Gazdálkodás és Szervezéstudományok Doktori Iskola

Doktori (PhD) értekezés tézisei

A FOGYASZTÓI MAGATARTÁS VIZSGÁLATÁNAK MÓDSZERTANI TOVÁBBFEJLESZTÉSE

Készítette: Ruff Ferenc

Gödöllő
2014

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

1. A munka előzményei, a kitűzött célok

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 klaszterfelosztások mellett. A szakirodalomban körüljártam egy ilyen eljárás [Tong, 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.

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, 2011], melynek módosítását hajtottam végre azért, mert az általuk létrehozott modell felállításának feltételrendszerét nem tartottam megalapozottnak. Az ő munkájuk is egy modell továbbfejlesztése [Fader, 2005]. Én is ezen utóbbi modellhez nyúlok vissza, azonban a fejlesztés iránya más, mint a van Oest [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, 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.)

2. Anyag és módszer

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

2.1.1. Az eddigi módszerek elméleti és empirikus elemzése

Dolgozatom első részének 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". Liu [2010] munkájában a klaszterszám meghatározása céljából végrehajtott vizsgálatának célja az volt, hogy megfigyeljék, hogy a vizsgált indexek (melyek segítségével a klaszterszámok meghatározhatók) pontosságára milyen hatással van az adatok szerkezete (zajos adatok, sűrűség különbségek, alcsoportok, aszimmetrikus eloszlás). Egy olyan index – az ún. S_Dbw index – volt a 11 között, mely mindegyik – az általuk elvégzett – szimulációs kísérletben helyes döntést hozott. Az eljárást Halkidi and Vazirgiannis [2001] dolgozta ki, mely a klaszterek közötti sűrűségkülönbségen alapszik. Ezt fejlesztette tovább Kim and Lee [2003] valamint Tong and Tan [2009] abba az irányba, hogy robusztusabb[1] legyen, valamint ne csak gömbszimmetrikus klasztereket ismerjen fel. A dolgozat ezen fejezetében kerül sor ezen index kritikai vizsgálatára elméleti és empirikus úton.

2.1.2. 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 esettel 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

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

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 1. táblázat mutatja a létrehozott adatbázisok paramétereit (klaszterek középpontja, szórása, elemszáma). Ezeken az adatbázisokon klaszterező eljárásokat futtatok le különböző paraméterbeállítások mellett, és a kapott klasztereken tesztelem az indexeket. Ezt az eljárást követték mindhárom cikkben, amelyek ennek az indexnek kidolgozásával foglalkoztak. Halkidi and Vazirgiannis [2001] valamint Tong and Tan [2009] elemzésében, többek között, az ún. DBSCAN [Ester, Kriegel, Sander, and Xu., 1996] algoritmust alkalmazták. Ez a módszer a sűrűségek vizsgálatán alapszik, és nagyon hatékony nem konvex, de jól szeparált 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, 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. 1. 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

1. 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

egyes klaszterelhelyezkedések esetében, illetve általánosan jobbnak tekinthető-e valamelyik index.

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

2.2.1. A BG/NBD modell bővítése (1)

Az irodalomfeldolgozásban bemutatott BG/NBD modell kibővítését készítette el van Oest [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.[2]

[2]Ld. előző (BG/NBD) modell.

3. Panaszmentes vásárlás esetén a vásárló q_p valószínűséggel válik inaktívvá.
4. q_p változékonysága béta eloszlást követ u_p és v_p paraméterekkel:

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

5. q_p és λ_p vásárlónként egymástól függetlenül változnak.
6. A vásárlás napján előforduló 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.
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 ($z_c \in \{0, 1\}$).

Ezen adatokból és feltételekből van Oest [2011] által megalkotott modell a dolgozat 3. fejezetében megtalálható. Vizsgálataik 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 második részében.

2.2.2. 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 tapasztalati 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. 6. old.). Természetesen ebben szerepe van még az általam a modellbe bevont egyéb hatásoknak, nevezetesen, hogy a vásárlás pozitívan elbírált panasszal ill. negatívan elbírált panasszal[3] 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.

A modellek teszteléséhez használt adatbázisok előállításához használt algoritmus a dolgozat mellékletében található.

2.2.3. 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

[3] Részletesebben lásd 3.2.1. alfejezet, 15.old.

[Cohen, 1960]. Ennek értékét a következő képlettel számolhatjuk:

$$\kappa = \frac{p_0 - p_e}{1 - p_e} \tag{2}$$

ahol
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 [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.

3. Eredmények

3.1. A klaszterezés eredményének ellenőrzése

3.1.1. 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 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[4]

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

megválasztásának módosítását javaslom. Az eredeti javaslat helyett a következőképpen definiálom az f^* függvényt, amelyet megkülönböztetésül f^{**}-nak nevezek:

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

ahol
$\boldsymbol{x}_i$: az i-edik megfigyelési egység,
$\boldsymbol{m}$ egy tetszőleges egyed,
$x_i^{(p)}$ az i-edik megfigyelési egység p-edik változójának értéke,
$m^{(p)}$ 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.

Ezt a módosított függvényt használva a $Dens_{bw}$ részindex helyett kapjuk a $Dens_{bw}^{**}$ részindexet, melyből a teljes index adódik:

$$S\_Dbw^{**}(c) = Dens_{bw}^{**}(c) + Scat(c) \tag{4}$$

3.1.2. 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 1. koordinátatengely 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 1. és 2. koordinátatengely menti 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 2. 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, 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.

2. 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

A 2. táblázat alapján vizsgálhatjuk a két részindexet is, melyek összegeként áll elő az előbb vizsgált index. A $Scat$ részindex méri a klasztereken 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 hányadosuk pedig csökken.

Ugyancsak a 2. táblázat alapján vizsgálhatjuk a másik, a $Dens_{bw}^{**}$ részindexet. 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 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 3. 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 esetén következett be a váltás.

3. 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 3. 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.[5]

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.

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

Ebben az alfejezetben az Anyag és módszer fejezetben bemutatott 8 féle adatbázison 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 csoportosítást választottam legjobbnak.

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

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

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 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. A dolgozat A.10. mellékletében 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.

3.2. A BG/NBD előrejelzési modell bővítése, és a tesztelések eredményei

3.2.1. 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[6]. 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.

3.2.2. A modell megalkotásának feltételei

1. Amíg a vásárló aktív, addig az egységnyi idő alatt bekövetkező vásárlások száma Poisson eloszlást követ, melynek paramétere λ.
2. λ változékonysága gamma eloszlást követ r és α paraméterekkel.
3. Panaszmentes vásárlás esetén a vásárló q_p valószínűséggel morzsolódik le.
4. q_p változékonysága béta eloszlást követ u_p és v_p paraméterekkel.
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 $0 < q_p, q_{c1}, q_{c2}, \mu, \epsilon < 1$.

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

3.2.3. 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.

3.2.4. 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 $\mathrm{E}\left(X(t)\right)$-t. Ez lesz az alapja annak, hogy a későbbiekben előrejelzést tudjunk adni a T-n túli időszakra.

A modell megalkotásának lépéseit a dolgozat 4. fejezete tartalmazza, itt csak a végeredményt közlöm:

$$\begin{aligned}\mathrm{E}(X(t)|\lambda, q_p, q_{c1}, q_{c2}, \mu, \epsilon) &= \lambda t \cdot \mathrm{P}(\tau > t) + \int_0^t \lambda x \cdot f(x)\,\mathrm{d}x = \\ &= \lambda t \cdot \mathrm{e}^{-\lambda c t} + \int_0^t \lambda x \cdot \lambda c\,\mathrm{e}^{-\lambda c x}\,\mathrm{d}x = \\ &= \frac{1 - \mathrm{e}^{-\lambda t[1-(1-\mu)(1-q_p)-\mu(1-\epsilon)(1-q_{c2})-\mu\epsilon(1-q_{c1})]}}{1-(1-\mu)(1-q_p)-\mu(1-\epsilon)(1-q_{c2})-\mu\epsilon(1-q_{c1})} \qquad (5)\end{aligned}$$

3.2.5. A vásárlásszám előrejelzése

A modellépítés célja annak meghatározása, 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

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

Itt is csak az eredmény levezetés nélküli közlére kerül sor (a levezetést a dolgozat 4. fejezete tartalmazza):

$$\mathrm{E}(Y(t)|\Theta, input) \approx \frac{1}{L(\Theta|input)} \frac{1}{N} \sum_{i=1}^{N} \left[\frac{1 - \mathrm{e}^{-\lambda_i t c_i}}{c_i} \cdot L_{\text{aktív}}(\Phi_i|input) \right] \quad (6)$$

ahol
N a véletlen minta elemszáma,
Θ a $r, \alpha, u_p, v_p, a, b, e, f, u_{c1}, v_{c1}, u_{c2}, v_{c2}$ paraméterek halmaza,
input a modell bemenő adatainak halmaza,
$c_i = 1 - (1 - \mu_i)(1 - q_{p_i}) - \mu_i(1 - \epsilon_i)(1 - q_{c2_i}) - \mu_i \epsilon_i (1 - q_{c1_i}), \quad 1 \leqslant i \leqslant N,$
továbbá
λ_i a $\Gamma(r, \alpha)$,
q_{p_i} a $\mathrm{B}(u_p, v_p)$,
μ_i a $\mathrm{B}(a, b)$,
ϵ_i a $\mathrm{B}(e, f)$,
q_{c1_i} a $\mathrm{B}(u_{c1}, v_{c1})$,
q_{c2_i} a $\mathrm{B}(u_{c2}, v_{c2})$ eloszlású valószínűségi változó i-edik véletlenszerűen kiválasztott értéke .

3.2.6. A vizsgálatba bevont modellek

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

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[7], 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, an-

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

nak a vásárlásszámát a megfigyelési időszak vásárlásszámához mérten számítottam ki (egyenes arányosságot feltételezve a vásárlásszám és az eltelt idő között).

3.2.7. 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. Minden egyes paraméterbeállítás esetén 10-10 modelleredményt átlagoltam és ezen értékekkel számoltam tovább. A kapott eredmények a dolgozat mellékletben találhatók. A Kappa statisztikák értékeit vizsgáltam az egyes modellek esetében, az eredményeket boxplot ábrán szemléltettem. Az 1. 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 $K3$-mal). Azonban az eltérés nem tekinthető statisztikailag igazoltnak, melyet alátámaszt a $K1$ és $K2$ eredményeken végrehajtott páros Wilcoxon próba[8], mely szerint 5%-os szignifikanciaszinten[9] nem vethető el a nullhipotézis, tehát a két átlagérték különbözősége nem igazolt ($p = 0{,}094$).

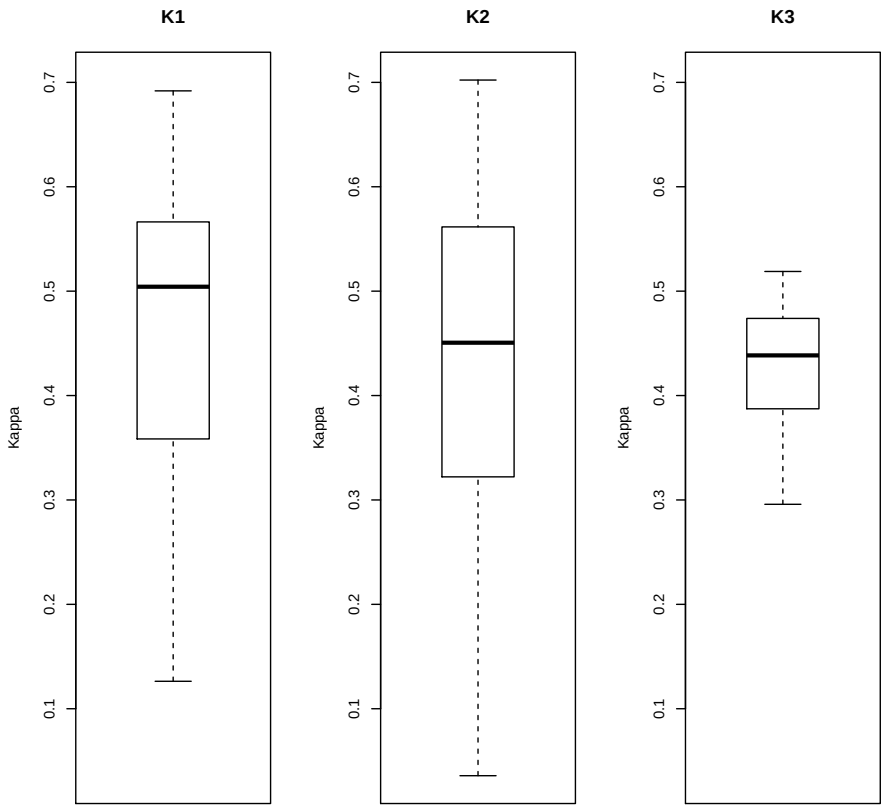
Ezt a képet azonban árnyalja, ha felbontjuk az egyes modellek eredményeit aszerint, hogy az előrejelzési időszak hányszorosa a megfigyelési időszaknak, vagyis t/T értékei (0,5; 1; 2) szerint három csoportot alkothatunk minden egyes modell esetében. Ezen eredmények a 2. ábrán láthatók. Összevetve az első 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. $t/T = 1$) az ábrán látható különbségek statisztikailag az első esetben kimutathatók, a második esetben viszont nem[10].
Vagyis a hosszabb távra szolgáló előrejelzés esetében a saját modell megbízhatóbbnak bizonyult, mint a BG/NBD modell.

A harmadik modellel való összevetés során az első szembetűnő különbség a szórásokban tapasztalható nagy különbség (2. ábra). Mivel az értékek nem

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

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

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

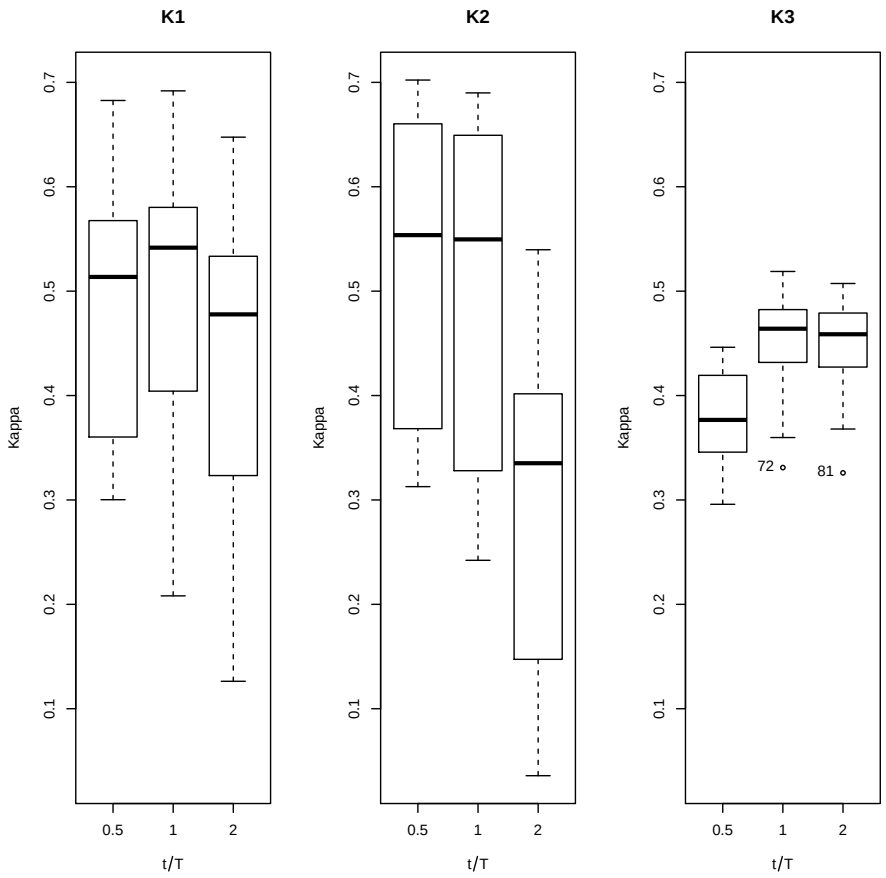


K1: a kappa statisztika értékei a saját modell esetében,
K2: a kappa statisztika értékei a BG/NBD modell esetében,
K3: a kappa statisztika értékei a heurisztikus modell esetében.

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

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 (1. ábra) kimutatható a különbség ($p = 0{,}006$), a t/T hányados szerint szétválogatott esetek közül az elsőben szintén kimutatható a különbség ($p = 6{,}3\text{e-}05$), 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($1-1$, $1-0$



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.

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

, $0-1$, $0-0$). Vannak olyan összehasonlító vizsgálatok [Persentili Batislam, 2007; Fader, 2005] 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.

3.2.8. A becsült és a tényleges vásárlásszám közötti különbségek összehasonlítása

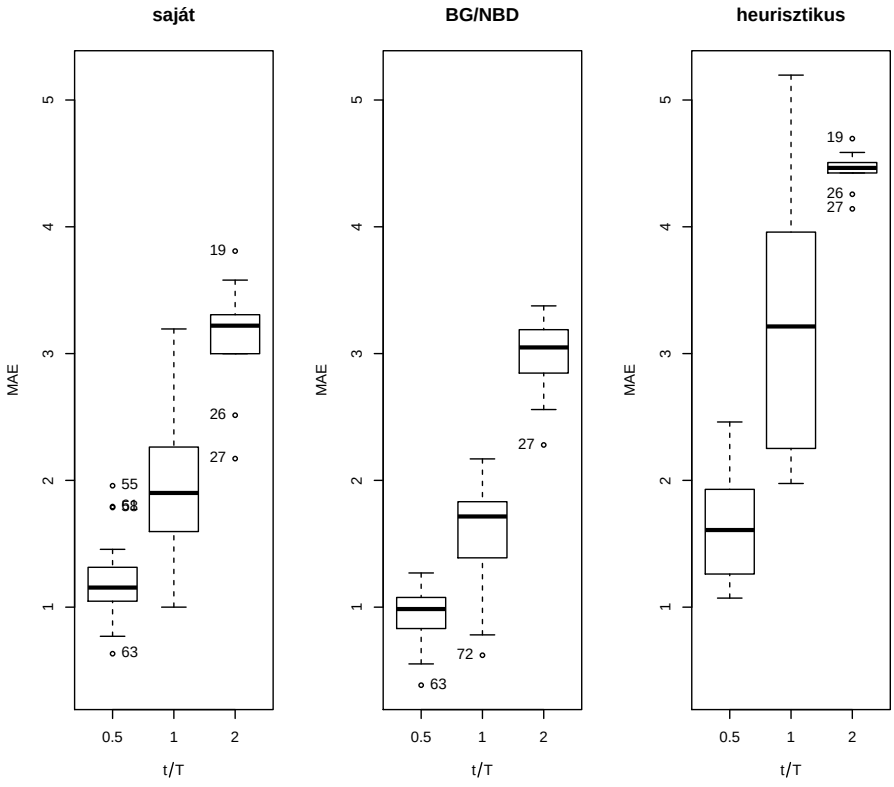
Ebben az alfejezetben olyan mutató alapján vizsgálom a modelleket, amely az egyes megfigyelési egységekhez tartozó találati pontatlanságok (eltérések) átlagos értékeit adja meg, és ezen értékeket hasonlítom össze. Erre több mód-

szer is adódik, melyek közül az egyik az átlagos abszolút eltérés (MAE = mean absolute error).

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_{\text{pred}} - y_{\text{val}}| \tag{7}$$

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

A MAE értékeket ismét boxplot diagramon ábrázoltam, és ismét a t/T arány, mint faktor szerinti csoportokra bontva (3. ábra). Ebben az esetben is megvizsgáltam, hogy az ábrán látható eltérések statisztikailag kimutathatók-e. A saját és a BG/NBD modellt hasonlítottam össze, az ábrából ugyanis



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

meggyőzően kiolvasható, hogy a heurisztikus modell ebben a vizsgálatban sokkal gyengébb eredményt adott a másik kettőhöz képest[11].

A két modell összehasonlításához itt a párosított t-próbát alkalmaztam, melyet mindhárom t/T hányados esetében elvégeztem. Megállapítottam, hogy az

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

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.

Az is megfigyelhető, hogy a BG/NBD modell sok esetben nem adott értékelhető eredményt a MAE indexre, 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.

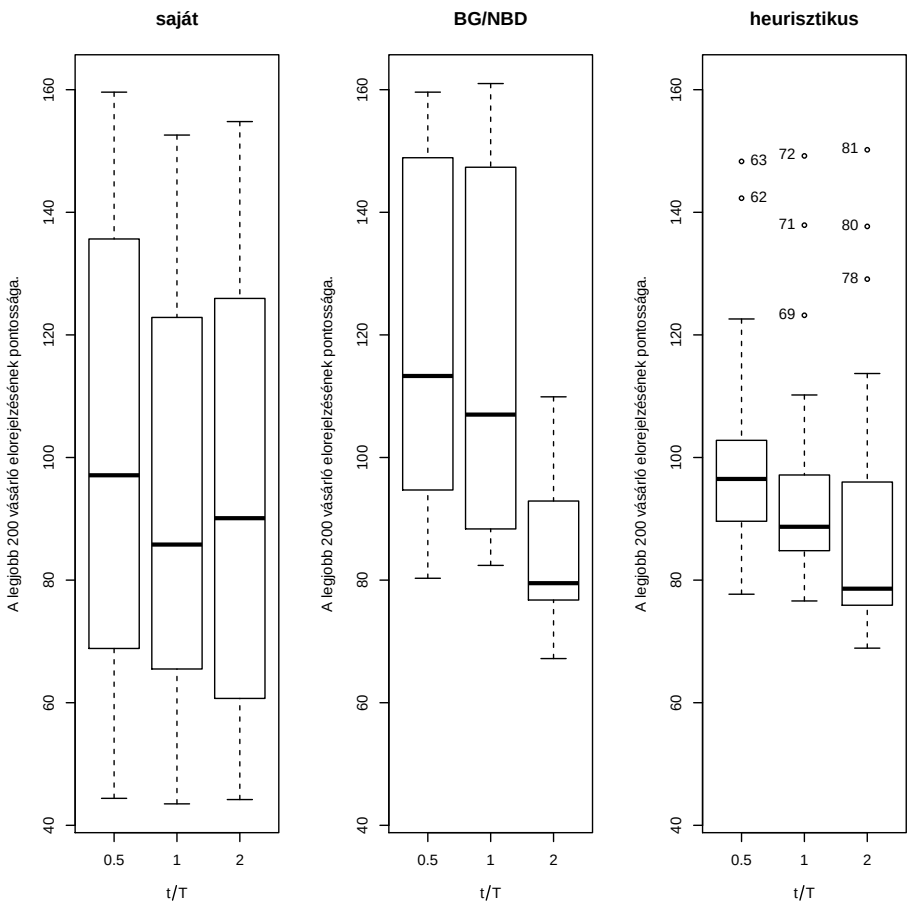
3.2.9. A jövőbeli legjobb vásárlók meghatározása

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[12]. Ezt támasztja alá pl. Homburg [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 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 (4. 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, va-

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



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

lamint a heurisztikus és a BG/NBD modell között a $t/T = 0{,}5$, és a $t/T = 1$ esetében. 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 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 mestersé-

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

3.3. Új és újszerű tudományos eredmények

1. A vásárlói csoportok elkülönítése, szegmentálása kapcsán végzett munkámban tapasztalati és elméleti elemzések segítségével megállapítottam, hogy a Tong [2009] által kidolgozott $\boldsymbol{m}_{ij}$ osztópont 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 ($Dens_{bw}$) számításában.

2. Megalkottam az f^{**} függvényt (3. 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 az S_Dbw^{**} indexet (4. 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 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.

4. 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[13] 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 választhatók 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

[13] Az adatbázisok létrehozásakor tértem ki ennek tárgyalására.

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 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[14]), 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 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.

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

Publikációs jegyzék

Tudományos cikk idegen nyelven

- Ruff Ferenc (2012): Empirical comparison of a model based and a non model based clustering methods. *Annals of The Polish Association of Agricultural and Agrobusiness Economists*. Vol. XIV. No.6. 242-246 p.
- Ruff Ferenc (2008): Methodological problems of classification and prediction in food marketing. *Annals of The Polish Association of Agricultural and Agrobusiness Economists*. Vol. X. No.5. 125-129 p. ISSN 1508-3535.

Tudományos cikk magyar nyelven

- Ruff Ferenc (2013): Klaszterszámok meghatározásának egy lehetséges megoldása. *Szigma*. XLIV. évf. 3-4. szám. 135-153. p.

Tudományos konferencián elhangzott előadás konferenciakiadványban megjelentetve, idegen nyelven

- Ruff Ferenc (2014): Clustering Methods for Ordinal Variables. Economics Questions, Issues and Problems. Komarno, Konferencia kiadvány 274-279 p. ISBN 978-80-89691-07-4. http://www.irisro.org/economics2014january/55RuffFerenc.pdf
- Ruff Ferenc, Szelényi László (2006): Environmental decision problems and operational research. X. Nemzetközi Agrárökonómiai Tudományos Napok. Gyöngyös, 2006. márc. 30-31. Konferencia CD: \Poszter \krf110. 1-6. p. ISBN 9632296230

Tudományos konferencián elhangzott előadás konferenciakiadványban megjelentetve, magyar nyelven

- Pitlik László, Ruff Ferenc (2011): Táplálkozási tanácsadó szimulátor fejlesztése, avagy modellezési stratégiák összehasonlító elemzése. IX. Magyar Biometriai, Biomatematikai és Bioinformatikai Konferencia. 2011. július 1., Budapest. Absztrakt: Program, Előadás- és poszterkivonatok, Résztvevők listája (konferencia kiadvány). 20. p.
- Szelényi László, Bedéné Szőke Éva, Ruff Ferenc, Vinogradov Szergej (2004): Agrárökonómiai elemzések többváltozós módszerekkel. XXX. Óvári Tudományos Napok. In: Gazdasági informatika szekció. Mosonmagyaróvár, 2004. október 7. Konferencia CD: aokonomia \ Szelenyi.pdf. 1-5 p.

– Szelényi László, Ruff Ferenc, Bedéné Szőke Éva (2004): Környezetvédelmi mutatók többváltozós elemzése. Környezetgazdálkodási szekció. IX. Nemzetközi Agrárökonómiai Tudományos Napok, Gyöngyös, 2004. március 25-26. Konferencia CD: 3.Környezetgazdálkodás\6\Szelényi, László - Ruff, Ferenc-Bedéné Szőke, Éva.doc. 1-6. p.
– Szelényi László, Bedéné Szőke Éva, Ruff Ferenc (2003): A vidékfejlesztés helyzetének többváltozós elemzése. Agrárgazdaság, Vidékfejlesztés és Agrárinformatika az évezred küszöbén /AVA nemzetközi konferencia 2003. április 01-02. Debrecen. Konferencia CD: cd\pdf\D098.pdf. 1-6. p.

Magyar nyelven megjelent könyvrészlet

– Ruff Ferenc (2002): A legjobban illeszkedő függvény-típus kiválasztása. 317-318, 537-539 p. In: Szűcs István (szerk): *Alkalmazott statisztika*. Agroinform Kiadó, Budapest. 551 p.

Egyéb publikáció

– Pitlik László, Ruff Ferenc (2011): Development of nutrition simulator or comparison modeling approaches. Magyar Internetes Agrárinformatikai Újság. 2011. No 160. 1-33. p. HU-ISSN-1419-1652
http://miau.gau.hu/miau/160/saltseer.doc
– Pitlik László, Ruff Ferenc (2008): „Konzisztencia-gyár", avagy stratégiai és operatív ajánlások a modellezés automatizálásához. Magyar Internetes Agrárinformatikai Újság. 2008. No 119. 1-36. p. HU-ISSN-1419-1652
http://miau.gau.hu/miau/119/cikk_plrf.doc

Kutatási jelentés

– Szűcs István, Farkasné dr. Fekete Mária, Széles Zsuzsanna, Ruff Ferenc: A földhasználat és a földjáradék összefüggései 43362 sz. OTKA kutatási téma zárójelentése 2007. 22 p.
– Szelényi László, Ruff Ferenc, Bedéné Szőke Éva, Vinogradov Szergej: A környezetvédelem jelenlegi helyzetének korszerű többváltozós ökonometriai módszerek felhasználásával történő elemzése és értékelése, a komplex összefüggések feltárása. Közcélú környezet- és természetvédelmi feladat, zárójelentés. Gödöllő, 2005. 55 p.
– Szelényi László, Szűcs István, Ruff Ferenc, Bedéné Szőke Éva, Szergej Vinogradov: Az agrárgazdaság prognosztizálását segítő programozási modellek és termelési függvények kidolgozása, A/0129/2003 sz. OKTK kutatási téma zárójelentése, SZIE, Gödöllő, 2004. 49 p.

– Szűcs István (szerk): Kedvezőtlen adottságú térségek lehatárolásának előkészítése. FVM tanulmány, Gödöllő, 2000. 80 p.