

ELMÉLET–MÓDSZERTAN

DR. MARTON ÁDÁM

Országos reprezentatív felvételek – (kis)területi becslések

A KSH által végrehajtott reprezentatív statisztikai adatfelvételek célja, hogy a társadalmi és gazdasági folyamatokról megfelelő szakmai részletezettségű, kellő megbízhatóságú (jó „minőségű”) becsléseket adjunk havonta, negyedévenként, évente. Az elvárt pontosság biztosításának legfontosabb feltétele a kielégítő mintanagyság és a jó mintavételi terv. A statisztikusoktól azonban sokszor azt is kéri, hogy az összegyűjtött adatokból az eredetileg tervezettnél részletesebb információkat is adjanak, például többdimenziós keresztáblákat termék és/vagy társadalmi csoportok szerint, vagy területi részletezést stb. (A gazdaságstatisztikában például az iparágak, a termékcsoportok, a vállalkozások nagysága, a társadalomstatisztikában a nem, korcsoport, iskolai végzettség, a családi állapot bontásban. Mindezeket az adatokat sokszor kéri megyék, régiók, más területi egységek szerint is.) Ilyen becslések ugyan formálisan előállíthatók (a cellákban adódik egy-két megfigyelt adat), azonban, néhány kivételtől eltekintve, nagy mintavételi hibájuk miatt nem használhatók. A minták nagysága tehát gátat szab a részletezéseknek, még akkor is, ha azok több tízezer elemből állnak.

Mivel a hivatalos statisztikusok egyre gyakrabban kerültek szembe azzal az igénnyel, hogy egy mintából részletesebb becsléseket készítsenek annál, mint amivel a mintavételi terv számolt, már a múlt század közepén kutatni kezdték az említett becslések megbízhatóságát növelő „olcsó” lehetőségeket.

Miről is van szó? Abból indulunk ki, hogy a rétegzett valószínűségi minta egyenletesen lefedi az egész országot, és minden területrész mintája az országoséval azonos struktúrájú. Minden kisterületen belül vannak megfigyelések és a súlyozáshoz (kalibráláshoz) is vannak adatok, tehát a becslés, bár nagy mintavételi hibával, minden kisterületre elvégezhető.

A kisterületi becslések pontosságát a minta nagyságának növelésével egyszerűen lehetne növelni. De ez az út általában anyagi vagy más korlátok miatt nem járható. Azt fogjuk vizsgálni, hogy könnyen elérhető külső információk és/vagy modellek segítségével miként növelhető a szóban forgó becslések megbízhatósága. Felhasználunk olyan információkat (cenzusok, adminisztratív regiszterek stb.), amelyek valamilyen kapcsolatba hozhatók a vizsgált változókkal.

Fontos, hogy az így előállított becslések „hibáját” ismerjük, vagy legalábbis biztosan állíthassuk, hogy azok megbízhatósága javult. Mint ismeretes, a hiba (error) kétféle lehet: véletlen és/vagy torzítás. A torzítás „szisztematikus”, a minta nagyságával nem csökkenthető. A hiba kedvező változása általában megítélhető, de számszerűsítése, ha egyáltalán lehetséges, nagyon nehéz.

Purcell és Kish korai, alapvető fogalmakat tisztázó tanulmányukban [1980] a feladatot általánosan fogalmazták meg: valamely nagy sokaságot részekre bontunk, ami lehet terület szerint, de szakmai összetevők szerint is. E kétféle megosztás elvileg különbözik: a kisterület az egésznek a része, azonos változókkal; míg a csoportok szerinti megosztás csak a jellemzők egy részére (nem, iskolai végzettség, korcsoport, iparág) koncentrál. A jelen összeállítás csak a területi dimenziót vizsgálja.

A kisterületek nagyságától nem független, milyen módszerrel próbáljuk a megoldást keresni. Az említett tanulmányban a részterületeket nagyságuk szerint négy kategóriába sorolták:

- Viszonylag nagyoknak nevezhetők azok a részterületek, amelyek az egész egytizedénél nagyobbak;
- kicsik, amelyek legalább egy századnyiak;
- minik, amelyek legalább egy ezrednyiek;
- míg „ritkák” az egy tízezred résznél kisebbek.

Az első esetben ki lehet indulni az egyszerű direkt becslésekből, a negyedik esetben egészen más módszereket kell keresni, míg a 2. és 3. esetben meg lehet kísérelni a bonyolultabb kisterületi becslési módszerek alkalmazását.

A kisterületi becslések módszereinek nagy külföldi szakirodalma van, és a nemzetközi konferenciák programjában is rendszeresen szerepel. (Lásd például az ISI¹ és az IASS² rendezvényeit, tanfolyamait. A Google-ban a „small area estimation” keresésre több mint 1 millió találat adódott!)

Az 1980-as évek első felében a KSH-ban is foglalkoztunk e becslési módszerekkel (Marton 1983, 1986, Elmeré 1987, Mihályffy 1993). A próbálkozások azonban nem vezettek széles körű alkalmazáshoz. Az elmúlt negyedszázad során lényegesen megváltozott körülmények (módszertani fejlesztés, növekvő igények) indokoltá teszik, hogy ismét felhívjuk a figyelmet e módszerekre.

A továbbiakban, az egyszerűektől a bonyolultabbak felé haladva ismertetjük röviden a már elterjedt és használt módszereket. A szóbeli ismertetésre tesszük a hangsúlyt, a formális matematika eszköztárat a lehető legkisebb mértékben használva.

A jelen összeállításban *J. N. K. Raonak*, e téma nemzetközileg elismert szakértőjének 2000-ben a baszk statisztikai hivatalban (Vitoria Gasteiz) megtartott előadása alapjául szolgáló jegyzetét és a 2003-ban megjelent *Small Area Estimation* c. könyvének gondolatmenetét követjük nagyon vázlatosan.

A könyv előszavában Graham Kalton felvázolja e probléma hátterét, jelentőségét.

A múlt században a modern reprezentatív felvételek adta lehetőségeket messze meghaladva nőtt az információigény, írja. Meggyorsult a módszertani fejlesztés is. Mindennek ellenére hamarosan kiderült, hogy a reprezentatív felvételek számának, a minták nagyságának növelésével sem lehet az egyes témák részleteit illető igényeket kielégíteni.

A fejlesztés, mint arról már volt szó, két irányban történt. Ahol lehetett, növelték a minta nagyságát, miáltal egyre több csoportonkénti (domain) vagy területi (area) mutató-

¹ ISI= International Statistical Institute – Nemzetközi Statisztikai Intézet.

² IASS= International Association of Survey Statisticians – Reprezentatív Felvételekkel Foglalkozó Statisztikusok Nemzetközi Társasága.

szám lett megfelelő pontossággal becsülhető. A részletezettség azonban mindenképpen korlátozott maradt. Új módszereket kellett tehát keresni. Az eredményeket kezdetben fenntartással fogadták, de olyan nagy volt az információigény, valamint a tapasztalatok is kedvezőek voltak, hogy ezek a módszerek egyre inkább elfogadottá váltak.

Kezdetben nagyon egyszerűek voltak ezek a módszerek. A számítástechnika fejlődése azonban nagyban hozzájárult az egyre bonyolultabb megoldások alkalmazásához, amelyek széles tárháza áll most már rendelkezésünkre. (Kalton szerint e technikák különösen fontosak lehetnek a közép- és kelet-európai országokban, ahol a központi irányítást a piaccgazdaság váltotta fel.)

A változók lehetnek összegek, megoszlások, quintilisek, illetve folyamatosak vagy kategóriák stb. A módszerek megválasztása függ az adott problémától, a vizsgálni kívánt paraméter jellegétől. A továbbiakban, az egyszerűség kedvéért, leginkább átlagok becslését vizsgáljuk.

Az egyszerű direkt becslés

Direkt becslésről akkor beszélünk, ha a mintának csak a szóban forgó területre vonatkozó adatait használjuk fel. Az lehet a minta elemeinek egyszerű vagy súlyozott, utólagosan rétegzett vagy rétegzés nélküli átlaga. A sokszor bonyolult mintavételi tervet adottnak vesszük, s annak megfelelően történik minden kisterületre vonatkozó számítás, például kalibrálás. A kisterületi becslések iránti igény sokszor utólag adódik, annak szempontjait a mintavételi terv elkészítésénél nem vették figyelembe. Ismétlődő felvételeknél az új minta célszerű megtervezése sokat segíthet a kisterületi becsléseknél.

Az egyszerű becslések esetében mindig fontos szerepet játszik a minta: annak struktúrája, nagysága. A jól ismert problémákon túl (rétegzés, allokáció stb.) javíthatja a kisterületi becsléseket, ha vannak hasonló témájú felvételekben azonos változók, „core” adatok, amelyek segítségével az állományok összevonhatók és együttesen elemezhetők. Ilyenek lehetnek a panelfelvételek, a guruló és ismétléses minták stb. (Brackstone 2002, Kish 1990, 1998, Schaible 1992, Sing 1994).

Az egyszerű becslés az adott területre vonatkozóan a jól ismert formula alapján:

$$\hat{Y}_k = \sum y_{ik} w_{ik} ,$$

ahol $\hat{Y}_k$ = valamilyen becslés,
 y_{ik} = a megfigyelt adat,
 w_i = a súlyarány.

Valójában a teljes minta részét képező kisterületek becsléseit úgy kezeljük, mintha azok önálló, egymástól független minták lennének. Ezek torzítatlan becslések, csak a minta terjedelmétől függően lehet, hogy nem kellően megbízhatók. A súlyok a mintavételi tervből adódnak (Horvitz–Thomson-becslések, rétegzett, többlépcsős stb. minták).

Lehet a kisterületi becsléseket javítani a kevésbé ismert, úgynevezett duális minták segítségével. Ezek lényege az, hogy az „alap”, a meglévő mintát kiegészítjük egy pótmintával, amelyen a kívánt adatokat viszonylag egyszerű módszerrel szerezzük meg.

(Például az interjúkon alapuló mintát telefonos megkérdezésekkel egészítjük ki, amikor csak néhány lényegesnek tartott kérdést teszünk fel.)

Demográfiai adatok tradicionális továbbvezetése

Népszámlálásokra, más teljes körű összeírásokra, amelyekből fontos és részletes adatok nyerhetők a sokaság szerkezetére, jellemzőire, nagyobb időközönként, általában öt-tízévenként kerül sor. Az évek során jelentős változások mennek végbe, amelyek szoros nyomon követése nem valósítható meg.

A köztes években számos, főként adminisztratív nyilvántartásokból származó információ áll rendelkezésre, melyek felhasználhatók a census adatok továbbvezetéséhez. Így viszonylag egyszerűen és jól elkészíthetők az úgynevezett census utáni, azt továbbvezető becslések (postcensal estimates).

Legyen egy census teljes népessége $N!$

$$N = \sum N_{ijk} ,$$

ahol $i, j, k, \dots$ a célsokaság valamely teljes körű szétosztását jelenti népességcsoportok és/vagy gazdasági tevékenységek és/vagy területek stb. szerint.

A censust követő években lezajló folyamatok nem arányosak a különböző csoportokban. Feladat az N_{ijk} mennyiségek továbbvezetése.

Lehetséges egyszerű megoldások: minden N_{ijk} csoportból mintát veszünk, s az abból becsült változással vezetjük tovább a census adatait. (Az 1960-as évek közepén a közönséges levélpostai küldemények számának becslése ilyen módszerrel történt: Marton 1966). Felhasználhatók külső forrásból származó adatok is: hányados- vagy regressziós becslések, szimptomatikus változók stb.

A Pfeffermann (2006) összeállításában található példa jól és könnyen érthetően jellemzi e módszert.

Az USA-ban néhány évtizeddel ezelőtt az 5–17 éves korú szegény gyermekek oktatásának támogatására fordított pénzek szétosztása megyénként (county), a népszámlálások adatai alapján történt. Mivel a két népszámlálás között eltelt 10 év alatt a szegénységben élő gyerekek száma nem arányosan változott, jobb, igazságosabb megoldást kellett találni. A népszámlálások adatainak továbbvezetésére kiváló megoldásnak kínálkozott az adminisztratív nyilvántartások felhasználásával a következő modell a k -adik kisterület (megye) becsléséhez:

$$Y_k = \alpha + \beta_1 x_{1k} + \beta_2 x_{2k} + \beta_3 x_{3k} + \beta_4 x_{4k} + \beta_5 x_{5k} + u_k + e_k ,$$

ahol x_{1k} = a családok adóbevallásaiban az érintett gyermekek száma;

x_{2k} = az élelmiszersegélyben (food stamps) részesülők száma;

x_{3k} = a 18 év alatti népesség becsült száma;

x_{4k} = a gyermekek száma az adóbevallásokból;

x_{5k} = az iskoláskorú szegény gyermekek száma a megelőző népszámláláskor;

u_k = modellhiba;

e_k = véletlen hiba.

A β paraméterek, valamint az u_k „modellhiba” a népszámlálás évében meglévő adatokból becsülhetők, az x -ek pedig folyamatosan rendelkezésre állnak.

Gyakran adódik olyan helyzet, hogy a teljes sokaság, vagy egy-két ismérv szerint megbontott részeinek továbbvezetett adatai („a marginálisok”) ismertek vagy jól becsülhetők (például N_{ij} , $N_{i.k}$, $N_{.jk}$), míg a további részleteké nem. Ekkor a $\cdot$ dimenzió-nak megfelelő megoszlást kell becsülni. Ebbe a körbe tartoznak az úgynevezett struktúrát megőrző, az angol rövidítés szerint SPREE, becslések. (További példákat illetően lásd Rao 1994!)

Indirekt kisterületi becslések

Szintetikus (egyszerű) becslések

Tipikus egyszerű kisterületi becslés. Akkor használható, ha az egyes kisterületek strukturális összetétele jelentősen különbözik, ugyanakkor egy-egy vizsgált változó értéke kisterületenként az országos átlag körül kisebb mértékben ingadozik, mint a másik változókhoz viszonyított különbség. (Egyik első ilyen alkalmazás volt, amikor az USA-ban a színesbőrű lakosság arányának különbözőségét használták ki az egyes államok átlagos egészségügyi helyzetének leírásához.)

A módszer lényegét a következő egyszerű példával illusztrálhatjuk. Legyen a mezőgazdasági dolgozók átlagbére országosan $\bar{X}_1$, az iparban dolgozóké $\bar{X}_2$, amelyek egymástól számottevően különböznek!

Ismerjük kisterületenként a mezőgazdaságban és az iparban dolgozók számát, illetve w_1 és w_2 arányát. Akkor a k -adik kisterület átlagos bérének becslése lehet:

$$\hat{\bar{Y}}_k = \sum w_1 \bar{x}_1 + w_2 \bar{x}_2,$$

amennyiben feltételezhető, hogy kisterületenként a mezőgazdasági, illetve ipari bérek szórása nem nagy. (A változókon belüli szórás kisebb, mint a változók közötti.)

Általánosan: az egyszerű szintetikus kisterületi becslés az egyes rétegek országos becsléseinek ($\hat{\bar{Y}}_k$) a kisterület tényleges szerkezetének megfelelően súlyozott átlaga:

$$\hat{\bar{Y}}_k = \sum W_{ik} \bar{Y}_{ik}.$$

Megjegyzés: ez a becslés megtartja, felhasználja a kisterület belső struktúráját, ami akár egy utólagos rétegzésnek is felfogható. (E nagyon egyszerű gondolat több irányban általánosítható. Akár felfogható úgy is, mint az imputálás speciális esete.)

Ez a becslés akkor is előállítható, ha a vizsgált kisterületen nincs megfigyelés, csak a demográfiai struktúrája ismert. Mivel abból indultunk ki, hogy minden kisterületen van megfigyelt adat, akkor jó megoldásnak látszik, ha az egyszerű direkt becslést a fenti szintetikus becsléssel kombináljuk.

Kombinált (composit) becslések

Nézzük azt az esetet, amikor vannak az országos becslés mellett előállított kisterületi egyszerű direkt becslések, valamint más adatok, információk segítségével is készítünk becsléseket! (Ez lehet az országos minta, az előzőekben említett szintetikus becslés, adminisztratív adatforrások, múltbeli adatok stb.) Több becslést is fel lehet használni. Kézenfekvő gondolat, hogy e becslések átlagát használjuk.

$$\hat{\bar{Y}}_k(C) = c\hat{\bar{Y}}_k + (1 - c)\hat{\bar{Y}},$$

ahol: c valamilyen súly.

A kulcskérdés: milyen c súlyokat adjunk az egyes becsléseknek? Nyilvánvalónak látszik, hogy annak a nagyobb, amelyik a viszonylag jobb. Ha van erre valamilyen objektív mérce, akkor egyszerű a dolog. A legtöbb esetben azonban nincs. Ekkor csak valamilyen szubjektív megítélés segít, valamint olyan szimuláció elvégzése, amikor a keresett becslést különböző feltételek mellett is kiszámítjuk.

A James–Stein becslés (JS) olyan kompozit becslőfüggvények használata, amelyeknél a komponensek súlyozását illetően speciális optimalizálási feltételek teljesülnek. Az alternatív becslés származhat magából a mintából vagy külső forrásból (L. Fay, Herriot 1979). Fontos tulajdonsága az, hogy a JS-becslések az *összes* kisterületre vonatkozóan jobbak, mint más formulák esetében, más szóval az MSE a legkisebb. Az egyes kisterületekre vonatkozóan azonban nem tudunk semmit mondani.

Kiegészítő információk birtokában használhatjuk az úgynevezett általánosított regressziós becslést. (Generalized Regression Estimator – GREG.)

$$\hat{\bar{Y}}_{GR} = \hat{\bar{Y}} + (X - \hat{X})^T \hat{B},$$

ahol: $X = (x_1, \dots, x_p)$ a kiegészítő információk regressziós változói,

B = a legkisebb négyzetek módszerével számított együtthatók.

Az utólagos rétegezés, kalibrálás felfogható a GREG-becslés speciális esetének.

Fontos megemlíteni, hogy Rao a munkáiban bemutatott konkrét példák esetében mindig értékeli a választott módszer hatékonyságát, hogy valójában azzal mennyire lehetett a minőséget javítani.

Modelleken alapuló (model-based) becslőfüggvények

A kisterületi becslések javításának egyre elterjedtebb módszere a vizsgált változó viselkedésének modellezése: annak felderítése, hogy milyen összefüggés van az ismert segédváltozók viselkedése és a vizsgált paraméter között. Erről Rao (2000) a következőket írja: (1) hatékony közvetett becslések készíthetők a feltárt modellek segítségével, (2) A modellek a minta adatai alapján validálhatók, (3) e módszerekkel komplex eseteket, mint például keresztmetszeti és idősoros adatsorokat lehet kezelni, (4) a becslések területspecifikus szóródása állítható elő, ami a szintetikus és kombinált becslések esetében nem lehetséges. Három modellel foglalkozik: a) „empirical best linear unbiased

prediction (EBLUP), b) empirical Bayes (EB), c) hierarchical Bayes (HB), amelyek felhasználják a területi szintű változókat (covariates).

E modellek közös tulajdonsága, hogy felhasználják az egyszerű direkt kistérségi becsléseket ($\hat{Y}_k$) és az ugyanarra a kistérsületre vonatkozó $\hat{Y}_k' = z_i\beta + v_i + e_i$ modell alapján előállított becslést. Ez tehát lényegét tekintve szintén kombinált becslés.

E témakörbe tartoznak az idősoros becslőfüggvények.

Röviden a következőkről van szó. Rendszeresen ismétlődő felvételek esetében adott időszakra vonatkozó kistérségi becslés javításához felhasználhatók a korábbi évek adatai is, feltéve, hogy érvényesül valamilyen hosszú távú trend, aminek ismerete javíthatja a konkrét becslést. Amennyiben például rendszeresen minden évben készítenénk kistérségi becsléseket, akkor a korábbi adatok idősorából lehetne következtetni (javítani) egy adott becslést.

E folyóirat oldalain néhány éve megjelent a kistérségi munkanélküliségi ráta becslési lehetőségeit vizsgáló tanulmány. Banai et al. (2000) a havonkénti és negyedévenkénti országos, illetve kistérségi becsléseket vizsgálták idősorok és a munkaügyi hivatalok regisztrációs adatainak felhasználásával. Mint ismeretes, az ILO munkanélküli-fogalma és a regisztrált munkanélküliek fogalma különböző, de a hozzájuk tartozó számok változása között van bizonyos kapcsolat, és kellő modell kialakításával várható, hogy a nagyon részletes regisztrációs adatok ismeretében a munkaerő-felmérés ILO szerinti mintán alapuló becsléseit javítani lehet.

Az idősorok felhasználásával kapcsolatban meg kell azonban jegyezni, hogy értelem-szerűen azok csak akkor tudják a becsléseket javítani, ha a feltárt tartós trend következetesen érvényesül. Vannak azonban olyan változások, amelyek – főként a kistérségekben – lökésszerűen következnek be. Így nagyon vitatható, hogy azok trendjével korrigálható-e a becsült adat. (Sem a statisztikusok, sem a felhasználók nem szeretik az egymás után következő adatok hektikus ingadozását. Szinte bármilyen becslési módszert alkalmazunk, az az adatok „kisimításához” vezet. Jelent ez valóban pontosítást is? Vagy jelent-e az minőségjavulást, ha azzal nyugtatjuk magunkat, hogy csináltunk valamit?) Félő, hogy a munkanélküliség kistérségi alakulása ilyen természetű folyamat. Nyilvánvalóan lényeges az is, hogy a kistérség mekkora, milyen időszakra (hónapra, negyedévre?) vonatkoznak a becslések.

Következtetések

A jelen összeállításnak figyelemfelkeltés a célja. E módszerek kínálta lehetőségeket nem szabad kihasználatlanul hagyni. Az anyagi erőforrások szűkössége, de egyéb tényezők is, mint például az adatszolgáltatói terhek csökkentése vagy az adatok minőségének javítása, megköveteli a reprezentatív felvételek és az adminisztratív nyilvántartások maximális kihasználását.

Gyakorlatilag ez úgy működhet, hogy a szakstatisztikus felismeri, megfogalmazza a problémát, aminek megoldásában a módszertanban járatos statisztikus közreműködik.

Látszólag bonyolult módszerekről van szó, ami azonban egyáltalán nincs így. A jelen összeállításnak nem volt célja a matematikai formulák, összefüggések tárgyalása. Az irodalomjegyzék érzékelteti, hogy igen bőséges irodalom áll rendelkezésre. De hangsúlyozni kell, hogy nincs általánosan alkalmazható legjobb kistérségi becslés, még akkor sem, ha a különböző megoldások alap gondolata azonos.

A korai magyar kísérletek nem keltették fel a szakstatistikusok érdeklődését. Ebben a jelen összeállítás szerzőjének tapasztalata szerint szerepet játszott az, hogy „idegen” adatokat használnánk fel, modellekre támaszkodnánk, amiknek természetesen megvannak a maguk hibaforrásai. Csökkentjük ugyan a mintavételi hibát, de növeljük a torzítást stb. A lényeg azonban az, hogy növeljük az adatok megbízhatóságát, felhasználhatóságát, azaz javítjuk a minőségét. Ez persze azt is jelenti, hogy a felhasználókat is meg kell győzni (bármilyen új statisztikai módszerről legyen is szó) arról, hogy jól használható információkat kapnak, melyek segítségével helyes következtetésekhez juthatnak. (Gondoljunk a múlt század első évtizedeinek nagy vitáira a reprezentatív mintavételek használhatóságáról. Ma már ez nem képezi vita tárgyát. A kistérségi becslési technikák továbbfejlesztést jelentenek. Ilyen továbbfejlesztés például a nemválaszolások kezelése, az imputálás, a statistical matching is.)

IRODALOM

- Banai M., Kovács A., Lázár Gy., Prisznyák M., Varga I. (2000): A kistérségi munkaügyi rendszer és alkalmazása. *Területi Statisztika*, 108–125. old.
- Brackstone, G. J. (2002): Strategies and Approaches for Small Area Statistics. *Survey Methodology*, 28. No.2. 117–123. old.
- Drew, D., Sing, M. P., Choudry, G. H. (1982): Evaluation of Small Area Techniques for the Canadian Labour Force Survey. *Survey Methodology*, No. 8. 17–47. old.
- Elmerné, Dr. Túri Magdolna. (1987): A kistérségi becslések alkalmazásának tapasztalatai Baranya megyében. *Területi Statisztika*, No. 1–2–3, 35–45. old.
- Falorsi, P. D., Falorsi, S., Russo, A. (1994): Empirical comparison of small estimation methods for the Italian Labour Force Survey. *Survey Methodology*, 20. 171–176. old.
- Fay, R. E., Herriot, R. A. (1979): Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74. 269–277. old.
- Ghangurde, D. D., Sing, M. P. (1977): Synthetic Estimators in Periodic Household Surveys. *Survey Methodology* 3., 152–181. old.
- Gosh, M., Rao, J. N. K. (1994): Small Area Estimation: An Appraisal. *Statistical Science* 9. 55–93. old.
- Kish, L. (1990): Rolling Samples and Censuses. *Survey Methodology*, 16. No.1, 63–71. old.
- Kish, L. (1995): Methods for Design Effects. *Journal of Official Statistics*, 11. 55–77. old.
- Kish, L. (1998): Space/Time Variations and Rolling Samples. *Journal of Official Statistics*, 14. 31–46. old.
- Marker, D. A. (2001): Producing Small Area Estimates From National Surveys: Methods for Minimizing use of Indirect Estimators. *Survey Methodology*, 27. No. 2. 183–188. old.
- Marton Á. (1966): Postaügyi reprezentatív felvétel. *Statisztikai Szemle*, 44. No. 7. 728–737. old.
- Marton Á. (1983): Kistérségek, alcsoporthoz tartozók paramétereinek becslési módszerei. *Statisztikai Szemle*, 61. No. 3. 261–279. old.
- Marton, A. (1986) Synthetic Estimates for small areas: Problems and results of a simulation experiment. *Statistical Journal of the United Nations ECE* 4. 71–80. old.

- Mihályffy, L.* et al. (1985): Területi és egyéb szempontok szerint részletezett statisztikai mutatószámok becslése (kistérségi becslések: módszertani tapasztalatok). KSH, Budapest
- Mihályffy, L.* (1993): Small Area Statistics in Hungary: Results and Problems. In: Kalton, G. et al.: Small Area Statistics and Survey Design Vol. II. Contributed Papers and Panel Discussion. Central Statistical Office, Warsaw
- Noble, A.* et al. (2002): Small Area Estimation via Generalized Linear Models. *Journal of Official Statistics*, 18. No. 1., 45–60. old.
- Piasecki, T.*: The Use of Small Area Methodology in Micro-enterprise Revenue Estimation. Conference on „Quality in Survey Statistics” Cardiff, UK. 2006 ápr. 24–26. keretében tartott előadás
- Pfeffermann, D.* (2006): Design Based and Model-dependent Small Area Estimation. Conference on „Quality in Survey Statistics” Cardiff, UK. 2006 ápr. 24–26. keretében tartott tanfolyam anyaga, 94. old.
- Purcell, N. J., Kish, L.* (1980): Postcensal Estimates for Local Areas (or Domains). *International Statistical Review*, 48, 3–18. old.
- Rao, J. N. K., Yu, M.* (1994): Small area estimation by combining time series and cross-sectional data. *Canadian Journal of Statistics*, 22., 511–528. old.
- Rao, J. N. K.* (1999): Some Recent Advances in Model-Based Small Area Estimation. *Survey Methodology*, 25. No. 2., 175–186. old.
- Rao, J. N. K.* (2000): Statistical Methodology for Indirect Estimations in Small Areas. Eustat, Vitoria-Gasteiz. 63 old.
- Rao, J. N. K.*, (2003): Small Area Estimation. Wiley-Interscience. 313 old.
- Schaible, W. L.* (1992): Use of small area statistics in US Federal Programs. In: Small Area Statistics and Survey Designs (Ed: Kalton, G. Kordos, J., Platek, J.) CSO Warsaw. 95–114. old.
- Sing, M. P., Gambino, J., Mantel, H. J.* (1994): Issues and Strategies for Small Area Data. *Survey Methodology*, 20. 3–22. old.
- Szép, K., Vigh, J.* (2004): A minőség a hivatalos statisztikában. *Statisztikai Szemle*, 82. No. 3. 773–798. old.
- US Department of Health, Education and Welfare (1978): State Estimators of Disability and Utilization of Medical Services, Rockville. 108 old.
- US Department of Health, Education and Welfare (1979): Small Area Estimation: an Empirical Comparison of Conventional and Synthetic Estimators for States. Hyattsville, 19 old.
- Woodruff, R. S.* (1966): Use of a Regression Techniques to Produce Area Breakdowns of the Monthly National Estimates of Retail Trade. *Journal of the American Statistical Association*, 496–504. old.

MELLÉKLET

Példák a szimulációs kísérletről

A már említett, a kezdeti próbálkozások kedvező eredményeinek kísérleti igazolására számításokat végeztünk, melyek részletes leírása megtalálható a Mihályffy (1985)-dolgozatban. Az alábbiakban nagyon vázlatosan ismertetünk egy-két részeredményt.

A kísérlet során olyan mutatókat becsültünk, amelyek „pontos” értéke ismert volt. Öt különböző becslési módszert próbáltunk ki és hasonlítottunk össze:

- közvetlen egyszerű becslés,
- szintetikus módszer,
- módosított szintetikus módszer,
- utólagos rétegzés és
- kombinált becslési eljárás.

A vizsgált sokaság az 1980-a népszámlálás lakossága volt, amelyből az akkori ELAR³-mintához hasonló szerkezetű és nagyságú mintát választottunk ki.

Összesen tíz változó megyei becsléseit vizsgáltuk: öt megoszlás- és öt átlagtípusút. (A kétfajta mutató kezelése valamelyest eltér.) A változónként kissé eltérő utólagos rétegzés nem, életkor, családi állapot, iskolai végzettség, foglalkozás szerint történt.

A négy kisterületi becslés hatékonysága nem azonos, de legtöbb esetben csökken a becslések és a tényadatok közötti eltérés. Ebből az következik, hogy szinte minden esetben található olyan eljárás, ami javítja az önmagában nem kielégítő pontosságú egyszerű becslések jószágát.

Az 1. mutató a családok megoszlása az aktív keresők száma szerint, míg a hatodik az egy lakásra jutó alapterület m²-ben.

Az eredményeket a relatív eltérések (ARD – átlagos relatív differencia) alapján értékeltük:

$$ARD = \frac{|\hat{X} - X|}{|X|}.$$

Az ARD kiszámítása a megoszlás- és átlagtípusú mutatók esetében valamelyest különbözik, aminek részletezésével nem foglalkozunk.

Az öt megoszlástípusú nem részletezett mutatóra végzett becslések összesített (átlagolt) relatív eltérései a következők voltak:

- ELAR-minta: 6,98%,
- Szintetikus becslés: 3,80%,
- Módosított szintetikus becslés: 6,03%,
- Utólagos rétegzés: 5,36%,
- Kombinált becslés: 3,70%.

Ugyanezen mutatók részletezett (nem, korcsoport stb.) becsléseinek relatív eltérései értelemszerűen sokkal nagyobbak voltak: 13,21%, 6,87%, 9,34%, 11,81% és 8,07%.

Az átlagtípusú mutatók becslése egyszerűbb, könnyebb, ezért, bár hasonló eredmények adódtak, az eltérések sokkal kisebbek voltak.

A relatív eltérések az átlagtípusú mutatók esetében a következők voltak:

- ELAR-minta: 2,93%,
- Szintetikus becslés: 2,97%,
- Módosított szintetikus becslés: 3,70%,
- Utólagos rétegzés: 3,10%,
- Kombinált becslés: 2,73%.

A részletezett mutatók esetében az eredmények a következők voltak: 9,09%, 4,16%, 6,44%, 7,99%, 4,99%.

Ezek az eredmények nagyon figyelemreméltóak. Azt jelzik ugyanis, hogy ott, ahol a becslések már eleve jók, ott ezek a módszerek nem javítanak, hanem még az is előfordul, hogy rontanak. Viszont ahol már nagy a bizonytalanság, ott lényeges javulás várható.

Mindkét esetben a szintetikus és a kombinált módszerek adták a legjobb eredményt. Mindez azonban nem jelenti azt, hogy minden egyes becslés javult, hanem csak azt, hogy összességükben jelentős a javulás. Egy-egy becslés, mint arról volt szó, még rosszabb is lehet, mint az egyszerű ELAR-becslés.

A következő négy táblázat áttekintést ad arról, milyen eredmények adódtak, és jól látható az is, hogy a kisterületi becslési eljárások miként javítják az eredményeket. Fontos azt is látni, a 6. számú mutató esetében egy megfelelő nagyságú ELAR-minta megenged még bizonyos részletezést is. A módszerek lehetőségei és korlátai tehát minden esetben mások és mások.

³ ELAR= egységes lakossági adatgyűjtési rendszer.

1. táblázat

A családok megoszlása az aktív keresők száma szerint, városok
Relatív eltérések a népszámlálási adatoktól, %-ban

Megyék	Közvetlen, egyszerű	Szintetikus	Módosított szintetikus	Utólagos rétegezéssel való	Kombinált
		becslés			
Baranya	17,0	1,8	8,2	18,8	7,0
Bács-Kiskun	15,4	2,6	11,0	1,6	1,8
Békés	22,4	2,2	11,4	11,0	3,0
Borsod-Abaúj-Zemplén	7,4	2,4	4,4	10,4	5,4
Csongrád	9,0	1,4	3,4	11,6	4,6
Fejér	9,0	6,8	9,8	9,4	4,8
Győr-Sopron	3,8	3,8	3,2	6,4	4,0
Hajdú-Bihar	7,6	4,0	9,4	9,2	5,4
Heves	11,8	2,8	3,0	5,2	1,4
Komárom	7,45	5,6	6,4	10,4	9,2
Nógrád	14,2	3,8	9,8	7,4	6,2
Pest	15,6	1,6	9,4	9,2	3,8
Somogy	20,0	3,4	7,2	13,8	4,6
Szabolcs-Szatmár	25,8	2,6	20,8	14,4	5,4
Szolnok	7,6	2,6	4,2	10,8	4,6
Tolna	8,4	7,2	3,8	8,2	6,2
Vas	12,6	4,6	14,0	12,8	4,0
Veszprém	4,4	3,4	2,6	7,4	5,2
Zala	10,0	3,0	3,2	9,4	5,0

2. táblázat

Az egy lakosra jutó alapterület, községek
Relatív eltérések a népszámlálási adattól, %-ban

Megyék	Közvetlen, egyszerű	Szintetikus	Módosított szintetikus	Utólagos rétegezéssel való	Kombinált
		becslés			
Baranya	1,4	0,0	1,4	0,0	0,0
Bács-Kiskun	0,0	4,7	1,6	3,1	1,6
Békés	3,0	11,9	11,9	3,0	3,0
Borsod-Abaúj-Zemplén	0,0	0,0	1,6	1,6	1,6
Csongrád	4,7	7,8	6,3	3,1	4,7
Fejér	1,6	8,2	6,6	1,6	1,6
Győr-Sopron	2,8	8,3	2,8	2,8	2,8
Hajdú-Bihar	1,7	1,7	3,4	0,0	0,0
Heves	4,6	1,5	1,5	6,2	3,1
Komárom	0,0	1,6	1,6	4,8	1,6
Nógrád	1,6	1,6	0,0	1,6	0,0
Pest	1,7	5,0	5,0	0,0	0,0
Somogy	1,6	3,1	4,7	1,6	1,6
Szabolcs-Szatmár	3,2	0,0	1,6	0,0	0,0
Szolnok	1,6	1,6	1,6	3,3	1,6
Tolna	1,5	3,0	4,5	1,5	0,0
Vas	1,5	8,8	7,4	2,9	2,9
Veszprém	0,0	3,1	0,0	1,5	1,5
Zala	0,0	1,6	1,6	1,6	0,0

3. táblázat

*A családok megoszlása az aktív keresők száma szerint
Hajdú-Bihar megyében*

Megnevezés						(Százalék)	
	0	1	2	3	4-x	Csak inatív	Össze- sen
	keresővel						
	<i>Községek</i>						
Közvetlen, egyszerű becslés	0,8	35,5	39,8	7,6	1,4	14,9	100,0
Szintetikus becslés	0,5	32,2	41,1	8,7	1,9	15,6	100,0
Módosított szintetikus becslés	0,4	31,5	42,1	8,8	2,2	15,0	100,0
Utólagos rétegezéssel való becslés	0,9	35,7	39,0	7,1	1,4	15,9	100,0
Kombinált becslés	0,5	33,5	40,5	8,2	1,5	15,8	100,0
Népszámlálási adat	0,9	33,9	39,2	7,8	1,8	16,4	100,0
	<i>Városok</i>						
Utólagos rétegezéssel való becslés	0,5	26,5	50,4	5,9	1,7	15,0	100,0
Kombinált becslés	0,3	28,4	49,0	6,1	1,3	14,9	100,0
Népszámlálási adat	0,4	30,6	46,3	6,3	1,3	15,1	100,0
	<i>Összesen</i>						
Közvetlen, egyszerű becslés	0,7	32,8	44,0	6,4	1,7	14,4	100,0
Szintetikus becslés	0,4	30,7	44,3	7,6	1,9	15,1	100,0
Módosított szintetikus becslés	0,4	31,7	45,9	7,5	1,8	12,7	100,0
Utólagos rétegezéssel való becslés	0,7	31,1	44,6	6,5	1,6	15,5	100,0
Kombinált becslés	0,4	31,0	44,7	7,2	1,4	15,3	100,0
Népszámlálási adat	0,6	32,2	42,7	7,1	1,6	15,8	100,0

4. táblázat

Az egy lakásra jutó alapterület Komárom megyében

Megnevezés				(m ²)
	Megye összesen	Községek	Városok	
Közvetlen, egyszerű becslés	60	63	58	
Szintetikus becslés	60	64	56	
Módosított szintetikus becslés	61	64	58	
Utólagos rétegezéssel való becslés	57	60	53	
Kombinált becslés	58	64	55	
Népszámlálási adat	59	63	55	

5. táblázat

Száz 15 éves és idősebb nőre jutó élve született gyermekek száma családi állapot szerint és város–község–összesen bontásban Győr-Sopron megyében

(Fő)			
Megnevezés	Házass	Özvegy	Elvált
<i>Községek</i>			
Közvetlen, egyszerű becslés	229,0	329,2	110,5
Szintetikus becslés	212,6	300,7	163,5
Módosított, szintetikus becslés	210,4	298,3	127,9
Utólagos rétegezéssel való becslés	229,2	328,8	95,6
Kombinált becslés	229,8	317,8	163,5
Népszámlálási adat	225,0	302,0	181,0
<i>Városok</i>			
Közvetlen, egyszerű becslés	179,9	237,9	121,4
Szintetikus becslés	180,0	227,0	151,4
Módosított szintetikus becslés	177,8	225,1	141,0
Utólagos rétegezéssel való becslés	183,2	241,1	90,7
Kombinált becslés	183,4	244,2	151,4
Népszámlálási adat	182,0	233,0	161,0
<i>Együtt</i>			
Közvetlen, egyszerű becslés	201,2	286,8	117,8
Szintetikus becslés	198,7	273,3	157,5
Módosított szintetikus becslés	196,2	270,6	141,7
Utólagos rétegezéssel való becslés	203,8	287,2	93,0
Kombinált becslés	204,0	285,4	160,8
Népszámlálási adat	202,0	269,0	166,0

Kulcsszavak: reprezentatív felvétel, kisterületi becslések, szintetikus becslés, kombinált becslés, modell alapú becslés, nem mintavételi hiba.

Resume

Sample size restricts the breakdown of representative survey data by, for example, territorial units, and there are only very rare opportunities to increase the sample. Therefore statisticians attempted – successfully – to increase the reliability of small area estimations by using certain internal interdependencies and external information. Based on a voluminous foreign literature of synthetic, combined, model-based and some other estimation methods, this study is a review of the possible applications of these methods, and illustrates the possible uses of these methods through examples.