Tóth Gergely

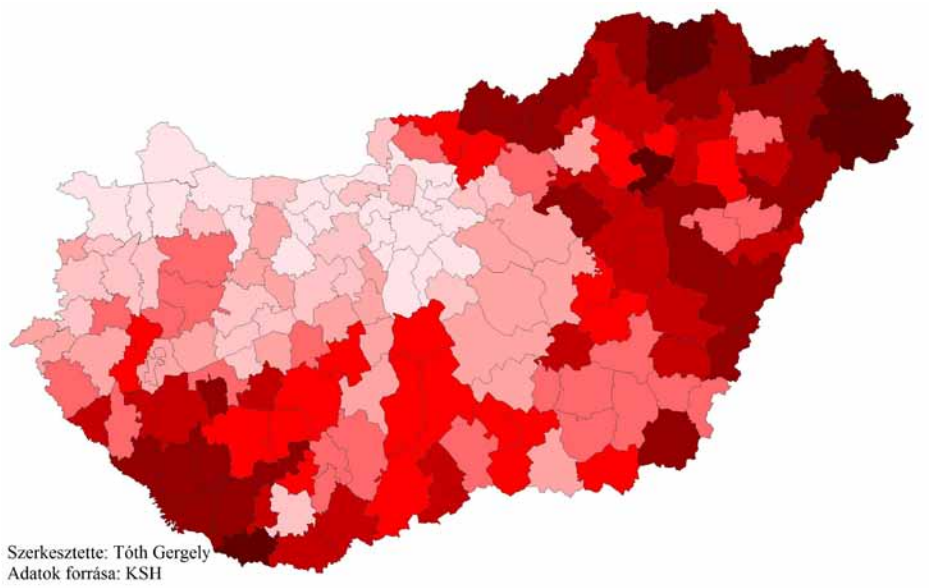
Az elveszett függetlenség nyomában – mintavételből származó társadalomtudományi adatok függetlenségének vizsgálata[1]

Bevezető

Jelen tanulmány alapkérdései már évekkel ezelőtt megfogalmazódtak bennem, mivel a kutatási terület, amivel foglalkozom, elsőrendűen a társadalom térbeli strukturálódása. A területi adatok egyik legfőbb jellegzetessége – amely részben a térfelosztás problematikájára és az adatgyűjtés hierarchikus jellegére is visszavezethető –, hogy a megfigyelési egységek (pl.: település, kistérség) egymástól vett függetlensége nem áll fenn, azaz a távolság/szomszédság függvényében az egyes megfigyelési egységek értékei hasonlóak, vagy más szempontból megközelítve a kérdést: az egyes megfigyelések értékei lokációs háttérváltozókkal magyarázhatóak. A legegyszerűbb, és többek számára ismert példa talán a munkanélküliségi eloszlása országosan (1. ábra), amely vitán felül képes rámutatni a területi szempontok fontosságára.

1 A tanulmány megírásában a szerzőt segítette a TAMOP 4.2.1./B-09/1/KMR-2010-0003. program keretében nyújtott támogatás.

1. ábra: A munkanélküliségi ráta – kistérségenként (2007)



A területi adatok elemzésének módszertana az utóbbi évtizedekben sokat fejlődött – bár tudományterületenként részben más és más módszertan és terminológia honosodott meg –, a területi elemzések egyik legalapvetőbb mutatószáma a területi autokorrelációs érték (ezek közül is talán az egyik legismertebb a Moran-féle I mutató). Ugyanakkor ennek a mutatónak a jelentősége jelen tanulmány számára csak annyiban érdekes, hogy interpretációs szempontból a legfontosabb jellemzőre, az autokorreláltság jelenségére mutat rá, amely az idősoroknál is használt autokorrelációhoz hasonlóan a megfigyelési egységek összefüggésének (interdependenciájának) egyfajta globális mu-

tatószáma.[2] A térbeli struktúrájú adatok elemzése esetén tehát fel sem merül kérdésként a kutatók számára az adatok függetlenségének problémaköre.[3]

Társadalomtudományi szempontból véleményem szerint ugyanakkor még egy fontos elméleti kérdéskörre képes a területi elemzések módszertana rámutatni: a működésben lévő társadalom, mint kommunikáción alapuló, valamint működésében kapcsolathálózatként felfogható rendszer állapotáról mint indikátor árulkodhat a térbeli struktúra, amely ugyanakkor erős stabilizáló/konzerváló visszahatással is bírhat a társadalmi folyamatokra (gondolva itt többek között kapcsolathálózati elemzések néhány alapvető elméleti elemére, így a gyenge és erős kötések jelentőségére [Granovetter 1973], de mindenekelőtt a homofília jelenségének fontosságára (a témát területi elemzéssel összekapcsolt módon lásd: Kmetty–Tóth 2011). Ezen gondolatok megfelelő konzisztens elméletbe foglalása ugyanakkor jelen tanulmánynak nem lehet célja, hanem csak annyiban kerül előtérbe, amennyiben részben mellékes célként szociológiai szempontú interpretációs magyarázattal is kíván szolgálni a tanulmányban bemutatott jelenségcsoportra.

A vizsgált kérdéskör

A bevezető gondolatok megalapozását kívánták adni a tanulmány fő kérdésfelvetésének: a társadalmi adatok elemzése esetén a társadalom szükségszerűen interdependens szövete mennyiben jelentkezik torzító tényezőként azon bevett módszertani eszközök használata során, amelyek a megfigyelési egységek függetlenségét feltételezik?

A jelen tanulmányban felhasznált kísérlet kialakításakor tehát arra a kérdésre kerestem a választ, hogy a többlépcsős mintavétel utolsó előtti lépcsőjét jelentő települési szintet vizsgálva tapasztalható-e klaszterződés az adatokban, azaz azt a kérdést tettem fel, hogy vajon a településeken belül megfigyelhető-e az interdependencia jelensége, és ha igen, akkor annak mértékére milyen

2 Ugyanakkor mind számításbeli, mind interpretációs eltérések is vannak a két jelenség között, amire jelen tanulmányban nem kívánok kitérni. (Magyarul áttekintő módon lásd Dusek [2004].)

3 Ugyanakkor fontos megjegyezni, hogy a társadalomtudományok területén belül – feltehetőleg belső tudományfejlődési okokra is visszavezethető módon – a területi elemzések fontossága, de legfőképpen speciális elemzési szempontjai sajnálatos módon a mai napig nem tudtak kellőképpen integrálódni, aminek révén a tanulmányokban a mai napig találkozhatunk statisztikai értelemben rosszul megalapozott területi elemzésekkel.

becslés adható. Azaz operacionalizálva a kérdést: a településeken belül az elemzési egységek jobban hasonlítanak-e egymásra annál, mint amit a teljes sokaság alapján várhatnánk?[4]

A kérdéskör ugyanakkor már ránézésre is összetett, hiszen egyrészt el kell különíteni a teljes sokaságra jellemző valós (azaz önmagában is létező) klaszterződési hatásokat a mintában találtaktól, másrészről számszerűen is ki kell tudni mutatni a mintavételi design által létrejött interdependencia mértékét is.

A kérdés kimerítő megválaszolását természetesen jelen tanulmány korlátai nem teszik számomra lehetővé, ezért csak a kérdés jogosságát kívánom néhány egyértelmű példán át igazolni. Így az elemzés csupán egy speciális – de a leginkább elterjedt – adatgyűjtő eljárás, a többlépcsős mintavételnél jelentkező hatások bemutatására szorítkozik, és a többváltozós elemzésekre gyakorolt hatások kérdéskörére már nem terjed ki.

Ugyanakkor viszont, annak érdekében, hogy állításomat azok is befogadhassák, akik a statisztika e speciális területén nem jártasak, a tanulmányban röviden bemutatásra kerülnek a megértéshez szükséges alapvető statisztikai módszerek is.

Az elemzéshez szerencsés módon több szükséges adatbázis is rendelkezésre állt. Így egyrészt a 2001-es népszámlálási adatok (illetőleg az azok alapján rekonstruált teljes sokasági adatbázis), valamint sikerült több olyan kérdőíves vizsgálat adatát is felhasználni, amelyek speciálisnak nevezhető módon tartalmazták a kérdezett lakhelyét is vagy pseudo településkódokkal, vagy valós ksh kód értékekkel.[5] Az adatok relatív bősége lehetővé tette a több szempontú összehasonlításokat, ellenőrzéseket és szimulációkat is, amelyek ilyen módon nagymértékben növelik az eredmények megbízhatóságát.

A vizsgálat során feltett hipotézisekre adott megerősítő válaszok jelentősége kettős: egyrészt amennyiben léteznek ezen településszintű összefüggések, akkor azok lényegében annak bizonyítékaként értelmezhetőek, hogy a megfigyeléseink településenkénti függetlensége nem teljesül, azaz az egyik igen fontos statisztikai alapfeltevés, a függetlenség hipotézise sérül. Másrészről viszont

4 Ugyanakkor mindenképpen fontos megemlíteni az interdependencia kapcsán, hogy statisztikai értelemben ez másképpen is értelmezhető, hiszen nem feltétlen a függetlenség hiánya jelenik meg ebben, hanem az lényegében csak az összefüggésnek a településen belüli erősebb volta, szemben azzal az esettel, amikor egy településről csak egy embert kérdeznénk meg, tehát a jelenség léte önmagában nem utal egyértelmű kauzális viszonyra.

5 Az elemzésben felhasznált adatbázisok köre: WVS – World Values Survey: TÁRKI, 2009 – 1007 fő; DKMKA – preelection adatfelvétel: IPSOS–Medián, 2010 – 1500 fő; EVS – European Values Study: Forsense, 2008/2009 – 1500 fő.

ezen összefüggések létezése rámutat arra is, hogy a többlépcsős mintavétel révén egy torzító design effectet viszünk az adatainkba, tehát az azokból származó becslések (minimálisan a szórások tekintetében) torzítást tartalmazhatnak.

Az irodalomban a csoporton belüli hasonlóság tesztelésére egyértelműen az ANOVA modellcsalád alkalmazását említik, amely eredeti kérdésfeltevését illetően nem is feltétlenül alkalmas a kérdés eldöntésére, de a modell logikáján alapulva az irodalomban fellelhetőek további mérőszámok is.

Alkalmazott eljárások

Ahhoz, hogy egyértelmű legyen az olvasók számára az elemzés során használt statisztikai mutatószámok értelmezése, mindenképpen célszerűnek tűnt számomra azokról röviden szót ejteni, tehát ebben az alfejezetben az ANOVA modellcsalád néhány elemét tekintem át:

1. ANOVA és az azon alapuló eta^2 hatásnagyság mutatószám (fix tényezős modell –Type I)
2. ICC – intraclass korrelációs együtthatók:
   a) ANOVA alapú ICC1 érték
   b) Random tényezős modellen alapuló ICC érték

One-way ANOVA és eta^2

Az ANOVA modellcsalád alapját a szórásfelbontás lehetősége jelenti, amennyiben egy kategoriális csoportváltozó alapján a célváltozó eloszlását vagyunk képesek vizsgálni. A fix tényezős modell esetén azt a hipotézist vizsgáljuk, hogy az egyes csoportok átlagai statisztikai értelemben megegyeznek-e egymással.

Az egytényezős modell esetén kettő vagy annál több sokasággal foglalkozunk, amelyekről feltesszük, hogy N(μi, σ^2) eloszlásúak, ahol μi az i-edik sokaság várható értéke, σ^2 pedig az i-edik sokaság varianciája. A tesztelés során arra a kérdésre keressük a választ, hogy a különböző csoportok között megfigyelt eltérések a véletlennek tulajdoníthatók-e, vagy már olyan mértékűek, ami alapján joggal feltételezhetjük, hogy különböznek. Ilyen módon a következő nullhipotézis elfogadásáról vagy elvetéséről szükséges döntenünk:

$$H0 : \mu i = \mu, \qquad \text{ahol } i = 1,2, \ldots k, \qquad k = \text{a csoportok száma}$$

A döntés alapját képező F statisztika értéke a következő módon számítandó:

$$F = \frac{MS_{Between}}{MS_{Within}}$$

,

ahol

$MS_{Between} = \frac{SS_{Between}}{df_{Between}}$, ahol $SS_{Between}$ a csoporton belüli átlagos négyzetes eltérés,

és

$MS_{Within} = \frac{SS_{Within}}{df_{Within}}$, ahol SS_{Within} a csoportok közötti átlagos négyzetes eltérés

(a df értékek az adott négyzetes eltérésekhez tartozó szabadságfokok, amelyek a csoport, illetőleg elemszámok függvényeként írhatóak fel).

Ha a kapott *F* érték nagyobb, mint az adott paraméterek mentén a kritikus Fp érték, akkor a nullhipotézist elvetjük, azaz elfogadjuk, hogy a csoportokon belül megfigyelt eltérések nem csak a véletlen ingadozás eredményeként jöttek létre.

Az egytényezős ANOVA használatának ugyanakkor van pár kritikus alapfeltevése:

– Függetlenség (independence of cases)
– Normál eloszlás (normality)
– Szóráshomogenitás (homoscedasticity).

Ezen feltételek teljesülése ugyanakkor erősen kérdéses a legtöbb társadalomtudományi adat esetén, hiszen lényegében jelen tanulmányban is épp ezen szempontok közül az egyik létezését kívánjuk megkérdőjelezni. Ezen feltételek megsértése (habár változó súllyal – Monday et al. 2005) torzítja az aszimptotikus becsléseket, amelyek kivédésére az irodalomban különféle korrekciós megoldási javaslatokat találhatunk. Így például:

1. Normalitás megsértése esetén (amire nagy esetszámoknál leginkább robusztus az ANOVA eljárása) a célváltozók matematikai transzformációkkal történő korrekcióját.

2. Szóráshomogenitás megsértése esetén a robusztusabb Brown–Welch-próbák alkalmazását.
3. Míg a függetlenség megsértése esetén súlyozással való kompenzációt, azaz az effektív mintanagyság csökkentését.
4. Ugyanakkor több szempontból talán a leginkább megbízható megoldásnak a permutációs tesztek használata javasolt, hiszen ezek által nem egy elméleti eloszláshoz viszonyítva vagyunk képesek a kapott eredmények kiértékelésére, hanem egy a vizsgált eloszlással azonos paraméterekkel rendelkezőhöz viszonyítva.

Ameddig az ANOVA alapmodell eredendően csak egy globális választ próbál adni a különböző csoportok azonosságára (szignifikánsnak tekinthetők-e az eltérések), addig az ANOVA modellből levezetve több, az alternatív modellek összehasonlítására alkalmas mutatószám is létezik, amelyek a hatásnagyságot (effect size) kívánják kifejezni. Az ANOVA alapú hatásnagyság-mérőszámok közül talán legismertebb az eta-négyzet (eta-squared) mutatószám.
Az eta-négyzet értékének könnyű interpretálhatósága abból ered, hogy lényegében az értéke megegyezik egy olyan regressziós modell R-négyzet értékével, amelyben N–1, a kategoriális változóból képzett dummy változó jelenti a független változók körét. Ilyen módon sokszor mint a modell által kifejezett százalékos magyarázóerőként szoktak rá hivatkozni. Ugyanakkor, hogy érthető lehessen számunkra az eta-négyzet kiemelt jelentősége, szükségszerűen érdemes áttekinteni kiszámításának módját is:

$$\eta^2 = \frac{SS_{Between\ (csoportok\ közötti\ eltérés\ négyzetösszegek)}}{SS_{Total\ (összes\ eltérés\ négyzetösszege)}}$$

Érdemes észrevenni, hogy logikailag egy tiszta hányadszámítással állunk szemben, hiszen a SSTotal = SSBetween + SSWithin. Azaz a képlet azt fejezi ki, hogy az átlagtól való eltérés-négyzetösszegek hány százaléka keletkezik abból adódóan, hogy a csoportok eltérnek egymástól. Ugyanakkor mindenképpen fontos felhívni a figyelmet arra, hogy az eta-négyzet értéke adott esetben alapvetően ugyanúgy szenved az ANOVA alapfeltevéseinek hiányosságaitól, mivel azok megsértése esetén a számított eta-négyzet értéke bizonytalan interpretálásúvá válik.

A képlet alapján ugyanakkor érdemes felhívni a mutatószám egy további fontos hiányosságára is a figyelmet, amennyiben láthatóvá válik, hogy a számítási módszer nem reflektál sem a csoportok számára, sem az egyes csoportokon belüli megfigyelések számára. Ebből adódóan az eta-négyzet értéke – főként kis esetszámok, illetőleg magas csoportszámok esetén – rendkívüli mértékben félrevezető lehet, amit Bliese és Halverson tanulmányukban látványosan prezentálnak (Bliese–Halverson 1998). Bliese-ék mesterséges adatokon bemutatják, hogy extrém esetben – diádok esetén – az eta-négyzet értéke akár tökéletes függetlenség esetén is 0,5 értéket vesz fel, azaz 50%-ra becsüli a csoportosítás hatását, akkor is, amikor mesterségesen létrehozott független adatok jelentik a számítások alapját.

Vizsgálatunk szempontjából a fix tényezős modell bemutatása ugyanakkor hiányosságai ellenére is mindenképpen alapvető jelentőségű, mivel elemzési szempontból egy a szociológusok körében általánosan bevett és ismert eljárásnak tekinthető, valamint statisztikailag is az alapját jelenti a további, alkalmazott mutatóknak.

ICC- intraclass korrelációs együttható

Az ICC mutató alapvetően a válaszadók egyfajta relatív következetességére próbál számszerűen utalni. Jelentősége elsőrendűen az olyan tudományterületeken kimagasló – és ezért jól kutatott –, ahol az adatokban meglévő csoportosítás – pl.: hierarchikus szervezetek elemzése – természetes jelenség. Az intraclass korreláció a többszintű (multilevel) modellek alkalmazása esetén annak egyik alapvető indikátoraként is használható, hogy a különböző szintek (levels) közül melyek azok, amelyek az adatok strukturális elemzésekor szerepet játszanak (random tényezős modellek szempontjából lásd: Pinheiro–Bates 2000).

Mindenképpen külön érdemes foglalkozni az ICC értékek magyarázatával is, mivel legalább két különböző kontextusban is értelmezhetőek: a csoportátlag megbízhatóságának mértékeként, illetőleg a függetlenség hiányának fokmérőjeként. Mint „megbízhatósági" (reliabilty) mutatószámra olyan értelemben utalhatunk, mint annak a mértékére, hogy a csoporton belüli válaszok mennyiben tekinthetőek konzisztensnek (Kozlowsky–Hattrup 1992), amennyiben a konzisztencia mértékét a saját átlagtól való relatív eltérés mértékeként értelmezzük (Bliese 2000).

Ugyanakkor, ami a jelen elemzés számára fontosabb, az ICC értékét a függetlenség hiányának mutatószámaként is értelmezhetjük. Ez esetben az értelmezési hangsúly áthelyeződik, és így fogal-

mazhatjuk meg a kérdést: vajon a vizsgálati egységre (egyénre) mekkora hatással bír a csoporttagság? Ilyen módon annak mértékére szeretnék becslést adni, hogy az egyén döntésére a csoport befolyása milyen mértékű. Szervezetelemzési szempontból Bliese a mögöttes oksági mechanizmus tekintetében több példával is szolgál (bővebben lásd: Bliese 2000), ugyanakkor szociológiai, kapcsolathálózati elemzési szempontból akár kevésbé formális rendszerekre vonatkozóan is találhatunk magyarázatokat. (Ilyen módon kiemelkedően érdekesek például Angelusz és Tardos négy falu vizsgálata alapján levont következtetései, amennyiben a személyek településen belüli kapcsolathálózati pozícióját az egyéni döntést befolyásoló tényezőként értelmezik (Angelusz–Tardos 2009).

Az említett két fogalmat, a „megbízhatóság"-ot és a „összefüggőség"-et talán gyakorlati jelentőségük szempontjából lehet leginkább elkülöníteni: amikor az ICC értékét, mint megbízhatósági mutatót tekintjük, elsőrendűen arra a kérdésre szeretnénk választ adni, hogy az adatok aggregálása után számíthatunk-e emergens jelenségek, azaz új típusú, aggregált elemzési szinten megjelenő összefüggések megjelenésére? (Így például míg az ICC 0 és 1 értéke esetén azonos eredményeket kapunk az aggregált és esetszintű változók korrelációjára, addig a köztes értékek esetén [0<ICC<1] várható az aggregált változók kapcsán emergens jelenségek megjelenése [Bliese 2000].)[6] Ezzel szemben, amikor az ICC értékét mint összefüggőségi értéket értelmezzük, lényegében arra a kérdésre keressük a választ, hogy vajon a célváltozó tekinthető-e a csoporttagság által befolyásoltnak?

Amíg longitudinális és geográfiai adatok esetében az összefüggőség mérőszámaként autokorrelációs értékeket tudunk számolni, amelyek révén bevett módon tesztelhetőek az egyes megfigyelések közötti korrelációs struktúrák, addig az olyan esetekben, amikor az interdependencia a csoporttagság révén várható, a leginkább bevett eljárásnak az ICC értékek használata nevezhető. Az ICC abszolút értékben 0 és 1 közötti értéke annak indikátora, hogy a csoporthoz tartozásnak az egyének tekintetében mekkora a befolyásoló szerepe.

Habár az ICC kérdésfelvetése az eddig elmondottak alapján elsőre látszólag nem különbözik az ANOVA modell kérdésfelvetésétől, az alkalmazott képletek alapján értelmezési szempontból egyértelműen elkülöníthető attól.

Az ICC intraclass korrelációs mutató számítására több képlet is létezik, és azok egymáshoz való viszonya nem minden esetben lineárisan függvényszerű. Mivel számunkra a vizsgálat szempontjá-

6 Ez utóbbi jelenség részletes vizsgálata véleményem szerint mind a módosítható területi egység problematikája (MAUP), mind akár a Simpson-paradoxon vizsgálata kapcsán is további vizsgálódásokat igényelne.

ból a mutatóra elsőrendűen mint a függetlenség tesztelésének eszközére tekintünk, ezért részletesen csak az úgynevezett ICC fő indikátorral foglalkozunk, amely az irodalomban többek között ICC(1), vagy ICC(1,1) néven is ismert. Byrk és Raudenbush (Byrk–Raudenbush 1992) az ICC értékét úgy definiálják, mint a teljes varianciából a csoporttagság által kifejezett hányad, és matematikailag a random tényezős modellből (hierarchikus lineáris modell) vezetik le. Ilyen módon:

$$ICC = \frac{\tau_{00}}{\tau_{00} + \sigma^2} \quad ,$$

ahol τ_{00} a csoportok közötti varianciát, míg a σ^2 a csoporton belüli varianciát jelenti. A random tényezős modell alapján számított ICC értékét interpretációs szempontból úgy is szokták értelmezni, mint azt az értéket, amely százalékosan kifejezi, hogy az adott individuális szintű célváltozóra mekkora befolyással bír a csoporttagság (Bliese 2000).

Ugyanakkor más szerzők (Bartko) közvetlenül az ANOVA modellből vezetik le az ICC értékét, és a következő definícióját adják:

$$ICC(1) = \frac{MS_{Between} - MS_{Within}}{MS_{Between} + [(k-1) * MS_{Within}]},$$

ahol az $MS_{Between}$ és az MS_{Within} az eddigieknek megfelelően a csoportok közötti és csoporton belüli átlagos négyzetes eltérést jelentik, valamint a k a csoportméretet. (Különböző méretű csoportok esetén a k értéke vagy az átlagos csoportmérettel helyettesítendő, vagy szélsőséges különbségek esetére Bliese és Halverson alapján Blalock kompenzációs képlete ajánlható (Bliese–Halverson 1998):

$$k = (df_{Within} + (df_{Between} + 1)) / (df_{Between} + 1)$$

A legfőbb különbség a két modell között [továbbiakban: ICC és ICC(1)] a felvehető értékek tartományában keresendő: amíg az ANOVA modell alapján számított ICC(1) mutató a [–1; +1] tartományban vehet fel értékeket, addig a random tényezős ICC modell a [0; +1] értéktartományban. Az ANOVA alapú ICC(1) modell a minimumát akkor veheti fel, ha a csoportátlagok azonosak (

$MS_{Between} = 0$), míg a random tényezős modell alapján számított ICC a minimumát értelemszerűen a $\tau_{00}\tau_{00}$ = 0 értéke mellett veszi fel.

Az elemzésben megtalálható számítások és szimulációk nagy része az R statisztikai programcsomagban készült.[7] azon belül is elsőrendűen a „multilevel" csomag elemeit alkalmaztam (Bliese 2008). A tanulmányban alkalmazott módszerek jól dokumentáltak, mindenki számára szabadon hozzáférhetőek és megismételhetőek.

Vizsgálat

A vizsgálat megtervezésekor alapvető célként az egymással összevethető és többszörösen ellenőrizhető design kialakítását tűztem ki célul.[8] Ahogy már fentebb utaltam rá, ehhez a célkitűzéshez több különböző adatbázist is felhasználtam, ilyen módon a 2001-es településsoros népszámlálási adatokat, valamint három különböző többlépcsős minta révén létrejött állományt (EVS, WVS, DKMK). Ahhoz, hogy a különböző állományok összehasonlíthatóak legyenek, több adatbázis-transzformációt is végre kellett hajtanom. Így először is létre kellett hoznom a településsoros népszámlálási állományból egy elemzésre alkalmas teljes populációt tartalmazó rekonstruált adatbázist. Ezt ugyanakkor természetesen csak olyan felbontásban tudtam megtenni, ahogyan a népszámlálási adatok rendelkezésre álltak. Ilyen módon az iskolai végzettség mint elemezendő paramétert lett kiemelve, amely többszörösen is előnyösnek ígérkezett:

– az iskolai végzettség egy kemény változó, amelyre a vonatkozó vizsgálati eredmények egyértelműen relevánsak, hiszen az iskolai végzettség a legtöbb társadalomtudományi elemzés egyik alaptényezője
– az iskolai végzettség garantáltan minden adatbázisban megtalálható információ
– az iskolai végzettség kódolása kismértékű torzítás mellett standardizálható akár magas mérési szinten is.

7 Fontos megjegyezni, hogy a rekonstruált népszámlálási adatok elemzésének egy részére memóriaproblémák miatt az R2.14.1 nem volt alkalmas. Ezen részek elemzését, illetőleg néhány ellenőrző számítást SPSS-ben végeztem el.

8 A vizsgálati design kialakításában Rudas Tamás volt segítségemre, aki jól feltett kérdéseivel és éleslátásával segítette elő a kutatást. Ezúton is szeretnék köszönetet mondani többszörös segítségéért!

Tehát jelen tanulmányban az iskolai végzettség településenkénti klasztereződési struktúráját vizsgáltam, olyan módon, hogy ehhez a feladathoz mind a négy adatbázisban létrehoztam az egységes „elvégzett iskolai osztályok száma" ('iskola') változót, illetőleg ezen dimenzió mentén rekonstruáltam a népszámlálási adatokból a teljes populációt.[9]

További alapkérdésként jelentkezett, hogy érdemes-e minden településre kiterjesztenem a vizsgálatot, vagy valamilyen szempontból érdemes szűkíteni a települések körét. A szűkítés mellett ugyanis több érv is szólt:

– Budapest és a megyeszékhelyek (illetve megyei jogú városok) a legtöbb mintában önreprezentálóak. Ebből adódóan az eredményekre gyakorolt esetleges torzítás feltehetőleg sokkal kevésbé érinti az ezen településről jövő válaszok körét.
– Amennyiben vizsgálatom során valójában a megfigyelések függetlenségének problémakörét szeretném a középpontba állítani, szemben valami globálisabb jellegű (például gazdasági) tényezővel, akkor elméletileg várható, hogy a kisebb településeken sokkal inkább a lokális viszonyok (pl.: faluközösség, civil társadalom, kapcsolathálózat) lehetnek hatással az egyéni döntésekre (így az iskolai végzettség tekintetében elvárt értékekre). Azaz elméleti szempontból a függetlenség hiányát elsősorban a kisebb települések esetén vártam.

Ilyen módon a vizsgálat során végül csak a kisebb településeket (azaz a városokat és a falvakat) tartalmazó adatállományokat használtam fel.

A) Az adatbázisok rendezése után minden adatbázisra lényegileg ugyanazon alapszámításokat végeztem el:

One-way ANOVA (F érték és a hozzá tartozó szignifikanciaszint)

Eta-négyzet értéke (mintavételi adatbázisok esetén súlyozva és súlyozatlan módon)

ICC(1) ANOVA modellen alapuló intraclass korrelációs együttható (mintavételi adatbázisok esetén súlyozva és súlyozatlan módon)

ICC – random tényezős modellen alapuló intraclass korrelációs együttható (programozástechnikai okokból csak súlyozatlan módon).

9 A népszámlálási adatokkal szemben esetlegesen felhozható ellenérvként, hogy időben relatíve távol esik a másik három adatfelvételtől. Habár tényszerűen az adott időszakban trendszerűen nőtt a népesség iskolázottsága, és várhatóan részben területileg is átalakult annak szerkezete, ismervén a hazai adatok nagyfokú stabilitását, teljes bizalommal kijelenthető, hogy a vizsgálat során kiszámolt eredmények várhatóan csak csekély eltérést mutatnának akár egy friss, 2011-es – sajnos még rendelkezésre nem álló – népszámlálási adatbázishoz viszonyítva.

B) Ahogyan az előzőekben felvázoltam, több tényező is amellett szólt, hogy az eredményeket permutációs próba segítségével is ellenőrizzem. Ilyen módon minden adatbázis esetén készültek szimulációs tesztek, amelyek során adottnak vettem az adatok eloszlását (iskolai végzettségek, településekről a megkérdezettek száma), de randomizálással pseudo- településkódokat hoztam létre, amelyekre újfent kiszámoltam a fenti mutatószámokat. Ezt a folyamatot minden adatbázis esetén 1000-szer megismételtem, majd az egyes szimulált adatokon kapott mutatószámok eloszlását vetettem össze az eredetileg mért értékekkel. Az eredmények könnyed áttekintése érdekében a különböző szimulációs eredmények eloszlásait egymásra vetítve grafikusan is megjelenítettem.

C) Az eredmények értelmezése közben felmerült a kérdés, hogy a kapott minták eredményei vajon valamiféleképpen specifikusaknak tekinthetőek-e, azaz mennyiben kapnánk hasonló eredményeket, ha a népszámlálási adatokon újra rekonstruálnánk ugyanazt a mintavételi designt, amit az egyes adatfelvételek során alkalmaztak. Ennek érdekében az egyik technikailag megfelelő adatbázist,[10] a DKMK-mintavétel eredményeit a népszámlálási adatok alapján megpróbáltam szimulációval rekonstruálni – a permutációhoz hasonló módon, 1000-szer megismételve a szimulált mintavételt. Ebből is származott tehát az egyes mutatók tekintetében egy-egy eloszlás, amit össze lehetett vetni a mért értékekkel, illetőleg azok permutációs eloszlásaival is.

Elemzés

A) Alapszámítások az egyes adatbázisokon

Ahogyan az 1b)–1d) táblázatokból látható, mindegyik F érték szignifikáns, tehát ezek alapján az iskolai osztály várható értéke tekintetében a települések különböznek egymástól. Az F értékek sorrendjének tekintetében (elméletileg az F = 1 érték a függetlenség indikátora) azt tapasztaljuk, hogy az EVS–DKMK–WVS–Népszámlálás sorrend alakult ki, tehát leginkább a népszámlálási adatok esetén várnánk erős összefüggést, míg az EVS esetén a legkevésbé.

10 A WVS esetén csak fiktív településkód állt rendelkezésre.

1a) táblázat: ANOVA tábla: Népszámlálási adatok

Népszámlálási adatok	Df	Sum.Sq.	Mean.Sq.	F value	Pr(>F)	Eta-square
Telep	3111	w2230395.157	716.9383	114.442	0.00E+00	0.0678
Residuals	4898164	30685212.22	6.264636			

1b) táblázat: ANOVA tábla: WVS

WVS	Df	Sum.Sq.	Mean.Sq.	F value	Pr(>F)	Eta-square
Telep	52	982.7	18.9	3.75	3.88E-15	0.2593
Residuals	557	2807	5.04			

1c) táblázat: ANOVA tábla: EVS

EVS	Df	Sum.Sq.	Mean.Sq.	F value	Pr(>F)	Eta-square
Telep	100	1148	11.48	1.818	6.60E-06	0.1767
Residuals	847	5349	9851			

1d) táblázat: ANOVA tábla: DKMK

DKMK	Df	Sum.Sq.	Mean.Sq.	F value	Pr(>F)	Eta-square
Telep	105	1230	11.712	2.381	1.97E-11	0.2420
Residuals	783	3851	4.919			

Ugyanakkor, ha megvizsgáljuk az ANOVA alapján számított eta-négyzet-értékeket, már más a sorrend, a Népszámlálás–EVS–DKMK–WVS sorrendet kapjuk. Ahogyan a korábbiakban szót ejtettem róla, az ANOVA modell és az eta-négyzet értékének megbízhatósága relatíve alacsony, hiszen többek között a csoportok mérete nagyon különbözik. Ebből adódóan a sorrendiség tekintetében

nem feltétlen értékelhetők az adott értékek, csupán csak figyelemfelkeltésül szolgálnak, amennyiben az eredmények arra utalnak, hogy az adatokban létezik a klasztereződés jelensége.

Annak eldöntésére, hogy valamiféleképpen megbízhatóbb becslést adjunk az interdependencia fokára, kiszámítottam a fent említett többi mutató értékét is (2. táblázat). A különböző mutatók értékei trendszerűen az eta-négyzet-mutatóval együtt mozogtak, tehát a fentebbi Népszámlálás–EVS–DKMK–WVS sorrend tűnik érvényesnek.

2. táblázat: Klasztereződési mutatók: népszámlálási adatok

	ANOVA-F	ANOVA-Pr	Eta-square	weighted-Eta-square	ICC1	weighted-ICC1	ICC-random-effects
Népszáml.	114.4421	0.00E+00	6.80%		6.70%		4.90%
WVS	3.7498	3.88E-15	25.90%	24.70%	19.30%	17.90%	19.50%
EVS	1.8177	6.60E-06	17.70%	20.40%	8.00%	11.00%	8.50%
DKMK	2.3812	1.97E-11	24.20%	25.10%	14.10%	15.20%	14.30%

Százalékosan tekintve, azaz arra a kérdésre keresve a választ, hogy a válaszok információtartalmának hány százaléka keletkezik a települések szintjén, a jelen adatokból még nem vonható le egyértelmű következtetés. Ami viszont talán egyértelműen kijelenthető, az az, hogy a teljes sokaság, azaz a népszámlálási adatok tekintetében is meglévő jelenséggel van dolgunk. További meglepőnek nevezhető jelenség, hogy a súlyozás alkalmazása révén egyértelmű eltérések tapasztalhatók az egyes paraméterek értékeiben, és ilyen módon sajnálatosnak mondható, hogy nem minden esetben volt lehetséges súlyozott becslések számítása.

B) A permutációs tesztek eredménye

A permutációs tesztek során minden adatbázis esetén 1000-szer elvégeztem a településértékek véletlenszerű összekeverését. Az eredményeket egyrészről táblázatokban is összefoglaltam [3a).–3c) táblázatok], amelyeket a 2. táblázattal összevetve tudunk számszerű következtetéseket levonni. Ugyanakkor ennél sokkal hatékonyabb megoldásnak ígérkezett a különböző eredmények grafikus összevetése, amelyek során egymásra vetítve láthatóak az egyes permutációs eloszlások hisztogramjai és az eredeti adatokban mért értékek (2a)–2f ábrák).

3a) táblázat: Permutációs eredmények: WVS

WVS	N	Range	Minimum	Maximum	Mean	Mean Std. Error	Std. Deviation	Variance
ANOVA-F	1000	1.350	0.501	1.851	1.013	0.007	0.214	4.590E-02
ANOVA-Pr	1000	0.998	0.000	0.999	0.490	0.009	0.295	8.701E-02
Etas-quare	1000	0.103	0.045	0.147	0.086	0.001	0.017	2.745E-04
weighted-Etasquare	1000	0.101	0.052	0.154	0.095	0.001	0.017	3.018E-04
ICC1	1000	0.114	-0.045	0.069	0.001	0.001	0.018	3.396E-04
weighted-ICC1	1000	0.112	-0.037	0.076	0.010	0.001	0.019	3.732E-04
ICC-random-effects	1000	0.068	0.000	0.068	0.007	0.000	0.011	1.304E-04

3b) táblázat: Permutációs eredmények: EVS

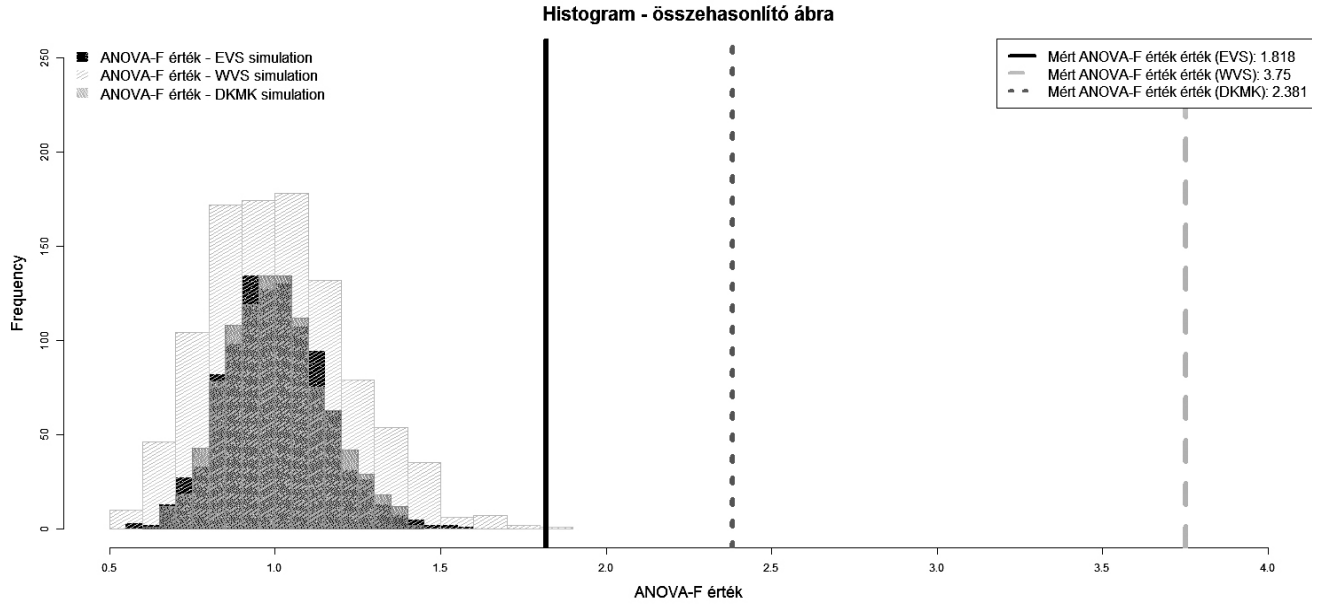
EVS	N	Range	Minimum	Maximum	Mean	Mean Std. Error	Std. Deviation	Variance
ANOVA-F	1000	1.027	0.556	1.582	0.999	0.005	0.151	2.292E-02
ANOVA-Pr	1000	0.999	0.000	1.000	0.504	0.009	0.289	8.336E-02
Etas-quare	1000	0.096	0.062	0.157	0.105	0.000	0.014	2.031E-04
weighted-Etasquare	1000	0.118	0.081	0.200	0.133	0.001	0.017	3.021E-04
ICC1	1000	0.108	-0.050	0.058	0.000	0.001	0.016	2.585E-04
weighted-ICC1	1000	0.133	-0.027	0.106	0.031	0.001	0.020	3.840E-04
ICC-random-effects	1000	0.059	0.000	0.059	0.006	0.000	0.009	8.578E-05

3c) táblázat: Permutációs eredmények: EVS

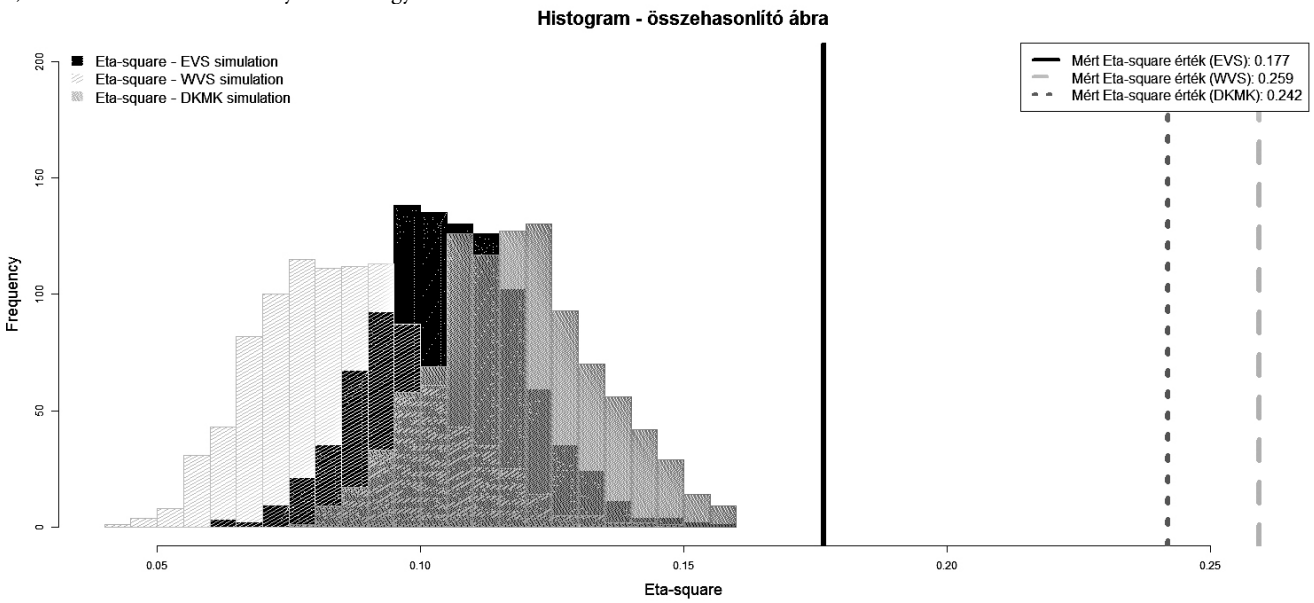
DKMK	N	Range	Minimum	Maximum	Mean	Mean Std. Error	Std. Deviation	Variance
ANOVA-F	1000	0.768	0.639	1.407	1.003	0.005	0.147	2.173E-02

ANOVA-Pr	1000	0.991	0.007	0.998	0.500	0.009	0.291	8.471E-02
Etas-quare	1000	0.080	0.079	0.159	0.118	0.000	0.015	2.342E-04
weighted-Etasquare	1000	0.106	0.095	0.201	0.142	0.001	0.018	3.245E-04
ICC1	1000	0.091	-0.045	0.046	0.000	0.001	0.018	3.063E-04
weighted-ICC1	1000	0.121	-0.026	0.095	0.027	0.001	0.021	4.240E-04
ICC-random-effects	1000	0.046	0.000	0.046	0.007	0.000	0.010	1.066E-04

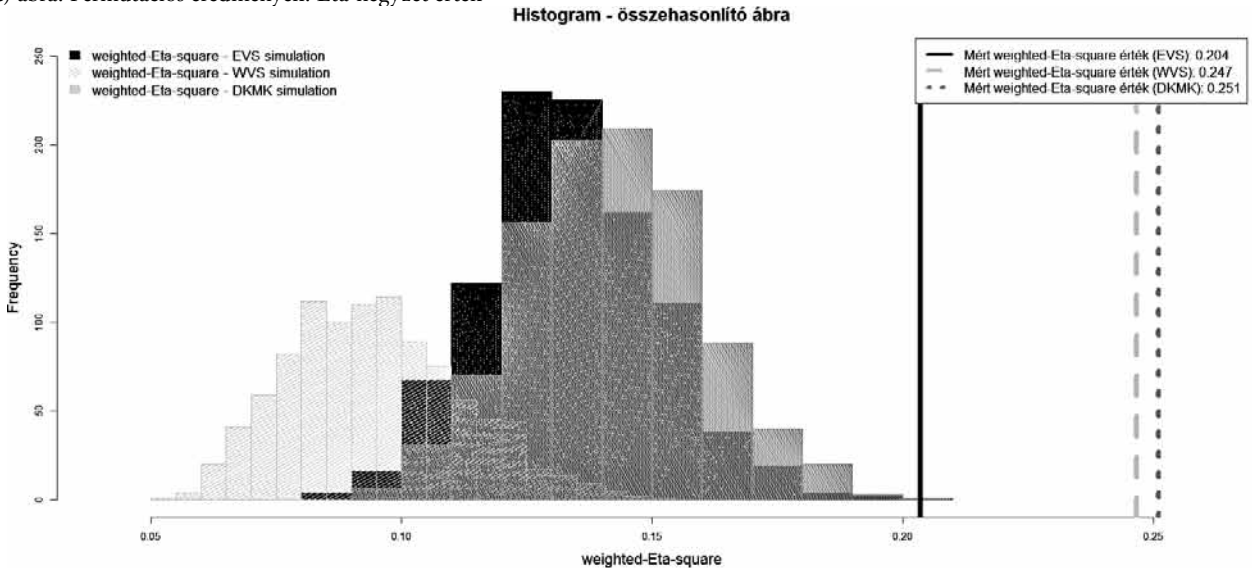
2a) ábra: Permutációs eredmények: ANOVA-F érték



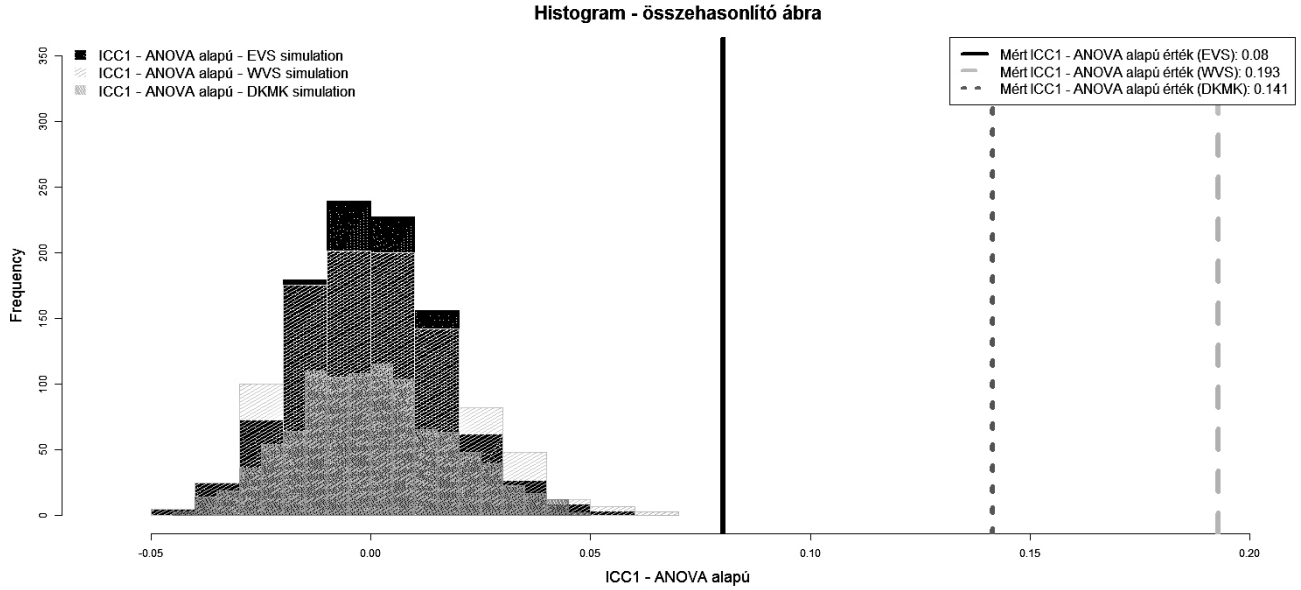
2b) ábra: Permutációs eredmények: Eta-négyzet-érték



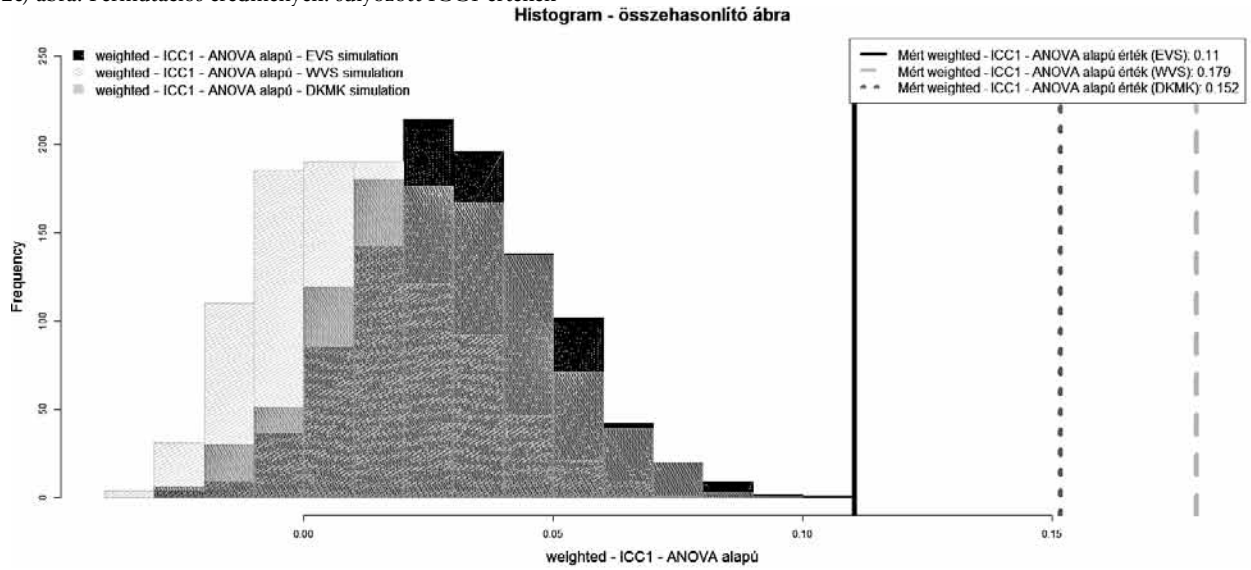
2c) ábra: Permutációs eredmények: Eta-négyzet érték



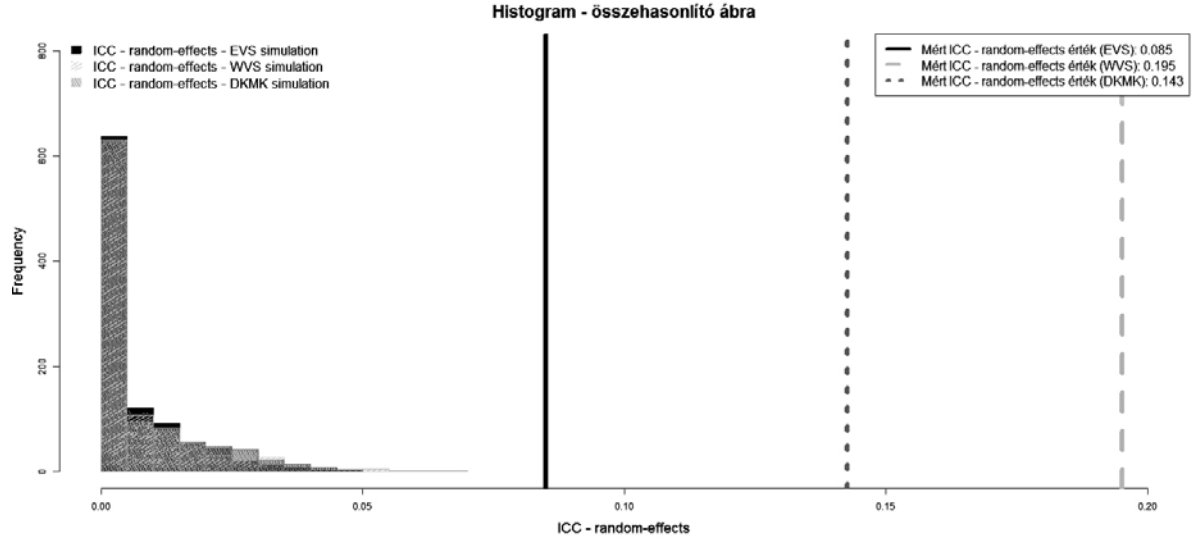
2d) ábra: Permutációs eredmények: ICC1 értékek



2e) ábra: Permutációs eredmények: súlyozott ICC1 értékek



2f) ábra: Permutációs eredmények: ICC értékek



A táblázatok és a grafikonok igazából túl sok magyarázatra nem szorulnak, hiszen rendkívül jól látszik, hogy a permutációs eredményekkel összevetve egyértelmű eltérés van a véletlen eloszlású adatok és a mért értékek között. (Ezeket az eltéréseket t-próbával is ellenőriztük, amelyeket itt terjedelmi okokból nem közlünk.) Az egyes mutatók permutációs szórása ugyanakkor nem azonos, és egyértelműen leginkább a random tényezős modellen alapuló ICC mutató diszkriminál. Az itt kapott eredményeket alapul véve, véleményem szerint, az ICC mutató értéke felel meg leginkább az eredeti kérdés, azaz az interdependencia mértékének becslésére.

C) A népszámlálási adatokon rekonstruált mintavétel – DKMK

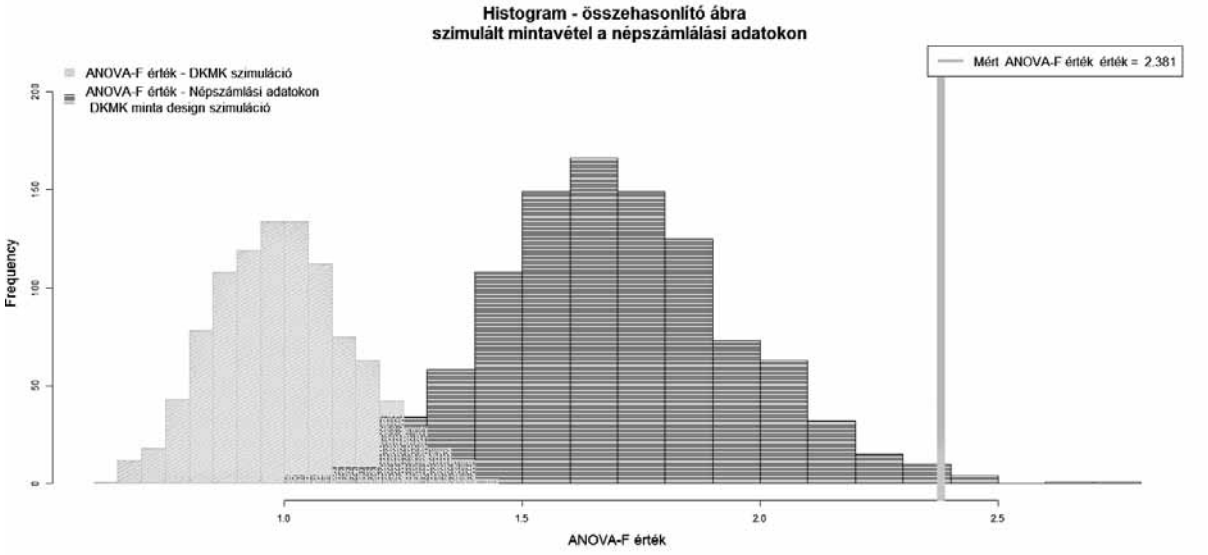
A jelen tanulmányban bemutatott utolsó vizsgálati eredmény annak a kérdéskörnek a megválaszolására irányult, hogy van-e a az alapadatokban meglévő klaszteresződésen túl bármilyen speciális effekt a mintaadatokon. Ehhez a DKMK adatfelvétel települési változóit használtam fel, ami olyan szempontból is szerencsés választásnak látszik, hogy köztes erősségű klaszterezettségi értékeket mutatott a másik két adatfelvételhez viszonyítva. Az eredményeket itt is kétféle módon adom meg: egyrészt táblázatos

formában (2., illetve 4. táblázat), illetve grafikusan (3a)–3d) ábrák). Hasonló módon itt is ellenőriztem t-próbával az eltéréseket, és itt is minden esetben szignifikáns eltéréseket kaptam, bár az előzőekkel összevetve kevésbé határozottan. Ilyen módon úgy tűnik számomra, hogy a többlépcsős mintavételi designnak további torzító hatásai is vannak, amelyek okára nézvést további vizsgálatok szükségesek.

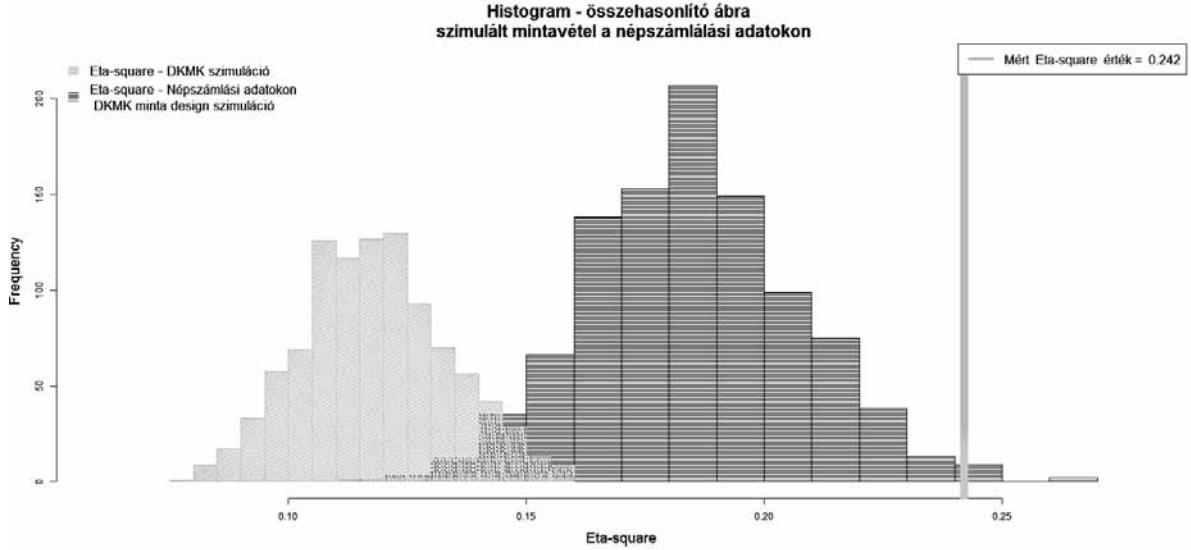
4. táblázat: Klaszterződési mutatók: rekonstruált szimulációs eredmények

DKMK minta design a népszámlálási adatokon	N	Range	Minimum	Maximum	Mean	Mean Std. Error	Std. Deviation	Variance
ANOVA-F	1000	1.694	1.012	2.706	1.700	0.008	0.251	6.275E-02
ANOVA-Pr	1000	0.452	0.000	0.452	0.006	0.001	0.026	6.997E-04
Eta-square	1000	0.147	0.120	0.266	0.185	0.001	0.022	4.891E-04
ICC1	1000	0.168	0.001	0.169	0.076	0.001	0.025	6.380E-04
ICC-random-effects	1000	0.168	0.003	0.171	0.073	0.001	0.026	6.727E-04

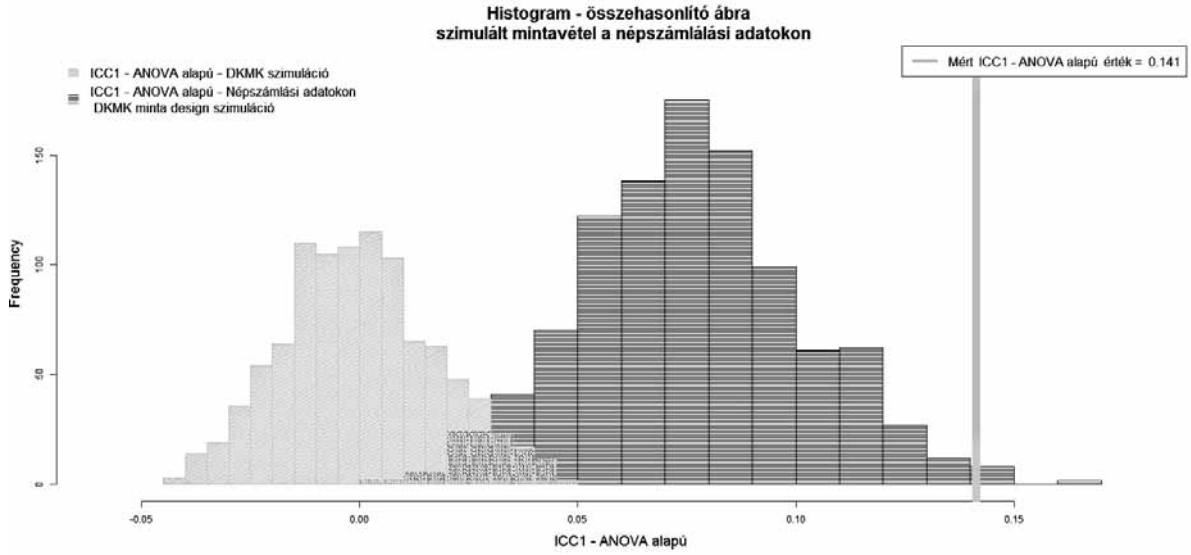
3a) ábra: Rekonstruált mintavétel: F értékek



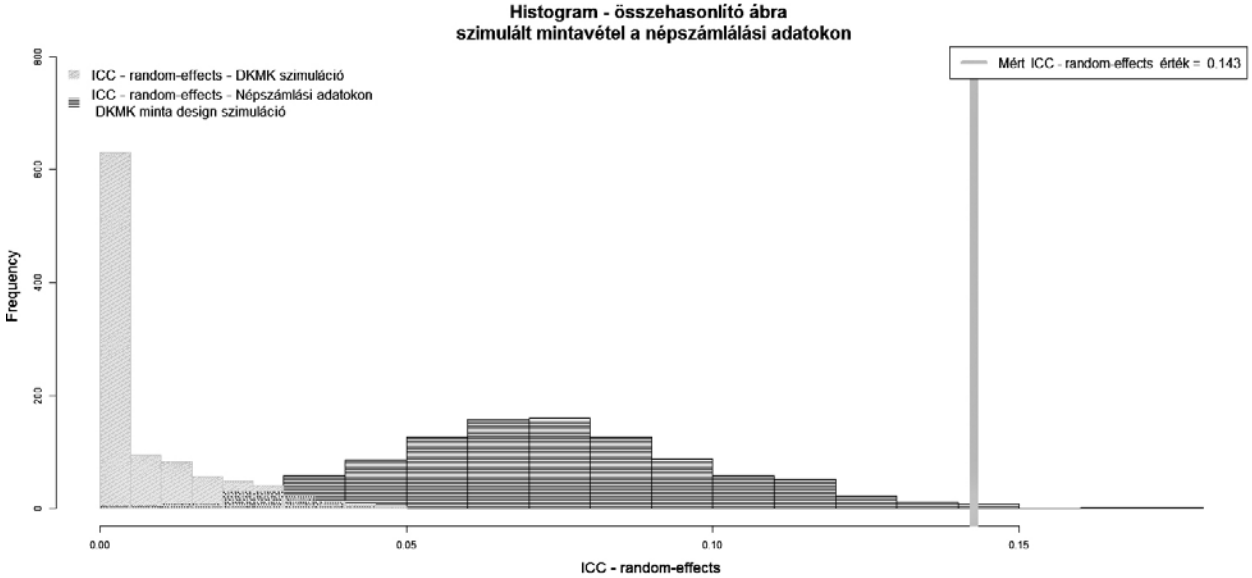
3b) ábra: Rekonstruált mintavétel: eta-négyzet-értékek



3c) ábra: Rekonstruált mintavétel: ICC1 értékek



3d) ábra: Rekonstruált mintavétel: ICC értékek



Összegzés

Az eredmények alapján egyértelműen beigazolódott az a hipotézisem, hogy a klaszterező dés jelensége jelen van mind a többlépcsős mintavételi, mint a népszámlálási adatokban – legalábbis a kiemelten fontosnak nevezhető iskolai végzettség mutató tekintetében. A randomizációs szimulációs tesztek révén láthatóvá vált, hogy habár a különböző mutatószámok és mintadesignok alapján eltérések láthatók az egyes mutatószámok között – többek között beigazolva Bliese-ék vizsgálati eredményeit az eta-négyzet-mutató gyengesége tekintetében –, az eltérések minden minta esetén olyan mértékűek, melyek nagyon nagy valószínűséggel nem származhatott a minták egyedi struktúrájából.

A felhasznált mutatószámok tekintetében az összes közül a random tényezős modellen alapuló ICC érték tűnik a leginkább stabilnak és robusztusnak, hiszen alig érzékeny a mintaszerkezetre, és a szimulációk során is leginkább konvergált az elvárt 0 értékhez

Az egyik minta, a DKMK népszámlálási adatok alapján történt rekonstrukciós szimulációja kimutatta, hogy a teljes sokaságban megfigyelthez képest a vizsgált minta jelentősebb klaszterződést mutat – legalábbis az iskolai végzettség településen belüli strukturálódását illetően. Ilyen módon tehát láthatóvá vált, hogy az adatok függetlenségének problematikája két szinten van jelen a többlépcsős adatokban:

– egyrészt, mint a forrásadatokban már önmagában meglévő adottság,

– másrészről, mint a speciálisan a többlépcsős mintában meglévő torzítás, amelynek oka jelen kutatás alapján nem azonosítható.

Az eredmények szociológiaelméleti értelemben is visszaigazolják a társadalom tagjainak függetlenségi szempontból elvárható hasonlóságát.

Jövőbeni célkitűzések

Mivel jelen vizsgálat egy a kérdéskör tekintetében mindenképpen hangsúlyos területre, a klasztereződés többváltozós modellekre gyakorolt hatására nem terjedt ki, ezért annak mindenképpen prioritást kell élveznie a jövőbeni kutatások során. Ugyanakkor a többváltozós kapcsolatokra gyakorolt hatások elemzése is több kérdéskörre bontható:

1. többváltozós többszintű modellek (mixed-effects models)
2. aggregált változókra gyakorolt hatások vizsgálata.

További módszertani szempontból ellenőrizendő kérdés, hogy a többlépcsős mintavételnek a klasztereződött társadalmi közegben történő alkalmazása vajon a mintavételi hiba tekintetében torzító hatással járhat-e, azaz a minták jóságát globálisan milyen mértékben veszélyeztetheti.

Mindezen említett területeken túl, vagy inkább azokkal párhuzamosan, ugyanakkor nem szabad egy további jelentős kérdésről, magáról a társadalmi térstrukturálódás kérdésköréről sem megfeledkezni, hiszen az adatok szisztematikus strukturálódása (lásd 1. ábra) egyértelműen nem csak települési szinten („mezo") lehet jelen, hanem nagytérségi („globális") trendek is megfigyelhetők (Kmetty–Tóth 2011). Ilyen módon a jövőbeni vizsgálatok során mindenképpen legalább két-, de akár három- vagy többszintű modellek vizsgálatát javasolnám, míg a térstruktúrák feltárása tekintetében várhatóan az aggregált adatok alapján történő térklaszteres eljárások alkalmazása tűnik számomra elsőrendűen célravezetőnek.

Irodalom

ANGELUSZ RÓBERT ÉS TARDOS RÓBERT (2009): A kapcsolathálózati szemlélet a társadalom- és politikatudományban. Politikatudományi Szemle, 18(2): 29–57.

BLIESE, PAUL D. ÉS HALVERSON, RONALD R. (1998): Group Size and Measures of Group-level Properties: An examination of eta-squared and ICC values. Journal of Management, 24: 157–172.

BLIESE, PAUL D. (2000): Within-group Agreement, Non-independence, and Reliability: Implications for Data Aggregation and Analysis. In: Multilevel Theory, Research, and Methods in Organizations. K. J. Klein– S. W. Kozlowski (szerk.). San Francisco, CA: Jossey-Bass, Inc. 349–381.

BLIESE, PAUL D. (2008): Multilevel: Multilevel Functions. R package version 2.3.

BRYK, ANTHONY S., RAUDENBUSH, STEPHEN W. (1992): Hierarchical Linear Models. Thousand Oaks, CA: Sage.

DUSEK TAMÁS (2004): A területi elemzések alapjai. Regionális Tudományi Tanulmányok 10. Budapest: ELTE TTK Regionális Földrajzi Tanszék.

GRANOVETTER, MARK (1973): The Strength of Weak Ties. American Journal of Sociology, 78(6): 1360–1380.

KMETTY ZOLTÁN ÉS TÓTH GERGELY (2011): A politikai részvétel három szintje. In: Részvétel, képviselet, politikai változás. Tardos Róbert, Enyedi Zsolt és Szabó Andrea (szerk.). Budapest: Demokrácia Kutatások Magyar Központja Alapítvány. 75–115.

KOZLOWSKI, STEVE W. J. ÉS HATTRUP, KEITH (1992): A Disagreement about within Group Agreement: Disentangling issues of consistency versus consensus. Journal of Applied Psychology, 77: 161–167.

MONDAY, DENNIS, KLEIN GARY ÉS LEE, SUNNI (2005): The Assumptions of ANOVA http://www-rohan.sdsu.edu/~cdlin/677/ANOVA_Assumptions.ppt.

PINHEIRO, JOSE C. ÉS BATES, DOUGLAS M. (2000): Mixed-effects Models in S and S-PLUS. New York: Springer-Verlag.