagoltam és ezen értékekkel számoltam tovább. A kapott eredmények az A.14. mellékletben találhatók. A Kappa statisztikák értékeit vizsgáltam az egyes modellek esetében, az eredményeket boxplot ábrán szemléltettem. A 7. ábra alapján úgy tűnik, hogy a legjobb átlagos eredményt a saját modell érte el (ennek kappa értékeit jelöltem $K1$-gyel, a BG/NBD modellét $K2$-vel, míg a heurisztikus modell kappa értékeit

[6]Ha a második $T/2$ időszakba nem vásároltak, akkor inaktívvá vált az első $T/2$ időszakban.

DOI: 10.14751/SZIE.2014.035

$K3$-mal). Azonban az eltérés nem tekinthető statisztikailag igazoltnak, melyet alátámaszt a $K1$ és $K2$ eredményeken végrehajtott páros Wilcoxon próba[7], mely szerint 5%-os szignifikanciaszinten[8] nem vethető el a nullhipotézis, tehát a két átlagérték különbözősége nem igazolt ($p = 0{,}094$).



K1: a kappa statisztika értékei a saját modell esetében,
K2: a kappa statisztika értékei a BG/NBD modell esetében,
K3: a kappa statisztika értékei a heurisztikus modell esetében.

7. ábra. A Kappa statisztika értékei a három modell esetében. Forrás: saját szerkesztés.

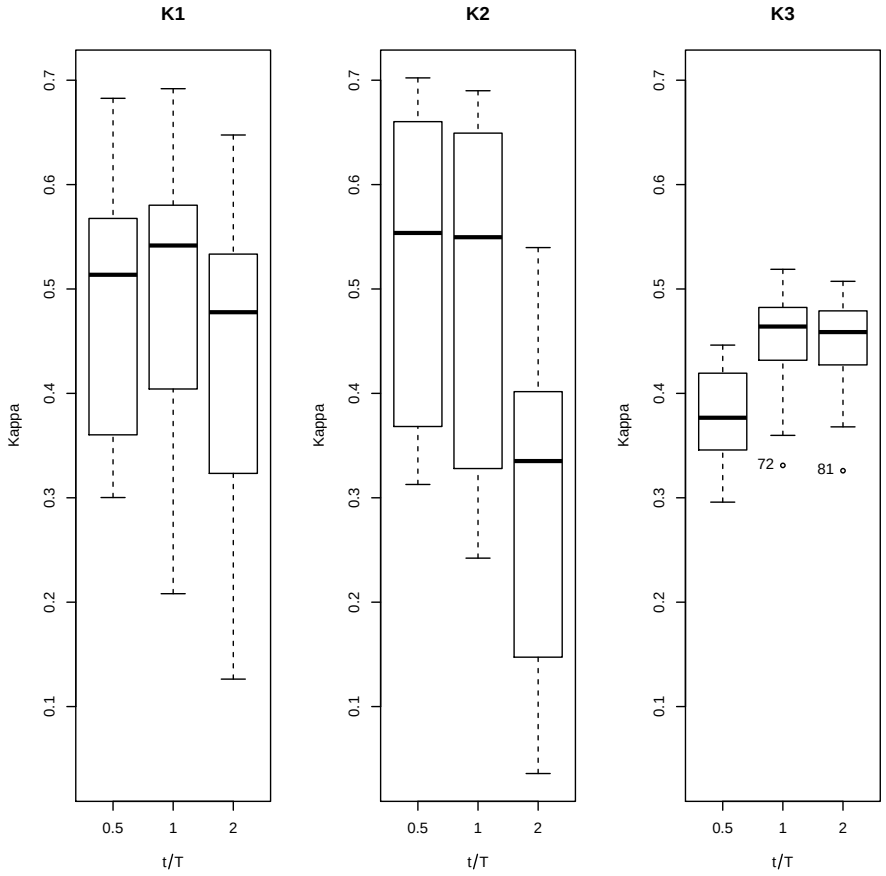
Ezt a képet azonban árnyalja, ha felbontjuk az egyes modellek eredményeit aszerint, hogy az előrejelzési időszak hányszorosa a megfigyelési időszaknak, vagyis t/T értékei (0,5; 1; 2) szerint három csoportot alkothatunk minden egyes modell esetében. Ezen eredmények a 8. ábrán láthatók. Összevetve az első két modellt (saját és BG/NBD) megfigyelhető, hogy a második modell teljesítménye a harmadik esetben, a $t/T = 2$ (azaz, ha az előrejelzési időszak duplája a megfigyelési időszaknak) paraméterbeállítás mellett nagyon lecsökkent. Az ábráról az olvasható le, hogy a két modell esetében a harmadik eredmények mediánja jelentősen eltér egymástól, melyet megerősít a Wilcoxon teszt eredménye ($p = 7{,}451\text{e-}08$). A másik kettő esetében ($t/T = 0{,}5$, ill.

[7] A párosított t-próba feltétele (a minta normális eloszlásból származása) nem teljesült, ezért alkalmaztam ezt a nem paraméteres próbát.

[8] A továbbiakban a szignifikancia szintet 5%-nak tekintem, ha ettől eltérés történik, akkor ezt külön jelzem.

DOI: 10.14751/SZIE.2014.035

$t/T = 1$) az ábrán látható különbségek statisztikailag az első esetben kimutathatók, a második esetben viszont nem[9].
Vagyis a hosszabb távra szolgáló előrejelzés esetében a saját modell megbízhatóbbnak bizonyult, mint a BG/NBD modell.



K1: a kappa statisztika értékei a saját modell esetében,
K2: a kappa statisztika értékei a BG/NBD modell esetében,
K3: a kappa statisztika értékei a heurisztikus modell esetében.

8. ábra. A Kappa statisztika értékei különböző t/T arányok mellett a három modell esetében. Forrás: saját szerkesztés.

A harmadik modellel való összevetés során az első szembetűnő különbség a szórásokban tapasztalható nagy különbség (8. ábra). Mivel az értékek nem mások, mint a Kappa statisztika értékei az inaktívvá válás előrejelzése kapcsán, így az mondható ki, hogy a heurisztikus modell kisebb szórása azt jelenti, viszonylag biztosan produkál egy gyenge közepes előrejelzést (Kappa $\in [0{,}3; 0{,}5]$). Ezzel szemben a másik két modell eredményei nagyon gyengétől (0,1) jóig (0,7) terjednek. Ha megvizsgáljuk az átlagok különbözőségét a saját és a heurisztikus modell esetében, akkor az összesített eredmények esetén (7. ábra) kimutatható a különbség ($p = 0{,}006$), a t/T hányados szerint szétválogatott esetek közül az elsőben szintén kimutatható a különbség ($p = 6{,}3\text{e-}05$),

[9] A páros Wilcoxon próbával kapott p értékek: $p = 3{,}1\text{e-}06$ ill. $p = 0{,}628$.

DOI: 10.14751/SZIE.2014.035

a második és a harmadik esetben viszont nem ($p = 0{,}229$ ill. $p = 0{,}878$). Az eredmények alapján a saját modell átlagosan jobb eredményt adott a heurisztikus modellnél.

Ez a vizsgálat arra irányult, hogy az egyes modellek mennyire képesek előre jelezni a vásárlók inaktívvá válását a megfigyelési időszak végére (vagyis, hogy az előrejelzési időszakban nem fog vásárolni). Itt természetesen nem csak az a fontos, hogy szám szerint mennyi a lemorzsolódók száma, hanem az is, hogy pontosan kik azok, akik le fognak morzsolódni. A Kappa statisztikát ugyanis éppen aszerint számoltam, hogy milyen kontingencia táblát kaptam az egyes egyedek besorolása és tényleges hovatartozása alapján($1-1$, $1-0$, $0-1$, $0-0$). Vannak olyan összehasonlító vizsgálatok [Persentili Batislam, Denizel és Filiztekin, 2007; Fader, Hardie és Lee, 2005a] ugyanis, melyek többek között (esetleg csak) csoport szintű összehasonlítást végeztek pl. oly módon, hogy darabszám szerint vetik össze az előrejelzett és tényleges vásárlások számát az egész csoport szintjén. Ezen mutató jó értéke nem feltétlen jelent jó megoldást, hiszen lehetséges, hogy a most vizsgált előrejelzés egyik értéket sem találta el az egyes megfigyelési egységek esetében (ki fog lemorzsolódni és ki nem), mégis csoportszinten jó eredmény születhet (a lemorzsolódó egyének száma közel azonos a ténylegessel). Ha célunk az *egyes* megfigyelési egységek (vásárlók) jövőbeli aktivitásának minél pontosabb előrejelzése, akkor szükséges az egyéni szintű mutatók használata.

4.2.4. A becsült és a tényleges vásárlásszám közötti különbségek összehasonlítása

Ebben az alfejezetben olyan mutató alapján vizsgálom a modelleket, amely az egyes megfigyelési egységekhez tartozó találati pontatlanságok (eltérések) átlagos értékeit adja meg, és ezen értékeket hasonlítom össze. Erre több módszer is adódik, melyek közül az egyik az átlagos abszolút eltérés (MAE = mean absolute error).

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_{\text{pred}} - y_{\text{val}}| \tag{4.39}$$

ahol
n: megfigyelések (objektumok) száma,
y_{pred}: előrejelzett érték (vásárlások száma) a t időtartamra,
y_{val}: tényleges érték (vásárlások száma) a t időtartamra.

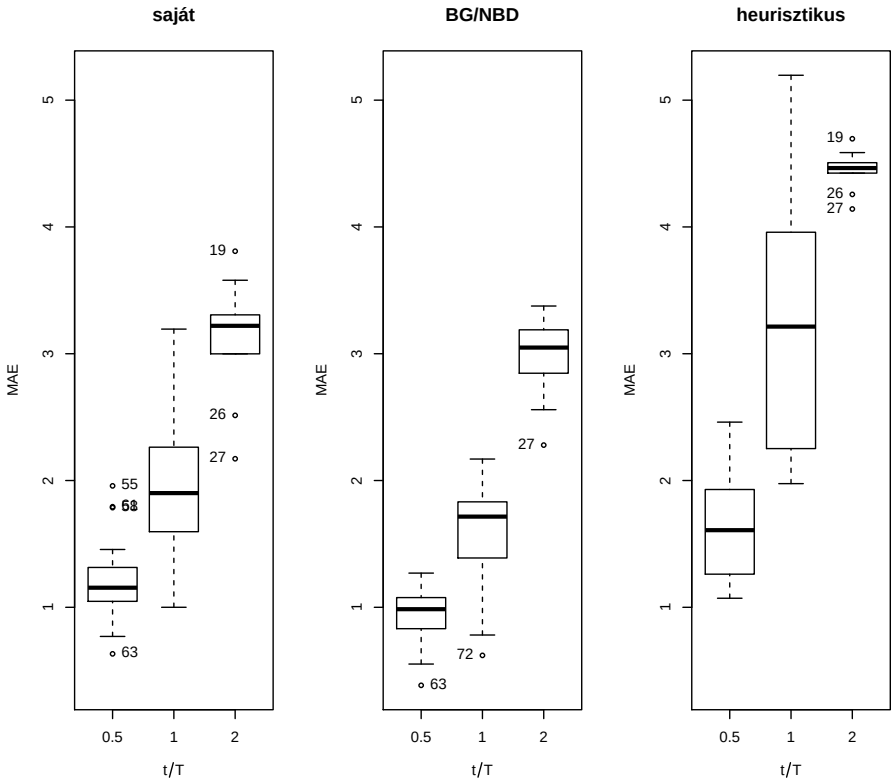
Felmerült még az átlagos abszolút százalékos eltérés ($MAPE$ = mean absolute percentage error) használata is, azonban, ha a tényleges érték (aminek a becslését végzi a modell) nulla (azaz a jövőbeli vásárlások száma 0), akkor

DOI: 10.14751/SZIE.2014.035

a mutató értelmezhetetlen, ami ezen adatbázis esetében is problémát okozott volna.

$$MAPE = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{y_{\text{pred}} - y_{\text{val}}}{y_{\text{val}}} \right| \tag{4.40}$$

A MAE értékeket (A.15. melléklet) ismét boxplot diagramon ábrázoltam, és ismét a t/T arány, mint faktor szerinti csoportokra bontva (9. ábra). Ebben az esetben is megvizsgáltam, hogy az ábrán látható eltérések statisztikailag kimutathatók-e. A saját és a BG/NBD modellt hasonlítottam össze, az áb-



9. ábra. A MAE index értékei különböző t/T arányok mellett a három modell esetében. Forrás: saját szerkesztés.

rából ugyanis meggyőzően kiolvasható, hogy a heurisztikus modell ebben a vizsgálatban sokkal gyengébb eredményt adott a másik kettőhöz képest[10].

A két modell összehasonlításához itt a párosított t-próbát alkalmaztam, melyet mindhárom t/T hányados esetében elvégeztem. Megállapítottam, hogy az első és a második esetben (vagyis, amikor t/T értéke 0,5 és 1) az átlagok különbözősége kimutatható (5%-os szignifikancia szinten), míg a harmadik esetben ($t/T = 2$ esetében), a próba alapján, az átlagok egyezőnek tekinthetők (a számítások eredményét az A.16. melléklet tartalmazza).

Az is megfigyelhető a táblázatban (A.15. melléklet), hogy a BG/NBD mo-

[10] A definícióból látszik, hogy az index nagyobb értéke pontatlanabb eredményt jelent.

DOI: 10.14751/SZIE.2014.035

dell sok esetben nem adott értékelhető eredményt a MAE indexre (ezeket helyettesítettem M-mel), ami azt jelenti, hogy sok esetben nagyon rossz becslést eredményezett. Ha megfigyeljük ezen eseteket, az a közös bennük, hogy mindegyik esetében a t/T hányados értéke 2. Ami azt jelenti, hogy a hosszabb távú előrejelzései bizonytalanok. Pontosabban, ha elfogadható eredményt ad ilyen esetben, akkor az hasonló az általam elkészített modell eredményéhez, de emellett sokszor (27 esetből 18-szor) értékelhetetlen eredményt adott.

Megállapítható tehát, hogy az általam elkészített modell gyengébb eredményeket adott a rövidebb távú előrejelzésekre, míg hosszabb távúra adott – lényegében az előzőekhez hasonló pontosságú – eredményeket nagy biztonsággal tudta előállítani.

4.2.5. A jövőbeli legjobb vásárlók meghatározása

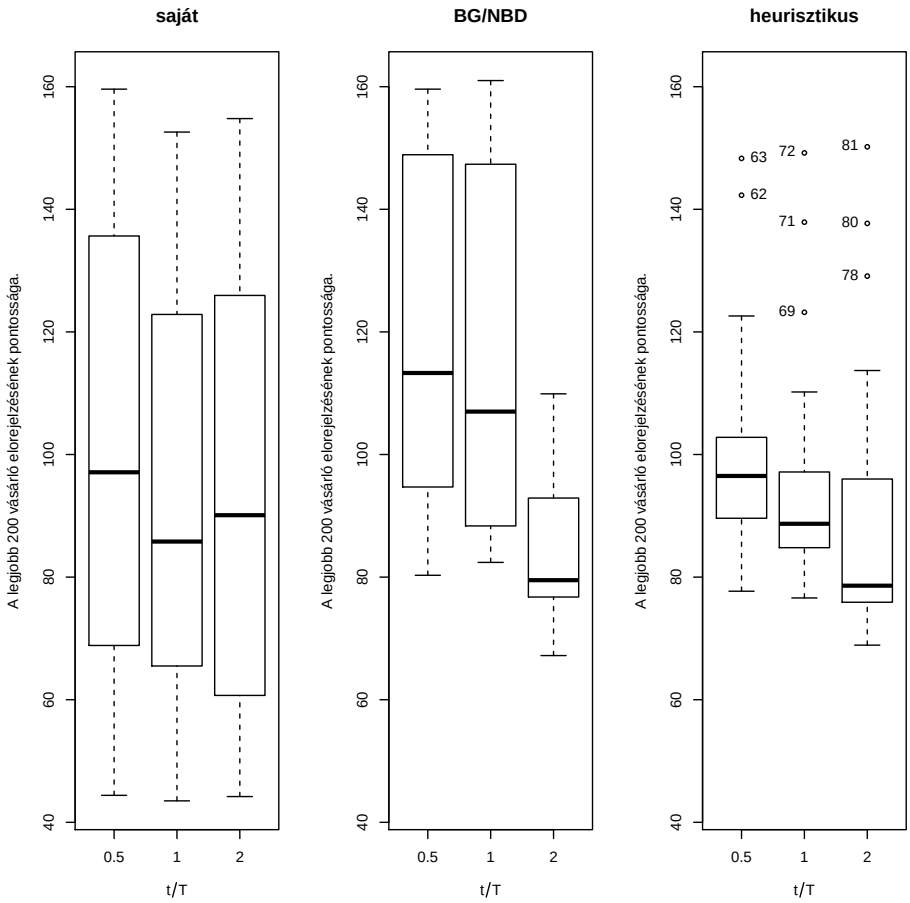
A harmadik összehasonlításban azt elemzem, hogy az egyes modellek mennyire képesek előre jelezni a jövőbeli legjobb 200 vevőt (vagyis a legjobb 20%-ot). Ebben az esetben legjobb alatt azt értem, hogy kik azok, akiknek az előrejelzési időszakban (t) a legtöbb számú vásárlásuk lesz. A vizsgálat jelentőségét az adja, hogy másként kezelendők az egyes vevők aszerint, hogy mennyire jövedelmezőek a cég számára[11]. Ezt támasztja alá pl. Homburg, Droll és Totzek [2008] cikke, melyben többek között az olvasható, hogy számításaik alapján a vevők megkülönböztetése növeli az átlagos jövedelmezőséget. Mivel ebben a modellben a vásárlásra fordított összeg nem szerepel, a legjobb vásárló az lesz, aki a legtöbbször vásárol egy megadott időszak (t) alatt.

Az összegyűjtött adatok (A.17. melléklet) tartalmazzák mindhárom modell esetében azon vevők számát, akiknek az előrejelzése sikeres volt, vagyis az előrejelzés szerint bekerültek a tényleges „top 200”-ba. Az adatokat ismét Boxplot ábrán szemléltetem (10. ábra) úgy, hogy mindhárom modell estében újra 3 csoportot hozok létre a t/T hányados értékei alapján.

Mivel az egyes csoportokban található adatok nem tekinthetők normál eloszlásból származónak (ennek ellenőrzésére ismét a Shapiro-Wilk tesztet alkalmaztam), ezért újból a páros Wilcoxon próbával hasonlítom össze a modelleket. A mediánok különbségét négy esetben lehetett statisztikailag kimutatni: a saját és a BG/NBD modell között a $t/T = 0{,}5$, és a $t/T = 1$ esetében, valamint a heurisztikus és a BG/NBD modell között a $t/T = 0{,}5$, és a $t/T = 1$ esetében (a próbák eredményei az A.18. mellékletben találhatók). Ez azt jelenti, hogy a BG/NBD modell a relatíve rövidebb előrejelzési időszakokra ($t/T = 0{,}5$ és $t/T = 1$) szignifikánsan jobb átlagos eredményt ért el, mint

[11] Mivel a modell nem tartalmazza a vásárlások értékét, ezért ebben a vizsgálatban a „jövedelmezőség” alatt csak a vásárlásszámok nagyságát érthetjük.

DOI: 10.14751/SZIE.2014.035



10. ábra. A legjobb 200 vásárló előrejelzésének találati értékei a t/T arányok mellett a három modell esetében. Forrás: saját szerkesztés.

az általam készített modell. A hosszabb távra történő előrejelzés viszont a saját modellem esetében jobb átlagos eredményt mutat (igaz, ez a különbség statisztikailag nem igazolható, $p = 0{,}1698$).

A vizsgálatnak mégis fontos eredménye a heurisztikus és a valószínűségi modellek összehasonlításából levonható következtetés. Huang [2012] cikkében éppen két ilyen modell előrejelző képességét vizsgálja (nevezetesen a heurisztikus, valamint a Pareto/NBD modelleket hasonlítja össze). Ő is sok mesterséges adatbázis esetében végzi el a számításokat, és megállapítja, hogy „a számítások többségében az egyszerű heurisztika teljesítménye felülmúlja azt a modellt, amely előállította az adatokat az előrejelzéshez". Számításaim azonban ezt az állítást nem támasztják alá. A saját modellem esetében a találatok átlaga nem rosszabb, mint a heurisztikus modellé, a BG/NBD modellé pedig két esetben is jobb.

Huang [2012] kiemeli, hogy a tapasztalati eredményeken alapuló mesterséges adatbázisok tulajdonsága, hogy „a múltban gyakoribb vásárlók valószínűleg a jövőben is gyakoribbak lesznek", és éppen ez a megfigyelés az alapja a heurisztikus eljárásnak is. A szórásokat megfigyelve látható, hogy a heurisztikus

DOI: 10.14751/SZIE.2014.035

kus eljárás robusztusabb, mint a másik kettő, megbízhatóbban hozza a 90/200 találati arány körüli értékeket.

DOI: 10.14751/SZIE.2014.035

5. fejezet

Új és újszerű tudományos eredmények

1. A vásárlói csoportok elkülönítése, szegmentálása kapcsán végzett munkámban tapasztalati és elméleti elemzések segítségével megállapítottam, hogy a TONG és TAN [2009] által kidolgozott $\boldsymbol{m}_{ij}$ osztópont (3.5. egyenlet) meghatározása azokban az esetekben, amikor a két klaszter elemszáma lényegesen különbözik egymástól, nem megfelelő, mert bizonyos esetekben nem olyan területre esik, amely alapján jól szétválasztható lenne a két klaszter. Ennek pedig fontos szerepe van a klaszteren belüli-, és azok közötti sűrűségek vizsgálatával összefüggő részindex ($Dens_{bw}$, 3.1. egyenlet) számításában.

2. Megalkottam az f^{**} függvényt (4.1. egyenlet), amely felelős azért, hogy mennyi megfigyelési egységet tartalmaz a kiválasztott pontok (a klaszterközéppontok ill. az $\boldsymbol{m}_{ij}$ pont) megadott környezete. Az f^{**} függvény segítségével kaptam az S_Dbw_{new} indexből (3.7. egyenlet) az S_Dbw^{**} indexet (4.2. egyenlet). Az indexek elméleti valamint szimulációs összehasonlító vizsgálatának eredményeként kimondható, hogy az általam konstruált index az egymást részben átfedő, egyenlőtlen elemszámú klaszterelrendezés esetén jobb eredményt adott, tehát alkalmasabb a döntéstámogatásra.

3. A BG/NBD modell továbbfejlesztéseként létrehoztam egy új, a vásárlások számának ill. a vásárlók lemorzsolódásának előrejelzésére alkalmas modellt, mely figyelembe veszi a vásárlással kapcsolatos panaszok előfordulását, valamint annak kezelését is, a vásárlások számának vizsgálatán túl. A kialakított saját modellt szimulációs tesztelésnek vetettem alá, melyet az R környezetben írt scriptek segítségével végeztem el, mesterségesen előállított adatbázisok alkalmazásával. Ezen tesztelések alapján megállapítottam, hogy az általam létrehozott modell a vizsgált adatbázisokon a hosszabb távú előrejelzésekben bizonyult pontosabbnak, ám a rövidebb távú előrejelzésekben hasonló vagy kicsit gyengébb eredményt produkált, mint a BG/NBD modell. A fejlesztés tehát a hosszútávú előrejelzések

DOI: 10.14751/SZIE.2014.035

területén jelent előrelépést.

4. A saját és a BG/NBD előrejelző modell eredményeit egy – a fogyasztói magatartást vizsgálatában gyakran használt – heurisztikus modellével összevetve megállapítottam, hogy a valószínűségi modellek előrejelzései fölülmúlják a heurisztikus modellét, főként a vásárlásszámok előrejelzésének esetében. Ezzel a valószínűségi modellek alkalmazhatóságát és az ilyen irányú kutatások fontosságát támasztottam alá.

DOI: 10.14751/SZIE.2014.035

6. fejezet

Következtetések és javaslatok

1. A klaszterszám meghatározását célzó vizsgálataimban azt elemeztem, hogy az eddigi (a vizsgált területen) legjobb megoldás képes-e „szélsőséges" körülmények között, vagyis különféle klaszterelrendezések (pl. egymást részben átfedő ill. egymáshoz közel álló klaszterek) esetében megfelelő támogatást nyújtani a döntéshozónak. Tapasztalatom az volt, hogy a szerzők nem fordítottak figyelmet ennek a vizsgálatára, vagy nem is tűzték ki ezt célul.

   Modellek teljesítményének empirikus vizsgálata esetében a következtetések levonásakor körültekintően kell eljárni, azaz fel kell tüntetni, hogy milyen adatbázison történt a tesztelés, mik az érvényesség keretei.

   Célom olyan adatbázisokon való alkalmazhatóság volt, amelyek nem teljesen szeparáltak, azonban az átfedés olyan mértékű legyen, hogy a klaszterező eljárások különbséget tudjanak tenni a két klaszter között, ne tekintse őket egynek (abban az esetben ugyanis a klaszterező eljárás több klaszterre bontás esetén szétvág(hat)ja ugyan ezt a képződményt, de nem feltétlenül helyesen). A mindennapi gyakorlatban előforduló adatbázisok ugyanis általában nem teljesen szeparált csoportokat tartalmaznak.

2. Az általam megalkotott index a vizsgált adatbázisokon jobb eredményt adott, mint az eddigi legjobbnak ítélt index, méghozzá a valósághoz közelebb álló klaszterelrendezések[1] esetében. Az eredmény azonban függ a kiválasztott klaszterező algoritmustól is. Dolgozatomban két különböző algoritmussal dolgoztam, és a legtöbb esetben mindkettő esetében ott volt a megoldások között a helyes besorolás is. Az én vizsgálatom arra irányult, hogy ezen megoldások közül ki tudjuk választani a valósághoz legközelebb állót. Ha azonban a klaszterező eljárás megoldásai között nincs ott a „tényleges" megoldás, akkor az általam megalkotott index ki fog ugyan választani egyet, azonban az nem lehet a tényleges, esetleg csak a választ-

[1] Az adatbázisok létrehozásakor tértem ki ennek tárgyalására.

DOI: 10.14751/SZIE.2014.035

hatók közül a „ténylegeshez legközelebb álló" megoldás (azonban ennek vizsgálatára dolgozatomban nem tértem ki).

Ebben a vizsgálatban kétváltozós adatbázissal dolgoztam, éppen a vizuális ellenőrizhetőség kedvéért (a megfigyelési egységek egy sík pontjaival azonosíthatók). Ha azonban a probléma három vagy több változós, az index meghatározása akkor is lehetséges, az általánosítás tehát megoldott (azonban a szemléletes megjelenítés nehezen vagy egyáltalán nem oldható meg).
Mivel az index számítása páros összehasonlításokon alapszik (klaszterpárok vizsgálata), ezért nagyon sok klaszter esetében a számításigény megnőne. Dolgozatomban a marketingkutatás területén való alkalmazást céloztam meg, ahol a nagyon sok klaszterből álló adatbázisok előfordulása nem jellemző, ezért ennek a problémának kezelésére nem tértem ki.

3. A BG/NBD és az abból fejlesztett saját modell összehasonlításából látszik, hogy az új modellbe bevont újabb változók csak részben eredményeztek teljesítményjavulást. Mint a dolgozat elején jeleztem, kérdéses, hogy újabb változók bevonása hasznos lesz-e, mert ugyan a több adat lehetőséget ad a valóság jobb megismerésére, ugyanakkor a modell bonyolódik, a meghatározandó paraméterek száma növekszik. Sok paraméter bevonása esetén a sok hatás eredőjeként létrejött eredményekből kell visszakövetkeztetni a hatások leírására használt eloszlások paramétereire, majd ezen paraméterek (eloszlások) ismeretében modellezni a jövőt. Azonban a sok eloszlás eredőjeként kialakult eredményből visszafejteni az egyes eloszlásokat nehezebb, mint kevés eloszlás esetén.

   Az adatbázisok előállítása a saját modell elmélete alapján történt, tehát feltételezhető volt, hogy a saját modell ezt jobban felismerve pontosabb előrejelzést ad. Nem így történt, tehát egy egyszerűbb modell lényegében ugyanolyan eredményes volt az előrejelzésben (rövidebb távon), annak ellenére, hogy kevesebb információt használt fel.

   Másrészt, a panaszok számát próbáltam reális tartományban tartani. Ennek kis értéke eredményezhette azt, hogy nem volt jelentős hatása az eredményre, vagyis az enélkül dolgozó BG/NBD modell hasonló eredményre vezetett. Ezért vizsgálat alá vontam a két modellt abból a célból, hogy a panaszok számának változása (az adatbázisok előállításához használt paraméterek módosítása révén) másként hat-e a két modell pontosságára. Ilyen összefüggés nem volt kimutatható.

4. A heurisztikus modell ill. valószínűségi modell körüli viták hatására elvégzett vizsgálatomban meglepően jól szerepelt a heurisztikus modell. Mivel

DOI: 10.14751/SZIE.2014.035

a számításokat 81 különböző adatbázison is elvégeztem (ezen belül mindegyik modellt 10-szer lefuttattam), a tudományos eredmények alfejezetben megfogalmazott állítás empirikusan lett megalapozva. Természetesen kérdés maradt, hogy van-e annyi plusz hozadéka a valószínűségi modellnek, amiért érdemes használni. A két modell között nagyon nagy a különbség (elvi nehézségek, gyakorlati nehézségek). Mivel az általam kidolgozott modell a vásárlások értékével nem foglakozott (csak a vásárlások darabszámával), így erre a kérdésre jelen dolgozat keretein belül nem lehet válaszolni. Az azonban biztos, hogy az empirikus vizsgálatok egyik része az egyik, másik része a másik modellt hozza ki győztesként. Mint látható volt, a valószínűségi modellek szórása két vizsgálat esetében is nagyobb volt, mint a heurisztikus modellé, így ha valaki egy adatbázis esetében lefuttatja azt, az eredmény tág határok között mozoghat. Egy ilyen vizsgálatból azonban messzemenő következtetést nem szabad levonni.

Ha a kutatónak egy adatbázisa van, és nem bízik eléggé a módszerben, akkor megoldható, hogy az egy adatbázisból többet csináljon (pl. bagging[2]), és ezen adatbázisok mindegyikén végrehajtja a számításokat, majd a kapott eredményeket értékelve hozhat döntést.

A 81 adatbázis mindegyike 1000 vásárló adatait tartalmazta. A mintát elegendően nagynak találtam ahhoz, hogy az eredményeket elfogadjam. Lehetett volna nagyobb objektumszámmal is dolgozni, de az általam létrehozott script (A.13. melléklet) így is túl lassan futott le és nagyon sok memóriát igényelt. A script optimalizálásával ezen lehetett volna módosítani, de jelen dolgozat szempontjából nem tartottam ezt lényegesnek.

[2]Véletlenszerű kiválasztással újabb adatbázisokat állítson elő a meglevő adatbázisból.

DOI: 10.14751/SZIE.2014.035

DOI: 10.14751/SZIE.2014.035

7. fejezet

Összefoglalás

A dolgozat célja a marketingkutatás során alkalmazott néhány kvantitatív módszer vizsgálata, ezek alkalmazási eredményeinek szakirodalmi áttekintése, valamint továbbfejlesztési lehetőségeinek keresése.

Elsőként olyan, a klaszteranalízissel kapcsolatos eljárást vizsgáltam, melynek segítségével a felkínált megoldások (a különböző klaszterszámokhoz tartozó csoportosítások) közül választható ki a legjobb. Ennek a megoldására többféle módszer is létezik. Kutatómunkám során az elemsűrűségekkel kapcsolatos algoritmusokat vizsgáltam, és azok közül választottam ki egyet (az ún. S_Dbw_{new} indexet), és ennek módosítását készítettem el. Az eljárás lényege, hogy a klaszterközéppontok, valamint a klaszterközéppontokat összekötő szakaszok egy megadott pontja körüli, egy előre megadott méretű tartományban található elemszámokból számított index segítségével tekintünk két klasztert különbözőnek, vagy egy klaszternek. A probléma jelentősége az, hogy ha az adatbázisunk 3-nál több változót tartalmaz, akkor nincs lehetőségünk vizuálisan ellenőrzést végezni, hanem valamilyen számítási módszerre hagyatkozhatunk.

A szakirodalomban található ilyen jellegű indexek vizsgálata során arra a megfigyelésre jutottam, hogy az eddig legjobbnak ítélt index is csak jól szeparált klaszterek esetében adott pontos válaszokat, kevésbé szeparált klaszterek esetén azonban már hibás döntésekhez vezetett. Ezért először az eredeti index szerkezetét vizsgáltam, mely két összetevőből áll. Az egyik a fent említett elemszámokból képzett részindex ($Dens_{bw}$), a másik pedig a klaszterek és a összes megfigyelési egység szórásából számított részindex ($Scat$). E két részindex konvex lineáris kombinációjaként áll elő az az index, mely alapján a klaszterszámokról döntést hozhatunk. Dolgozatomban az első részindex módosítását készítettem el, aminek segítségével sikerült olyan eredményre jutni, mely már az előbb hiányolt esetekben is jobb döntést eredményezett.

A módosítás – ugyanúgy, mint az eredeti munkák esetében – kétváltozós szimulált adatbázisokon lett tesztelve, ám ez nem jelent megszorítást, leszűkí-

DOI: 10.14751/SZIE.2014.035

tést, hiszen az index bármilyen változószám esetén számolható. A kétváltozós tesztelés az egyszerűbb ellenőrizhetőség miatt volt célszerű. Az adatbázisok, melyeken az összehasonlításokat végeztem, szemléletesen mutatják a különbséget azon adatbázisokhoz képest, melyeken az eddigi indexeket tesztelték.

A kapott elméleti és empirikus eredmények alapján kimondható, hogy az általam alkalmazott módosítások szélesebb körben teszik alkalmazhatóvá ezt az indexet.

A másik kutatás a vásárlók múltbeli vásárlási szokásainak megfigyelése által a jövőben várható vásárlási minták (megadott időszak alatt bekövetkező vásárlások számának) előrejelzésével foglalkozik. Erre már sokféle módszer létezik, melyek közül a valószínűségi modellek segítségével történő előrejelzést elemeztem. Ezen módszerek lényege, hogy a rendelkezésre álló adatbázist valószínűségeloszlások segítségével próbáljuk leírni, és keressük ezen eloszlások azon paramétereit, melyek esetében legnagyobb annak a valószínűsége, hogy az adott adatbázis a kapott modellből származhat (mint véletlen minta). Ezek között is figyelemmel kísértem egy modell (az ún. Pareto/NBD) születését és annak továbbfejlesztési lehetőségeit, és ezen továbbfejlesztések egy újabb módosítását dolgoztam ki, majd teszteltem véletlenszerűen kialakított adatbázisokon (szimulációs kísérletek segítségével).

Az összehasonlításba belevontam még egy ún. Heurisztikus modellt is, mely az egyik legáltalánosabban használt módszer jövőbeli viselkedések leírására a mindennapi gyakorlatban. Ebben az esetben a szakértők hoznak meg szabályokat, hogy a ténylegesen megfigyelt adatokból milyen előrejelzéseket lehet tenni. Természetesen ennek alapja is a múltbeli adataik vizsgálata, hozzátéve a szakértői tudásukat, intuíciójukat.

A vizsgált BG/NBD, valamint ennek módosításával született saját modell feltevései: a vásárlások között eltelt idő exponenciális eloszlást követ (melynek paramétere λ), minden vásárlás után a vásárló p valószínűséggel lemorzsolódik. Mindkét paraméter vásárlóról - vásárlóra változik, λ gamma eloszlás, p geometriai eloszlás szerint. Először a múltbeli adatok alapján ezen eloszlások paramétereinek értékét számolják ki, majd ezen paraméterek ismeretében lehet elvégezni az előrejelzéseket egy megadott időtartamra. A saját modellem elődjéhez képest több inputot vesz figyelembe: vizsgálja, hogy a vásárlással kapcsolatban felmerült-e probléma (panasz), ha igen, akkor az hogyan végződött (kezelt ill. nem kezelt). A kialakított modellem szerint ezen esetekben különböző lesz a lemorzsolódás valószínűsége. Ezen valószínűségi változók is különböznek a vásárlók esetében, melyek változékonyságát béta eloszlással írtam le. Ezáltal több eloszlás szükséges a leíráshoz, aminek eredménye az, hogy sokkal több paraméter értékét kell meghatározni a múltbeli adatok alapján.

DOI: 10.14751/SZIE.2014.035

A szimulált 81 adatbázison való tesztelés alapján azt a következtetést lehetett levonni, hogy a módosított modell a hosszabb távú előrejelzések esetén eredményezett lényeges javulást a BG/NBD modellhez képest, míg a rövidebb távú előrejelzésekben nem tudott pontosabb eredményt adni.

A Heurisztikus modellel való összehasonlításból viszont egyértelműen kimutatható a valószínűségi modellek hatékonysága, főként az előrejelzett vásárlásszámokat tekintve. Több alkalommal is megkérdőjelezték már a heurisztikus modellek helyett alkalmazott tudományos modellek létjogosultságát. Ebben a vizsgálatban a sok adatbázison végzett empirikus eredmények ennek ellenkezőjét mutatták. Természetesen a két módszer alkalmazása nem egyforma nehézségű. A valószínűségi modell csak akkor alkalmazható széleskörűen, ha egy felhasználóbarát szoftver formájában érhető el. Ezzel szemben, felhasználói szinten, a heurisztikus modell könnyen alkalmazható, és a módosítások lehetőségét is magában foglalja az előbbivel szemben.

DOI: 10.14751/SZIE.2014.035

DOI: 10.14751/SZIE.2014.035

8. fejezet

Summary

The aim of the study is the examination of some quantitative marketing research methods and the literature review of their results of the applications as well as opportunities for improvement.

First, a procedure related to the cluster analysis has been examined, which can be used to select the best solution of the possible different cluster groupings. There are several methods to resolve this. In the first part of the research algorithms related to the densities (of elements) have been tested, and one of them (the so-called S_Dbw_{new} index) has been chosen, analyzed and modified. The essence of the method is to numerate the observation units around the cluster centers and around a given point between two cluster centers (in a range with predefined size). From these numbers an index is calculated, and with the help of this index the two clusters are considered as different clusters, or not. The significance of the problem is that if the database contains more than three variables, there is no opportunity to carry out a visual inspection, but rely on a calculation method.

It is concluded by the examination of such indices in the literature that the index has been considered the best also have led to accurate answers given only for well separated clusters, but in the case of less separated clusters it led to incorrect decisions. Therefore, first, the structure of the original index has been examined, which consists of two components. The first sub-index is qualified from the (above-mentioned) number of observation units around the predefined points (the so-called $Dens_{bw}$) and the second one is calculated from the standard deviations of the clusters and the standard deviation of all the observation units ($Scat$). The whole index is derived as the convex linear combination of these two sub-indices, and on the base of this a decision can be made on the number of clusters. In this thesis a modification of the first sub-index has been created, which helps to get a better result in cases when the clusters are not well separated.

The amendment – in the same way as for the original work – has been tested

DOI: 10.14751/SZIE.2014.035

with simulated bivariate databases, but this is not a restriction because for any number of variables the index can be calculated. The bivariate testing has been appropriate due to the simple verifiability. The databases on which the comparisons have been carried out, clearly show the difference to the databases on which the index has been tested previously.

Because of the theoretical and the empirical results can be stated that the modification of the index makes it more applicable.

The second study is concerned with the forecast of customers' expected shopping patterns in the future (number of purchases occurring during the specified time) by the observation of the past purchase habits. A wide variety of methods exist for this purpose, and in this thesis the probabilistic prediction models have been analyzed. The essence of these methods is that the available database has been modeled by probability distributions (i.e. determination of the parameters of these distributions) such that let the probability – to choose randomly the given database from this probability distribution – be maximal. Such a model has been selected (the so-called Pareto/NBD) with its further development and in this thesis a new modification of this models has been developed and tested. Testing has been done by randomly selected databases (simulation experiments).

The (so-called) Heuristic model has also been taken into the comparisons of the models, which is one of the most commonly used methods to describe the future behaviour of customers in everyday practice. In this case, the rules are made by experts which determine the predictions can be made by the actual observed data. Of course, this analysis is based on historical data as well, adding the expert knowledge and intuition to the facts.

The common assumptions of the analyzed BG/NBD model and the own model (generated by the modification of BG/NBD model) are the following: the time between purchases follows an exponential distribution (with parameter λ), the customer drop out after every purchase with probability p. Both parameters vary from customer to customer according to a gamma distribution and geometric distribution by order of succession. First, the values of the parameters of these distributions are calculated on the basis of data of previous period, and then knowing these parameters, forecasts can be done for a specified future periods. The own new model considers more inputs in comparison with its predecessor: check for problems incurred about the purchase (complaint), if so, how it ended (resolved or not). In the case of the developed model the probabilities of dropping out are different in these cases. Furthermore, these probabilities are different for the customers, so they can be characterized by random variables with beta distribution (with different para-

DOI: 10.14751/SZIE.2014.035

meters of course). Since more distributions are required, so more parameters have to be determined by previous data.

It has been concluded – based on the results of 81 simulated tests – that the modified (new) model has resulted significantly better forecasts for long-term projections in comparison with the BG / NBD model, while the short-term forecasts could not be more accurate than the short-term forecasts of BG/NBD model.

Compared with the heuristic model, however, clearly demonstrated the effectiveness of probabilistic models, especially considering the projected number of purchases. On several occasions has been queried the legitimacy of probability models compared to the heuristic models. In this study a lot of simulations on different databases have shown just the opposite. Of course, the difficulty of the usage the two methods are not the same. The probability model can only be applied widely, as if it is the part of a user-friendly software. In contrast, at the user level, the heuristic model is easier to apply, and also includes the possibility of changes compared to the probability model.

DOI: 10.14751/SZIE.2014.035

DOI: 10.14751/SZIE.2014.035

Mellékletek

DOI: 10.14751/SZIE.2014.035

Irodalomjegyzék

Agresti A. (2010): *Analysis of Ordinal Categorical Data*, John Wiley & Sons, 396 pp.

Baumgartner H., Pieters R. (2003): The Structural Influence of Marketing Journals: A Citation Analysis of the Discipline and Its Subareas over Time, *The Journal of Marketing*, vol. 67, pp. 123–139.

Berger P.D., Nasr N.I. (1998): Customer lifetime value: Marketing models and applications, *Journal of Interactive Marketing*, vol. 12, pp. 17–30.

Berry M.J., Linoff G.S. (2004): *Data Mining Techniques For Marketing, Sales, and Customer Relationship Management*, Wiley Publishing, Inc., 2nd ed., 643 pp.

Blattberg R.C., Getz G., Thomas J.S. (2001): *Customer Equity: Building and Managing Relationships As Valuable Assets*, Harvard Business Review Press, 1st ed., 228 pp.

Bodon F. (2010): Adatbányászati algoritmusok, URL `http://www.cs.bme.hu/~bodon/magyar/adatbanyaszat/tanulmany/adatbanyaszat.pdf`, (Letöltve: 2012.01.15.).

Boriah S., Chandola V., Kumar V. (2008): Similarity Measures for Categorical Data: A Comparative Evaluation, in: *SIAM Data Mining Conference*, Atlanta, Georgia, pp. 243–254.

Breiman L. (1996): Bagging Predictors, *Machine Learning*, vol. 24, pp. 123–140.

Caples J. (1932): *Tested Advertising Methods*, Harper & Row Pub., 276 pp.

Cha S. (2007): Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions, *International Journal of Mathematical Models and Methods in Applied Sciences*, vol. 1, pp. 300–307.

Chakrapani C. (2010): Introduction to Multivariate Analysis, in: *Applied Research Methods All-Tutorial Training (AMA)*, Philadelphia,

DOI: 10.14751/SZIE.2014.035

April 19-21, URL http://www.marketingpower.com/Calendar/Pages/AppliedResearchMethods2010.aspx.

Choi S., Cha S., Tappert C. (2010): A Survey of Binary Similarity and Distance Measures, *Journal on Systemics, Cybernetics and Informatics*, vol. 8, pp. 43–48.

Cohen J. (1960): A Coefficient of Agreement for Nominal Scales, *Educational and Psychological Measurement*, vol. 20, pp. 37–46.

Crone S.F., Lessmann S., Stahlbock R. (2006): The impact of preprocessing on data mining: An evaluation of classifier sensitivity in direct marketing, *European Journal of Operational Research*, vol. 173, pp. 781–800.

Ehrenberg A.S.C. (1959): The Pattern of Consumer Purchases, *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 8, pp. 26–41.

Ester M., Kriegel H.P., Sander J., Xu. X. (1996): A density based algorithm for discovering clusters in large spatial databases with noise., in: *Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining*, Portland, OR, pp. 226–231.

Everitt B.S., Landau S., Leese M., Stahl D. (2011): *Cluster Analysis*, Wiley, 5th ed., 346 pp.

Fader P., Hardie B., Lee K.L. (2006): More than Meets the Eye, *Marketing Research*, vol. 18, pp. 9–14.

Fader P.S., Hardie B.G. (2009): Probability Models for Customer-Base Analysis, *Journal of Interactive Marketing*, vol. 23, pp. 61 – 69.

Fader P.S., Hardie B.G.S., Lee K.L. (2005a): „Counting Your Customers" the Easy Way: An Alternative to the Pareto/NBD Model, *Marketing Science*, vol. 24, pp. 275–284.

Fader P.S., Hardie B.G.S., Lee K.L. (2005b): RFM and CLV: Using Iso-Value Curves for Customer Base Analysis, *Journal of Marketing Research*, vol. 42, pp. 415–430.

Füstös L., Kovács E. (1989): *A számítógépes adatelemzés statisztikai módszerei*, Tankönyvkiadó, Budapest, 380 pp.

Füstös L., Kovács E., Meszéna G., Simonné N.M. (2004): *Alakfelismerés*, Új Mandátum Könyvkiadó, 644 pp.

Goldstein D.G., Gigerenzer G. (2009): Fast and frugal forecasting, *International Journal of Forecasting*, vol. 25, pp. 760 – 772.

DOI: 10.14751/SZIE.2014.035

Gupta S., Hanssens D., Hardie B., Kahn W., Kumar V., Lin N., Ravishanker N., Sriram S. (2006): Modeling Customer Lifetime Value, *Journal of Service Research*, vol. 9, pp. 139–155.

Gupta S., Lehmann D.R., Stuart J.A. (2004): Valuing Customers, *Journal of Marketing Research*, vol. 41, pp. 7–18.

Hair J.F., Black W.C., Babin B.J., Anderson R.E. (2009): *Multivariate Data Analysis*, Prentice Hall, 7th ed., 816 pp.

Hajdu O. (2003): *Többváltozós statisztikai számítások*, Központi Statisztikai Hivatal, 457 pp.

Halkidi M., Vazirgiannis M. (2001): Clustering validity assessment: finding the optimal partitioning of a data set, in: *ICDM 2001, Proceedings IEEE International Conference on Data Mining*, IEEE, pp. 187–194.

Hallberg G., Ogilvy D. (1995): *All Consumers Are Not Created Equal: The Differential Marketing Strategy for Brand Loyalty and Profits*, Wiley, 1st ed., 336 pp.

Hartigan J.A., Wong M.A. (1979): A K-means clustering algorithm, *Applied Statistics*, vol. 28, pp. 100–108.

Hoek J., Gendall P., Esslemont D. (1996): Market segmentation: A search for the Holy Grail?, *Journal of Marketing Practice: Applied Marketing Science*, vol. 2, pp. 25–34.

Hogan J.E., Lemon K.N., Libai B. (2003): What Is the True Value of a Lost Customer?, *Journal of Service Research*, vol. 5, pp. 196–208.

Homburg C., Droll M., Totzek D. (2008): Customer Prioritization: Does It Pay Off, and How Should It Be Implemented?, *Journal of Marketing*, vol. 72, pp. 110–130.

Hopkins C.C. (2010): *Scientific Advertising*, Cosimo, Inc., 90 pp.

Howard J.A., Sheth J.N. (1969): *The theory of buyer behavior*, Wiley, 458 pp.

Huang C.Y. (2012): To model, or not to model: Forecasting for customer prioritization, *International Journal of Forecasting*, vol. 28, pp. 497–506.

Hussey M., Hooley G. (1995): The diffusion of quantitative methods into marketing management, *Journal of Marketing Practice: Applied Marketing Science*, vol. 1, pp. 13–31.

DOI: 10.14751/SZIE.2014.035

Jain A.K., Dubes R.C. (1988): *Algorithms for Clustering Data*, Prentice Hall College Div, 1st ed., 304 pp.

Jain A.K., Murty M.N., Flynn P.J. (1999): Data clustering: a review, *ACM Computing Surveys*, vol. 31, pp. 264–323.

Jay J.J., Eblen J.D., Zhang Y., Benson M., Perkins A.D., Saxton A.M., Voy B.H., Chesler E.J., Langston M.A. (2012): A Systematic Comparison of Genome Scale Clustering Algorithms, in: J. Chen, J. Wang, A. Zelikovsky (eds.) *Bioinformatics Research and Applications*, vol. 6674, Springer Berlin Heidelberg, pp. 416–427.

Kaufman L., Rousseeuw P.J. (2005): *Finding groups in data: an introduction to cluster analysis*, Wiley, Hoboken, N.J.

Kim Y., Lee S. (2003): A Clustering Validity Assessment Index, in: K.Y. Whang, J. Jeon, K. Shim, J. Srivastava (eds.) *Advances in Knowledge Discovery and Data Mining*, vol. 2637 of *Lecture Notes in Computer Science*, Springer Berlin / Heidelberg, pp. 562–562.

Kleinberg J. (2003): *An Impossibility Theorem for Clustering*, MIT Press, Cambridge, MA, pp. 446–453.

Kotler P. (1967): *Marketing management : analysis, planning, and control*, Prentice-Hall, Englewood Cliffs, NJ, 628 pp.

Kotler P., Armstrong G. (2010): *Principles of Marketing*, Pearson Education, 637 pp.

Kotler P., Keller K.L. (2012): *Marketingmenedzsment*, Akadémiai Kiadó, 893 pp.

Kovács E., Füstös L., Meszéna G. (2007): *Alakfelismerés: Sokváltozós statisztikai módszerek*, Új Mandátum Könyvkiadó, 660 pp.

Legány C., Juhász S., Babos A. (2006): Cluster validity measurement techniques, in: *Proceedings of the 5th WSEAS International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases*, World Scientific and Engineering Academy and Society (WSEAS), Stevens Point, Wisconsin, pp. 388–393.

Lemmens A., Croux C. (2006): Bagging and boosting classification trees to predict churn, *Journal of Marketing Research*, vol. 18, pp. 276–286.

Liu Y., Li Z., Xiong H., Gao X., Wu J. (2010): Understanding of Internal Clustering Validation Measures, in: *Proceedings of the 2010 IEEE Inter-*

DOI: 10.14751/SZIE.2014.035

national Conference on Data Mining, ICDM '10, IEEE Computer Society, Washington, DC, USA, pp. 911–916.

Ma S.H., Liu J.L. (2007): The MCMC Approach for Solving the Pareto/NBD Model and Possible Extensions, in: *Third International Conference on Natural Computation*, vol. 2, pp. 505–512.

Maex D. (2009): Math Marketing: The New Landscape of Marketing Analytics, URL `http://www.wpp.com/wpp/marketing/marketing/math-marketing.htm`, (Letöltve: 2012.01.15.).

Magee J.F. (1960): Operations Research in Making Marketing Decisions, *The Journal of Marketing*, vol. 25, pp. 18–23.

Mahalanobis P.C. (1936): On the generalised distance in statistics, in: *Proceedings National Institute of Science*, vol. 2, pp. 49–55.

Malhotra N.K. (2002): *Marketingkutatás*, KJK-KERSZÖV Jogi és Üzleti Kiadó Kft., Budapest, 904 pp.

Malthouse E.C., Blattberg R.C. (2005): Can we predict customer lifetime value?, *Journal of Interactive Marketing*, vol. 19, pp. 2–16.

Mao J., Jain A.K. (1996): A self-organizing network for hyperellipsoidal clustering (HEC), *IEEE Transactions on Neural Networks*, vol. 7, pp. 16–29.

McCarty J.A., Hastak M. (2007): Segmentation approaches in datamining: A comparison of RFM, CHAID, and logistic regression, *Journal of Business Research*, vol. 60, pp. 656 – 662.

Miglautsch J.R. (2000): Thoughts on RFM scoring, *The Journal of Database Marketing*, vol. 8, pp. 67–72.

Moutinho L., Meidan A. (2003): Quantitative methods in marketing, in: M.J. Baker (ed.) *The Marketing Book*, Butterworth-Heinemann, 5th ed., pp. 197–245.

Mulhern M.G. (2010): Market Segmentation Basics, in: *Applied Research Methods All-Tutorial Training (AMA)*, Philadelphia, April 19-21, URL `http://www.marketingpower.com/Calendar/Pages/AppliedResearchMethods2010.aspx`.

Nakata C., Huang Y. (2005): Progress and promise: the last decade of international marketing research, *Journal of Business Research*, vol. 58, pp. 611 – 618.

DOI: 10.14751/SZIE.2014.035

Ngai E., Xiu L., Chau D. (2009): Application of data mining techniques in customer relationship management: A literature review and classification, *Expert Systems with Applications*, vol. 36, pp. 2592–2602.

Persentili Batislam E., Denizel M., Filiztekin A. (2007): Empirical validation and comparison of models for customer base analysis, *International Journal of Research In Marketing*, vol. 24, pp. 201–209.

Pfeifer P.E., Haskins M.E., Conroy R.M. (2005): Customer Lifetime Value, Customer Profitability, and the Treatment of Acquisition Spending, *Journal of Managerial Issues*, vol. XVII, pp. 11–25.

R Core Team (2013): R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, URL `http://www.R-project.org/`.

Razali N.M., Wah Y.B. (2011): Power Comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling Tests, *Journal of Statistical Modeling and Analytics*, vol. 2, pp. 21–33.

Reichheld F.F., Teal T. (2001): *The Loyalty Effect: The Hidden Force Behind Growth, Profits, and Lasting Value*, Harvard Business School Press, 352 pp.

Schmittlein D.C., Morrison D.G., Colombo R. (1987): Counting Your Customers: Who-Are They and What Will They Do Next?, *Management Science*, vol. 33, pp. 1–24.

Schmittlein D.C., Peterson R.A. (1994): Customer Base Analysis: An Industrial Purchase Process Application, *Marketing Science*, vol. 13, pp. 41–67.

Schwarz G. (1978): Estimating the Dimension of a Model, *Annals of Statistics*, vol. 6, pp. 461–464.

Seo D., Ranganathan C., Babad Y. (2008): Two-level model of customer retention in the US mobile telecommunications service market, *Telecommunications Policy*, vol. 32, pp. 182–196.

Sharma S., Kumar A. (2006): Cluster Analysis and Factor Analysis, in: R. Grover, M. Vriens (eds.) *The Handbook of Marketing Research*, SAGE Publications, Inc, pp. 365–393.

Shawand E.H., Jones D. (2005): A history of schools of marketing thought, *Marketing Theory*, vol. 5, pp. 239 –281.

DOI: 10.14751/SZIE.2014.035

Simon J. (2006): A klaszterelemzés alkalmazási lehetőségei a marketingkutatásban, *Statisztikai Szemle*, vol. 7, pp. 627–650.

Sneath P.H. (2005): Numerical Taxonomy, in: D.J. Brenner, N.R. Krieg, J.T. Staley, G.M. Garrity (eds.) *Bergey's Manual of Systematic Bacteriology*, Springer US, Boston, MA, pp. 39–42.

Swift R.S. (2000): *Accelerating Customer Relationships: Using CRM and Relationship Technologies*, Prentice Hall, 1st ed., 512 pp.

Theodoridis S., Koutroumbas K. (2003): *Pattern recognition*, Academic Press. , 2nd ed., 689 pp.

Tong J., Tan H. (2009): Clustering validity based on the improved S-Dbw* index, *Journal of Electronics (China)*, vol. 26, pp. 258–264.

Tsiptsis K., Chorianopoulos A. (2010): *Data Mining Techniques in CRM: Inside Customer Segmentation*, Wiley, 1st ed., 372 pp.

van Oest R., Knox G. (2011): Extending the BG/NBD: A simple model of purchases and complaints, *International Journal of Research in Marketing*, vol. 28, pp. 30 – 37.

Ward J. H. J. (1963): Hierarchical Grouping to Optimize an Objective Function, *Journal of the American Statistical Association*, vol. 58, pp. 236–244.

Wübben M., Wangenheim F.V. (2008): Instant Customer Base Analysis: Managerial Heuristics Often "Get It Right", *Journal of Marketing*, vol. 72, pp. 82–93.

Wilkie W.L., Moore E.S. (2003): Scholarly Research in Marketing: Exploring the "4 Eras" of Thought Development, *Journal of Public Policy & Marketing*, vol. 22, pp. 116–146.

Xu R., Wunsch D.C. (2008): *Clustering*, John Wiley & Sons, 400 pp.

Yang M. (1993): A survey of fuzzy clustering, *Mathematical and Computer Modelling*, vol. 18, pp. 1–16.

DOI: 10.14751/SZIE.2014.035

Ábrák jegyzéke

DOI: 10.14751/SZIE.2014.035

Táblázatok jegyzéke

DOI: 10.14751/SZIE.2014.035

Jelölések, rövidítések jegyzéke

$\forall$	„minden” (univerzális kvantor)
$(a, b)^T$	vektor transzponáltja
AC	Acquisition Cost
BG/NBD	béta-geometriai/negatív binomiális eloszlásokon alapuló valószínűségi modell
$\mathrm{B}(a, b)$	Béta függvény
CE	Customer Equity
CLV	Customer Lifetime Value
CRM	Customer Relationship Management
f	folytonos valószínűségi változó sűrűségfüggvénye
${}_2F_1(a, b; c; z)$	Hipergeometriai függvény
$\Gamma(s)$	Gamma függvény
κ	Cohen-féle kappa mutató
L	Likelihood függvény
lim	határérték
LL	Log-likelihood függvény
$\mathrm{M}(\xi)$	a ξ valószínűségi változó várható értéke
$\mathbb{N}$	Természetes számok halmaza
$\boldsymbol{m}_{ij}$	az i-edik és j-edik klaszter középpontját elválasztó osztópont
$m^{(p)}$	az $\boldsymbol{m}$ pont p-edik komponense
$\mathrm{P}(A)$	az A esemény bekövetkezési valószínűsége
$\mathrm{P}(A\|B)$	az A esemény bekövetkezési valószínűsége, a B esemény bekövetkezése mellett
SVM	Support Vector Machine
T	a megfigyelési időszak időtartama
$\Phi(x)$	Standard normál eloszlású valószínűségi változó eloszlásfüggvénye
$\mathbb{R}$	Valós számok halmaza
σ	szórás
σ^2	variancia

DOI: 10.14751/SZIE.2014.035

t_x	az utolsó vásárlás időpontja a megfigyelési időszakban
x	vásárlások száma a megfigyelési időszakban
$\boldsymbol{x}_i$	az i-edik megfigyelési egység változóinak értékét tartalmazó vektor
$X(t)$	a t időpontig bekövetkező vásárlások száma
$Y(t)$	vásárlások számának becslése a megfigyelési időszakon túli t időtartamra
$\boldsymbol{v}_i$	az i-edik klaszter középpontja

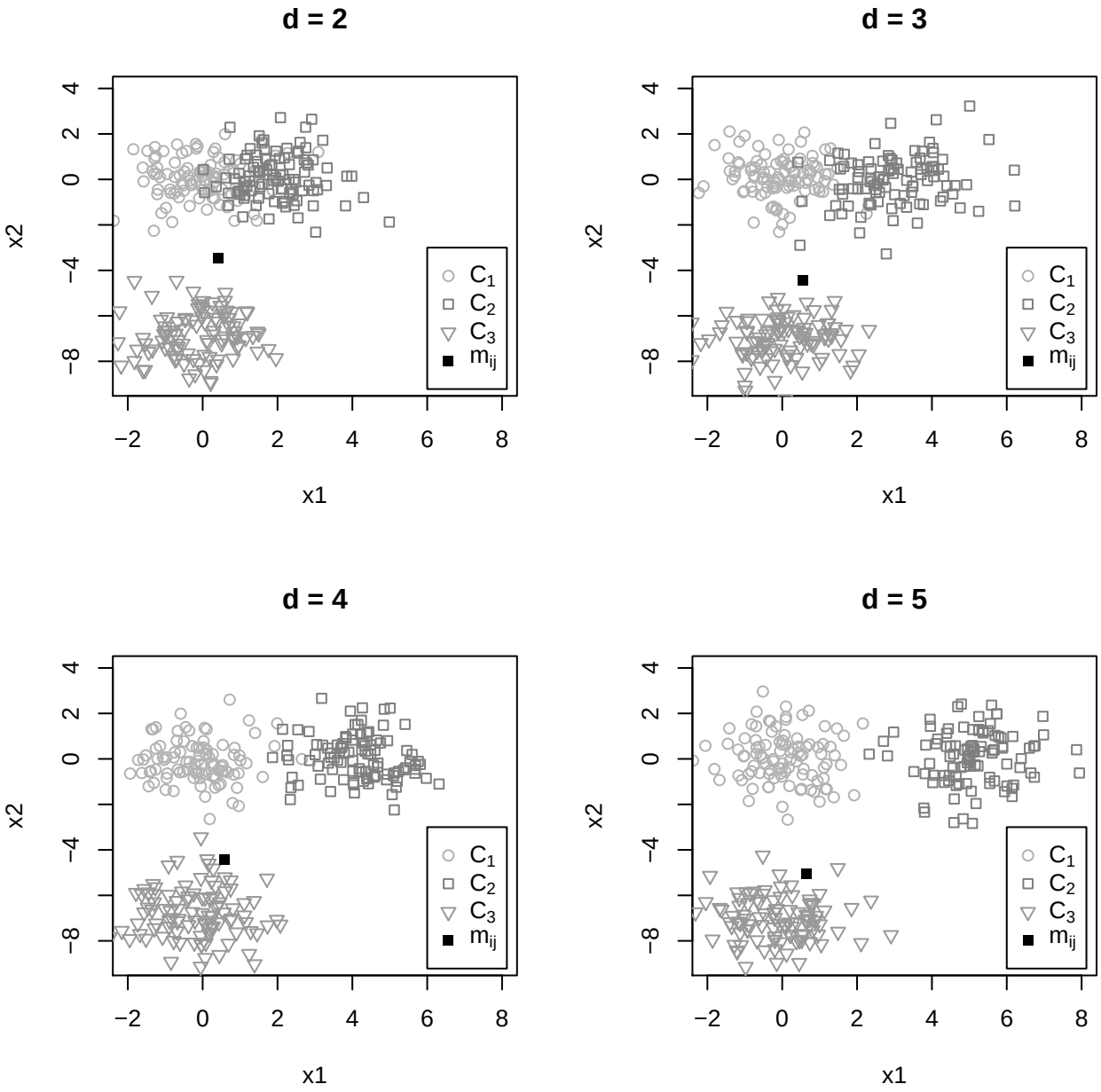
DOI: 10.14751/SZIE.2014.035

A.1. Az $ind = 1$ megoldásainak ábrázolása az α függvényében (R kód).

```
szAx <- 1
szAy <- 1
szBx <- 1
szBy <- 1
eredm <- data.frame()
szamlalo <- 1
for (alfa in seq(0.1, 3, by = 0.01)){
lepeskoz <- 0.05
A <- c(0,0)
tabla <- data.frame()
i <- 1
for (b in seq(0, 7, by = lepeskoz)){
B <- c(b,0)
K <- (A+B)/2
szx <- min(szAx,szBx)
szy <- min(szAy,szBy)
alfainv <- 1-2*pnorm(c(alfa*szy), mean=0, sd=szy, lower.tail=FALSE)
PA <- (1-2*pnorm(c(0+alfa*szx), mean=0, sd=szAx, lower.tail=FALSE))*alfainv
PB <- (1-2*pnorm(c(b+alfa*szx), mean=b, sd=szBx, lower.tail=FALSE))*alfainv
PKA <- pnorm(c(K[1]-alfa*szx), mean=0, sd=szAx, lower.tail=FALSE)*alfainv-
pnorm(c(K[1]+alfa*szx),
mean=0, sd=szAx, lower.tail=FALSE)*alfainv
PKB <- pnorm(c(K[1]+alfa*szx), mean=b, sd=szBx, lower.tail=TRUE)*alfainv-
pnorm(c(K[1]-alfa*szx),
mean=b, sd=szBx, lower.tail=TRUE)*alfainv
PK <- PKA+PKB
ind <- PK/(max(PA,PB))
tabla[i,1] <- b
tabla[i,2] <- ind
tabla[i,3] <- PA
tabla[i,4] <- PB
tabla[i,5] <- PK
tabla[i,6] <- abs(1-tabla[i,2])
i <- i+1
}
colnames(tabla) <- c("b","ind","PA","PB","PK","y=1")
plot(tabla[,1],tabla[,2], xlab="klaszterközéppontok közötti távolság", ylab="ind")
cim <- paste("alfa =", alfa)
title(main = cim )
m <- apply(tabla[1:nrow(tabla),],2,min)
a <- as.numeric(m[6])
y1 <- subset(tabla[,1],tabla[,6]==a)
eredm[szamlalo,1] <- alfa
eredm[szamlalo,2] <- y1
szamlalo <- szamlalo+1
}
colnames(eredm) <- c("alfa","ind = 1 megoldásai")
eredm
plot(eredm)
```

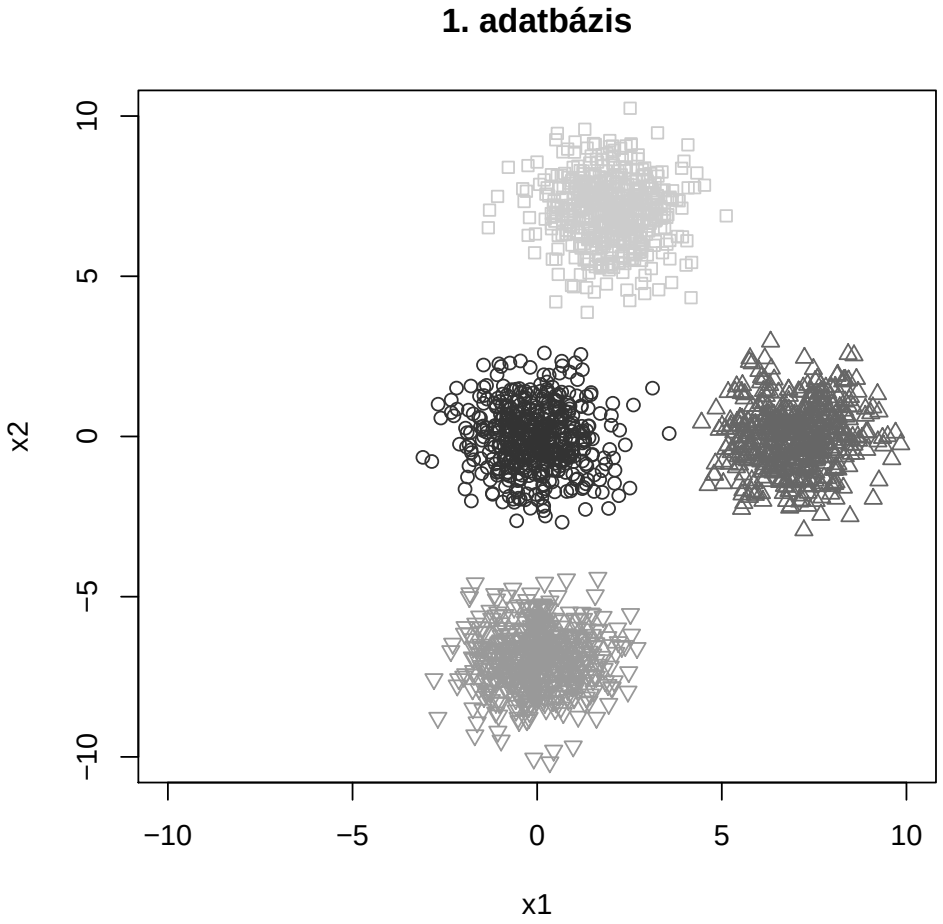
DOI: 10.14751/SZIE.2014.035

A.2. Az m_{ij} osztópont helyzete a C_1 és a C_2 klaszter távolítása esetén $(nc = 2)$.



DOI: 10.14751/SZIE.2014.035

A.3. Az indexek összehasonlítására használt 1. adatbázis (4.1.4. alszakasz).

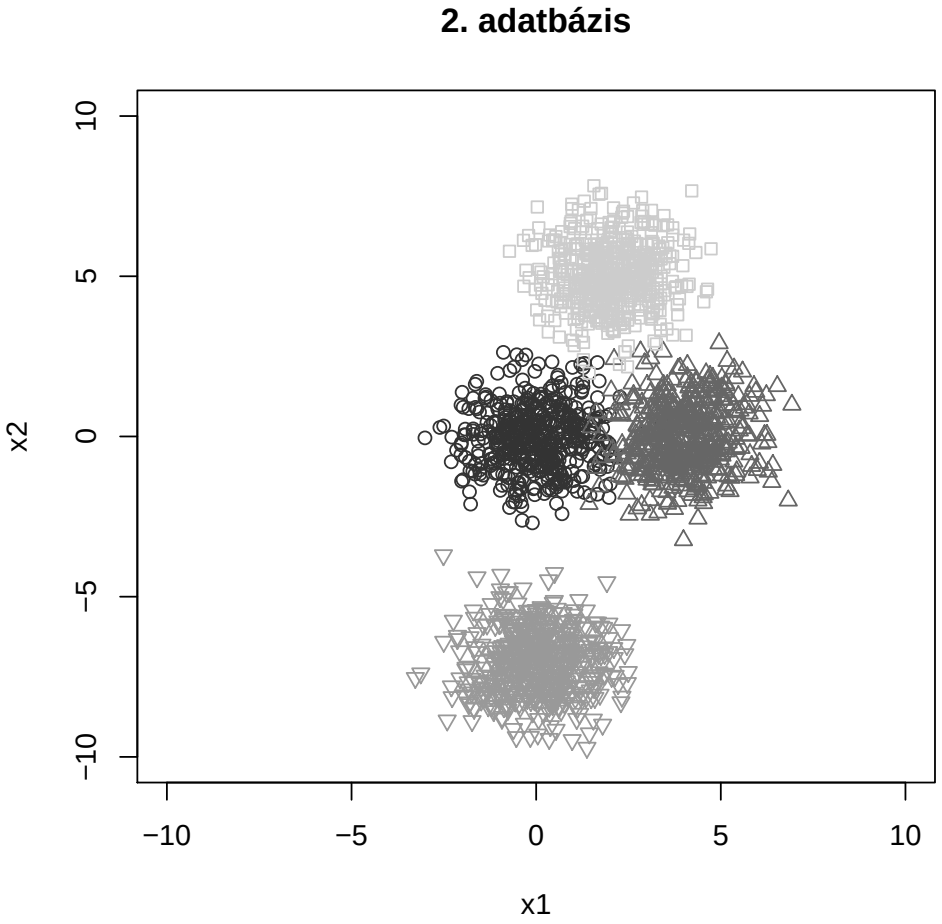


	K_1	K_2	K_3	K_4
$\boldsymbol{v}_i$	$(0,0)$	$(7,0)$	$(0,-7)$	$(2,7)$
$\boldsymbol{\sigma}_i$	$(1,1)$	$(1,1)$	$(1,1)$	$(1,1)$
n_i	1000	1000	1000	1000

ahol $\boldsymbol{v}_i$: az i-edik klaszter középpontja, $\boldsymbol{\sigma}_i$: az i-edik klaszter x és y irányú szórása, n_i: az i-edik klaszter elemszáma.

DOI: 10.14751/SZIE.2014.035

A.4. Az indexek összehasonlítására használt 2. adatbázis (4.1.4. alszakasz).

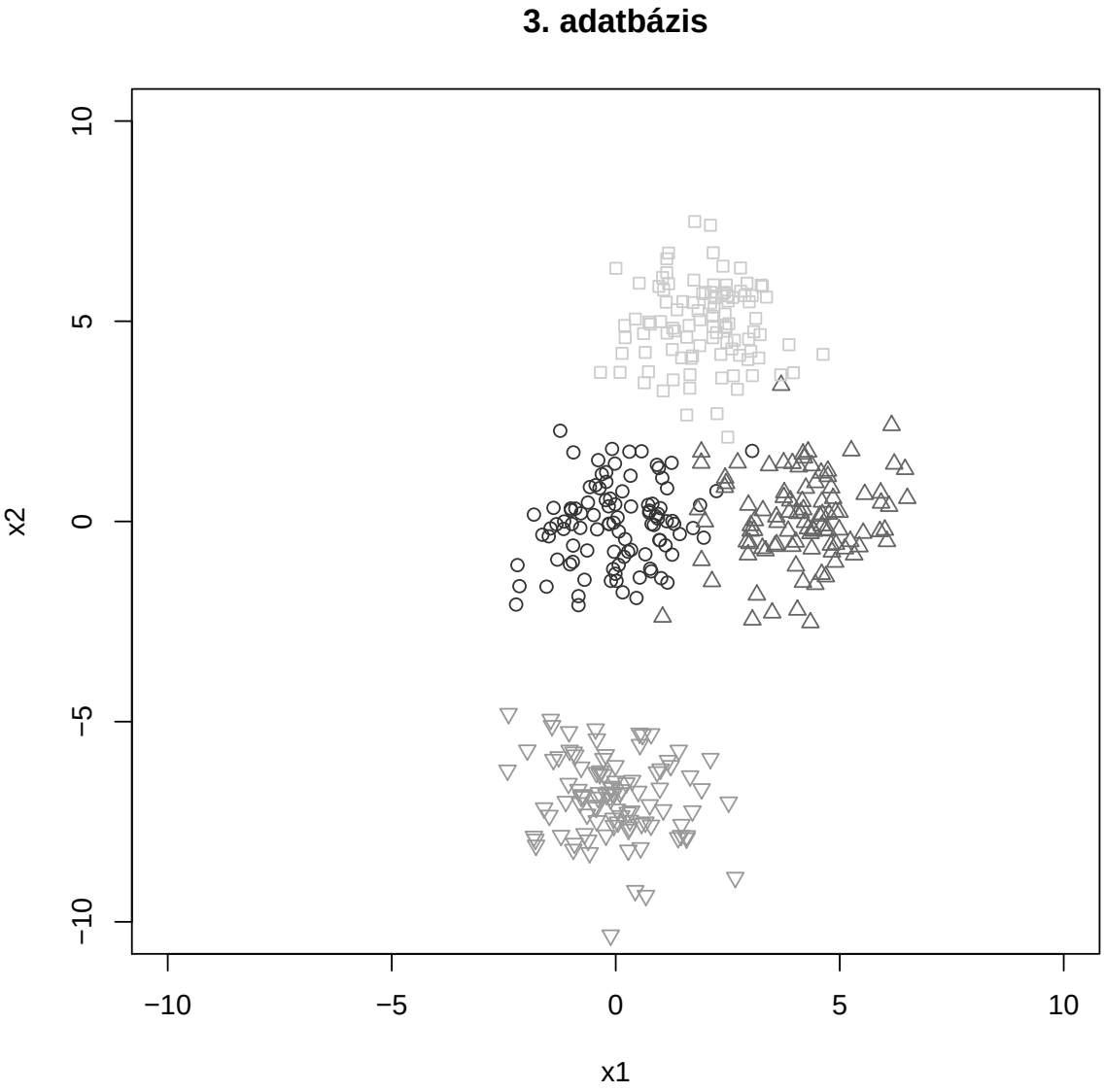


	K_1	K_2	K_3	K_4
$\boldsymbol{v}_i$	$(0,0)$	$(4,0)$	$(0,-7)$	$(2,5)$
$\boldsymbol{\sigma}_i$	$(1,1)$	$(1,1)$	$(1,1)$	$(1,1)$
n_i	500	500	500	500

ahol $\boldsymbol{v}_i$: az i-edik klaszter középpontja, $\boldsymbol{\sigma}_i$: az i-edik klaszter x és y irányú szórása, n_i: az i-edik klaszter elemszáma.

DOI: 10.14751/SZIE.2014.035

A.5. Az indexek összehasonlítására használt 3. adatbázis (4.1.4. alszakasz).

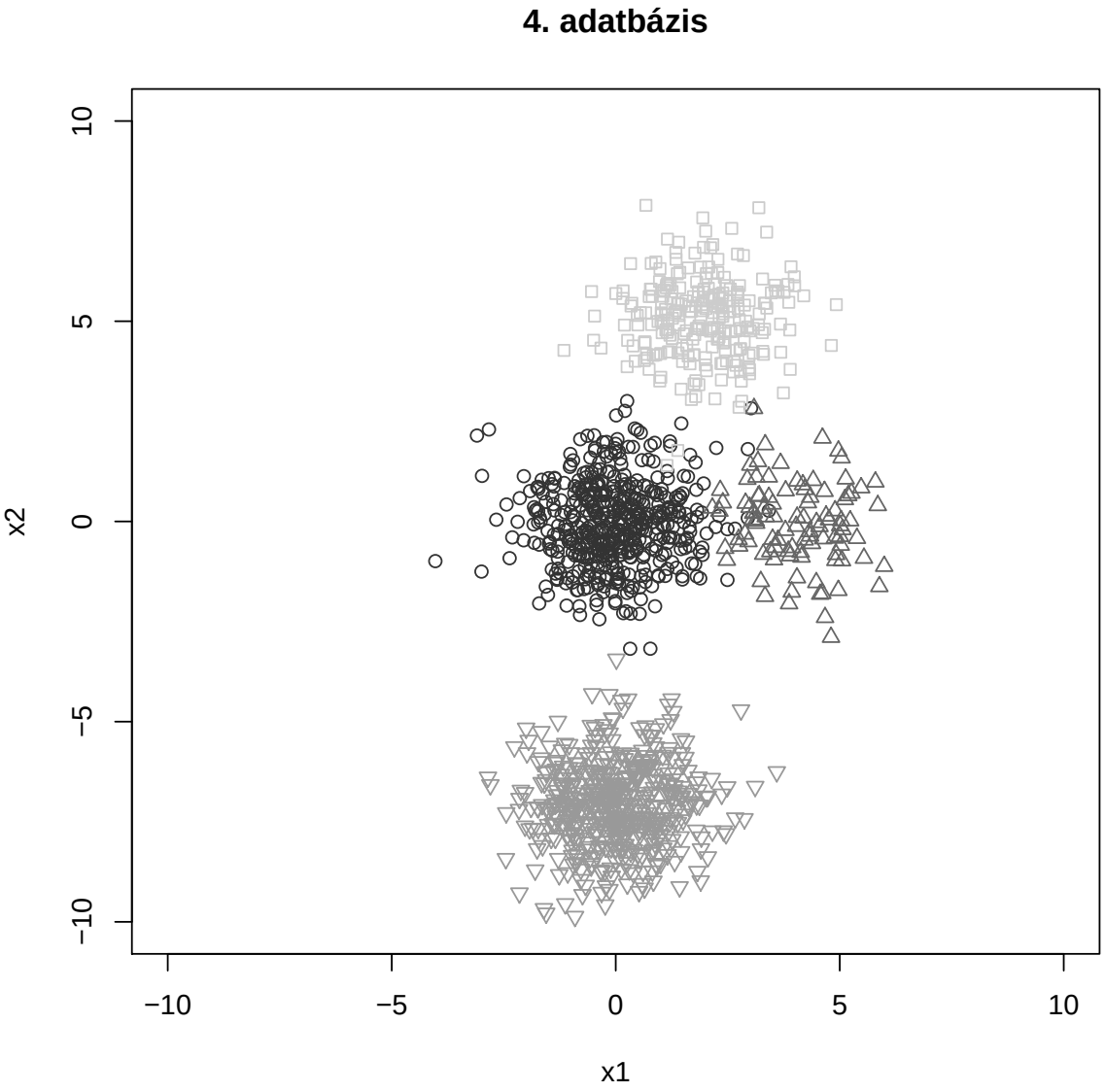


	K_1	K_2	K_3	K_4
$\boldsymbol{v}_i$	$(0,0)$	$(4,0)$	$(0,-7)$	$(2,5)$
$\boldsymbol{\sigma}_i$	$(1,1)$	$(1,1)$	$(1,1)$	$(1,1)$
n_i	100	100	100	100

ahol $\boldsymbol{v}_i$: az i-edik klaszter középpontja, $\boldsymbol{\sigma}_i$: az i-edik klaszter x és y irányú szórása, n_i: az i-edik klaszter elemszáma.

DOI: 10.14751/SZIE.2014.035

A.6. Az indexek összehasonlítására használt 4. adatbázis (4.1.4. alszakasz).

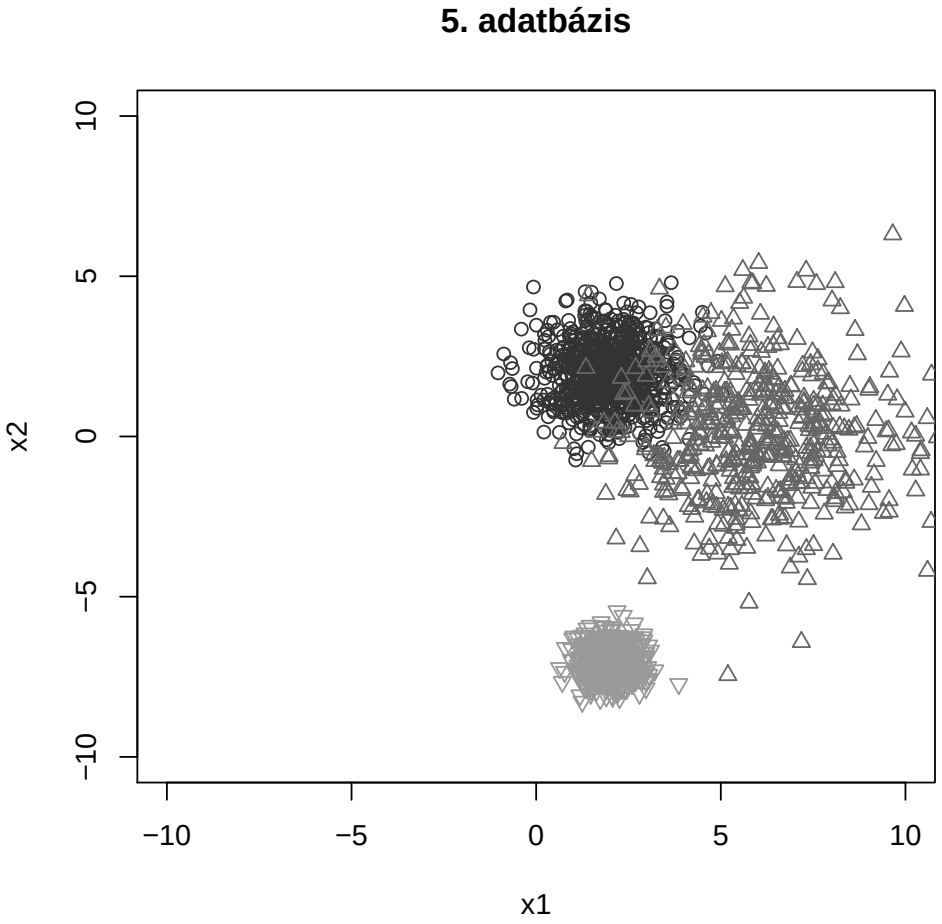


	K_1	K_2	K_3	K_4
$\boldsymbol{v}_i$	$(0,0)$	$(4,0)$	$(0,-7)$	$(2,5)$
$\boldsymbol{\sigma}_i$	$(1,1)$	$(1,1)$	$(1,1)$	$(1,1)$
n_i	500	100	500	250

ahol $\boldsymbol{v}_i$: az i-edik klaszter középpontja, $\boldsymbol{\sigma}_i$: az i-edik klaszter x és y irányú szórása, n_i: az i-edik klaszter elemszáma.

DOI: 10.14751/SZIE.2014.035

A.7. Az indexek összehasonlítására használt 5. adatbázis (4.1.4. alszakasz).

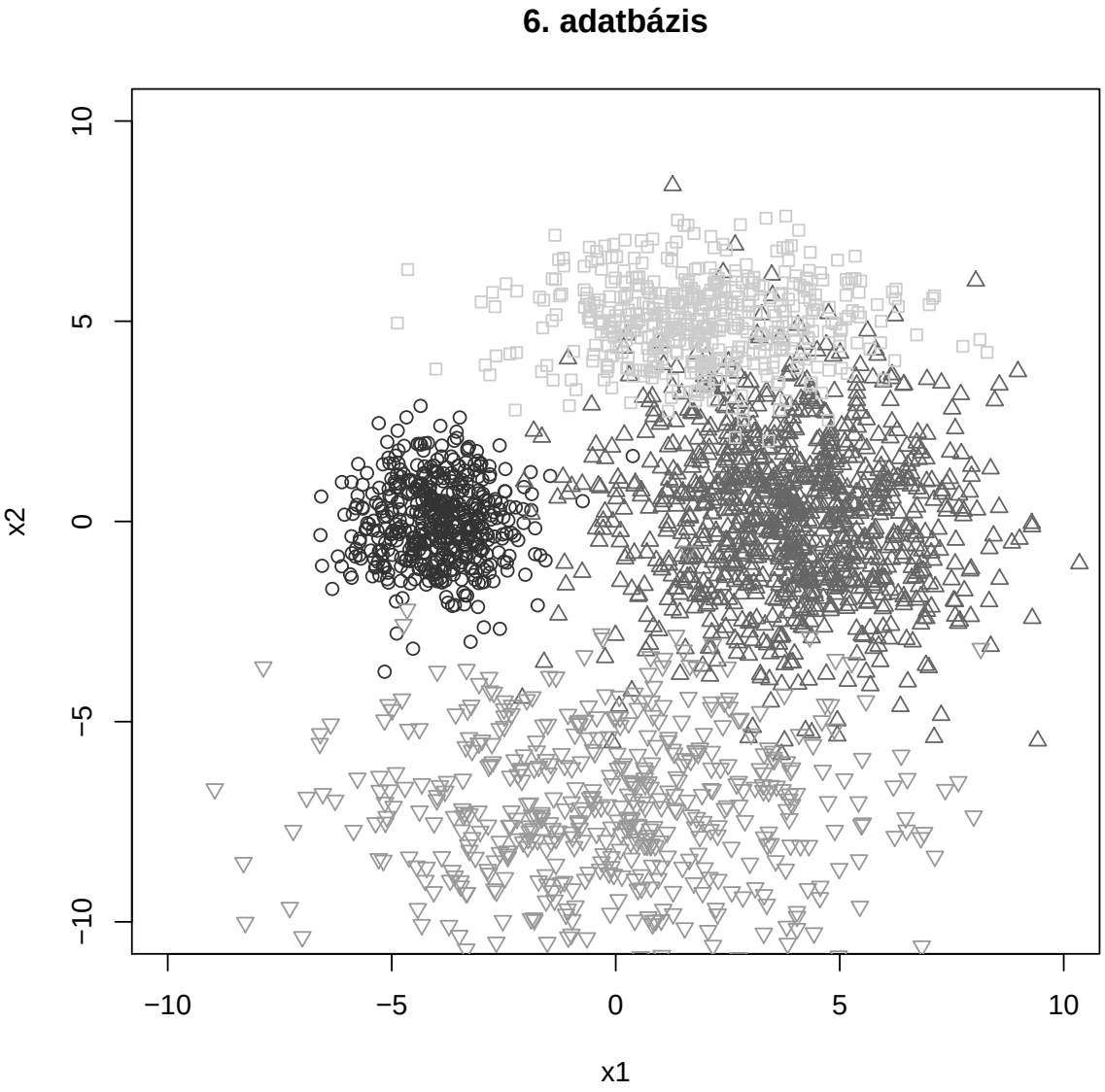


	K_1	K_2	K_3
$\boldsymbol{v}_i$	$(2, 2)$	$(6, 0)$	$(2, -7)$
$\boldsymbol{\sigma}_i$	$(1, 1)$	$(2, 2)$	$(0{,}5, 0{,}5)$
n_i	750	500	500

ahol $\boldsymbol{v}_i$: az i-edik klaszter középpontja, $\boldsymbol{\sigma}_i$: az i-edik klaszter x és y irányú szórása, n_i: az i-edik klaszter elemszáma.

DOI: 10.14751/SZIE.2014.035

A.8. Az indexek összehasonlítására használt 6. adatbázis (4.1.4. alszakasz).

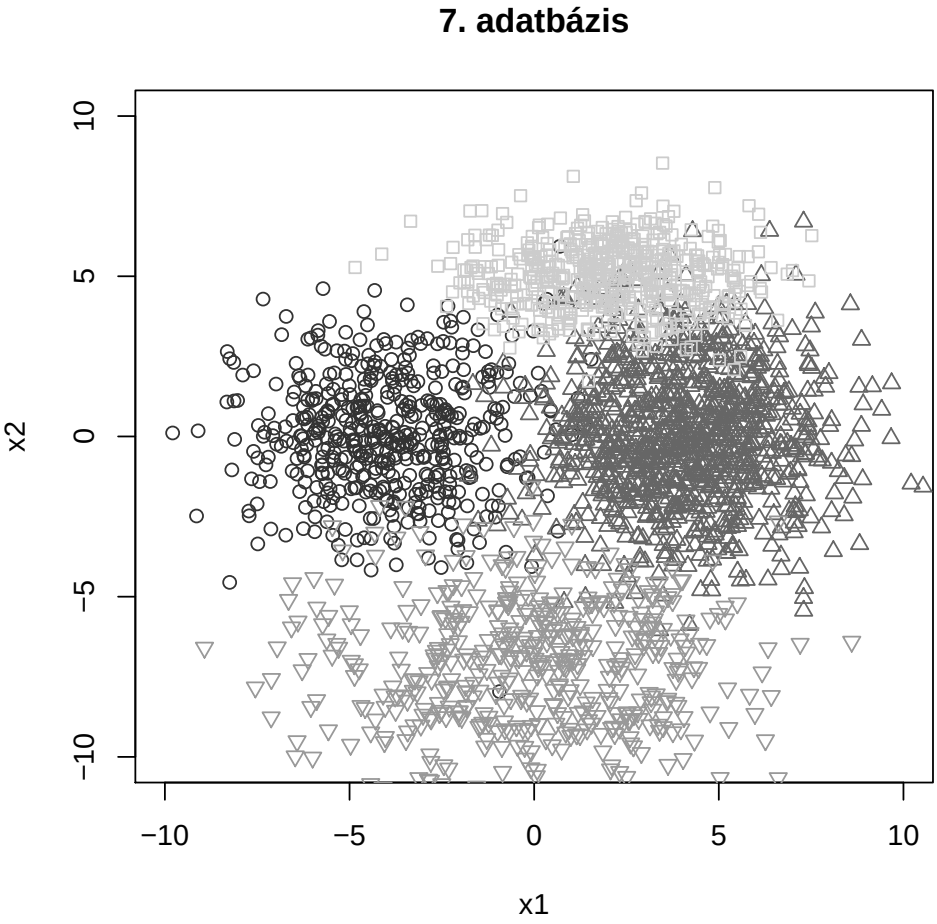


	K_1	K_2	K_3	K_4
$\boldsymbol{v}_i$	$(-4, 0)$	$(4, 0)$	$(0, -7)$	$(2, 5)$
$\boldsymbol{\sigma}_i$	$(1, 1)$	$(2, 2)$	$(3, 2)$	$(2, 1)$
n_i	500	1000	500	500

ahol $\boldsymbol{v}_i$: az i-edik klaszter középpontja, $\boldsymbol{\sigma}_i$: az i-edik klaszter x és y irányú szórása, n_i: az i-edik klaszter elemszáma.

DOI: 10.14751/SZIE.2014.035

A.9. Az indexek összehasonlítására használt 7. adatbázis (4.1.4. alszakasz).

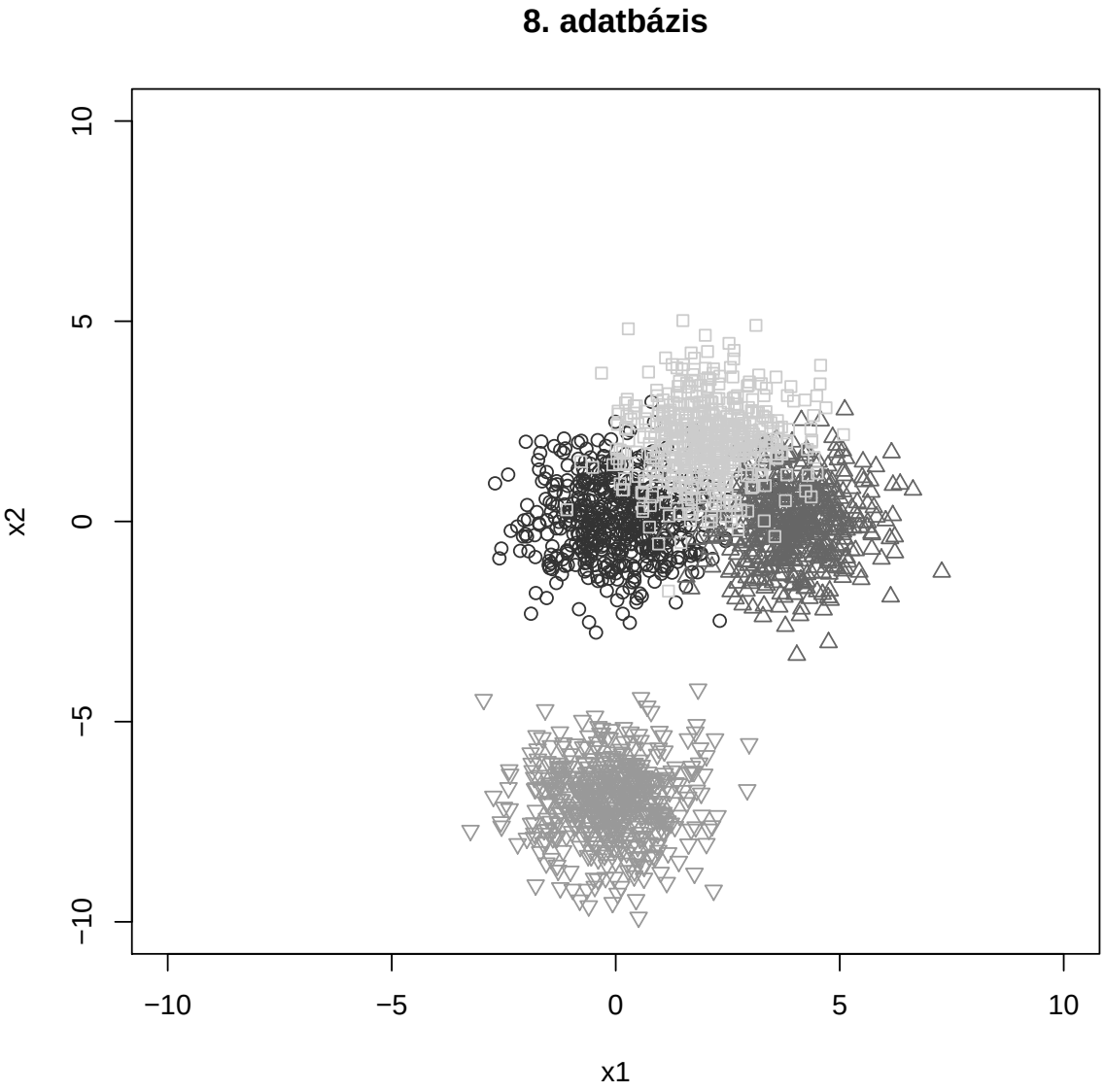


	K_1	K_2	K_3	K_4
$\boldsymbol{v}_i$	$(-4, 0)$	$(4, 0)$	$(0, -7)$	$(2, 5)$
$\boldsymbol{\sigma}_i$	$(2, 2)$	$(2, 2)$	$(3, 2)$	$(2, 1)$
n_i	500	1000	500	500

ahol $\boldsymbol{v}_i$: az i-edik klaszter középpontja, $\boldsymbol{\sigma}_i$: az i-edik klaszter x és y irányú szórása, n_i: az i-edik klaszter elemszáma.

DOI: 10.14751/SZIE.2014.035

A.10. Az indexek összehasonlítására használt 8. adatbázis (4.1.4. alszakasz).



	K_1	K_2	K_3	K_4
$\boldsymbol{v}_i$	$(0,0)$	$(4,0)$	$(0,-7)$	$(2,2)$
$\boldsymbol{\sigma}_i$	$(1,1)$	$(1,1)$	$(1,1)$	$(1,1)$
n_i	500	500	500	500

ahol $\boldsymbol{v}_i$: az i-edik klaszter középpontja, $\boldsymbol{\sigma}_i$: az i-edik klaszter x és y irányú szórása, n_i: az i-edik klaszter elemszáma.

DOI: 10.14751/SZIE.2014.035

A.11. Az indexek összehasonlítása (R kód).

```
#### adatok #####
maxkl <- 7  # klaszterek számának maximuma
valkl <- 3  #klaszterek tényleges száma
eh <- 0.4 # a szórás hányad részét vegye figyelembe a környezet meghatározásakor
library(mclust)
g <- list()
g[[1]] <- data.frame(x1=rnorm(3000, mean=0, sd=1),x2=rnorm(3000, mean=2, sd=1))
g[[2]] <- data.frame(x1=rnorm(2000, mean=6, sd=2),x2=rnorm(2000, mean=0, sd=2))
g[[3]] <- data.frame(x1=rnorm(1000, mean=2, sd=0.5),x2=rnorm(1000, mean=-7, sd=0.5))
plot(g[[1]], col="gray20", pch=1, xlim=c(-3,13), ylim=c(-10,8))
points(g[[2]], col="gray40", pch=2)
points(g[[3]], col="gray60", pch=6)
title(main="5. adatbázis")
savePlot(filename = "adatbazis_5",type = "pdf")
dat <- data.frame()
for (i in 1:valkl){
dat <- data.frame(rbind(dat,g[[i]]))
}
eredm <- data.frame(matrix(0,nrow=maxkl-1, ncol=6))
#### klaszterek előállítása  ####
### Kmeans ###
for (nc in 2:maxkl){
klasz <- kmeans(dat,nc)
dat[,3] <- klasz\$cluster
f <- list()
for (j in 1:nc){
f[[j]] <- subset(dat[,1:2],dat[,3]==j)
}
### klaszterek ábrázolása
plot(f[[1]], col=1, xlim=c(-10,10), ylim=c(-10,10))
title("Kmeans")
for (i in 2:nc){
points(f[[i]], col=i)
}
if (nc==valkl){savePlot(filename = "kmean_jo_klaszrerszam",type = "pdf")}

### KLASZTEREK KÖZÖTT (Dens_bw) új

## density
cmean <- data.frame(matrix(0, ncol=2, nrow=nc))
std <- data.frame(matrix(0, ncol=2, nrow=nc))
for (i in 1:nc){
cmean[i,1:2] <- colMeans(data.frame(f[i]))
std[i,1:2] <- sapply(data.frame(f[i]),sd)
stdm <- c(min(std[,1]), min(std[,2]))
}
s <- 0
tabla <- data.frame(matrix(0,nrow=(nc*(nc-1))/2, ncol=10))
for (a in 1:(nc-1)){
for (b in (a+1):nc){
ind <- 0
sza <- 0
for (k in 1:nrow(f[[a]])) {
if (f[[a]][k,1]>as.numeric(cmean[a,1]-eh*std[a,1]) &
f[[a]][k,1]<as.numeric(cmean[a,1]+eh*std[a,1]) &
f[[a]][k,2]>as.numeric(cmean[a,2]-eh*std[a,2]) &
```

DOI: 10.14751/SZIE.2014.035


```
f[[a]][k,2]<as.numeric(cmean[a,2]+eh*std[a,2]))
{ind = 1} else {ind=0}
sza <- sza+ind
}
#print(sza)
ind <- 0
szb <- 0
for (k in 1:nrow(f[[b]])) {
if (f[[b]][k,1]>as.numeric(cmean[b,1]-eh*std[b,1]) &
f[[b]][k,1]<as.numeric(cmean[b,1]+eh*std[b,1]) &
f[[b]][k,2]>as.numeric(cmean[b,2]-eh*std[b,2]) &
f[[b]][k,2]<as.numeric(cmean[b,2]+eh*std[b,2]))
{ind = 1} else {ind=0}
szb <- szb+ind
}
#print(szb)
## saját erdmény
egyutt <- data.frame(rbind(f[[a]],f[[b]]))
szm <- 0
ind <- 0
if (sza+szb==0) {szab =1} else {szab = sza + szb}
m <- 0.7*(nrow(f[[a]])*cmean[b,1:2]+nrow(f[[b]])*
cmean[a,1:2])/(nrow(f[[a]])+nrow(f[[b]]))+0.3*
(szb*cmean[b,1:2]+sza*cmean[a,1:2])/(szab)   # Tong féle kp

for (k in 1:nrow(egyutt)) {
if (egyutt[k,1]>as.numeric(m[1]-eh*stdm[1]) &
egyutt[k,1]<as.numeric(m[1]+eh*stdm[1]) &
egyutt[k,2]>as.numeric(m[2]-eh*stdm[2]) &
egyutt[k,2]<as.numeric(m[2]+eh*stdm[2]) )
{ind = 1} else {ind=0}
szm <- szm+ind
}
#print(szm)
s <- s+1
tabla[s,1] <- a
tabla[s,2] <- b
tabla[s,3] <- sza
tabla[s,4] <- szb
tabla[s,5] <- szm
if (sza+szb==0) {tabla[s,6] <- szm/1} else {tabla[s,6] <- szm/(max(sza,szb))}
## Tong eredménye
stdm <- c((std[a,1]+std[b,1])/2, (std[a,2]+std[b,2])/2)
ind <- 0
sza <- 0
for (k in 1:nrow(f[[a]])) {
if (f[[a]][k,1]>as.numeric(cmean[a,1]-1.96*std[a,1]/sqrt(nrow(f[[a]]))) &
f[[a]][k,1]<as.numeric(cmean[a,1]+1.96*std[a,1]/sqrt(nrow(f[[a]]))) &
f[[a]][k,2]>as.numeric(cmean[a,2]-1.96*std[a,2]/sqrt(nrow(f[[a]]))) &
f[[a]][k,2]<as.numeric(cmean[a,2]+1.96*std[a,2]/sqrt(nrow(f[[a]]))))
{ind = 1} else {ind=0}
sza <- sza+ind
}
ind <- 0
szb <- 0
for (k in 1:nrow(f[[b]])) {
if (f[[b]][k,1]>as.numeric(cmean[b,1]-1.96*std[b,1]/sqrt(nrow(f[[b]]))) &
f[[b]][k,1]<as.numeric(cmean[b,1]+1.96*std[b,1]/sqrt(nrow(f[[b]]))) &
f[[b]][k,2]>as.numeric(cmean[b,2]-1.96*std[b,2]/sqrt(nrow(f[[b]]))) &
```

DOI: 10.14751/SZIE.2014.035


```
f[[b]][k,2]<as.numeric(cmean[b,2]+1.96*std[b,2]/sqrt(nrow(f[[b]])))))
{ind = 1} else {ind=0}
szb <- szb+ind
}
if (sza+szb==0) {szab =1} else {szab = sza + szb}
#points(mT[1],mT[2],col="red", pch=19)
mT <- 0.7*(nrow(f[[a]])*cmean[b,1:2]+nrow(f[[b]])*
cmean[a,1:2])/(nrow(f[[a]])+nrow(f[[b]]))+0.3*(szb*cmean[b,1:2]+sza*
cmean[a,1:2])/(szab)   # Tong féle középpont

ind <- 0
szmT <- 0
for (k in 1:nrow(egyutt)) {
if (egyutt[k,1]>as.numeric(mT[1]-1.96*stdm[1]/sqrt(nrow(f[[b]])+nrow(f[[a]])))
& egyutt[k,1]<as.numeric(mT[1]+1.96*stdm[1]/sqrt(nrow(f[[b]])+nrow(f[[a]])))
& egyutt[k,2]>as.numeric(mT[2]-1.96*stdm[2]/sqrt(nrow(f[[b]])+nrow(f[[a]])))
& egyutt[k,2]<as.numeric(mT[2]+1.96*stdm[2]/sqrt(nrow(f[[b]])+nrow(f[[a]]))) )
{ind = 1} else {ind=0}
szmT <- szmT+ind
}
#szmT
tabla[s,7] <- sza
tabla[s,8] <- szb
tabla[s,9] <- szmT
if (sza+szb==0){tabla[s,10] <- szmT/1} else {tabla[s,10] <- szmT/(max(sza,szb))}
}
}
###  KLASZTREREN BELÜL Scat
t <- 0
tabla2 <- data.frame()
for (a in 1:nc){
Scat <- 0
stdS <- sapply(dat,sd)
Scat <- ((nrow(dat)-nrow(f[[a]]))/nrow(dat)*sqrt(as.matrix(std[a,1:2]^2)
%*%t(as.matrix(std[a,1:2]^2)))/sqrt(t(as.matrix(stdS^2))%*%as.matrix(stdS^2)))
t <- t+1
tabla2[t,1] <- Scat
}
colnames(tabla) <- c("a","b","sza","szb","szm","Dens_sajat","szaT","szbT","szmT","Dens_T")
print(tabla)
print(tabla2)
### eredmenyek
Dens_bw_sajat <- 1/(nc*(nc-1))*sum(tabla[,6])
Dens_bw_sajat
Dens_bw_Tong <- 1/(nc*(nc-1))*sum(tabla[,10])
Dens_bw_Tong
Scat <- 1/(nc-1)*sum(tabla2[,1])
### INDEX
index_sajat <- Dens_bw_sajat + Scat
index_Tong <- Dens_bw_Tong + Scat
eredm[nc-1,1] <- nc
eredm[nc-1,2] <- Dens_bw_sajat
eredm[nc-1,3] <- Scat
eredm[nc-1,4] <- index_sajat
eredm[nc-1,5] <- Dens_bw_Tong
eredm[nc-1,6] <- index_Tong
}
colnames(eredm) <-
 c("klaszterszám","Dens_bw_sajat","Scat","index_sajat","Dens_bw_Tong","index_Tong")
```

DOI: 10.14751/SZIE.2014.035


```
eredm_Kmeans <- eredm

### Mclust ###
for (nc in 2:maxkl){
#Mclust - modell alapú
Klasz <- Mclust(dat,G=nc)
Klasz$modelName
dat[,4] <- Klasz$classification
f <- list()
for (j in 1:nc){
f[[j]] <- subset(dat[,1:2],dat[,4]==j)
}
### klaszterek ábrázolása
plot(f[[1]], col=1, xlim=c(-10,10), ylim=c(-10,10))
title("Mclust")
for (i in 2:nc){
points(f[[i]], col=i)
}
if (nc==valkl){savePlot(filename = "mclust_jo_klaszrerszam",type = "pdf")}
### KLASZTEREK KÖZÖTT (Dens_bw) új
## density
cmean <- data.frame(matrix(0, ncol=2, nrow=nc))
std <- data.frame(matrix(0, ncol=2, nrow=nc))
for (i in 1:nc){
cmean[i,1:2] <- colMeans(data.frame(f[i]))
std[i,1:2] <- sapply(data.frame(f[i]),sd)
stdm <- c(min(std[,1]), min(std[,2]))
}
s <- 0
tabla <- data.frame(matrix(0,nrow=(nc*(nc-1))/2, ncol=1))
for (a in 1:(nc-1)){
for (b in (a+1):nc){
ind <- 0
sza <- 0
for (k in 1:nrow(f[[a]])) {
if (f[[a]][k,1]>as.numeric(cmean[a,1]-eh*std[a,1]) &
f[[a]][k,1]<as.numeric(cmean[a,1]+eh*std[a,1]) &
f[[a]][k,2]>as.numeric(cmean[a,2]-eh*std[a,2]) &
f[[a]][k,2]<as.numeric(cmean[a,2]+eh*std[a,2]))
{ind = 1} else {ind=0}
sza <- sza+ind
}
#print(sza)
ind <- 0
szb <- 0
for (k in 1:nrow(f[[b]])) {
if (f[[b]][k,1]>as.numeric(cmean[b,1]-eh*std[b,1]) &
f[[b]][k,1]<as.numeric(cmean[b,1]+eh*std[b,1]) &
f[[b]][k,2]>as.numeric(cmean[b,2]-eh*std[b,2]) &
f[[b]][k,2]<as.numeric(cmean[b,2]+eh*std[b,2]))
{ind = 1} else {ind=0}
szb <- szb+ind
}
#print(szb)
### saját erdmény
egyutt <- data.frame(rbind(f[[a]],f[[b]]))
szm <- 0
ind <- 0
if (sza+szb==0) {szab =1} else {szab = sza + szb}
```

DOI: 10.14751/SZIE.2014.035


```
m <- 0.7*(nrow(f[[a]])*cmean[b,1:2]+nrow(f[[b]])*cmean[a,1:2])/
(nrow(f[[a]])+nrow(f[[b]]))+0.3*(szb*cmean[b,1:2]+sza*cmean[a,1:2])/(szab)
for (k in 1:nrow(egyutt)) {
if (egyutt[k,1]>as.numeric(m[1]-eh*stdm[1]) &
egyutt[k,1]<as.numeric(m[1]+eh*stdm[1]) &
egyutt[k,2]>as.numeric(m[2]-eh*stdm[2]) &
egyutt[k,2]<as.numeric(m[2]+eh*stdm[2]) )
 {ind = 1} else {ind=0}
szm <- szm+ind
}
#print(szm)
s <- s+1
tabla[s,1] <- a
tabla[s,2] <- b
tabla[s,3] <- sza
tabla[s,4] <- szb
tabla[s,5] <- szm
if (sza+szb==0) {tabla[s,6] <- szm/1} else {tabla[s,6] <- szm/(max(sza,szb))}
## Tong eredménye
stdm <- c((std[a,1]+std[b,1])/2, (std[a,2]+std[b,2])/2)
ind <- 0
sza <- 0
for (k in 1:nrow(f[[a]])) {
if (f[[a]][k,1]>as.numeric(cmean[a,1]-1.96*std[a,1]/sqrt(nrow(f[[a]]))) &
f[[a]][k,1]<as.numeric(cmean[a,1]+1.96*std[a,1]/sqrt(nrow(f[[a]]))) &
f[[a]][k,2]>as.numeric(cmean[a,2]-1.96*std[a,2]/sqrt(nrow(f[[a]]))) &
f[[a]][k,2]<as.numeric(cmean[a,2]+1.96*std[a,2]/sqrt(nrow(f[[a]])))
{ind = 1} else {ind=0}
sza <- sza+ind
}
ind <- 0
szb <- 0
for (k in 1:nrow(f[[b]])) {
if (f[[b]][k,1]>as.numeric(cmean[b,1]-1.96*std[b,1]/sqrt(nrow(f[[b]]))) &
f[[b]][k,1]<as.numeric(cmean[b,1]+1.96*std[b,1]/sqrt(nrow(f[[b]]))) &
f[[b]][k,2]>as.numeric(cmean[b,2]-1.96*std[b,2]/sqrt(nrow(f[[b]]))) &
f[[b]][k,2]<as.numeric(cmean[b,2]+1.96*std[b,2]/sqrt(nrow(f[[b]])))
{ind = 1} else {ind=0}
szb <- szb+ind
}
if (sza+szb==0) {szab =1} else {szab = sza + szb}
#points(mT[1],mT[2],col="red", pch=19)
mT <- 0.7*(nrow(f[[a]])*cmean[b,1:2]+nrow(f[[b]])*cmean[a,1:2])/
(nrow(f[[a]])+nrow(f[[b]]))+0.3*(szb*cmean[b,1:2]+sza*cmean[a,1:2])/(szab)
ind <- 0
szmT <- 0
for (k in 1:nrow(egyutt)) {
if (egyutt[k,1]>as.numeric(mT[1]-1.96*stdm[1]/sqrt(nrow(f[[b]])+nrow(f[[a]])))
& egyutt[k,1]<as.numeric(mT[1]+1.96*stdm[1]/sqrt(nrow(f[[b]])+nrow(f[[a]])))
& egyutt[k,2]>as.numeric(mT[2]-1.96*stdm[2]/sqrt(nrow(f[[b]])+nrow(f[[a]])))
& egyutt[k,2]<as.numeric(mT[2]+1.96*stdm[2]/sqrt(nrow(f[[b]])+nrow(f[[a]]))))
 {ind = 1} else {ind=0}
szmT <- szmT+ind
}
#szmT
tabla[s,7] <- sza
tabla[s,8] <- szb
tabla[s,9] <- szmT
if (sza+szb==0){tabla[s,10] <- szmT/1} else {tabla[s,10] <- szmT/(max(sza,szb))}
```

DOI: 10.14751/SZIE.2014.035


```
}
}
###  KLASZTRERENBELÜL Scat
t <- 0
tabla2 <- data.frame()
for (a in 1:nc){
Scat <- 0
stdS <- sapply(dat,sd)
Scat <- ((nrow(dat)-nrow(f[[a]]))/nrow(dat)*sqrt(as.matrix(std[a,1:2]^2)
%*%t(as.matrix(std[a,1:2]^2)))/sqrt(t(as.matrix(stdS^2))%*%as.matrix(stdS^2)))
t <- t+1
tabla2[t,1] <- Scat
}
colnames(tabla) <- c("a","b","sza","szb","szm","Dens_sajat","szaT","szbT","szmT","Dens_T")
print(tabla)
### eredmenyek
Dens_bw_sajat <- 1/(nc*(nc-1))*sum(tabla[,6])
Dens_bw_sajat
Dens_bw_Tong <- 1/(nc*(nc-1))*sum(tabla[,10])
Dens_bw_Tong
Scat <- 1/(nc-1)*sum(tabla2[,1])

#stdm <- c(min(std[,1]), min(std[,2]))
#abline(v=as.numeric(mT[1]-stdm[1]), col="red", lty=3)
#abline(v=as.numeric(mT[1]+stdm[1]), col="red", lty=3)
#abline(h=as.numeric(mT[2]-stdm[2]), col="red", lty=3)
#abline(h=as.numeric(mT[2]+stdm[2]), col="red", lty=3)
### INDEX
index_sajat <- Dens_bw_sajat + Scat
index_Tong <- Dens_bw_Tong + Scat
eredm[nc-1,1] <- nc
eredm[nc-1,2] <- Dens_bw_sajat
eredm[nc-1,3] <- Scat
eredm[nc-1,4] <- index_sajat
eredm[nc-1,5] <- Dens_bw_Tong
eredm[nc-1,6] <- index_Tong
}
colnames(eredm) <-
c("klaszterszám","Dens_bw_sajat","Scat","index_sajat","Dens_bw_Tong","index_Tong")
eredm_Mclust <- eredm
eredm_Kmeans
eredm_Mclust
```

DOI: 10.14751/SZIE.2014.035

A.12. Az $\mathrm{E}(X(t)|\Phi)$ várható érték levezetése (78. old., 4.31. egyenlet)

$$\mathrm{E}(X(t)|\Phi) = \lambda t \cdot \mathrm{e}^{-\lambda ct} + \int_0^t \lambda x \cdot \lambda c\, \mathrm{e}^{-\lambda cx}\, \mathrm{d}x =$$

$$= \lambda t \cdot \mathrm{e}^{-\lambda ct} + \lambda^2 c \int_0^t x\, \mathrm{e}^{-\lambda cx}\, \mathrm{d}x =$$

$$= \lambda t \cdot \mathrm{e}^{-\lambda ct} + \lambda^2 c \left(\left[x \frac{\mathrm{e}^{-\lambda cx}}{-\lambda c} \right]_0^t - \int_0^t \frac{\mathrm{e}^{-\lambda cx}}{-\lambda c}\, \mathrm{d}x \right) =$$

$$= \lambda t \cdot \mathrm{e}^{-\lambda ct} + \lambda^2 c \left(\left(\frac{t}{-\lambda c}\, \mathrm{e}^{-\lambda ct} \right) + \frac{1}{\lambda c} \left[\frac{\mathrm{e}^{-\lambda cx}}{-\lambda c} \right]_0^t \right) =$$

$$= \lambda t \cdot \mathrm{e}^{-\lambda ct} + \lambda^2 c \left(\frac{t}{-\lambda c}\, \mathrm{e}^{-\lambda ct} + \frac{1}{\lambda c} \left(\frac{\mathrm{e}^{-\lambda ct}}{-\lambda c} - \frac{1}{-\lambda c} \right) \right) =$$

$$= \lambda t \cdot \mathrm{e}^{-\lambda ct} + \lambda^2 c \left(-\frac{t}{\lambda c}\, \mathrm{e}^{-\lambda ct} - \frac{1}{\lambda^2 c^2}\, \mathrm{e}^{-\lambda ct} + \frac{1}{\lambda^2 c^2} \right) =$$

$$= \frac{1 - \mathrm{e}^{-\lambda ct}}{c} =$$

$$= \frac{1 - \mathrm{e}^{-\lambda t[1-(1-\mu)(1-q_p)-\mu(1-\epsilon)(1-q_{c2})-\mu\epsilon(1-q_{c1})]}}{1 - (1-\mu)(1-q_p) - \mu(1-\epsilon)(1-q_{c2}) - \mu\epsilon(1-q_{c1})}$$

DOI: 10.14751/SZIE.2014.035

A.13. Az inaktívvá válás előrejelzésének összehasonlítása (R kód).

```
library(epicalc) # kappa statisztika

szamlalo <- 0
osszesito <- data.frame()
eredmeny <- data.frame()

Ti <- 2
avk <- as.vector(c(2,3,6))
ucl <- as.vector(c(4,6,8))
for (jj in 1:3){
alp <- c(0.5,1,2)
Ej <- Ti*alp
for (kk in 1:3){
for (ll in 1:3){
szamlalo <- szamlalo+1
aa <- 0
repeat{
aa <- aa+1
if (aa > 10) {break}
rv <- 2
alphav <- 5
upv <- 2
vpv <- 25
av <- avk[kk]
bv <- 10
uc1v <- ucl[ll]
vc1v <- 20
uc2v <- ucl[ll]
vc2v <- 15
ev <- 5
fv <- 2
parameters <- data.frame()
parameters[1:14,1] <- c("r","alpha","up","vp","a","b","uc1","vc1","uc2",
"vc2","e","f","A","B")
parameters[1:12,2] <- c(rv,alphav,upv,vpv,av,bv,uc1v,vc1v,uc2v,vc2v,ev,fv)
T_min <- Ti  # a vizsgált időtartam minimuma
T_max <- Ti # a vizsgált időtartam maximuma (egyenletes eloszlás)
max_vasar <- 1000   # vásárlások maximális száma személyenként
E <- Ej[jj]  # előrejelzés időtartama

## adatmátrix kezd
sor <- 1000
oszlop <- 9
dat <- as.data.frame(matrix(nrow = sor, ncol = oszlop))
colnames(dat)[1] <- "ID"
ID <- c(1:sor)
dat[,1] <- ID
colnames(dat)[2] <- "lambda"
GammaSamples <- as.data.frame(matrix(rgamma(sor*1, shape=rv, scale=alphav), ncol=1))
dat[,2] <- GammaSamples
colnames(dat)[3] <- "qp"
BetaSamples <- as.data.frame(matrix(rbeta(sor*1, shape1=upv, shape2=vpv), ncol=1))
dat[,3] <- BetaSamples
colnames(dat)[4] <- "mu"
BetaSamples <- as.data.frame(matrix(rbeta(sor*1, shape1=av, shape2=bv), ncol=1))
```

DOI: 10.14751/SZIE.2014.035


```
dat[,4] <- BetaSamples
colnames(dat)[5] <- "eps"
BetaSamples <- as.data.frame(matrix(rbeta(sor*1, shape1=ev, shape2=fv), ncol=1))
dat[,5] <- BetaSamples
colnames(dat)[6] <- "qc1"
BetaSamples <- as.data.frame(matrix(rbeta(sor*1, shape1=uc1v, shape2=vc1v), ncol=1))
dat[,6] <- BetaSamples
colnames(dat)[7] <- "qc2"
BetaSamples <- as.data.frame(matrix(rbeta(sor*1, shape1=uc2v, shape2=vc2v), ncol=1))
dat[,7] <- BetaSamples
colnames(dat)[8] <- "T"
UniformSamples <- as.data.frame(matrix(runif(sor*1, min=T_min, max=T_max), ncol=1))
dat[,8] <- round(UniformSamples, digits=2)
l1 <- list()
l11 <- list()
for (i in 1:sor) {
 l1[i] <- as.data.frame(matrix(rbinom(1*max_vasar, size=1, prob=dat$mu[i]), ncol=1))
 l11[i] <- as.data.frame(matrix(rbinom(1*max_vasar, size=1, prob=dat$eps[i]), ncol=1))
}
m <- vector()
l2 <- list()
for (i in 1:sor){
for (j in 1:max_vasar){
if (l1[[i]][j]==0) { m[j] <- as.numeric(matrix(rbinom(1*1, size=1, prob=dat$qp[i]), ncol=1))}
else { if (l11[[i]][j]==0) {m[j] <-
as.numeric(matrix(rbinom(1*1, size=1, prob=dat$qc2[i]), ncol=1))}
else {m[j] <- as.numeric(matrix(rbinom(1*1, size=1, prob=dat$qc1[i]), ncol=1))} }
}
l2[i] <- as.data.frame(m)
}
for (i in 1:sor){
sz <- 0
for (j in 1:max_vasar){
if (l2[[i]][j]==1) {tlive <- sz+1
break}
else {sz <- sz+1}
}
dat[i,9] <- tlive
}
colnames(dat)[9] <- "tlive"
for (i in 1:sor) {
j <- 0
k <- 0
t_ij <- 0
ipt <- 0
repeat{
repeat{
ipt <- rexp(1*1, rate=dat[i,2])
if (ipt <= dat$T[i]) {break}
}
j <- j+1
k <- k+1
t_ij <- t_ij + ipt
  if (t_ij < dat$T[i] & j <= dat$tlive[i]) {
dat[i,10] <- round(t_ij, digits=2)
dat[i,11] <- j         # EREDETILEG J
  }
  if (t_ij < E+dat$T[i] & k <= dat$tlive[i]) {
dat[i,12] <- k-dat[i,11]
```

DOI: 10.14751/SZIE.2014.035


```
  } else {break}

  #cat("ipt =" , ipt , " , t_ij =" , dat[i,9] , "\n")
}
}
colnames(dat)[10] <- "tx"  # utolsó vásárlás időpontja
colnames(dat)[11] <- "x"   # vásárlások száma
colnames(dat)[12] <- "y"   # jövöbeli vásárlások száma
## adatmátrix vég


for (i in 1:sor){
j <- 0
n <- vector()
repeat{
j <- j+1
if (l1[[i]][j]==1 & j < dat$x[i]) {
  if (l11[[i]][j]==0) {n[j] <- 0 } else {n[j] <- 1}
 }
 else
 { if (l1[[i]][j]==0 & j < dat$x[i]){n[j] <- -1} else {break}}
}
dat[i,13] <- sum(n==1)
dat[i,14] <- sum(n==0)
}
colnames(dat)[13] <- "xc1"  # kezelt panaszok száma
colnames(dat)[14] <- "xc2"  # nem kezelt panaszok száma
z <- vector()
z1 <- vector()
z2 <- vector()
for (i in 1:sor){
if (l1[[i]][dat$x[i]]==0) {z[i] <- 1; z1[i] <- 0 ; z2[i] <- 0 } else {z[i] <- 0}
if (l11[[i]][dat$x[i]]==0 & l1[[i]][dat$x[i]]==1) {z2[i] <- 1 ; z1[i] <- 0 }
if (l11[[i]][dat$x[i]]==1 & l1[[i]][dat$x[i]]==1) {z1[i] <- 1 ; z2[i] <- 0 }
dat[i,15] <- z[i]
dat[i,16] <- z1[i]
dat[i,17] <- z2[i]
}
colnames(dat)[15] <- "z"
colnames(dat)[16] <- "z1"
colnames(dat)[17] <- "z2"

###### Likelihood függvény
#param : r, alpha, up, vp, a, b, uc1, vc1, uc2, vc2, e, f  (sajat)
x <- dat$x
T <- dat$T
tx <- dat$tx
xc1 <- dat$xc1
xc2 <- dat$xc2
z <- dat$z
z1 <- dat$z1
z2 <- dat$z2
param <- vector()
sajatll <- function(param){
r <- param[1]
alpha <- param[2]
up <- param[3]
vp <- param[4]
a <- param[5]
```

DOI: 10.14751/SZIE.2014.035


```
b <- param[6]
uc1 <- param[7]
vc1 <- param[8]
uc2 <- param[9]
vc2 <- param[10]
e <- param[11]
f <- param[12]

L_aktiv <- gamma(r+x)*alpha^r/(gamma(r)*(alpha+T)^(r+x))*beta(up,x-1-xc1-xc2+vp+z)/
beta(up,vp)*beta(xc1+xc2+a,x-1-xc1-xc2+b)/beta(a,b)*beta(uc1,xc1+vc1+z1)/
beta(uc1,vc1)*beta(uc2,xc2+vc2+z2)/beta(uc2,vc2)*beta(xc1+e,xc2+f)/beta(e,f)

L_inaktiv <- gamma(r+x)*alpha^r/(gamma(r)*(alpha+tx)^(r+x))*beta(up+z,x-1-xc1-xc2+vp)/
beta(up,vp)*beta(xc1+xc2+a,x-1-xc1-xc2+b)/beta(a,b)*beta(uc1+z1,xc1+vc1)/
beta(uc1,vc1)*beta(uc2+z2,xc2+vc2)/beta(uc2,vc2)*beta(xc1+e,xc2+f)/beta(e,f)

#cat("L_aktiv =" , L_aktiv ,  "\n")
#cat("L_inaktiv =" , L_inaktiv ,  "\n")
ll1 <-sum(log(L_aktiv + L_inaktiv))
return(-ll1)
}
fit1 <- optim(c(5,5,5,5,5,5,5,5,5,5,5,5), sajatll)
parameters[1:12,3] <- round(as.data.frame(fit1$par),digits=2)

# előrejelzés (saját)
r <- fit1$par[1]
alpha <- fit1$par[2]
up <- fit1$par[3]
vp <- fit1$par[4]
a <- fit1$par[5]
b <- fit1$par[6]
uc1 <- fit1$par[7]
vc1 <- fit1$par[8]
uc2 <- fit1$par[9]
vc2 <- fit1$par[10]
e <- fit1$par[11]
f <- fit1$par[12]
sor2 <- 1000
oszlop2 <- 6
dat2 <- as.data.frame(matrix(nrow = sor2, ncol = oszlop2))
# a priori  sűrűségfüggvényekből számolt paraméterek
colnames(dat2)[1] <- "lambda"
GammaSamples <- as.data.frame(matrix(rgamma(sor2*1, shape=r, scale=alpha), ncol=1))
dat2[,1] <- GammaSamples
colnames(dat2)[2] <- "qp"
BetaSamples <- as.data.frame(matrix(rbeta(sor2*1, shape1=up, shape2=vp), ncol=1))
dat2[,2] <- BetaSamples
colnames(dat2)[3] <- "mu"
BetaSamples <- as.data.frame(matrix(rbeta(sor2*1, shape1=a, shape2=b), ncol=1))
dat2[,3] <- BetaSamples
colnames(dat2)[4] <- "eps"
BetaSamples <- as.data.frame(matrix(rbeta(sor2*1, shape1=e, shape2=f), ncol=1))
dat2[,4] <- BetaSamples
colnames(dat2)[5] <- "qc1"
BetaSamples <- as.data.frame(matrix(rbeta(sor2*1, shape1=uc1, shape2=vc1), ncol=1))
dat2[,5] <- BetaSamples
colnames(dat2)[6] <- "qc2"
BetaSamples <- as.data.frame(matrix(rbeta(sor2*1, shape1=uc2, shape2=vc2), ncol=1))
dat2[,6] <- BetaSamples
```

DOI: 10.14751/SZIE.2014.035


```
for (i in 1:sor){
cikl <- 0
repeat{
cikl <- cikl+1
L_aktiv2 <- gamma(r+x[i])*alpha^r/(gamma(r)*(alpha+T[i])^(r+x[i]))*
beta(up,x[i]-1-xc1[i]-xc2[i]+vp+z[i])/beta(up,vp)*beta(xc1[i]+xc2[i]+
a,x[i]-1-xc1[i]-xc2[i]+b)/beta(a,b)*beta(uc1,xc1[i]+vc1+z1[i])/beta(uc1,vc1)*
beta(uc2,xc2[i]+vc2+z2[i])/beta(uc2,vc2)*beta(xc1[i]+e,xc2[i]+f)/beta(e,f)

L_inaktiv2 <- gamma(r+x[i])*alpha^r/(gamma(r)*(alpha+tx[i])^(r+x[i]))*beta(up+z[i],x[i]-1-
xc1[i]-xc2[i]+vp)/beta(up,vp)*beta(xc1[i]+xc2[i]+a,x[i]-1-xc1[i]-xc2[i]+b)/beta(a,b)*
beta(uc1+z1[i],xc1[i]+vc1)/beta(uc1,vc1)*beta(uc2+z2[i],xc2[i]+vc2)/beta(uc2,vc2)*
beta(xc1[i]+e,xc2[i]+f)/beta(e,f)

L <-L_aktiv2+L_inaktiv2
C <- 1-(1-dat2$mu)*(1-dat2$qp)-dat2$mu*(1-dat2$eps)*(1-dat2$qc2)-dat2$mu*
dat2$eps*(1-dat2$qc1)
L_aktiv1 <- (dat2$lambda)^x[i]*exp(-dat2$lambda*T[i])*(dat2$mu)^(xc1[i]+xc2[i])*
(1-dat2$mu)^(x[i]-1-xc1[i]-xc2[i])*(dat2$eps)^xc1[i]*(1-dat2$eps)^xc2[i]*
(1-dat2$qp)^(x[i]-1-xc1[i]-xc2[i]+z[i])*(1-dat2$qc1)^(xc1[i]+z1[i])*
(1-dat2$qc2)^(xc2[i]+z2[i])
M <- mean((1/C-(1/C)*exp(-dat2$lambda*C*E))*L_aktiv1)
exp_val <- M/L
dat[i,18] <- round(exp_val,digits=0)
if (exp_val != "NaN") {break}
if (cikl > 1000) { print("ciklus probléma")
break}
}
}
colnames(dat)[18] <- "pred1y"


############################  param : r, alpha, a, b (BG/NBD)
x <- dat$x
T <- dat$T
tx <- dat$tx
param2 <- vector()

bgll <- function(param2,dat){
r <- param2[1]
alpha <- param2[2]
A <- param2[3]
B <- param2[4]

A1 <- gamma(r+x)*alpha^r/gamma(r)
A2 <- gamma(A+B)*gamma(B+x)/(gamma(B)*gamma(A+B+x))
A3 <- (1/(alpha+T))^(r+x)
A4 <- (A/(B+x-1))*(1/(alpha+tx))^(r+x)
for (i in 1:sor) {
if (x[i]>0) {A4[i] <- A4[i]} else {A4[i] <- 0}
}
ll2 <- sum(log(A1*A2*(A3+A4)))

return(-ll2)
}
fit2 <- optim(c(5,5,5,5), bgll)
parameters[1:2,4] <- round(as.data.frame(fit2$par[1:2]),digits=2)
parameters[13:14,4] <- round(as.data.frame(fit2$par[3:4]),digits=2)
```

DOI: 10.14751/SZIE.2014.035


```
colnames(parameters) <- c("név", "valódi", "saját", "BG/NBD")
parameters[13:14,2:3] <- ""
parameters[3:12,4] <- ""


########## előrejelzés (BG/NBD)
library(hypergeo)
r <- fit2$par[1]
alpha <- fit2$par[2]
A <- fit2$par[3]
B <- fit2$par[4]

for (i in 1:sor){
B1 <- (A+B+x[i]-1)/(A-1)
B2 <- 1-((alpha+T[i])/(alpha+T[i]+E))^(r+x[i])*Re(hypergeo(r+x[i],B+x[i],A+B+x[i]-1,E/
(alpha+T[i]+E)))
if (x[i]>0) {B3 <- 1+(A/(B+x[i]-1))*((alpha+T[i])/(alpha+tx[i]))^(r+x[i])} else {B3 <- 1}
dat[i,19] <- round(Re(B1*B2/B3))
}
colnames(dat)[19] <- "pred2y"
###### naive (annyi, amennyi a vizsgált időszakból következik)
subdat <-dat[dat$y==0,]
hiatus <- mean(subdat$tx)
for (i in 1:sor){
if (dat$tx[i]<hiatus) {dat$naive[i] <- 0} else {dat$naive[i] <-
round(dat$x[i]*E/dat$T[i],digits=0)}
}
teny <- vector()
pred1 <- vector()
pred2 <- vector()
naive <- vector()
for (i in 1:sor){
teny[i]<- if (dat$y[i]==0) {0} else {1}
pred1[i]<- if (dat$pred1y[i]==0) {0} else {1} # elpártoltnak tekinti-e
pred2[i]<- if (dat$pred2y[i]==0) {0} else {1}
naive[i]<- if (dat$naive[i]==0) {0} else {1}
}
tab1 <- table(teny,pred1)
tab2 <- table(teny,pred2)
tab3 <- table(teny,naive)
osszesito[aa,1] <- mean(dat$tx)
osszesito[aa,2] <- mean(dat$x)
osszesito[aa,3] <- mean(dat$T)
osszesito[aa,4] <- E
osszesito[aa,5] <- mean(dat$xc1)
osszesito[aa,6] <- mean(dat$xc2)
osszesito[aa,7] <- tryCatch(kap(tab1)$kappa, error=function(e) NA)
osszesito[aa,8] <- tryCatch(kap(tab2)$kappa, error=function(e) NA)
osszesito[aa,9] <- tryCatch(kap(tab3)$kappa, error=function(e) NA)
}
eredmeny <- rbind(eredmeny,colMeans(osszesito,na.rm=TRUE))
print(szamlalo)
}
}
}
colnames(eredmeny) <- c("tx","x","T","E","xc1","xc2","K1","K2","K3")
eredmeny
```

DOI: 10.14751/SZIE.2014.035

A.14. A Kappa statisztikák értékei az egyes modellek esetében az egyes vásárlók megfigyelési időszakra vonatkozó paramétereinek függvényében. Forrás: saját számítás.

	t_x	x	T	t	x_{c1}	x_{c2}	$K1$	$K2$	$K3$
1	0,33	4,03	0,50	0,25	0,33	0,13	0,30	0,31	0,33
2	0,33	3,95	0,50	0,25	0,29	0,11	0,31	0,33	0,34
3	0,32	3,86	0,50	0,25	0,27	0,09	0,32	0,35	0,35
4	0,33	3,98	0,50	0,25	0,45	0,17	0,32	0,32	0,33
5	0,32	3,81	0,50	0,25	0,39	0,15	0,34	0,36	0,36
6	0,31	3,66	0,50	0,25	0,35	0,12	0,36	0,37	0,35
7	0,32	3,83	0,50	0,25	0,71	0,27	0,32	0,34	0,34
8	0,31	3,56	0,50	0,25	0,62	0,22	0,36	0,36	0,33
9	0,30	3,37	0,50	0,25	0,54	0,19	0,38	0,38	0,35
10	0,33	4,04	0,50	0,50	0,33	0,12	0,21	0,24	0,47
11	0,33	3,99	0,50	0,50	0,28	0,11	0,26	0,26	0,48
12	0,32	3,79	0,50	0,50	0,26	0,10	0,32	0,30	0,47
13	0,33	3,95	0,50	0,50	0,44	0,17	0,30	0,26	0,49
14	0,32	3,90	0,50	0,50	0,41	0,15	0,31	0,29	0,47
15	0,31	3,68	0,50	0,50	0,36	0,13	0,37	0,32	0,48
16	0,32	3,79	0,50	0,50	0,70	0,27	0,35	0,32	0,48
17	0,31	3,53	0,50	0,50	0,59	0,22	0,44	0,33	0,46
18	0,29	3,38	0,50	0,50	0,54	0,19	0,43	0,37	0,45
19	0,33	4,08	0,50	1,00	0,34	0,12	0,13	0,04	0,47
20	0,33	3,93	0,50	1,00	0,28	0,10	0,19	0,09	0,48
21	0,32	3,82	0,50	1,00	0,26	0,10	0,24	0,04	0,46
22	0,33	3,97	0,50	1,00	0,45	0,17	0,26	0,07	0,49
23	0,32	3,82	0,50	1,00	0,38	0,15	0,21	0,14	0,48
24	0,31	3,70	0,50	1,00	0,35	0,13	0,22	0,15	0,48
25	0,32	3,82	0,50	1,00	0,69	0,27	0,29	0,12	0,47
26	0,31	3,55	0,50	1,00	0,60	0,23	0,36	0,15	0,46
27	0,29	3,41	0,50	1,00	0,55	0,20	0,38	0,25	0,45
28	0,61	6,14	1,00	0,50	0,53	0,20	0,55	0,55	0,45
29	0,59	5,88	1,00	0,50	0,47	0,17	0,57	0,56	0,44
30	0,57	5,51	1,00	0,50	0,42	0,14	0,56	0,56	0,43
31	0,60	5,94	1,00	0,50	0,72	0,28	0,47	0,55	0,44
32	0,56	5,46	1,00	0,50	0,61	0,23	0,47	0,56	0,43
33	0,55	5,28	1,00	0,50	0,55	0,20	0,55	0,56	0,42
34	0,57	5,47	1,00	0,50	1,10	0,41	0,54	0,55	0,42
35	0,53	4,99	1,00	0,50	0,93	0,34	0,51	0,55	0,39
36	0,50	4,56	1,00	0,50	0,79	0,28	0,54	0,55	0,38
37	0,61	6,14	1,00	1,00	0,55	0,20	0,58	0,54	0,52
38	0,59	5,79	1,00	1,00	0,45	0,17	0,59	0,55	0,50
39	0,58	5,65	1,00	1,00	0,41	0,14	0,46	0,54	0,49
40	0,60	5,98	1,00	1,00	0,73	0,28	0,54	0,53	0,49
41	0,57	5,51	1,00	1,00	0,62	0,23	0,57	0,54	0,48
42	0,55	5,25	1,00	1,00	0,53	0,19	0,55	0,56	0,46
43	0,57	5,47	1,00	1,00	1,11	0,41	0,58	0,55	0,48
44	0,53	4,96	1,00	1,00	0,93	0,33	0,52	0,55	0,45
45	0,50	4,56	1,00	1,00	0,79	0,28	0,55	0,57	0,43
46	0,62	6,20	1,00	2,00	0,55	0,21	0,45	0,37	0,51
47	0,59	5,83	1,00	2,00	0,47	0,17	0,50	0,36	0,49
48	0,58	5,61	1,00	2,00	0,40	0,15	0,49	0,43	0,48
49	0,60	5,99	1,00	2,00	0,74	0,27	0,46	0,40	0,50
50	0,56	5,46	1,00	2,00	0,61	0,23	0,53	0,48	0,47
51	0,55	5,21	1,00	2,00	0,54	0,19	0,49	0,27	0,47

DOI: 10.14751/SZIE.2014.035

	t_x	x	T	t	x_{c1}	x_{c2}	$K1$	$K2$	$K3$
52	0,57	5,51	1,00	2,00	1,13	0,40	0,54	0,26	0,48
53	0,53	4,99	1,00	2,00	0,94	0,35	0,48	0,38	0,45
54	0,51	4,63	1,00	2,00	0,81	0,29	0,46	0,45	0,43
55	1,02	8,69	2,00	1,00	0,80	0,31	0,51	0,70	0,44
56	0,96	8,14	2,00	1,00	0,68	0,25	0,63	0,70	0,40
57	0,92	7,56	2,00	1,00	0,57	0,20	0,68	0,68	0,40
58	0,98	8,23	2,00	1,00	1,04	0,40	0,47	0,70	0,42
59	0,90	7,42	2,00	1,00	0,87	0,32	0,64	0,69	0,38
60	0,85	6,77	2,00	1,00	0,72	0,27	0,63	0,67	0,37
61	0,90	7,43	2,00	1,00	1,58	0,59	0,60	0,67	0,37
62	0,80	6,37	2,00	1,00	1,24	0,45	0,52	0,65	0,33
63	0,75	5,68	2,00	1,00	1,02	0,37	0,59	0,64	0,30
64	1,03	8,97	2,00	2,00	0,82	0,31	0,52	0,68	0,46
65	0,97	8,00	2,00	2,00	0,66	0,25	0,55	0,69	0,44
66	0,93	7,66	2,00	2,00	0,58	0,22	0,59	0,67	0,42
67	0,98	8,35	2,00	2,00	1,07	0,41	0,48	0,68	0,43
68	0,90	7,38	2,00	2,00	0,83	0,32	0,64	0,67	0,41
69	0,85	6,85	2,00	2,00	0,74	0,26	0,61	0,65	0,38
70	0,91	7,44	2,00	2,00	1,59	0,59	0,69	0,66	0,40
71	0,81	6,40	2,00	2,00	1,23	0,45	0,55	0,65	0,36
72	0,74	5,64	2,00	2,00	1,02	0,35	0,60	0,63	0,33
73	1,02	8,64	2,00	4,00	0,80	0,30	0,52	0,39	0,45
74	0,96	8,13	2,00	4,00	0,66	0,24	0,55	0,40	0,43
75	0,93	7,63	2,00	4,00	0,56	0,21	0,61	0,39	0,42
76	0,98	8,31	2,00	4,00	1,07	0,40	0,48	0,04	0,44
77	0,92	7,52	2,00	4,00	0,90	0,31	0,60	0,32	0,41
78	0,86	6,91	2,00	4,00	0,74	0,27	0,64	0,34	0,39
79	0,91	7,42	2,00	4,00	1,58	0,59	0,65	0,54	0,40
80	0,80	6,32	2,00	4,00	1,21	0,45	0,48	0,53	0,37
81	0,74	5,70	2,00	4,00	1,04	0,37	0,60	0,45	0,33

$K1$: a saját modell Kappa statisztikái
$K2$: a BG/NBD modell Kappa statisztikái
$K3$: a tapasztalati modell Kappa statisztikái

DOI: 10.14751/SZIE.2014.035

A.15. A MAE indexek ill. ezekből számolt mutatók értékei az egyes modellek esetében az egyes vásárlók megfigyelési időszakra vonatkozó paramétereinek függvényében - 4.2.4. alszakasz, 84. old. Forrás: saját számítás.

	T	t	x	x_{c1}	x_{c2}	t_x	Saját	BG/NBD	Heuriszt.	jobb-e	szign	sz.j.	sz. r.
1	0,50	0,25	4,06	0,34	0,12	0,34	1,25	1,132	1,30	0,20	0,80	0,10	0,70
2	0,50	0,25	4,01	0,29	0,11	0,33	1,21	1,093	1,27	0,20	0,70	0,00	0,70
3	0,50	0,25	3,87	0,26	0,10	0,32	1,14	1,050	1,23	0,20	0,80	0,10	0,70
4	0,50	0,25	3,99	0,44	0,18	0,33	1,12	1,092	1,26	0,40	0,40	0,20	0,20
5	0,50	0,25	3,86	0,39	0,15	0,32	1,07	1,032	1,23	0,30	0,40	0,10	0,30
6	0,50	0,25	3,75	0,35	0,13	0,32	1,03	0,985	1,18	0,10	0,30	0,10	0,20
7	0,50	0,25	3,87	0,71	0,27	0,32	1,28	1,025	1,20	0,20	0,70	0,10	0,60
8	0,50	0,25	3,67	0,63	0,23	0,31	1,26	0,903	1,14	0,20	0,40	0,10	0,30
9	0,50	0,25	3,45	0,56	0,19	0,29	0,85	0,822	1,07	0,30	0,40	0,10	0,30
10	0,50	0,50	4,06	0,34	0,12	0,34	2,05	1,904	2,29	0,50	0,70	0,20	0,50
11	0,50	0,50	3,97	0,30	0,11	0,33	1,99	1,848	2,26	0,20	0,60	0,00	0,60
12	0,50	0,50	3,89	0,27	0,10	0,32	2,00	1,740	2,19	0,10	0,70	0,00	0,70
13	0,50	0,50	4,00	0,46	0,17	0,33	1,90	1,856	2,24	0,40	0,50	0,20	0,30
14	0,50	0,50	3,92	0,41	0,15	0,32	1,85	1,740	2,22	0,20	0,70	0,20	0,50
15	0,50	0,50	3,79	0,37	0,13	0,32	1,83	1,620	2,14	0,30	0,70	0,10	0,60
16	0,50	0,50	3,88	0,72	0,26	0,32	2,37	1,741	2,25	0,50	0,70	0,20	0,50
17	0,50	0,50	3,63	0,63	0,22	0,31	1,54	1,581	2,12	0,70	0,40	0,30	0,10
18	0,50	0,50	3,44	0,54	0,21	0,30	1,34	1,401	1,98	0,80	0,40	0,40	0,00
19	0,50	1,00	4,11	0,34	0,13	0,34	3,81	3,376	4,70	0,30	0,70	0,00	0,70
20	0,50	1,00	3,97	0,30	0,10	0,33	3,31	3,189	4,51	0,30	0,70	0,20	0,50
21	0,50	1,00	3,92	0,28	0,10	0,33	3,22	3,168	4,50	0,40	0,60	0,30	0,30
22	0,50	1,00	4,04	0,46	0,17	0,33	3,25	3,271	4,59	0,60	0,80	0,50	0,30
23	0,50	1,00	3,90	0,42	0,14	0,32	3,08	2,974	4,46	0,40	0,90	0,40	0,50
24	0,50	1,00	3,79	0,36	0,13	0,31	3,00	2,846	4,42	0,10	0,40	0,00	0,40
25	0,50	1,00	3,87	0,72	0,28	0,32	3,58	3,049	4,45	0,20	0,40	0,00	0,40
26	0,50	1,00	3,61	0,61	0,23	0,31	2,51	2,559	4,26	0,70	0,40	0,30	0,10
27	0,50	1,00	3,46	0,55	0,19	0,29	2,17	2,279	4,14	0,70	0,70	0,60	0,10
28	1,00	0,50	6,46	0,58	0,22	0,62	1,46	1,270	1,77	0,00	0,70	0,00	0,70
29	1,00	0,50	6,11	0,51	0,19	0,59	1,32	1,135	1,67	0,10	0,60	0,00	0,60
30	1,00	0,50	5,94	0,43	0,16	0,58	1,16	1,055	1,61	0,20	0,60	0,10	0,50
31	1,00	0,50	6,22	0,77	0,30	0,60	1,29	1,190	1,72	0,40	0,50	0,20	0,30
32	1,00	0,50	5,92	0,66	0,25	0,58	1,15	1,061	1,67	0,20	0,60	0,10	0,50
33	1,00	0,50	5,57	0,58	0,21	0,54	1,06	0,933	1,61	0,30	0,60	0,00	0,60
34	1,00	0,50	5,76	1,18	0,44	0,57	1,31	1,014	1,62	0,30	0,60	0,10	0,50
35	1,00	0,50	5,27	1,00	0,36	0,53	0,94	0,841	1,52	0,30	0,60	0,00	0,60
36	1,00	0,50	4,83	0,85	0,31	0,49	0,77	0,697	1,36	0,10	0,60	0,00	0,60
37	1,00	1,00	6,46	0,57	0,22	0,62	2,45	2,169	3,46	0,00	0,60	0,00	0,60
38	1,00	1,00	6,19	0,50	0,18	0,60	2,28	2,069	3,40	0,20	0,80	0,10	0,70
39	1,00	1,00	5,85	0,44	0,16	0,57	2,11	1,815	3,19	0,20	0,60	0,10	0,50
40	1,00	1,00	6,32	0,78	0,30	0,60	2,29	2,041	3,46	0,10	0,50	0,10	0,40
41	1,00	1,00	5,85	0,66	0,24	0,57	1,84	1,755	3,27	0,20	0,40	0,00	0,40
42	1,00	1,00	5,52	0,56	0,21	0,55	1,80	1,586	3,12	0,40	0,40	0,10	0,30
43	1,00	1,00	5,80	1,21	0,44	0,57	2,24	1,758	3,25	0,00	0,90	0,00	0,90
44	1,00	1,00	5,30	1,01	0,37	0,52	1,60	1,377	3,02	0,30	0,90	0,20	0,70
45	1,00	1,00	4,90	0,87	0,31	0,49	1,34	1,120	2,75	0,30	0,80	0,10	0,70
46	1,00	2,00	6,49	0,58	0,22	0,62	4,35	M	7,55	0,60	0,70	0,60	0,10
47	1,00	2,00	6,09	0,50	0,17	0,59	3,77	M	7,15	0,80	0,80	0,80	0,00
48	1,00	2,00	5,94	0,43	0,15	0,58	3,48	M	7,00	0,90	1,00	0,90	0,10
49	1,00	2,00	6,24	0,77	0,29	0,60	4,34	M	7,33	1,00	1,00	1,00	0,00
50	1,00	2,00	5,77	0,64	0,24	0,56	3,21	M	6,83	1,00	1,00	1,00	0,00

DOI: 10.14751/SZIE.2014.035

	T	t	x	x_{c1}	x_{c2}	t_x	Saját	BG/NBD	Heuriszt.	jobb-e	szign	sz.j.	sz. r.
51	1,00	2,00	5,48	0,58	0,20	0,54	2,75	M	6,58	1,00	1,00	1,00	0,00
52	1,00	2,00	5,85	1,21	0,46	0,57	3,64	M	7,15	1,00	1,00	1,00	0,00
53	1,00	2,00	5,30	1,00	0,38	0,52	2,27	M	6,40	0,90	1,00	0,90	0,10
54	1,00	2,00	4,82	0,85	0,30	0,49	1,87	M	5,77	1,00	1,00	1,00	0,00
55	2,00	1,00	9,28	0,85	0,32	1,00	1,96	1,131	2,46	0,20	0,80	0,00	0,80
56	2,00	1,00	8,64	0,71	0,26	0,93	1,36	0,954	2,29	0,10	0,80	0,10	0,70
57	2,00	1,00	8,15	0,60	0,22	0,88	1,41	0,847	2,25	0,00	0,90	0,00	0,90
58	2,00	1,00	8,71	1,12	0,42	0,94	1,79	0,952	2,36	0,10	0,90	0,00	0,90
59	2,00	1,00	7,96	0,94	0,34	0,86	0,97	0,783	2,16	0,20	0,60	0,00	0,60
60	2,00	1,00	7,50	0,80	0,29	0,83	1,08	0,673	2,03	0,20	0,40	0,00	0,40
61	2,00	1,00	7,83	1,69	0,62	0,85	1,79	0,763	2,16	0,10	0,50	0,00	0,50
62	2,00	1,00	6,89	1,36	0,51	0,76	0,83	0,552	1,83	0,30	0,70	0,00	0,70
63	2,00	1,00	6,02	1,10	0,39	0,68	0,63	0,385	1,60	0,30	0,80	0,10	0,70
64	2,00	2,00	9,37	0,87	0,32	0,99	3,19	2,009	5,20	0,10	0,70	0,00	0,70
65	2,00	2,00	8,75	0,73	0,28	0,94	2,63	1,715	4,96	0,00	0,90	0,00	0,90
66	2,00	2,00	8,21	0,62	0,23	0,88	2,47	1,431	4,63	0,10	0,80	0,00	0,80
67	2,00	2,00	8,73	1,12	0,42	0,94	2,24	1,661	4,86	0,10	0,70	0,00	0,70
68	2,00	2,00	7,95	0,93	0,33	0,87	1,60	1,335	4,51	0,10	0,50	0,10	0,40
69	2,00	2,00	7,30	0,79	0,29	0,80	1,30	1,093	4,10	0,20	0,40	0,00	0,40
70	2,00	2,00	7,91	1,69	0,63	0,87	1,85	1,310	4,53	0,10	0,30	0,00	0,30
71	2,00	2,00	6,75	1,32	0,46	0,75	1,54	0,781	3,81	0,10	0,80	0,10	0,70
72	2,00	2,00	5,98	1,10	0,39	0,68	1,00	0,621	3,21	0,20	0,80	0,20	0,60
73	2,00	4,00	9,43	0,86	0,32	1,00	4,70	M	11,60	0,90	0,90	0,80	0,10
74	2,00	4,00	8,63	0,71	0,26	0,93	3,38	M	10,68	0,90	0,80	0,80	0,00
75	2,00	4,00	8,28	0,64	0,22	0,89	3,61	M	10,18	0,90	0,90	0,90	0,00
76	2,00	4,00	8,81	1,12	0,43	0,94	3,62	M	10,79	1,00	1,00	1,00	0,00
77	2,00	4,00	7,96	0,92	0,34	0,87	3,12	M	9,53	0,90	1,00	0,90	0,10
78	2,00	4,00	7,49	0,80	0,29	0,82	2,32	M	9,07	0,90	0,80	0,80	0,00
79	2,00	4,00	7,96	1,71	0,62	0,86	2,06	M	9,62	1,00	1,00	1,00	0,00
80	2,00	4,00	6,93	1,35	0,52	0,78	1,81	M	8,23	0,90	0,90	0,90	0,00
81	2,00	4,00	6,01	1,10	0,38	0,68	1,14	M	6,75	1,00	1,00	1,00	0,00

Saját: a saját modell MAE értékei
BG/NBD: a BG/NBD modell MAE értékei
Heuriszt.: a heurisztikus modell MAE értékei
jobb-e: a saját modell a 10 ismétlés hányad részében bizonyult jobbnak a BG/NBD modellnél.
szign.: a 10 kísérlet hányad részében mutatható ki különbség a saját és a BG/NBD modell között.
sz.j.: a 10 kísérlet hányad részében bizonyult szignifikánsan jobbnak a saját modell.
sz.r.: a 10 kísérlet hányad részében bizonyult szignifikánsan rosszabbnak a saját modell.
M: az érték nagyon nagy (több nagyságrend eltérés).

DOI: 10.14751/SZIE.2014.035

A.16. A párosított t-próba, valamint annak feltételeit ellenőrző próbák eredményei (R output) - 4.2.4. alszakasz, 84. old. Forrás: saját számítás.

1. Shapiro-Wilks teszt
   (a) $t/T = 0,5$
   data: sub1$sajat, W = 0,9517, p-value = 0,2352
   data: sub1$BG.NBD, W = 0,945, p-value = 0,1616
   (b) $t/T = 1$
   data: sub2$sajat, W = 0,9812, p-value = 0,889
   data: sub2$BG.NBD, W = 0,9339, p-value = 0,08634
   (c) $t/T = 2$
   data: sub3$sajat, W = 0,948, p-value = 0,6684
   data: sub3$BG.NBD, W = 0,918, p-value = 0,3761
2. F-próba
   (a) $t/T = 0,5$
   data: sub1$sajat and sub1$BG.NBD, F = 2,2663, num df = 26, denom df = 26, p-value = 0,04151
   (b) $t/T = 1$
   data: sub2$sajat and sub2$BG.NBD, F = 1,5932, num df = 26, denom df = 26, p-value = 0,2417
   (c) $t/T = 2$
   data: sub3$sajat and sub3$BG.NBD, F = 2,0075, num df = 8, denom df = 8, p-value = 0,3441
3. t-próba
   (a) $t/T = 0,5$
   data: sub1$BG.NBD and sub1$sajat, t = -5,138, df = 26, p-value = 2,336e-05
   (b) $t/T = 1$
   data: sub2$BG.NBD and sub2$sajat, t = -5,8191, df = 26, p-value = 3,931e-06
   (c) $t/T = 2$
   data: sub3$BG.NBD and sub3$sajat, t = -1,895, df = 8, p-value = 0,09469

DOI: 10.14751/SZIE.2014.035

A.17. A legjobb 200 vásárló előrejelzése az egyes modellek esetében az egyes vásárlók megfigyelési időszakra vonatkozó paramétereinek függvényében - 4.2.5. alszakasz, 86. old. Forrás: saját számítás.

	legjobb1	legjobb2	legjobb3	T	E	x	x_{c1}	x_{c2}	t_x
1	66,90	84,30	83,80	0,50	0,25	4,06	0,34	0,12	0,34
2	44,40	81,30	79,70	0,50	0,25	4,01	0,29	0,11	0,33
3	55,90	80,30	77,70	0,50	0,25	3,87	0,26	0,10	0,32
4	70,60	80,50	78,50	0,50	0,25	3,99	0,44	0,18	0,33
5	65,60	81,90	80,20	0,50	0,25	3,86	0,39	0,15	0,32
6	57,40	92,40	91,70	0,50	0,25	3,75	0,35	0,13	0,32
7	70,10	83,70	81,00	0,50	0,25	3,87	0,71	0,27	0,32
8	67,60	98,70	94,90	0,50	0,25	3,67	0,63	0,23	0,31
9	61,10	97,00	87,50	0,50	0,25	3,45	0,56	0,19	0,29
10	60,00	84,30	83,90	0,50	0,50	4,06	0,34	0,12	0,34
11	50,70	82,40	81,00	0,50	0,50	3,97	0,30	0,11	0,33
12	43,50	86,70	82,90	0,50	0,50	3,89	0,27	0,10	0,32
13	60,60	83,10	79,80	0,50	0,50	4,00	0,46	0,17	0,33
14	54,30	89,70	85,70	0,50	0,50	3,92	0,41	0,15	0,32
15	55,90	90,70	86,50	0,50	0,50	3,79	0,37	0,13	0,32
16	72,70	87,00	83,00	0,50	0,50	3,88	0,72	0,26	0,32
17	66,40	84,10	76,60	0,50	0,50	3,63	0,63	0,22	0,31
18	70,20	84,00	81,90	0,50	0,50	3,44	0,54	0,21	0,30
19	44,20	79,80	75,60	0,50	1,00	4,11	0,34	0,13	0,34
20	49,80	77,80	75,90	0,50	1,00	3,97	0,30	0,10	0,33
21	55,40	79,00	76,40	0,50	1,00	3,92	0,28	0,10	0,33
22	58,50	83,80	79,40	0,50	1,00	4,04	0,46	0,17	0,33
23	50,30	79,50	75,40	0,50	1,00	3,90	0,42	0,14	0,32
24	47,80	79,10	71,30	0,50	1,00	3,79	0,36	0,13	0,31
25	62,90	80,70	75,90	0,50	1,00	3,87	0,72	0,28	0,32
26	58,10	76,70	68,90	0,50	1,00	3,61	0,61	0,23	0,31
27	71,30	79,70	71,90	0,50	1,00	3,46	0,55	0,19	0,29
28	77,00	108,10	98,30	1,00	0,50	6,46	0,58	0,22	0,62
29	82,60	110,70	99,40	1,00	0,50	6,11	0,51	0,19	0,59
30	98,00	116,90	101,40	1,00	0,50	5,94	0,43	0,16	0,58
31	89,30	109,60	96,50	1,00	0,50	6,22	0,77	0,30	0,60
32	106,40	114,20	94,20	1,00	0,50	5,92	0,66	0,25	0,58
33	97,10	109,30	91,90	1,00	0,50	5,57	0,58	0,21	0,54
34	97,00	113,30	94,00	1,00	0,50	5,76	1,18	0,44	0,57
35	100,20	121,20	96,80	1,00	0,50	5,27	1,00	0,36	0,53
36	118,40	120,20	92,40	1,00	0,50	4,83	0,85	0,31	0,49
37	83,40	105,20	91,40	1,00	1,00	6,46	0,57	0,22	0,62
38	73,70	103,80	91,40	1,00	1,00	6,19	0,50	0,18	0,60
39	85,80	107,00	88,70	1,00	1,00	5,85	0,44	0,16	0,57
40	79,70	104,40	92,30	1,00	1,00	6,32	0,78	0,30	0,60
41	101,30	107,20	88,00	1,00	1,00	5,85	0,66	0,24	0,57
42	104,10	107,90	88,60	1,00	1,00	5,52	0,56	0,21	0,55
43	64,60	106,50	86,70	1,00	1,00	5,80	1,21	0,44	0,57
44	94,20	114,30	86,50	1,00	1,00	5,30	1,01	0,37	0,52
45	114,10	120,40	89,20	1,00	1,00	4,90	0,87	0,31	0,49
46	70,20	78,60	78,50	1,00	2,00	6,49	0,58	0,22	0,62
47	73,10	76,30	78,60	1,00	2,00	6,09	0,50	0,17	0,59
48	86,70	76,80	81,00	1,00	2,00	5,94	0,43	0,15	0,58
49	90,10	73,50	77,30	1,00	2,00	6,24	0,77	0,29	0,60
50	97,90	72,00	80,10	1,00	2,00	5,77	0,64	0,24	0,56

DOI: 10.14751/SZIE.2014.035

	legjobb1	legjobb2	legjobb3	T	E	x	x_{c1}	x_{c2}	t_x
51	98,00	68,40	78,00	1,00	2,00	5,48	0,58	0,20	0,54
52	75,50	67,20	75,10	1,00	2,00	5,85	1,21	0,46	0,57
53	97,00	76,30	76,00	1,00	2,00	5,30	1,00	0,38	0,52
54	119,90	77,20	86,90	1,00	2,00	4,82	0,85	0,30	0,49
55	123,90	142,00	101,00	2,00	1,00	9,31	0,86	0,32	0,99
56	134,70	152,10	104,20	2,00	1,00	8,52	0,69	0,26	0,92
57	147,20	157,80	106,90	2,00	1,00	8,19	0,62	0,23	0,87
58	136,60	147,50	97,40	2,00	1,00	8,82	1,14	0,42	0,95
59	159,60	159,60	104,90	2,00	1,00	7,92	0,92	0,34	0,86
60	150,30	154,30	122,60	2,00	1,00	7,47	0,81	0,29	0,82
61	139,90	157,80	115,90	2,00	1,00	7,78	1,66	0,62	0,84
62	137,90	150,30	142,30	2,00	1,00	6,86	1,36	0,50	0,75
63	158,00	157,20	148,30	2,00	1,00	5,98	1,07	0,39	0,68
64	117,90	138,50	95,70	2,00	2,00	9,27	0,86	0,32	0,99
65	127,80	149,60	93,40	2,00	2,00	8,60	0,72	0,27	0,93
66	116,30	159,90	103,30	2,00	2,00	8,14	0,63	0,23	0,88
67	135,00	148,20	98,60	2,00	2,00	8,92	1,16	0,44	0,95
68	143,50	161,00	108,30	2,00	2,00	7,86	0,93	0,33	0,85
69	152,60	159,70	123,20	2,00	2,00	7,43	0,81	0,28	0,81
70	150,20	159,30	110,20	2,00	2,00	8,00	1,72	0,64	0,86
71	137,20	150,90	137,90	2,00	2,00	6,76	1,34	0,49	0,75
72	138,90	146,50	149,20	2,00	2,00	6,07	1,11	0,39	0,69
73	114,80	107,10	93,50	2,00	4,00	9,43	0,86	0,33	1,00
74	132,10	109,90	98,50	2,00	4,00	8,77	0,72	0,26	0,94
75	114,40	96,80	99,70	2,00	4,00	8,22	0,62	0,22	0,90
76	134,80	100,70	91,60	2,00	4,00	8,90	1,17	0,43	0,95
77	141,50	97,70	103,60	2,00	4,00	8,16	0,96	0,35	0,88
78	153,20	92,40	129,10	2,00	4,00	7,36	0,79	0,29	0,81
79	154,80	93,40	113,70	2,00	4,00	7,78	1,68	0,64	0,84
80	132,00	100,90	137,70	2,00	4,00	6,90	1,36	0,49	0,76
81	132,40	92,40	150,20	2,00	4,00	6,11	1,11	0,40	0,69

legjobb1: a legjobb 200 találati eredményei a saját modell esetében
legjobb2: a legjobb 200 találati eredményei a BG/NBD modell esetében
legjobb3: a legjobb 200 találati eredményei a heurisztikus modell esetében

DOI: 10.14751/SZIE.2014.035

A.18. A Wilcoxon teszt eredményei (R output) - 4.2.5. alszakasz, 86. old. Forrás: saját számítás.

1. $t/T = 0,5$
   data: sub1$legjobb1 and sub1$legjobb2, V = 1, p-value = 9,904e-06
2. $t/T = 0,5$
   data: sub1$legjobb1 and sub1$legjobb3, V = 198,5, p-value = 0,8288
3. $t/T = 0,5$
   data: sub1$legjobb2 and sub1$legjobb3, V = 378, p-value = 1,49e-08
4. $t/T = 1$
   data: sub2$legjobb1 and sub2$legjobb2, V = 0, p-value = 5,93e-06
5. $t/T = 1$
   data: sub2$legjobb1 and sub2$legjobb3, V = 188, p-value = 0,9906
6. $t/T = 1$
   data: sub2$legjobb2 and sub2$legjobb3, V = 374, p-value = 9,303e-06
7. $t/T = 2$
   data: sub3$legjobb1 and sub3$legjobb2, V = 247, p-value = 0,1698
8. $t/T = 2$
   data: sub3$legjobb1 and sub3$legjobb3, V = 225, p-value = 0,3997
9. $t/T = 2$
   data: sub3$legjobb2 and sub3$legjobb3, V = 160,5, p-value = 0,5011