Szociológiai Szemle 2006/3, 3–18.

MÓDSZERTAN

AZ AGGREGÁLT ADATOK FELBONTÁSÁNAK LEHETŐSÉGEI (AZ ÖKOLÓGIAI KAPCSOLAT PROBLÉMÁJA)

MÉSZÁROS József–MÉRŐ Csaba–MOLNÁR D. László

Budapesti Műszaki és Gazdaságtudományi Egyetem
H-1111 Budapest, Műegyetem rkp. 3–9.; e-mail: mj@eik.bme.hu

ELTE Társadalomtudományi Kar
H-1117 Budapest, Pázmány Péter sétány 1/a.; e-mail: mero.csaba@gmail.com

Szociomed Kft.
H-1204 Budapest, Szent Imre h. u. 52.; e-mail: molnar@szociomed.hu

Összefoglaló: Társadalomtudományi vizsgálatok során gyakorta nem egyéni szintű adatokból szeretnénk a cselekvő egyének magatartására következtetéseket levonni. A szakirodalomban ezt az ökológiai problémának nevezik. Az utóbbi évtizedben jelentős módszertani újítások történtek e területen. Tanulmányunkban bemutatjuk az irodalomban szokásos módszereket, rávilágítunk azok erényeire és fogyatékosságaira. A módszereket a 2004 évi kettős népszavazás adatain ismertetjük. Bemutatjuk azt, hogy a különböző szintekre aggregált adatok más-más eredményeket adnak, azaz adatelemzés során döntően fontos a kellő körültekintés és lehetőség szerint az adatok aggregálásának elkerülése.

Kulcsszavak: ökológiai kapcsolatok, Simpson paradoxon

BEVEZETÉS

Az aggregált adatokból az egyéni magatartásra történő következtetés vizsgálata majd száz évre nyúlik vissza. Durkheim *Öngyilkosság* című könyvében már hasonló problémafelvetéssel él, hiszen különböző nemzetiségi, vallási adatokból kívánt az egyének magatartására következtetéseket levonni. Durkheim számos különböző franciaországi közigazgatási területre bontott térképet készített el és vizsgált meg részletesen. Különböző hipotéziseket elemzett. Az elmebetegség és az öngyilkosság kapcsolatát, az öngyilkosság és az alkoholizmus kapcsolatát vizsgálta. Durkheim rendelkezésére különböző területi egységre összevont statisztikai adatok álltak és ezekből az adatokból kívánt az egyes egyének öngyilkossági magatartására következtetéseket

levonni, azaz mai szóhasználattal élve, az ökológiai kapcsolat problémájának úttörő elemzője volt.

A kérdés vizsgálata a második világháború után kapott új lendületet, Leo Goodmann munkássága nyomán, aki a weimari Németország szavazói magatartását vizsgálta. Az 1930-as németországi választásokon a korábban viszonylag jelentéktelen náci párt a voksok 18,3%-át kapta, mely arány 37,3%-ra növekedett az utolsó választáson. A kortársak közül is, de a második világháború után is sokan kérdezték, hogy kik szavaztak a náci pártra, a lecsúszó középosztály, esetleg a protestánsok, esetleg a katolikusok, netán az értékesítési nehézségekkel küszködő gazdálkodók, stb. A kérdésre nem egyszerű a válasz, hiszen a választási eredmények választó körzetenként állnak rendelkezésre és hasonlóan a népszámlálási adatok is, miközben a szavazás egyéni cselekvés. Így a problémánk a következő: az összegzett adatokból szeretnénk az egyes egyének magatartására következtetni. Lehetséges-e ez egyáltalán?

A kutatásoknak újabb lökést adott az Egyesült Államok választási törvényének 1965. évi módosítása, mely megtiltja az etnikai hovatartozás szerinti diszkriminációt. Amennyiben ez mégis előfordulna, a bíróság elrendelheti a szavazási körzetek újraalakítását. A törvény szerint diszkrimináció akkor keletkezik, amennyiben az adott etnikai kisebbség az adott területen a többi szavazótól egyértelműen különböző módon szavaz, és a többségi szavazók megakadályozzák a kisebbségi szavazókat állampolgári jogaik érvényesítésében. A fenti tényállás bizonyítása igen nehéz, hiszen választási adatok az Egyesült Államokban is (mint bárhol a világon, ahol titkos szavazási jog van) csak szavazókör szinten állnak rendelkezésre. Tehát újra az ökológiai probléma megoldásához jutunk.

Klasszikus eset a fentiek illusztrálására az 1990-es Louisana Államban történt vizsgálat a fekete lakosság szavazási magatartásáról. Feltevések szerint a fekete lakosság a fehéreknél kisebb arányban szavaz a republikánusokra. Amennyiben ez a hipotézis igaznak bizonyul, és egyes szavazókörökben egy-egy népcsoport túlsúlya megállapítható, akkor az 1965-ös jogszabályt alkalmazni kell. A 64 választókörzet több mint felében fehérek voltak többségben, tehát amennyiben a feketék valóban kisebb arányban szavaztak volna a republikánusokra, akkor a republikánusok győzelme valószínűsíthető volt. Abban az esetben, ha más tényezők is meghatározó jelentőségűek, akkor a végeredmény már nem következtethető ki ilyen egyszerűen.

SIMPSON PARADOXON

A tanulmányban tárgyalt jelenség bizonyos értelemben vett megfordítottja az ún. Simpson paradoxon (Simpson 1951). A jelenséget már Pearlson is leírta 1949-ben, de a szakirodalomban (legalább is a statisztikában és a szociológiában) Simpson nevével szokás nevezni az alábbi jelenséget. Az elemzések során gyakorta különböző aggregáltsági szintű adatokból vonunk le következtetéseket. Gyakorta adataink egymásra nézve nem feltétlen függetlenek, így a különböző aggregáltsági szintű adatokból történő elemzés gyakorta különböző eredményekre vezet. Érdemes tehát megfontolnunk, hogy következtetéseink levonásában lehetőség szerint csak az elemi adatokat használjuk. Gyakorta sajnos ez nem lehetséges, hiszen már aggregált adatok állnak rendelkezésünkre, melyet az elemzés céljára magunk is összegzünk. A Simpson

paradoxon jelensége azonban felhívja a figyelmet arra, hogy nagy körültekintéssel kell ezekben az esetekben eljárni. A paradoxont egy egyszerű példával mutatjuk be.

DOHÁNYZÁS ÉS TÚLÉLÉSI VALÓSZÍNŰSÉGEK

1972-ben egészségügyi vizsgálatot végeztek Newcastle upon Tyne környékén Angliában a nők egészségi állapotáról, 1994-ben ugyanazon személyek megkeresésével megismételték a vizsgálatot (Appleton et al. 1996). Az alábbi táblázatok ebből a vizsgálatból származnak.

Vizsgáljuk meg a dohányzás és az egészségi állapot kapcsolatát!

1. táblázat 55–64 korosztály

	55-64 korosztály	
	Meghalt	Él
Dohányzó	51=44%	64=56%
Nem dohányzó	40=33%	81=67%

2. táblázat 65–74 korosztály

	65-74 korosztály	
	Meghalt	Él
Dohányzó	29=80%	7=20%
Nem dohányzó	101=78%	28=22%

Korosztályonként vizsgálva úgy tűnik, hogy a dohányzás káros az egészségre. Mi történik azonban, ha összegezzük (aggregáljuk) a táblákat?

3. táblázat 55–74 korosztály összegezve

	55-64 korosztály	
	Meghalt	Él
Dohányzó	80=53%	71=47%
Nem dohányzó	141=56%	109=44%

Úgy tűnik, hogy a dohányosok tovább élnek!

Mi történt? A dohányosok nagy része meghalt még az idős kor elérése előtt, így a nagyobb számú nem dohányzó idősek természetesen magas időskori halandósága megfordította az összefüggést!

Magyarázat:
A = {él}, *B* ={nem dohányzó}, *C* = {idő t1},
A' ={halott}, *B'* ={dohányzó}, *C*' = {idő t2},

$$P(A\backslash B) < P(A\backslash B')$$
és $$P(A\backslash BC) = P(A\backslash B'C)$$
$$P(A\backslash BC') = P(A\backslash B'C')$$

A feltételes valószínűségek:

$$P(A\backslash B) = \{P(C\backslash B)\} *. P(A\backslash BC) + P(C'\backslash B) \,.\, P(A\backslash BC')$$
$$P(A\backslash B') = \{P(C\backslash B)\} \,.\, P(A\backslash B'C) + P(C'\backslash B') \,.\, P(A\backslash B'C')$$

Az előbbi példában:

$$P(A\backslash B) = 0{,}44 < P(A\backslash B') = 0{,}47$$
$$P(A\backslash BC) = 0{,}67 > P(A\backslash B'C) = 0{,}56$$
$$P(A\backslash BC') = 0{,}22 > P(A\backslash B'C') = 0{,}20$$

$$0{,}44 = 0{,}67 \,.\, 0{,}484 + 0{,}22 \,.\, 0{,}516$$
$$0{,}47 = 0{,}56 \,.\, 0{,}761 + 0{,}20 \,.\, 0{,}239$$

Azaz a paradoxon oka az hogy B és C nem független események.

A paradoxon rávilágít arra, hogy aggregált adatokból történő következtetések esetében gyakorta tévkövetkeztetéseket állíthatunk, ha nem vagyunk eléggé körültekintőek.

A RÉSZBEN MEGFIGYELT KERESZTTÁBLÁK PROBLÉMÁJA

Az ökológiai kapcsolat problémájának legegyszerűbb esetét egy egyszerű keresztáblával illusztrálhatjuk. A keresztábla peremeit (marginálisait) ismerjük csak, azonban a cellákat nem. Tekintsünk egy egyszerű esetet. A választási adatokat (mivel a szavazás titkos) csak jelöltek szintjén összesítve ismerjük, egy-egy szavazókörre. Az adott szavazókörre esetleg más adatforrásokból (népszámlálás, stb.) ismerjük a szavazók összetételét, és különböző csoportokba tudjuk sorolni őket. Érdekel bennünket az egyes csoportok szavazói magatartása, azaz az adott keresztábla egyes celláit kívánjuk becsülni. Tekintsük a következő keresztáblát!

4. táblázat Az egyes csoportok szavazói magatartásának keresztáblája

	A jelölt	B jelölt	C jelölt	D jelölt	Összesen
1. csoport					N_1
2. csoport					N_2
3. csoport					N_3
Összesen	N^1	N^2	N^3	N^4	N

A valóságban gyakran előfordulnak részben megfigyelt keresztáblák. Tipikusan ilyen adatok például a választási eredmények. Ekkor ismerjük a jelöltekre eső szavazatok számát esetleg más forrásból különböző társadalmi csoportok lélekszámát az adott területen. Egy keresztáblából nem sok információt tudunk kifacsarni, csak annyit,

hogy milyen tartományban mozoghatnak a belső cellák. A választási adatokban viszont egyszerre nagyon sok kereszttábla áll rendelkezésre, szavazókörönként egy. Ez lehetőséget ad arra, hogy a nem ismert cellákra – különböző eloszlásokat feltételezve – becsléseket adjunk.

A feladatot az aggregációs torzítás teszi nehézzé, az adatokból nehezen lehet következtetni arra, hogy az aggregálás után fennálló összefüggés az aggregálás előtt is létezett-e. Ez az ökológiai tévkövetkeztetés veszélye.

Összefoglalva a fentieket, a társadalomtudományi kutatásokban gyakorta olyan adatokkal vagyunk kénytelenek számolni, amelyek nem kísérleti vizsgálatok eredményei és így csak aggregálva vagy átlagolva állnak rendelkezésünkre.

MÓDSZEREK ÁTTEKINTÉSE

A kérdésre több megoldási javaslat született, Goodman (1953) regressziót, Duncan és Davis (1953) a határok módszerét javasolta. A két legújabb modell King (1997) és Wakefield (2003) bayesi alapú likelihood-becslést alkalmaz.

Az alábbiakban az egyszerűség kedvéért 2 x 2-es táblázatokon mutatjuk be a módszereket, de e módszerek értelemszerűen esetén általánosíthatóak nagyobb táblázatokra is.

5. táblázat Módszerek bemutatása 2 x 2-es táblázatokon

Jelölje:

i szavazókör	C_i^1	C_i^2	
R_i^1	p_i^1	$1\text{-}p_i^1$	r_i^1
R_i^2	p_i^2	$1\text{-}p_i^2$	r_i^2
	c_i^1	c_i^2	

ahol:

– R_i, C_i jelöli az N_i szavazóból az adott csoport számát.

– p_i az I szavazókörben az R_i. csoporton belül a C_i^1-t választók aránya

– c_i is az adott alternatívát választók aránya.

(a normál betűkkel jelölt mennyiségeket megfigyeljük, míg a *dőlt* betűkkel jelölt valószínűségeket kívánjuk becsülni)

6. táblázat Példa a fentiekre

	Kórházprivatizáció: Igen	Kórházprivatizáció: Nem	Összesen
Kettős állampolgárság: Igen	P^1	$1\text{-}p^1$	1 521 271
Kettős állampolgárság: Nem	P^2	$1\text{-}p^2$	1 436 049
Összesen	1 922 680	1 034 640	2 957 320

Számos, a peremeknek megfelelő eloszlás képzelhető el az adott táblára. Így a probléma nem oldható meg egyértelműen. Másképpen fogalmazva az aggregált adatokból nem lehetséges általánosságban egyértelműen az egyéni adatokat kinyerni, holott bennünket igazából az egyének magatartása érdekel.

Ez nem jelenti azt, hogy ne tegyünk a feladat megoldására valamilyen kísérletet.

MEGOLDÁSI JAVASLATOK

1. megoldás: regressziós módszerek

Tekintsük a következő kereszttáblát:

7. táblázat Példa kereszttábla

i szavazókör	C_i^1	C_i^2	
R_i^1	p_i^1	$1-p_i^1$	r_i^1
R_i^2	p_i^2	$1-p_i^2$	$1-r_i^1$
	c_i^1	$1-c_i^1$	

Ahol az alábbi azonosságok igazak:

$$c_i^1 = p_i^1 * r_i^1 + p_i^2 * r_i^2 \quad (1)$$

$$c_i^2 = (1-p_i^1) * r_i^1 + (1-p_i^2) * r_i^2 \quad (1^*)$$

A nehézség az, hogy minden i megfigyelés esetén, két paramétert kell becsülnünk: p_i^1-t és p_i^2-t.

Az első megoldás az úgynevezett regressziós módszerek (ezen belül az *„ökológiai regresszió”*, vagy *„Goodman regresszió”*).

Tegyük fel, hogy mindegyik p_i^1 és mindegyik p_i^2.ugyanaz minden szavazókörben. Példánkkal illusztrálva: ez az jelenti, hogy a kettős állampolgárságra eső igen szavazatok aránya állandó a szavazókörökben, és hasonlóan a nem szavazatok aránya is.

A megoldás feltevései nagyon erősek, ráadásul gyakorta a becslés kívülesik a [0,1] intervallumon.

Részletesebben kifejtve:

(1) $c_i^1 = p_i^1 * r_i^1 + p_i^2 * r_i^2$

$c_i^1 = p_i^1 * r_i^1 + p_i^2 * (1 - r_i^1)$

tegyük fel, hogy p_i^1-t és p_i^2-t két független részre bonthatjuk, az egyik rész állandó minden szavazókörre míg a másik r^1 lineáris függvénye:

(2) $p_i^1 = b^1{}_i + b^2 * r_i^1$

$p_i^2 = b^3{}_i + b^4 * r_i^1$

átrendezve:

(3) $p_i^1 = b^1{}_i + b^2 * r_i^1 + e_i^1$

$p_i^2 = b^3{}_i + b^4 * r_i^1\ e_i^2$

ahol e_i fehér zaj.

(3) –t (1)-be helyettesítve:

$c_i^1 = (b^1{}_i + b^2 * r_i^1 + e_i^1) * r_i^1 + (b^3{}_i + b^4 * r_i^1\ e_i^2) * r_i^2$

átrendezve:

(4) $c_i^1 = b^3 + (b^1 - b^3 + b^4) * r_i^1 + (b^2 - b^4) * (r_i^1)\,2 + [e_i^2 + (e_i^1 - e_i^2) * r_i^1]$

$[e_i^2 + (e_i^1 - e_i^2) * r_i^1]$ hibatag N (0, ó) eloszlással

jelölje:

$a^1 = b^3$

$a^2 = (b^1 - b^3 + b^4)$

$a^3 = (b^2 - b^4)$

ekkor a jelöléssel:

$c_i^1 = a^1 + a^2 * r_i^1 + a^3 * (r_i^1)\,2 + N\ (0, \sigma)$

így 4 paraméterünk van: b^1, b^2, b^3, b^4, az egyenletekből csak 3 paramétert tudunk kifejezni.

Ezért feltevésekkel kell élnünk:

1. $b^2 = b^4 = 0$ *Goodman*

Feltevés: $E(p_i^1) = b^1$, $E(p_i^2) = b^3$ minden i-re.

Ekkor:

$c_i^1 = b^3 + (b^1 - b^3) * r_i^1$

2. $b^1 = b^3 = b^1$ és $b^2 = , b^4$ *Linear neighbourhood model*

Feltevés: $E(p_i^1) = E(p_i^2)$ minden i-re.

Ekkor:

$c_i^1 = b^1 + b^2 * r_i^1$

2. megoldás: „határok módszere“

A Duncan és Davis módszere (1953, idézi Achen–Shively1995: 8. fejezet) abból indul ki, hogy a sor- és oszlopmarginálisok eloszlását adottnak véve p_i^1 és p_i^2 értékei egyértelműen meghatározzák egymást.

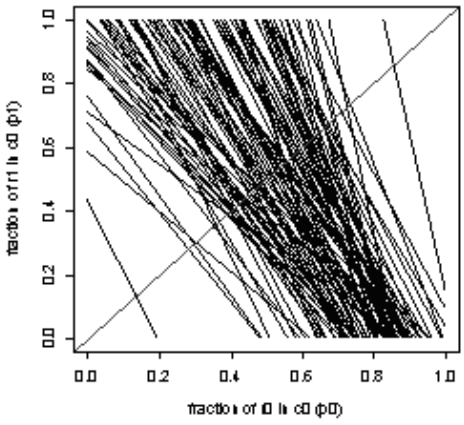
A p_i-k korlátait az alábbi egyenlőtlenségek fogalmazzák meg:

$max\ (0, (C_i - (1 - R_i)) / R_i) = p_i^1 = min\ (C_i / R_i, 1)$

$max\ (0, (C_i - R_i) / (1 - R_i)) = p_i^2 = min\ (C_i / (1 - R_i), 1)$

A fenti egyenleteket átrendezve::

$p_i^2 = C_i/(1 - R_i) - (R_i/(1 - R_i))\ p_i^1$



1. ábra Tomográfia ábra

Az összefüggést foglalja össze az úgynevezett tomográfia ábra. Mindegyik vonal egy táblázat lehetséges p^1, p^2 kombinációit jelöli. Önmagukban a határok semmitmondóak, túl tág határok közé lövik be a becsülni kívánt paramétereket. A becslésre a legegyszerűbb módszer az, ha a felezőpontokat jelöljük ki becslésnek. A határok szűkítésére egy lehetőség (Achen–Shively 1995), hogy a vizsgált adatokhoz egyéni feltételezéseket csatolhatunk.

(Például – a kettős népszavazás esetében – az a feltételezés, hogy $p^1 > p^2$ nem tűnik túl erősnek, mivel a két nagy párt az igen-igen illetve a nem-nem mellett kampányolt. A feltételezést persze nem lehet ellenőrizni, viszont a határokat ezzel elég erősen lecsökkentettük. Mindent egybevetve ez a módszer elég erősen hagyatkozik heurisztikus, kvalitatív megállapításokra.)

3. megoldás: King modellje

King (1977) az utóbbi időben gyakorta (talán túlságosan is) idézett megoldási javaslata három feltevéssel él:

1. feltevés: p_i^1 és p_i^2 az R_i,-re feltételes levágott kétváltozós normáliseloszlásból származik, azaz:

$P(p_i^1, p_i^2) = TN\ (p_i^1, p_i^2 | E, \Sigma)$,

2. feltevés: p_i^1 és p_i^2 várható értéke független R_i -től.

3. feltevés: C_i elemi értékei függetlenek R_i. mint feltétel rögzítése mellett

Ez a módszer a tomográfia vonalakon bayesi likelihood becslést végez a legvalószínűbb p^1 and p^2 párra. Első körben C^1 és C^2 feltételes eloszlását r^1-re illetve r^2-re binomiálisnak feltételezi. Második körben olyan apriori eloszlásokat feltételez p^1 és p^2-re, hogy az eloszlásuk egyenletes legyen a (0,1) intervallumon. Ezután a megfigyelt sor- és oszlopmarginálisokból MCMC algoritmussal kiszámítja p^1 and p^2 aposzteriori eloszlását.

$C^1i \mid r^1i \sim B(r^1i * N_i, p_i^1)$
$C^2i \mid r^2i \sim B(r^2i * N_i, p_i^2)$

p^1 *és* p^2 apriori eloszlása*:*
$p_i^1 \mid a^1, b^1 \sim Beta(a^1, b^1)$
$p_i^2 \mid a^2, b^2 \sim Beta(a^2, b^2)$

a^1, b^1, a^2, b^2 apriori eloszlása:
$a^1 \sim Exp(0.5)$
$b^1 \sim Exp(0.5)$
$a^2 \sim Exp(0.5)$
$b^2 \sim Exp(0.5)$

(King honlapján *http://gking.harvard.edu/ei/ei.html* elérhető egy ingyenesen letölthető DOS alapú program, amely felhasználható szimulációk elkészítésére.)

King módszerének kritikája

Szeretnénk hangsúlyozni, hogy King eljárása csak az alapfeltevések fennállása esetén működőképes. Ezeket az alapfeltevéseket egyszerűbb nyelvre lefordítva a következőkről van szó. Az első alapfeltétel pedig az eloszlás típusáról szól. A második alapfeltevés lényegében az „aggregációs torzítás" hiányát fogalmazza meg, azaz az aggregációs változó mentén a becsülni kívánt valószínűségek függetlenek. A harmadik alapfeltevés lényegében azt fogalmazza meg, hogy az adatokban nincs területi autókorreláció, azaz az egyes területi egységeken tapasztalható előfordulások egymástól függetlennek tekinthetőek. Szeretnénk hangsúlyozni, hogy ezek a feltevések nagyon erősek. A gyakorlati életben többnyire nem teljesülnek. Az elemzésbe vont adatok többnyire autókorreláltak és az aggregációs torzítás is előfordul. Az alapfeltevések nem teljesülésének következményeit elemzi Cho tanulmánya (Cho 1998).

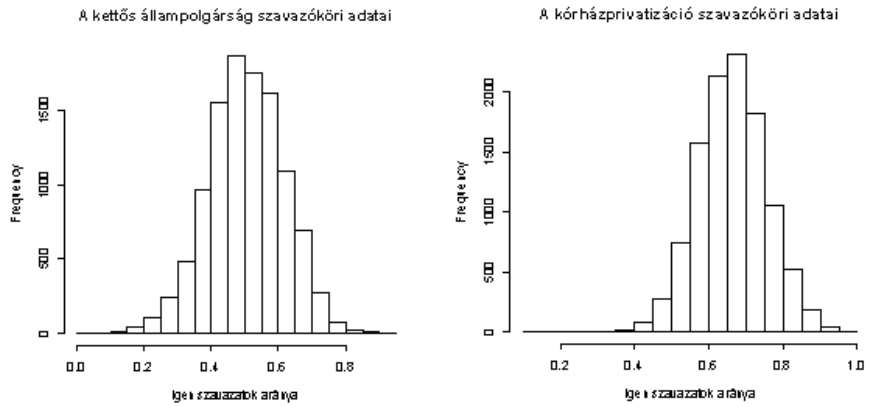
A 2004 DECEMBER 5-I KETTŐS NÉPSZAVAZÁS EREDMÉNYEINEK AZ ELEMZÉSE

2004 decemberében a választásra jogosultak két kérdéssel kapcsolatosan nyilváníthattak véleményt:

8. táblázat 2004 decemberi kettős népszavazás kérdései

Kérdés sorszáma	Kérdés szövege (a népszavazás tárgya)
1	Egyetért-e Ön azzal, hogy az egészségügyi közszolgáltató intézmények, kórházak maradjanak állami, önkormányzati tulajdonban, ezért az Országgyűlés semmisítse meg az ezzel ellentétes törvényt? (továbbiakban: *kórház privatizáció*)
2	Akarja-e, hogy az Országgyűlés törvényt alkosson arról, hogy kedvezményes honosítással - kérelmére - magyar állampolgárságot kapjon az a magát magyar nemzetiségűnek valló, nem Magyarországon lakó, nem magyar állampolgár, aki magyar nemzetiségét a 2001. évi LXII. tv. 19. § szerinti "Magyar igazolvánnyal" vagy a megalkotandó törvényben meghatározott egyéb módon igazolja? (továbbiakban: *kettős állampolgárság*)

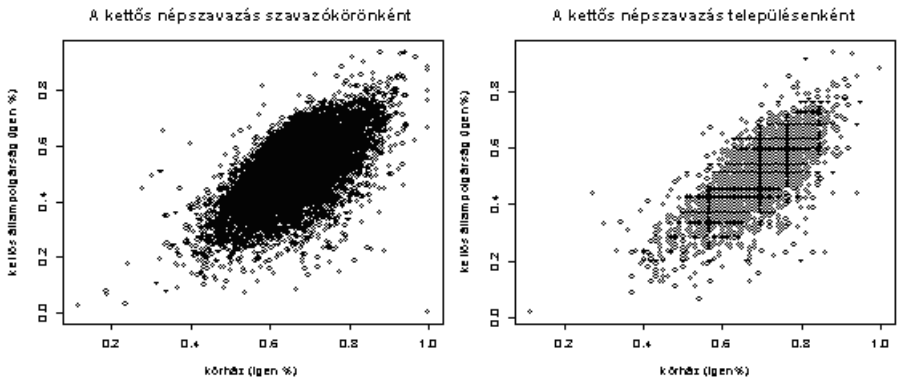
A kérdések megfogalmazása miatt az igen szavazatok a kórház privatizáció elutasítását, valamint a kettős állampolgárság megadásával történő egyetértést jelentették. A következőkben a korábban ismertetett elméleti apparátus használatát, használhatóságát mutatjuk be a népszavazás adatain.



2. ábra A 2004 decemberi kettős népszavazás szavazóköri adatai

A kettős állampolgárságra adott igen szavazatok aránya: 0,516
A kórház privatizációra adott igen szavazatok aránya: 0,65
A kettős állampolgárságra adott igen szavazatok arányának szórása: 0,114
A kórház privatizációra adott igen szavazatok arányának szórása: 0,093

A szavazókörök nagysága összefügg a szavazatokkal, a kisebb szavazókörökben az országos átlagnál több igen volt a kórház privatizációra, és kevesebb igen a kettős állampolgárságra.



3. ábra A 2004 decemberi kettős népszavazás igen szavazatainak korrelációja

A kettős állampolgárságra adott igen szavazatok és a kórházprivatizációra adott igen szavazatok korrelációja a szavazókörök szintjén: 0,6618267, települések szintjén 0,7341489. Ez erős kapcsolatnak tűnik első ránézésre. A kérdés az, hogy ezt a kapcsolatot csak az aggregálációs torzítás szülte, vagy az egyének szintjén is fennálló összefüggésről van szó. Elképzelhető-e hogy ez az összefüggés a szavazókörök aggregálása során alakult ki, a szavazókörökön belül nem áll fenn? Ha települési szinten néztük az összefüggést, az erősebb volt, mint ha szavazókörök szintjén.

A szavazókörök aggregált adatában között viszont erős kapcsolat mutatkozik C és R között. Ez a kapcsolat csak az aggregált adatok között jelentkezik, de amikor a valóságban szembesülünk ilyen adatokkal, nem tudhatjuk, hogy ez az összefüggés fennáll-e az egyének szintjén is.

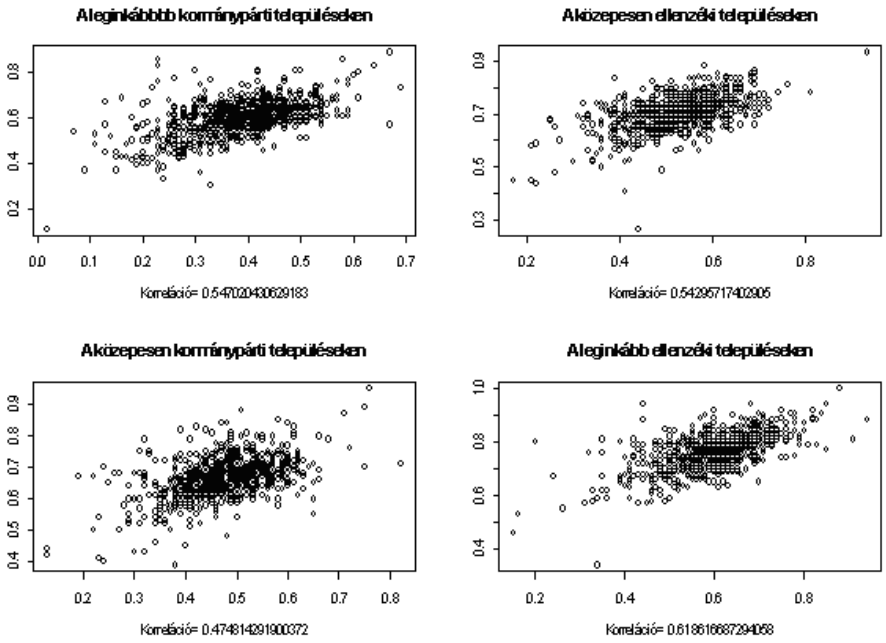
Megfogalmazódhat az a sejtésünk, hogy a szavazókörök összetétele esetleg nem azonos. Tehát van egy olyan kontextuális változónk, ami C-re és R-re is hat, és ennek a változónak az eloszlása változik a szavazókörönként. Az is előfordulhat, hogy a választók eloszlása a szavazókörökben R és C szerint teljesen véletlenszerű, és így az aggregációs torzítás véletlen hatása.

Egy külföldi megfigyelő számára úgy tűnhet, hogy a kettős állampolgárság és a kórházprivatizáció egymástól annyira távol eső kérdések, hogy a köztük jelentkező esetleges ökológiai kapcsolat kizárólag az aggregációs torzításnak tudható be. Külföldi megfigyelőnk megdöbbenten tapasztalta volna, hogy a népszavazást megelőző kampányban szinte kizárólag az igen-igen vagy a nem-nem mellett érveltek politikusaink. (Egyszer-egyszer elhangzott ugyan egy nem-igen is, de ez nem halatszott a nagy hangzavarban.) Ezzel az információval felvértezve úgy módosítaná az álláspontját, hogy az aggregációs torzításból a legnagyobb részt magyarázó kontextuális változó minden valószínűség szerint a pártszimpátia volt, de ezen túl valószínűleg egyéni szintű összefüggést is találhatunk.

Ha úgy gondoljuk, hogy a két választás közötti kapcsolat kizárólag a pártszimpátia okozta aggregációs torzítás műve, akkor azt várnánk, hogy a pártszimpátia szerint hasonló körzetekben független lesz a kettős állampolgárságra és a kórházprivatizációra

adott szavazatok eloszlása. (Ahogy a bevezetőben említett példában is eltűnik az aggregációs torzítás, ha a körzeteket csapadékmennyiség szerinti csoportokra bontjuk.)

Az Európa Parlamenti választások adataival kontrollálva a kettős népszavazás eredményeit, úgy tekinthetjük, hogy az igen-igen mellett a Fidesz, a nem-nem mellett az MSZP és az SZDSZ kampányolt.



4. ábra A 2004 decemberi kettős népszavazás eredményei az Európa Parlamenti választások adataival kontrollálva

Az eredmény az, hogy az Európa Parlamenti választások pártpreferenciái szerinti bontás nem változtatta meg a kapcsolat irányát, de az erősségét csökkentette. Tehát, bár a pártszimpátia okozott némi aggregációs torzítást, ezen túl is marad még összefüggés a kettős népszavazás két kérdése között.

KING MODELLJÉNEK ALKALMAZÁSA

Települési szintű becslés

Választási eredményeket általában nem lehet 2 x 2-es táblázatokkal leírni. Két választás között 4 év telik el, ennyi idő alatt egy választókörzet népességének legalább 10%-a elhalálozás, költözés miatt lecserélődik, a választási rész vevők is változnak. A

jelenség nem elhanyagolható, nem ugyanazok az emberek szavaznak a két alkalommal. Ezért választási eredményeket modellezni csak úgy lehet, ha a 2 x 2-es táblázathoz hozzáfűzünk még egy oszlopot és sort arra az esetre, ha a szavazó – bármilyen okból kifolyólag – távolmaradt. A kettős népszavazás esetében viszont a két szavazás egy időben folyt, és a szavazóköri adatok alapján kevés olyan ember volt, aki megjelent, de csak az egyik kérdésre adott le érvényes szavazatot – szavazókörönként 1-2 fő, azaz kb. a szavazók 0,5%-a. Ennyi ember miatt nem éri meg egy új kategóriát bevezetni, őket nyugodtan vehetjük úgy, mintha a többséggel szavaztak volna a másik kérdésben.

A települési szintre aggregált adatokból King programjának használatával készült becslések:

A becslés eredménye az, hogy azoknak, akik igennel szavaztak a kettős állampolgárságra, 96%-a kórházprivatizációra is igennel szavazott. A kettős állampolgárságra nemmel szavazóknak pedig kb. 60%-a a kórházprivatizációra is nemmel szavazott.

Az adataink szavazóköri szinten állnak rendelkezésre (szeretnénk azt hangsúlyozni, hogy a szavazóköri adatok is már aggregálás eredményei), mikor település vagy választókerület szintjén végzünk elemzéseket, gyakorta olyan következtetéseket vonhatunk le, melyek az aggregációs torzítás eredményei, azaz az adatokban lévő szórást és információt lényegében kiátlagolja. (Szeretnénk emlékeztetni a tanulmány elején ismertetett Simpson Paradoxonra.) Gyakorta az elemzők egyik legnehezebb feladata az, hogy eldöntse, hogy valójában a jelenség esetén aggregációs torzításról beszélhetünk, vagy van egyéb oka is a jelenségnek.

Érdemes néhány jellemző esetet konkrétan vizsgálnunk. Három települést kiválasztva: Budapest II. kerületét, Debrecent és Pécset, vizsgáljuk meg a becsléseket ezekre a településekre. A települési szinten aggregált adatokból a becslés a következő lett:

9. táblázat Települési szinten aggregált becslés

	Kórházprivatizáció	Kettős állampolgárság	p0 (becslés)	p1 (becslés)
Budapest II. ker.	63%	63%	0,87	0,22
Debrecen	70%	56%	0,96	0,38
Pécs	59%	51%	0,93	0,24

Budapest II. (települési szintű becslés)	Kórházprivatizáció: Igen	Kórházprivatizáció: Nem	
Kettős állampolgárság: Igen	az összes II. kerületi 55%-a	az összes II. kerületi 8%-a	63%
Kettős állampolgárság: Nem	az összes II. kerületi 8%-a	az összes II. kerületi 29%-a	37%
	63%	37%	

Debrecen (települési szintű becslés)	Kórházprivatizáció: Igen	Kórházprivatizáció: Nem	
Kettős állampolgárság: Igen	az összes debreceni 54%-a	az összes debreceni 2%-a	56%
Kettős állampolgárság: Nem	az összes debreceni 17%-a	az összes debreceni 27%-a	44%
	70%	30%	

Pécs (települési szintű becslés)	Kórházprivatizáció: Igen	Kórházprivatizáció: Nem	
Kettős állampolgárság: Igen	az összes pécsi 47%-a	az összes pécsi 4%-a	51%
Kettős állampolgárság: Nem	az összes pécsi 12%-a	az összes pécsi 37%-a	49%
	59%	41%	

Szavazóköri szintű becslés

A szavazóköri szintű adatok több információt hordoznak, így elvárnánk, hogy pontosabb becslést adjanak, mint a települési szintre aggregált adatok. A modell becsléseit tovább rontja a települési szintre aggregált adatok esetében az, hogy a településen belüli eloszlást konstansnak vettük, és egy város szavazatainak az eloszlását ugyanolyan mértékben tettük függővé a szomszédos települések eredményeitől, mint az ország másik felében levőkétől.

A három kiválasztott körzetben komoly eltérések jelentkeztek a szavazóköri és a települési szinten becsült eredmények között, ezzel is bizonyítva, hogy a települési szintre való aggregálás közben jelentős torzítások keletkeztek.

10. táblázat Szavazóköri szintű becslés

Budapest II.	p0	p1
Települési szintű becslés	0,87	0,22
Szavazóköri becslés átlaga	0,95	0,09

Budapest II. (települési szintű becslés)	Kórházprivatizáció: Igen	Kórházprivatizáció: Nem	
Kettős állampolgárság: Igen	az összes II. kerületi 60%-a	az összes II. kerületi 3%-a	63%
Kettős állampolgárság: Nem	az összes II. kerületi 3%-a	az összes II. kerületi 34%-a	37%
	63%	37%	100%

Debrecen	p0	p1
Települési szintű becslés	0,96	0,38
Szavazóköri becslés átlaga	0,8	0,6

Debrecen (települési szintű becslés)	Kórházprivatizáció: Igen	Kórházprivatizáció: Nem	
Kettős állampolgárság: Igen	az összes debreceni 45%-a	az összes debreceni 11%-a	56%
Kettős állampolgárság: Nem	az összes debreceni 26%-a	az összes debreceni 18%-a	44%
	70%	30%	100%

Pécs	p0	p1
Települési szintű becslés	0,93	0,24
Szavazóköri becslés átlaga	0,84	0,34

Pécs (települési szintű becslés)	Kórházprivatizáció: Igen	Kórházprivatizáció: Nem	
Kettős állampolgárság: Igen	az összes pécsi 43%-a	az összes pécsi 8%-a	51%
Kettős állampolgárság: Nem	az összes pécsi 17%-a	az összes pécsi 32%-a	49%
	59%	41%	100%

A három vizsgált körzetet összehasonlítva elmondható, hogy a szavazóköri szintű becslések szerint Budapest II. kerületében volt a legpolarizáltabb az eredmény, az összes szavazónak csupán 5%-a nem a nagy pártok kampányának megfelelően szavazott. (Igen-igen, illetve nem-nem.) A debrecenieknél és a pécsieknél ez az arány 37 illetve 25% volt.

A két vidéki városban a szavazók közül jóval többen vették számításba külön-külön a népszavazás két kérdését, mint a Budapest II. kerületében élők, akik gyakorlatilag egy kormány és ellenzék közötti szavazásként értelmezték a kettős népszavazást. Ennek a jelenségnek a magyarázata részletes elemzést igényelne, melyre e cikkben nem vállalkozhatunk

EREDMÉNYIENK ÉS ÍGY KING MÓDSZERÉNEK KRITIKÁJA

King második és harmadik feltevése azt fogalmazza meg, hogy nincs „aggregációs hiba" és az adatok területi homogenitást mutatnak. Jól láthatóan mindkét feltevés nagyon erős (Cho 1998). Az első követelmény problematikus voltára az adatok elemzésének elején rámutattunk, a területi homogenitás feltevésével kapcsolatban pedig tekintsük a következő táblázatot. A táblázat adatai az országgyűlési választások adatainak területi inhomogenitását mutatják be tömbönként (Mészáros–Szakadát 2005). A táblázat adatai az egyes politikai tömbökre számított autokorrelációs mértékek.

11. táblázat Autokorrelációs mértékek a 176 választókerületre politikai blokkonként

Év	„Jobboldali" pártok				„Baloldali" pártok				Liberális pártok			
	Moran I	p	Geary c	p	Moran I	p	Geary c	p	Moran I	p	Geary c	p
1990	04,5	0,001	0,574	0,001	0,405	0,001	0,574	0,001	0,404	0,001	0,587	0,001
1994	0,419	0,001	0,551	0,001	0,419	0,001	0,551	0,001	0,293	0,001	0,685	0,001
1998	0,374	0,001	0,596	0,001	0,374	0,001	0,596	0,001	0,312	0,001	0,661	0,001
2002	0,451	0,001	0,528	0,001	0,451	0,001	0,528	0,001	0,327	0,001	0,639	0,001

Az adatok egyértelműen mutatják King módszere 3. feltevésének problematikusságát.

Összefoglalva megállapíthatjuk, hogy a az ökológiai probléma feltevések nélkül nem megoldható, általánosságban a feltevések alkalmazását gyakorlati vizsgálatnak kell az adott adathalmazon megelőznie. A módszer individuális mintából származó adatokkal javítható (Wakefield 2004), de ez már egy másik tanulmány tárgya lehet.

IRODALOM

Achen, Ch.–Shively W.Ph. (1995): *Cross Level Inference*. Chicago and London: The University of Chicago Press.

Appleton, D.R.–French, J.M.–Vanderpump, M.P. (1996): Ignoring a covariate: An example of Simpson's Paradox. *American Statistician*, 50: 340–341.

Cho, W.T. (1998): If the Assumption Fits... : A Comment on the King Ecological Inference Solution. *Political Analysis,* 7: 143–163.

Fotheringham, A.S.–Wong, D.W.S. (1991): The Modifiable Areal Unit Problem in Multivariate Statistical Analysis. *Environment and Planning*, 23: 1025–1045.

Duncan O.D.–Davis B. (1953): An Alternative to Ecological Correlation. *American Sociological Review*, 18: 665–666.

Goodman, L. (1953): Ecological Regressions and the Behavior of Individuals. *American Sociological Review*, 18: 663–665.

Goodman, L. (1959): Some Alternatives to Ecological Regression. *American Journal of Sociology,* 64: 610–624.

Hox, J.J. (1994): *Applied Multilevel Analysis.* Amsterdam: TT-Publikaties.

King, G. (1997): *A Solution to the Problem of Ecological Inference*. Princeton, New Jersey: Princeton University Press.

King, G. (2000): Geography, Statistics, and Ecological Inference. *Annals of the Association of American Geographers*.

King, G.–Rosen, O.–Tanner, M.A. (2004): *Ecological Inference.* Cambridge: Cambridge Univ. Press.

King, G. (é.n.): A Program for Ecological Inference.

Mészáros J.–Szakadát I. (2005): *Magyarország Politikai Atlasza.*

Robinson, W.S. (1950): Ecological Correlation and the Behavior of Individuals. *American Sociological Review*, 351–357.

Simpson, E.H. (1951): The Interpretation of Interaction in Contingency Tables. *Journal of the Royal Statistical Society,* 13: 238–241.

Wakefield, J. (2003): Ecological Inference in 2 × 2. Tables. *Journal of the Royal Statistical Society.*