# STATISZTIKAI SZEMLE

A KÖZPONTI
STATISZTIKAI HIVATAL
FOLYÓIRATA



82. ÉVFOLYAM 8. SZÁM 2004. AUGUSZTUS

E SZÁM SZERZŐI:

*Bauer Péter*, a KSH fogalmazója; *Csereháti Zoltán*, a KSH tanácsosa; *Éltető Ödön*, KSH ny. főosztályvezető-helyettese; *Erdei Virág*, a KSH fogalmazója; *Földesi Erika*, a KSH tanácsosa; *György Erika*, a KSH fogalmazója; *Horváth Roland*, a KSH fogalmazója; *Dr. Szép Katalin* kandidátus, a KSH főosztályvezető-helyettese; *Dr. Telegdi László* kandidátus, a KSH statisztikai tanácsadója; *Dr. Vigh Judit*, a KSH főtanácsosa.

*

*Balogh András* kandidátus, a KSH főtanácsosa; *Dévai Péter*, a KSH Könyvtár és Dokumentációs Szolgálat munkatársa; *Földházi Erzsébet*, a KSH Népességtudományi Kutató Intézet tudományos kutatója; *Nádudvari Zoltán*, a KSH főtanácsosa.

ISSN 0039 0690

Megjelenik havonta egyszer
Főszerkesztő: dr. Hunyadi László
Osztályvezető: Dobokayné Szabó Orsolya
Kiadja: a Központi Statisztikai Hivatal
A kiadásért felel: dr. Pukli Péter
4105 – Akadémiai Nyomda
Martonvásár, 2004
Felelős vezető: Reisenleitner Lajos

Szerkesztők: Várady Soma, Visi Lakatos Mária
Tördelőszerkesztők: Bartha Éva, Simonné Káli Ágnes

Szerkesztőség: Budapest II., Keleti Károly utca 5–7. Postacím: Budapest, 1525. Postafiók 51.
Telefon: 345-6908, 345-6546 Telefax: 345-6594
Internet: www.ksh.hu/statszemle
E-mail: statszemle@ksh.gov.hu
Kiadóhivatal: Központi Statisztikai Hivatal, Budapest II., Keleti Károly utca 5–7.
Postacím: Postafiók 51. Budapest, 1525. Telefon: 345-6000
Előfizetésben terjeszti a Magyar Posta Rt. Hírlap Üzletág. Előfizethető közvetlen a postai kézbesítőknél, az ország bármely postáján, Budapesten a Hírlap Ügyfélszolgálati Irodákban és a Központi Hírlap Centrumnál
(Budapest VIII., Orczy tér 1., Telefon: 06-1-477-6300; Postacím: Budapest 1900)
További információ: 06-80-444-444; hirlapelofizetes@posta.hu
Előfizetési díj: fél évre 3000 Ft, egy évre 5400 Ft
Beszerezhető a KSH Könyvesboltban. Budapest II., Keleti Károly u. 10. Telefon: 212-4348

# TARTALOM

# A MINTAVÉTELI ÉS MÓDSZERTANI OSZTÁLYON FOLYÓ MŰHELYMUNKA

A Központi Statisztikai Hivatalban – a hivatal történetében már nem először – 2001 őszén központi módszertani osztály alakult. Az osztály feladata a módszertani koordináció, kutatás és fejlesztés, valamint a szakfőosztályok eseti problémáinak megoldásában nyújtott, matematikai statisztikai módszereket igénylő segítség. A *Statisztikai Szemle* jelen száma az osztály tevékenységét kívánja bemutatni.

A fejlett országok többségének statisztikai hivatalában működik egy olyan központi módszertani, fejlesztési, kutatási szervezeti egység, melynek tevékenységi köréhez a mintavétel, becslés, hibaszámítás, editálás, imputálás, idősorelemzés – szezonális kiigazítás, statisztikai adatvédelem, felvételek tervezése, ár- és volumenváltozás mérése és elemzési módszerek tartoznak. Ez a kör az utóbbi évtizedekben a statisztikai minőségfejlesztés, minőségirányítás koordinációs feladataival is bővült. A magasan képzett, matematikai statisztikai ismeretekkel rendelkező, az adott munkaszakasz (például mintavétel, becslés, idősorelemzés) szakértői a központi módszertani osztályon minden olyan szakstatisztikai területen támogatást tudnak nyújtani, ahol erre az igény felmerül. Egy ilyen osztály révén biztosítható az adott munkaszakasz hivatalon belüli átláthatósága, a horizontális információáramlás és a tapasztalatok hasznosulása, továbbá nagyobb esély van a területhez kötődő tudományos eredmények követésére és – megfelelő kapacitás esetén – a kutatásra, fejlesztésre is.

A KSH Statisztikai kutatási és oktatási főosztályához tartozó Mintavételi és módszertani osztály célja, hogy járuljon hozzá a statisztikai munka minőségének javításához a statisztikai módszertani fejlesztés és összehangolás területén. E tevékenység fő elemei a következők:

– reprezentatív mintavételi tervek, becslés, hibaszámítás készítése, oktatása;
– a statisztikai munka egyes szakaszaira vonatkozó, több szakstatisztikai területen alkalmazható módszertani standardok, jó gyakorlatok, módszertani ellenőrzőlisták kialakítása;
– a statisztikai munkát segítő módszertani kutatások végzése, fejlesztések elméleti és kísérleti megalapozása;
– az egyes szakstatisztikák területén végzett fejlesztésekben való részvétel;
– eseti problémák megoldásához matematikai statisztikai módszereket igénylő segítségnyújtás.

A KSH-ban a mintavétel, a becslés és a mintavételi hibaszámítás azok a tevékenységek, amelyeket a rendszeres felvételek esetében hagyományosan a statisztikatudomány erre szakosodott művelői végeznek. Ők alkották az osztály magját megalakulásakor. Mellettük matematikus, közgazdász, tanár és informatikus végzettségű fiatalok kezdték sta-

tisztikusi pályájukat, lehetővé téve új témák (például adatvédelem, indexszámítás) kutatását, fejlesztését. Az ugyancsak osztályunkon működő *Józan Péter* és csoportja által végzett demográfiai, egészségügyi kutató és elemző munkákat (például a dohányzás és az alkohol hatása az egészségre) a KSH kiadványaiból már jól ismerhetik az olvasók.

Együttműködünk a szakfőosztályok és az Informatikai főosztály szakértőivel, szem előtt tartva, hogy a statisztikai munkát segítő eredmények csak így érhetők el.

Korábbi tevékenységeit az osztály fokozatosan fejleszti és bővíti új feladatokkal. Ez a munka tükröződik az itt bemutatott cikkekben is, bár ezúttal az osztály egyes fontos tevékenységeiről (például a turizmus határon történő megfigyeléséről) nem közlünk tanulmányt.

A lakossági felvételek összehangolásának szükségessége már a hetvenes években nyilvánvalóvá vált. A KSH-ban, a Népszámlálási főosztállyal együttműködve, a szakértők kidolgozták és működtették az Egységes Lakossági Adatszolgáltatási Rendszert, az ELAR-t, amelyről a korábbi évtizedekben többször is beszámoltak a *Statisztikai Szemle* hasábjain és a KSH módszertani kiadványában. A lakossági felvételek mintái a Hivatalban többnyire a népszámlálások eredményeire épülnek, mintavételi keretként az összeírt címállományt használják. Ezért egy-egy új népszámlálás egyben lehetőség is a mintavételi és becslési módszerek megújítására. Ez történt a 2001. évi népszámlálást követően, a 2003. évtől bevezetésre került új mintákkal is. A lakossági felvételeknél egyre nagyobb gond az adatszolgáltatási hajlandóság romlása, amely egyes rétegeknél már-már az adatok megbízhatóságát veszélyezteti. A Háztartási Költségvetési Felvétel az első kísérlet arra, hogy az egyes rétegek válaszolási hajlandóságában megmutatkozó különbségeket már a mintavételi eljárásnál figyelembe vegyük, és így kisebb költséggel érjük el a kívánt pontosságú eredményeket. Ez alkalommal *Éltető Ödön* ismerteti a Háztartási Költségvetési Felvétel (HKF) új mintavételi és becslési eljárását. A munkaerőfelvétel új mintájáról a *Statisztikai Szemle* 2003. évi 12. számában, a hibaszámítási módszerről pedig a 9. angol nyelvű számban *Mihályffy László* cikkét már olvashatták.

A gazdasági szervezetekre vonatkozó felmérésekben az integrált gazdaságstatisztikai megfigyelési rendszer kialakítása – az Iparstatisztikai főosztály szervezésében – a kilencvenes évek második felében értelemszerűen az egységes elveken nyugvó mintavételi, becslési és mintavételi hibaszámítási eljárások kialakításával járt. A gazdaság koncentrációja és a megbízható statisztikák előállításának követelménye vezetett oda, hogy a nagyvállalatokat teljes körűen megfigyeljük, és a mintavétel csak a kisebb vállalkozásokra korlátozódik. A gazdasági szervezetek teljesítményére vonatkozó éves és évközi (havi, negyedéves) statisztikai felvételek mintavételi és becslési gyakorlata már többször szerepelt a *Statisztikai Szemle* hasábjain és a KSH módszertani kiadványaiban; ebben a számban *Telegdi László* a kisszervezeteknek a 2000-es években végzett integrált reprezentatív évközi megfigyeléséről számol be.

A statisztikai munka egyes szakaszaira vonatkozó módszertani koordináció és fejlesztés, a meglévő források korlátozottsága miatt, csak fokozatosan alakítható ki. Ilyen jellegű fejlesztési igény az osztály megalakulásakor először a szezonális kiigazítás területén merült fel. *Berki Natália* és *Fábián László* kezdte el a helyzetértékelést és a módszertani fejlesztést. A működési rendszer kialakítása és kísérleti működtetése után kialakított gyakorlatot mutatják be közös cikkükben *Bauer Péter* és *Földesi Erika*, a szakterület jelenlegi felelősei. A dinamikusan növekvő igények következtében a KSH-ban ma már több

mint száz idősor szezonális kiigazítása történik rendszeresen. A szakfőosztályok és a módszertani osztályon dolgozó munkatársak közötti munkamegosztás és együttműködés következményeként a publikálásra kerülő eredmények megfelelnek mind a matematikai statisztikai, mind a szakstatisztikai szempontoknak. A problémák, tapasztalatok, a kapcsolódó tudományos eredmények és nemzetközi események rendszeres évi megbeszélése, az osztályon végzett kutatás és fejlesztés, valamint az igény esetén tartott oktatás biztosítja a folyamatos minőségi fejlesztést ezen a szakterületen.

A másik terület az adatvédelem, ahol hasonló módszertani fejlesztő, koordináló munka iránti igény merült fel. Az adatvédelmet törvény szabályozza, de adatszolgáltatóink bizalmának megőrzése a KSH-nak is alapvető érdeke. Miközben egyre több és részletesebb statisztikai információval látjuk el a felhasználókat, meg kell védenünk adatszolgáltatóinkat, azaz biztosítanunk kell, hogy egy-egy adatszolgáltatóval közvetlenül kapcsolatba hozható adat ne legyen feltárható. Ennek a – talán az olvasók többségének is új – területnek a bemutatását vállalta el *Horváth Roland* és *Erdei Virág*, a Népszámlálási főosztály munkatársa közösen írt cikkükben.

Az osztályon folyó kutató és elemző munkából két cikk nyújt ízelítőt, mindkettő témájából a szerzők sikeres előadást tartottak 2004 májusában, Mainzban a „Quality and Methodology in Official Statistics" című konferencián. *Csereháti Zoltán* az outlierek felismerésének és kezelésének terén elért eredményeit mutatja be. A minta alapján történő becslés megbízhatósága erősen függ attól, hogy mennyire tudjuk elválasztani a sokaságra valóban jellemző mintaelemeket az atipikus, valójában csak önmagukat képviselő elemektől. A mintába bekerült, de valójában igen ritka, speciális egységet jellemző érték, ha nem szűrjük ki, nagyon „elviheti" a becslést.

Hasonlóképpen veszélyt hordoz a nemválaszolás is. Ha a nem elérhető, választ megtagadó egységek a felvétel célja szempontjából a válaszolóktól eltérő sajátosságokkal rendelkeznek, akkor a felvétel eredménye torzított lehet. A nemválaszolás vizsgálata a további felvételeknél is segíthet a sajátosságok, okok feltárásában, és így a nemválaszolás megelőzésében és csökkentésében. A lakossági felvételeknél az okozza a nehézséget, hogy a nemválaszolókról általában nincs információnk, nem ismerjük jellemzőiket. Kivételes, de a nemzetközi gyakorlatban nem ismeretlen lehetőség, hogy a nemválaszolók jellemzésére a népszámlálási adatokat használjuk fel, ha a felvétel a népszámlálást követi. Ilyen jellegű vizsgálatra a KSH gyakorlatában – ismereteink szerint – ez ideig még nem került sor, így *György Erikának* az ezzel a kérdéssel foglalkozó tanulmánya úttörő jellegű.

A statisztika minőségének mérésével, javításával kapcsolatos fejlesztő, koordináló munkát is szeretnénk szervezett formában végezni. Az ennek megalapozására végzett irodalmi feldolgozás és helyzetelemzés összefoglalójának tekinthető *Vigh Judittal* közösen írt tanulmányunk. A minőség divatos szó, elég általános fogalom ahhoz, hogy mindenütt hivatkozzanak rá. Cikkünkben azt szeretnénk bemutatni, hogy a minőség definíciója hogyan transzformálható mérhető kategóriákká, a minőség javítása megvalósítható feladatokká. Reményeink szerint azoknak is hasznos ez az áttekintés, akik tudják, hogy ez nem valami „új csoda", hiszen eddig is ez vezérelte a hivatalban dolgozó statisztikusok munkáját, de a rendszerszerű megközelítés új eredményeket hozhat majd a gyakorlatban.

*Szép Katalin*

# AZ ÚJ HKF-MINTA KIVÁLASZTÁSI ELJÁRÁSA ÉS A 2003. ÉVI TAPASZTALATOK

ÉLTETŐ ÖDÖN

Az Egységes Lakossági Adatfelvételi Rendszer (ELAR) 1976. évi bevezetése óta e rendszer része a Háztartási Költségvetési Felvétel (HKF). Mintája azóta a népszámlálások adatain alapul, s így a népszámlálások után mindig új HKF-minta készült. A legutóbbi, a 2001. évi népszámlálás után is így történt. Bár az elmúlt évtizedekben sok tekintetben kialakultak a HKF-minta kiválasztásánál alkalmazandó standard eljárások, a 2002. évi minta kiválasztása során több új, előzetesen különböző fórumokon megvitatott elemet is tartalmazott a mintavételi eljárás.

A tanulmány dokumentálja a 2003-tól bevezetett új HKF-minta kiválasztásánál alkalmazott alapelveket és módszereket, egyúttal beszámol az első év tapasztalatai alapján a háztartások közreműködési készségéről, illetve a megtagadási és egyéb meghiúsulási arányokról.

TÁRGYSZÓ: Mintavétel. Meghiúsulások.

A 2003. januártól működő HKF-minta kiválasztási eljárásának kidolgozása során az alábbi *alapelvekből*, illetve feltevésekből indultunk ki.

*1*. Az új HKF-minta nagysága gyakorlatilag nem változik, és továbbra is körzetenként hat elsődleges cím kerül kiválasztásra, ami egyúttal azt is jelenti, hogy országos szinten a kiválasztott háztartások száma sem változik.

*2*. Eltérően a munkaerő felmérés új mintájától, az új HKF-minta kiválasztása továbbra is két, illetve három lépcsőben történik, mert a HKF működéséhez feltétlenül szükség van a háztartások kiválasztás előtti rétegbe sorolására, ez pedig, a legkisebb települések kivételével, gyakorlatilag csak a lakások valamilyen kisebb csoportjainak kiválasztása révén valósítható meg.

*3*. Célszerűnek tűnt a településeket a mintavétel előtt továbbra is nagyság szerint rétegezni, de a rétegezés az új mintánál a népességszám helyett a települések *lakásszáma* alapján történt, hiszen egyrészt a kiválasztási valószínűségek meghatározásának alapja is a lakásszám, másrészt a népességszám tartalmazza a felvétel körébe nem tartozó intézeti népességet is.

*4*. Az előző időszak HKF-mintája kiválasztási eljárásához hasonlóan alapkövetelmény volt, hogy a minta kiválasztása során az elsődleges mintavételi egységeket, illetve ahol vannak, a másodlagos, de nem végső mintavételi egységeket is *nagysággal arányos valószínűséggel* (nav), a végső mintavételi egységeket – a lakásokat – pedig az elsődleges, il-

letve másodlagos egységeken belül az egyes háztartási rétegekből *egyenlő valószínűséggel (ev*) válasszuk ki. Ennek megfelelően a mintába kiválasztott nem önreprezentáló településekről egy-egy rétegen – településnagyság-kategórián – belül azonos számú másodlagos mintavételi egységnek kellett bekerülnie a mintába, és a lakás körzetekből is azonos számú lakást kell majd az elkövetkező évek során kiválasztani.

*5*. Mivel a HKF-nél továbbra is alkalmazni kívánják kétévenként a címbejárást, továbbá a címpótlásokhoz feltétlenül szükség van a háztartások előzetes rétegezésére, ehhez településen belül feltétlenül valamilyen kisebb kiválasztási egységet kell alkalmazni. Ez a kisebb kiválasztási egység továbbra is csak a *népszámlálási számlálókörzet* lehet.

*6*. Fontos követelménynek tekintettük, hogy bizonyos lakásszám felett a települések *önreprezentálók* legyenek, azaz mind kerüljenek be a mintába. Ezen települések esetében tehát a minta *két lépcsőben* került kiválasztásra, elsődleges mintavételi egységek a népszámlálási körzetek voltak. 2002-ig bezárólag a 20 ezer vagy annál több lakosú városok voltak önreprezentálók a HKF-mintában. E tekintetben csak annyi változás történt, hogy a nagysághatárt - miként a többi település nagyságkategória esetében is - nem a népesség, hanem a *lakott lakások* (és egyéb lakóegységek) száma határozta meg. A népszámlálás adatai szerint, ha az új HKF-mintánál az önreprezentáló településeket úgy határozzuk meg, hogy azok a települések tartozzanak ide, ahol a lakott lakások száma ($L$) legalább 7000, azaz $L \geq 7000$ , akkor a 68 ilyen város közül mindössze hétben kevesebb a népesség száma 20 ezernél, és ezek közül csak három (Komárom, Budaörs és Szigetszentmiklós) nincs benne az 1993-tól 2002-ig bezárólag működött HKF-mintában. Ezért az új HKF-mintában a legalább 7000 lakású települések az önreprezentálók.

*7*. Az eredeti terv szerint a legkisebb település nagyságkategóriában – ahol $L < 250$ – szintén két lépcsőben történt volna a minta kiválasztása oly módon, hogy a lakások ezen községek teljes lakásállományából kerültek volna kiválasztásra, elhagyván a körzet kiválasztást. Annak érdekében azonban, hogy a lakások kiválasztására szolgáló program egységes lehessen, formailag itt is körzetenként hat lakás kerül kiválasztásra. Ez a mintanagyságot és a minta struktúráját alig érintette, mivel az ebbe a kategóriába kiválasztott települések zömében amúgy is éppen két önálló körzet van.

*8*. Fontos követelmény az új HKF-mintával szemben, hogy a következő tízéves időszak (2003–2012) alatt minél kevesebb körzetet kelljen cserélni a miatt, hogy elhasználódtak a címek, bár a HKF új mintakerete bizonyos esetekben fog tartalmazni ún. tartalékkörzeteket is (lásd később). Figyelembe véve a feltevések szerint lényegében változatlanul maradó rotációs szkémát, számításaink szerint legalább negyven lakást kell egy körzetnek tartalmaznia ahhoz, hogy tíz évig mintakeretül szolgálhasson a HKF-mintához. Ez azt jelenti, hogy a negyvennél kevesebb lakást tartalmazó körzeteket össze kellett vonni úgy, hogy az összevont körzetek együttes lakásszáma legalább negyven legyen. A körzetek összevonását lehetőleg az azonos jellegű (belterületi, egyéb belterületi, külterületi) körzetek összevonásával igyekeztünk megoldani, erre azonban nem mindig volt lehetőség. Az összevonást természetesen – miként az 1992 és 2002 közötti években működő minta kiválasztásánál is történt, ahol a legalább huszonöt lakásos körzeteket tekintettük önálló körzeteknek – a *körzetek kiválasztása előtt* végeztük el, a kiválasztásnál ezek az összevont körzetek egy mintavételi egységként szerepeltek.

*9*. Az új HKF-minta kiválasztási eljárásának megtervezése során abból indultunk ki, hogy a kiválasztási arányoknak az egyes településnagyság-kategóriákban nem feltétlenül

azonosnak, de egyértelműen meghatározhatónak kell lenniük, továbbá egy-egy kategórián és háztartásrétegen belül minden lakás kiválasztási valószínűségének azonosnak kell lennie.

## A RÉTEGEZÉS

A *települések nagyság szerinti rétegezésére* az alábbi kategóriákat alakítottuk ki:

| Réteg | Lakásszám |
|---|---|
| *1.* | – 249 |
| *2.* | 250 – 499 |
| *3.* | 500 – 799 |
| *4.* | 800 – 1299 |
| *5.* | 1300 – 2499 |
| *6.* | 2500 – 3799 |
| *7.* | 3800 – 6999 |
| *8.* | 7000 – |

A 6. alapelvben leírtaknak megfelelően a 8. rétegbe kerülő települések az önreprezentáló települések.

Az elmúlt évtizedben a megvalósult HKF-minta összetétele a *háztartások rétegei* szerint minden évben elég jelentősen különbözött a kiválasztott minta összetételétől. Ez elsősorban abból adódott, hogy az idősebb, nyugdíjas háztartások szívesebben működtek közre a felvételben, mint a fiatalabb, aktív háztartások. Különösen nagy volt a nemválaszolási arány a vállalkozók, illetve a felsőfokú végzettségű aktív keresők háztartásai esetében. Ezért felmerült az igény, hogy eltérő kiválasztási arányt alkalmazva kíséreljük meg elérni, hogy a megvalósult minta struktúrája jobban közelítsen a kiválasztott mintáéhoz. Először a körzetekből próbáltunk meg rétegeket kialakítani oly módon, hogy a zömmel jobb közreműködési készségű háztartásokból álló körzetek alkottak volna egy külön réteget, egy másikat azok a körzetek, ahol nagy arányban vannak kisebb, de még elfogadható mértékű válaszadási arányú háztartások, egy harmadik réteg pedig a zömmel legkritikusabb közreműködési készségű háztartásokat tartalmazó körzetekből állt volna. A számításokból azonban kiderült, hogy nemigen vannak homogén körzetek, a körzetek többsége a tekintetbe vett ismérvek szerint meglehetősen inhomogén. Ezért a körzetek rétegezése helyett járhatóbb útnak bizonyult magukat a háztartásokat rétegezni, és e háztartásrétegeknél alkalmazni eltérő kiválasztási arányokat. Sok réteget természetesen nem lehetett kialakítani, hiszen a kiválasztási eljárás megszabta, hogy minden körzetből évente azonos számú, hat háztartás kerüljön be a mintába. A felvételt irányító osztály véleményét is figyelembe véve végül is az alábbi háztartásrétegeket alkalmaztuk.

Nem önreprezentáló települések: *1*. réteg: a háztartásfő 60 éves vagy idősebb
*2*. réteg: a háztartásfő fiatalabb 60 évnél

Önreprezentáló települések: *1*. réteg: a háztartásfő 60 éves vagy idősebb
*2*. réteg: a háztartásfő fiatalabb 60 évnél és legfeljebb középfokú végzettségű
*3*. réteg: a háztartásfő fiatalabb 60 évnél és felsőfokú végzettségű

Természetesen felmerül az a probléma, hogy az előbbi rétegzés a háztartásokra vonatkozik, a végső kiválasztási egység viszont a lakás, nem a háztartás. Az adatok azon-

ban azt mutatják, hogy ma már csak elvétve fordul elő olyan lakás, ahol több háztartás él. Így a lakás és a háztartás elvi különbözősége a gyakorlatban nem okoz problémát.

A kiválasztási valószínűségeket oly módon differenciáltuk, hogy az *1*. rétegben a kiválasztási arány fele, a *3*. rétegben (az önreprezentáló településeken belül) pedig kétszerese legyen a településen belüli átlagos kiválasztási aránynak.

A következőkben leírjuk a minta rétegzésének elméleti hátterét.

*Jelölések*

$T_r$ = az *r*-edik rétegben a *települések* száma ($r = 1,2,\ldots,8$)

$t_r$ = az *r*-edik rétegből a mintába *kiválasztandó települések* száma

$K_{ri}$ = az *r*-edik réteg *i*-edik településén a körzetek száma

$k_r$ = az *r*-edik rétegből kiválasztott településeken a mintába *kiválasztandó körzetek* száma (az 1–7 rétegekben rétegen belül minden minta településén azonos)

$L_r$ = az *r*-edik réteg *teljes lakásszáma:*

$$L_r = \sum_{i=1}^{T_r} \sum_{j=1}^{K_{ri}} L_{rij}$$

$L_{ri}$ = az *r*-edik réteg *i*-edik településén a *lakások* száma

$L_{rij}$ = az *r*-edik réteg *i*-edik településén a *j*-edik körzet *lakásainak* száma

$l_r$ = az *r*-edik rétegből a mintába kiválasztott lakások száma, $l_r = 6t_rk_r$ ($r = 1,\ldots,7$)

$P_r$ = az *r*-edik rétegben egy lakás kiválasztásának átlagos valószínűsége

$$P_r = \frac{l_r}{L_r} = \frac{6t_rk_r}{L_r} \qquad (r = 1,\ldots,7)$$

$P_{ri}$ = az *r*-edik rétegben az *i*-edik település kiválasztási valószínűsége. Az alapelvek 4. pontja értelmében $P_{ri} = \gamma_r L_{ri}$ ($r = 1,\ldots,7$)

$P_{rij}$ = az *r*-edik réteg *i*-edik településén a *j*-edik körzet kiválasztási valószínűsége. Az előző fejezet 4. pontja szerint $P_{rij} = \lambda_r L_{rij}$

$l_{ri}^{(1)}$ = az *r*-edik nagyságkategóriából a mintába bekerült *i*-edik településen a *kiválasztott* körzetekben összesen az 1. háztartásrétegbe tartozó lakások száma ($r = 1,\ldots,8$)

$l_{ri}^{(2)}$ = az *r*-edik nagyságkategóriából a mintába került *i*-edik településen a *kiválasztott* körzetekben összesen a *2*. háztartásrétegbe tartozó lakások száma ($r = 1,\ldots,8$)

$l_{8i}^{(3)}$ = az *i*-edik önreprezentáló település *kiválasztott* körzeteiben összesen a *3*. háztartásrétegbe tartozó lakások száma

$l_{ri} = l_{ri}^{(1)} + l_{ri}^{(2)}$ = az *r*-edik nagyságkategóriából a mintába került *i*-edik településen a kiválasztott körzetekben a lakások száma ($r = 1,\ldots,7$)

$l_{8i} = l_{8i}^{(1)} + l_{8i}^{(2)} + l_{8i}^{(3)}$ = az *i*-edik önreprezentáló település kiválasztott körzeteiben a lakások száma összesen

*Kiválasztási valószínűségek*

Az *r*-edik település rétegben egy lakás kiválasztásának $P_r$ átlagos valószínűsége három valószínűség szorzataként írható fel:

$$P_r = P_{ri} \cdot P_{rj|i} \cdot P_{rk|ij},$$

ahol $P_{rj|i}$ annak valószínűségét jelöli, hogy az *r*-edik réteg *i*-edik *kiválasztott* településén a *j*-edik körzet bekerül a mintába, $P_{rk|ij}$ pedig annak átlagos valószínűségét, hogy az *r*-edik réteg *i*-edik kiválasztott településéből kiválasztott *j*-edik körzetből a *k*-adik lakás kerül bele a mintába. A fenti feltételes valószínűségeket az alábbi módon lehet meghatározni:

mivel, $\sum_{i=1}^{T_r} P_{ri} = t_r$ , azaz $\sum_{i=1}^{T_r} \gamma_r L_{ri} = \gamma_r \sum_{i=1}^{T_r} L_{ri} = \gamma_r L_r = t_r$ , ebből $\gamma_r = \frac{t_r}{L_r}$ , $(r = 1,\ldots,7)$.

Hasonló megfontolással

$$\sum_{j=1}^{K_{ri}} P_{rj|i} = k_r \text{ , azaz } \sum_{j=1}^{K_{ri}} \lambda_r L_{rij} = \lambda_r \sum_{j=1}^{K_{ri}} L_{rij} = \lambda_r L_{ri} = k_r \quad \text{amiből} \quad \lambda_r = \frac{k_r}{L_{ri}} .$$

Így $P_{ri} = \gamma_r L_{ri} = \frac{t_r L_{ri}}{L_r}$ és $P_{rj|i} = \lambda_r L_{rij} = k_r \frac{L_{rij}}{L_{ri}}$ $(r = 1,\ldots,7)$.

A *8*. nagyságcsoportban, az önreprezentáló települések közül nincs kiválasztás, ezért $P_{ri} = 1$. Viszont a mintába kerülő körzetek száma nem konstans, településenként változik, ezért $r = 8$ esetén a fenti formulákban $k_r$ helyett $k_{8i}$ írandó.

Mint előzőleg említettük, az *1*. háztartásrétegre a kiválasztási valószínűség fele az *i*-edik településre vonatkozó átlagos kiválasztási valószínűségnek, azaz

$$\pi^{(1)}_{rk|i} = \frac{1}{2}\pi_{rk|i} = \frac{1}{2}\frac{6k_r}{l_{ri}} = \frac{3k_r}{l_{ri}} .$$

Ha $n^{(1)}_{ri}$ jelöli az *r*-edik nagyságcsoport $(r = 1,\ldots,7)$ *i*-edik településén az *1*. háztartásrétegből ily módon kiválasztott lakások számát, azaz

$$n^{(1)}_{ri} = l^{(1)}_{ri} . \pi^{(1)}_{rk|i} = 3k_r \frac{l^{(1)}_{ri}}{l_{ri}} \quad \text{(egészre kerekítve)},$$

akkor a nem önreprezentáló településeken a *2*. háztartásrétegből

$$n^{(2)}_{ri} = 6k_r - n^{(1)}_{ri} = 6k_r - 3k_r \frac{l^{(1)}_{ri}}{l_{ri}} = 3k_r \left(2 - \frac{l^{(1)}_{ri}}{l_{ri}}\right)$$

lakás választandó ki, s így a *2*. háztartásrétegben egy lakás kiválasztási valószínűsége:

$$\pi^{(2)}_{rk|i} = \frac{n^{(2)}_{ri}}{l^{(2)}_{ri}} = \frac{3k_r\left(2 - l^{(1)}_{ri}/l_{ri}\right)}{l^{(2)}_{ri}} = 3k_r \left(\frac{1}{l_{ri}} + \frac{1}{l^{(2)}_{ri}}\right) .$$

A településről kiválasztandó $n_{ri}^{(1)}$ lakást a mintába került körzetek között az egyes körzetekben található, az *1*. háztartásrétegbe tartozó lakások arányában kell elosztani, természetesen egész számra kerekítve. A *2*. háztartásrétegből pedig körzetenként annyi lakást kell egyenlő valószínűséggel kiválasztani, hogy a két rétegből összesen 6 lakás kerüljön bele a mintába. Ritkán előfordulhat, hogy olyan körzet kerül bele a mintába, amelyikben csak az egyik háztartásrétegbe tartozó lakások vannak. Ilyen esetben először ebből a körzetből kell kiválasztani a 6 lakást egyenlő valószínűséggel, majd, ha ez mondjuk az *1*. háztartásrétegből történt, akkor az $n_{ri}^{(1)} - 6$ lakást kell elosztani a maradék $k_r$–1 körzetre az *1*. rétegbe tartozó lakások arányában.

Az önreprezentáló településeken annyiban bonyolultabb a helyzet, hogy a körzetenkénti *6* lakást három háztartásréteg között kell elosztani. Első lépésben ekkor is az *1*. háztartásrétegből kiválasztandó lakások számát kell meghatározni és elosztani a körzetek között a fentiekhez analóg módon. A következőkben a *3*. háztartási rétegre vonatkozó kiválasztási valószínűséget kell kiszámítani úgy, hogy az *kétszerese* legyen a településre vonatkozó átlagos kiválasztási aránynak, azaz

$$P_{8i}^{(3)} = 2P_{8i} = 12\frac{k_{8i}}{l_{8i}}$$

A *2*. háztartás rétegben pedig a kiválasztási valószínűség

$$P_{8i}^{(2)} = \frac{6k_{8i} - n_{8i}^{(1)} - n_{8i}^{(3)}}{l_{8i} - l_{8i}^{(1)} - l_{8i}^{(3)}} = 3k_{8i}\frac{2l_{8i} - l_{8i}^{(1)} - 4l_{8i}^{(3)}}{l_{8i}\left(l_{8i} - l_{8i}^{(1)} - l_{8i}^{(3)}\right)},$$

ami lehet kisebb is, nagyobb is $P_{8i}$-nél a három réteg súlyától függően.

A fenti meggondolások a tekintetben nem teljesen pontosak, hogy a települések, illetve körzetek teljes lakásszámai szerepelnek a formulákban, holott a kiválasztási eljárás szerint a háztartások két, illetve három rétegében különböző a kiválasztási arány. A $\frac{6k_r}{l_{ri}}$ valószínűség annak a $P_{rj|t}.P_{rk|ij} = \frac{6k_r}{L_{ri}}$ valószínűségnek a szerepét veszi át, amely az *r*-edik kategória *i*-edik települése valamelyik lakásának a mintába kerülésével kapcsolatos, ha ez a település biztosan bekerül a mintába. $L_{ri}$ $l_{ri}$-vel történő helyettesítése azt jelenti, hogy a település összes lakása helyett csak azokkal számolhatunk, amelyek a mintakörzetekhez tartoznak, hiszen a többiekre nem ismerjük, mely háztartásréteghez tartoznak.

## A MINTA ELOSZTÁSA MEGYÉKRE ÉS TELEPÜLÉSNAGYSÁG-KATEGÓRIÁKRA

Az új HKF-minta elosztásához rendelkezésre állt a lakott lakások és egyéb lakott lakóegységek száma a 2001. évi népszámlálás adatai szerint Budapesten kerületenként, a megyékben településenként.

*A számítások menete*

Első lépésben a településeket be kellett sorolni a kialakított nagyságkategóriákba, majd a jelzett tagolásban az egyes kategóriákba eső települések és lakások száma és százalékos megoszlása került meghatározásra. Összehasonlítva ezeket a 2002 végéig működő minta megfelelő számaival kiderült, hogy a régi minta struktúrája jelentősen eltért a legutóbbi népszámlálás adataiból kirajzolódó struktúrától, ami önmagában nem lett volna baj, ha az eltérés következetesen a nagyságkategóriánként eltérő kiválasztási arányokból adódott volna.

Az eltérések azonban sok esetben esetlegesek voltak, ami részben a minta többszöri nem következetes átalakításának következménye volt, részben pedig abból adódott, hogy az elmúlt tíz év alatt nem elhanyagolható mértékben módosult az ország település- és lakásstruktúrája. A fővárosi kerületek vonatkozásában is elég jelentős és indokolatlan eltérések mutatkoztak a régi minta és a népszámlálási adatok között.

A számítások következő lépésében dönteni kellett arról, milyen kiválasztási arányokat alkalmazzunk az új HKF-mintánál az egyes nagyságkategóriákban, illetve a főváros egyes kerületeiben. A globális $f$ kiválasztási arányt a lényegében változatlan nagyságú minta és a népszámlálás során összeírt 3 720 578 háztartás hányadosa adja, ami nagyjából 3 ezrelék, pontosabban

$$f = \frac{11646}{3720578} = 0{,}0031302\,.$$

Tekintettel a fővárosban hosszú évek óta tapasztalt alacsonyabb közreműködési készségre, indokoltnak tűnt Budapesten a mintát körülbelül 11 százalékkal növelni annak érdekében, hogy a megvalósult minta jobban reprezentálja a főváros háztartásait. A fővároson belül pedig azokban a kerületekben alkalmazni nagyobb kiválasztási arányt, ahol az átlagosnál nagyobb a nemválaszolási arány. A fővárosi kerületeknél alkalmazott $f$ százalékos kiválasztási arányokat és a kiválasztott körzetek $k$ számát az 1. tábla mutatja. Mivel körzetenként továbbra is hat lakás kerül kiválasztásra, a minta elemszámait kerületenként $6k$ adja.

1. tábla

*Kiválasztási arányok és a mintába került körzetek száma Budapesten, kerületenként*

| Kerület | $f$ (százalék) | $k$ | Kerület | $f$ (százalék) | $k$ |
|---|---|---|---|---|---|
| 01 | 0,329 | 8 | 13 | 0,304 | 28 |
| 02 | 0,361 | 25 | 14 | 0,351 | 34 |
| 03 | 0,356 | 31 | 15 | 0,318 | 18 |
| 04 | 0,302 | 20 | 16 | 0,323 | 14 |
| 05 | 0,323 | 9 | 17 | 0,317 | 14 |
| 06 | 0,320 | 12 | 18 | 0,303 | 22 |
| 07 | 0,317 | 18 | 19 | 0,339 | 14 |
| 08 | 0,302 | 20 | 20 | 0,320 | 14 |
| 09 | 0,324 | 19 | 21 | 0,303 | 15 |
| 10 | 0,319 | 17 | 22 | 0,322 | 10 |
| 11 | 0,348 | 37 | 23 | 0,323 | 5 |
| 12 | 0,355 | 21 | Budapest | 0,327 | 425 |

Ami a különböző nagyságkategóriákon belül alkalmazott kiválasztási arányokat illeti, az önreprezentáló városok esetén ez megegyezik a globális kiválasztási aránnyal, azaz $f_8 = 0{,}0031285$. A többi nagyságkategóriánál a kiválasztási arányok enyhén növekednek, az $f_1 = 0$, 285 százaléktól $f_7 = 0{,}308$ százalékig. A kiválasztási arányokat természetesen befolyásolja az is, hogy a nagysággal arányos valószínűségű kiválasztás megvalósításához egy-egy kategórián belül településenként azonos egész számú körzetet kell a mintába kiválasztani, s így a mintaelemszámok is szükségszerűen a hat egészszámú többszöröseinek kell lenniük.

Az előzők szerint a nem önreprezentáló települések kategóriáiban az alábbi összefüggés áll fenn az $f_r$ kiválasztási arány, a $t_r$ településszám, a településeken belül azonos $k_r$ körzetszám, valamint a kategória $L_r$ teljes lakásszáma között:

$$f_r = \frac{6t_r k_r}{L_r},$$

ahonnan $f_r$ és $L_r$ ismeretében $t_r$ a $k_r$ függvényében meghatározható. A településstruktúra, valamint korábbi tapasztalatok figyelembe vételével az alábbi, a 2. táblában látható eredményre jutottunk.

2. tábla

Kiválasztási arány és mintaelemszám nagyságkategóriánként

| Mintaelem | Nagyságkategória | | | | | | | | | Budapest | *Összesen* |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1. | 2. | 3. | 4. | 5. | 6. | 7.1. | 7.2. | 8. | | |
| | Kiválasztási arány (százalék) | | | | | | | | | | |
| | 0,285 | 0,297 | 0,304 | 0,305 | 0,306 | 0,311 | 0,312 | 0,316 | 0,317 | 0,327 | 0,312 |
| | Mintaelemszám | | | | | | | | | | |
| Település | 36 | 40 | 32 | 28 | 34 | 16 | 16 | 2 | 68 | 1 | *274* |
| Körzet | 81 | 120 | 128 | 140 | 170 | 96 | 135* | 20 | 626 | 425 | *1 941* |
| Lakás | 486 | 720 | 768 | 840 | 1 020 | 576 | 810 | 120 | 3 756 | 2 550 | *11 646* |

* Szervezési okok miatt Tolna megyében két településre lett elosztva a kilenc körzet.

Az új HKF-minta tízzel több települést tartalmaz, mint az előző, viszont a körzetek száma és egyúttal a lakások száma is valamivel kevesebb (42 körzettel, illetve 252 lakással).

Végül elkészült az új minta elosztása megyékre, illetve régiókra is. Az egyes megyékben a nagyságkategóriákra vonatkozó kiválasztási arányok természetesen csak közelítően érvényesülnek, mert egy adott kategóriára vonatkozó településenkénti körzetszámok miatt a minta elosztása nem lehet egészen pontos, hiszen egy település bevonása a mintába vagy elhagyása a mintából a 6 egészszámú többszörösével növeli vagy csökkenti a mintába kerülő lakások számát. A mintába került települések és körzetek számát régiónként és nagyságkategóriánként a 3. tábla tartalmazza. A Közép-Magyarország régióban a 3. tábla külön mutatja Budapest és Pest megye adatait. A lakásszámok természetesen a körzetszámok hatszorosai.

3. tábla

*A kiválasztott települések (t) és körzetek (k) száma régiónként és nagyságkategóriánként*

| Régió | *t, k* | Nagyságkategória | | | | | | | | | *Összesen* |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1. | 2. | 3. | 4. | 5. | 6. | 7.1 | 7.2 | 8. | |
| Közép-Magyarország | | | | | | | | | | | |
| Budapest | *t* | – | – | – | – | – | – | – | – | 1 | *1* |
| | *k* | – | – | – | – | – | – | – | – | 425 | *425* |
| Pest megye | *t* | – | 2 | 2 | 3 | 7 | 2 | 4 | 1 | 10 | *31* |
| | *k* | – | 6 | 8 | 15 | 35 | 12 | 36 | 10 | 55 | *177* |
| Közép-Dunántúl | *t* | 6 | 6 | 3 | 3 | 4 | 2 | 1 | – | 11 | *36* |
| | *k* | 14 | 18 | 12 | 15 | 20 | 12 | 9 | – | 90 | *190* |
| Nyugat-Dunántúl | *t* | 10 | 8 | 5 | 2 | 1 | 1 | 1 | – | 7 | *36* |
| | *k* | 21 | 24 | 20 | 10 | 5 | 6 | 9 | – | 82 | *177* |
| Dél-Dunántúl | *t* | 10 | 6 | 5 | 3 | 3 | – | 3 | – | 8 | *38* |
| | *k* | 21 | 18 | 20 | 15 | 15 | – | 18* | – | 74 | *181* |
| Észak-Magyarország | *t* | 8 | 9 | 7 | 6 | 4 | 1 | 2 | 1 | 8 | *46* |
| | *k* | 20 | 27 | 28 | 30 | 20 | 6 | 18 | 10 | 83 | *242* |
| Észak-Alföld | *t* | 2 | 6 | 6 | 6 | 8 | 4 | 3 | – | 9 | *44* |
| | *k* | 5 | 18 | 24 | 30 | 40 | 24 | 27 | – | 104 | *272* |
| Dél-Alföld | *t* | – | 3 | 4 | 5 | 7 | 6 | 2 | – | 15 | *42* |
| | *k* | – | 9 | 16 | 25 | 35 | 36 | 18 | – | 138 | *277* |
| Ország összesen | *t* | 36 | 40 | 32 | 28 | 34 | 16 | 16 | 2 | 69 | *274* |
| | *k* | 81 | 120 | 128 | 140 | 170 | 96 | 135 | 20 | 1051 | *1940* |

* Szervezési okok miatt Tolna megyében két településre lett elosztva a kilenc körzet.

Ha megyénként vetjük össze a 2002. év végéig működött minta megfelelő adatait az új mintáéval, kitűnik, hogy Nógrád megye kivételével az összes megyében – és Budapesten is – kisebb-nagyobb mértékben változtak a HKF-minta elemszámai. Tíz megyében csökkent a mintanagyság – legjelentősebben Jász-Nagykun-Szolnok megyében, ahol 612-ről 486-ra csökkent a mintába évenként kiválasztandó lakások száma –, nyolc megyében viszont nőtt a minta nagysága, Bács-Kiskun megyében például közel 10 százalékkal. Mivel Budapesten jelentősen nagyobb a minta, mint volt előzőleg, a megyékben összességében szükségszerűen kisebb lett. A megyei minták nagysága nemcsak a megye nagyságától, a megyében található lakott lakások számától függ, hanem a megye településszerkezetétől is, hiszen a kiválasztási arányok nem azonosak az egyes településnagyság-kategóriákban. A mintavételi terv értelmében a településeket az új mintába nagysággal arányos valószínűséggel választottuk ki megyénként és nagyságkategóriánként, a minta településszámainak megfelelően. A mintába nem kerülhettek be azok a kis, *1*., *2*. és *3*. nagyságkategóriájú települések, amelyek benne voltak az előző HKF-mintában, hiszen ezek címanyaga az évek során teljesen vagy erősen kimerült. A kiválasztás után a megyéknek lehetőségük volt esetenként cserét javasolni, de csak nagyságkategórián belül, elsősorban az összeírók szervezésére tekintettel, figyelembe véve a munkaerő-felvétel új mintájába kiválasztott kisebb települések szerkezetét. A nem jelentős számú javasolt cserék nagyobb részét elfogadtuk, és ennek megfelelően alakult ki a végleges településminta.

*A minta körzeteinek kiválasztása*

Mint a dolgozat elején leírt alapelvek *8*. pontjában már utaltunk rá, a mintába került településeken belül a kisebb körzeteket kiválasztás előtt oly módon vontuk össze, hogy az összevont körzetek legalább 40 lakást tartalmazzanak. Ahol lehetőség volt rá, azonos jellegű (belterületi, egyéb belterületi, illetve külterületi körzetek) és azonos településrészen levő körzeteket vontunk össze, de erre nem minden esetben volt lehetőség, így előfordult, hogy egy belterületi körzethez kellett csatolni egy vagy több külterületi körzetet. A körzetek kiválasztása is visszatevés nélküli nagysággal arányos valószínűségű mintavétellel történt. Azokon a településeken, amelyek mind a régi, mind az új HKF-mintában szerepelnek, a kiválasztható körzetek közül kimaradtak azok, amelyekben jelentős számban (tíz vagy annál több) voltak olyan lakások, amelyek az 1999 és 2002 évek HKF-mintájában szerepeltek (ugyanis lakásszinten nem lehetett azonosítani a régi mintában is szereplő mintaelemeket).

A körzetek előzetes rétegezéséhez nem álltak rendelkezésre megfelelő adatok azon kívül, hogy a körzetszám általában automatikus rétegezést jelentett, legalábbis a körzetek jellege szerint. Ennek megfelelően a nagysággal arányos valószínűségű kiválasztáshoz, a körzetek lakásszámát a körzetszám szerint sorba rendezett listán kumuláltuk. Az így végrehajtott kiválasztás többnyire biztosította, hogy a különböző jellegű körzetekből a mintába kerülő lakások jelleg szerinti megoszlása hasonló legyen az alapsokaságbeli megoszláshoz. Annak érdekében azonban, hogy a két megoszlás teljesen hasonló legyen, néhány esetben cserét kellet végrehajtani a különböző jellegű körzetek között. Összesen 27 körzetcserére került sor, ami 16 megyét érintett (Budapesten és három megyében nem volt szükség cserére).

Mivel a munkaerő-felmérés mintájának kiválasztása során nem került sor körzetek kiválasztására, előzetesen biztosítani kellett, hogy ha olyan település valamelyikében, amelyek mind a HKF, mind a MEF mintájában szerepelnek, mégis kimerül a HKF-minta valamelyik körzetének címanyaga, az ilyen körzetet pótolni lehessen más körzettel.

Ezért a két minta közös településein a HKF-mintához *tartalékkörzeteket* is kiválasztottunk, mégpedig 5-7 mintakörzetenként egy tartalékkörzetet. A tartalékkörzetek címanyaga, hasonlóan a mintakörzetekéhez, nem választhatók be a MEF-mintába, sem most, sem a következő tíz évben.

A 2001. évi népszámlálás végrehajtói a kiválasztáshoz nekünk átadott állományon időközben javították a címek rendeltetési kódját, ennek következtében akadtak olyan kiválasztott körzetek, amelyekben mégsem volt meg a határként megszabott 40 lakás. Emiatt 13 körzet cseréjére volt szükség.

## A HÁZTARTÁSOK ÉS A PÓTCÍMEK KIVÁLASZTÁSÁRA SZOLGÁLÓ PROGRAM

Az 1993 és 2002 közötti időszakban működő HKF-nél a mintába bekerülő elsődleges és pótcímeket az igazgatóságok választották ki egy központilag kidolgozott program segítségével. Az elsődleges címek esetében ez viszonylag egyszerű volt, hiszen minden körzetből hat elsődleges címet kellett kiválasztani egyenlő valószínűséggel. Az új HKF-mintára vonatkozó mintavételi terv szerint azonban a különböző háztartási rétegeknél el-

térő kiválasztási arányt kell alkalmazni, ami bonyolultabbá teszi a címek kiválasztására szolgáló programot. A program elkészítéséhez az alábbi, a minta háztartásrétegek elosztására szolgáló eljárást adtuk át az Informatikai főosztálynak.

*Nem önreprezentáló települések*

Első lépésben mindig a településszinten az *1*. háztartásrétegre (idős háztartásfőjű háztartások) jutó minta elemszámát kell meghatározni.

Ha $l = l^{(1)} + l^{(2)}$ jelöli a mintába került körzetekben az *1*., illetve *2*. rétegbe tartozó lakások összes számát és $k$ az adott településről a mintába került körzetek számát, akkor a településre vonatkozó átlagos kiválasztási arány:

$P = \frac{6k}{l}$ és így az *1*. rétegre vonatkozó kiválasztási arány $P_1 = \frac{1}{2}P = \frac{3k}{l}$. Ennek alapján a településről $n_1 = P_1 \cdot l$ lakás (egészre kerekítve) választandó ki az *1*. rétegből és $n_2 = 6k - n_1$ lakás a *2*. rétegből.

Az $n_1$ számú lakást a település körzetei között az *1*. rétegbe tartozó lakások számának arányában kell elosztani, azaz $n_{1j} = l_{1j}/l_1$, természetesen itt is egészre kerekítve. A kerekítések miatt előfordulhat, hogy az így meghatározott körzetenkénti elemszámok összege nem adja ki pontosan a már kiszámított $n_1$ települési elemszámot, ekkor a szokásos módon egy-egy körzetben növelni vagy csökkenteni kell az $n_{1j}$ elemszámot, hogy végeredményben a

$$\sum_{j=1}^{k} n_{1j} = n_1$$

egyenlőség pontosan fennálljon. Ez már biztosítja azt is, hogy $\sum_{j=1}^{k} n_{2j} = n_2$ legyen, azaz a településről $n_1 + n_2 = 6k$ elsődleges cím kerüljön a mintába.

*Önreprezentáló települések*

Az eljárás a három réteg miatt kissé bonyolultabb, s szélsőséges rétegstruktúra esetén előfordulhat, hogy a kívánt elemszámokat és ezek megfelelő körzetekre osztását csak iterálással lehet elérni.

Itt is alapelv, hogy minden körzetből hat elsődleges cím kerül kiválasztásra, és először a településre jutó rétegelemszámokat kell meghatározni, majd ezekhez ragaszkodva, elosztani a rétegek elemszámát a körzetekre. Első lépésben most is az *1*. háztartásréteg elemszámát kell meghatározni és elosztani a körzetekre az előző pontban leírt módon úgy, hogy a körzetelemszámok pontosan kiadják a településre meghatározott $n_1$ elemszámot. A következő lépésben a *3*. háztartásrétegre (fiatalabb, felsőfokú végzettségű háztartásfők) kell elvégezni a számításokat. Itt azonban már előfordulhat, hogy az *1*. rétegre történő elosztás és az $n_{1j}+n_{2j}+n_{3j} = 6$ megkötés miatt – és nem csak a kerekítések következtében – az első lépésben nem adják ki a körzetenkénti elemszámok a település egészére meghatározott $n_3$ elemszámot.

Ha

$$\sum_{j=1}^{k} n_{3j} < n_3 ,$$

akkor a hiányzó $n_3 - \sum_{j=1}^{k} n_{3j}$ elemet azokba a körzetekbe kell elosztani, ahol a legnagyobbak az $l_3$ lakásszámok, csökkentve egyúttal, ha kell, e körzetekben a már előzőleg meghatározott $n_{1j}$ elemszámokat. Ha már ily módon a körzetelemszámok összege mind az *1*., mind a *3*. rétegben pontosan kiadják az előzőleg meghatározott $n_1$ és $n_3$ település szintű elemszámokat, akkor a *2*. rétegre vonatkozó körzetelemszámok egyszerűen a $6 - n_{1j} - n_{3j}$ összefüggés alapján adódnak. Az ily módon meghatározott $n_{2j}$ elemszámok természetesen nem feltétlenül lesznek pontosan arányosak a *2*. réteg $l_{2j}$ lakásszámaival.

Példaképpen bemutatjuk a leírt eljárást Nyíregyháza esetében. Itt $k = 21$, a lakás rétegszámok a három rétegben a mintába került körzetekben:
$l_1 = 467$, $l_3 = 370$ és $l_2 = 1167$, a 21 körzet összesen 2004 lakást tartalmaz. Ennek megfelelően az átlagos kiválasztási arány $P = 126/2004 = 0{,}062874$. Így $P_1 = 1/2\,P = 0{,}031437$ és $P_3 = 2P = 0{,}12575$. Következésképpen az *1*. rétegből $n_1 = P_1 l_1 = 0{,}031437 \cdot 467 = 15$ cím választandó ki összesen, a *3*. rétegből pedig $n_3 = P_3 l_3 = 0{,}12575 \cdot 370 = 47$ cím. A *2*. rétegből kiválasztandó címek száma kivonással adódik:

$$n_2 = n - n_1 - n_3 = 126 - 15 - 47 = 64.$$

Az előbbi rétegminta-elemszámok elosztását a körzetekre a következő séma szemlélteti.

| Kiválasztott körzetek száma | Elsődleges elosztás: $n_{1j} = \frac{n_1 l_{1j}}{l_1}$ | $\sum n_{1j} = 15$ feltétel biztosítása | A *2*. iterációs lépés a *3*. réteg elemszámának biztosításához | Elsődleges elosztás $n_j = 6$ feltétel mellett | $\sum n_{3j} = 47$ feltétel biztosítása | $n_{2j} = 6 - n_{1j} - n_{3j}$ |
|---|---|---|---|---|---|---|
| 0042 | 0,578 ~ 1 | 1 | 1 | 6 – 1 = 5 | 5 | 6 – 6 = 0 |
| 0152 | 0,578 ~ 1 | 1 | 0 | 6 – 1 = 5 | 6 | 6 – 6 = 0 |
| 0062 | 0,867 ~ 1 | 1 | 1 | 6 – 1 = 5 | 5 | 6 – 6 = 0 |
| 0010 | 1,028 ~ 1 | 1 | 1 | 6 – 1 = 5 | 5 | 6 – 6 = 0 |
| 0186 | 0,514 ~ 1 | 0 | 0 | 2,668 ~ 3 | 3 | 6 – 3 = 3 |
| 0262 | 0,578 ~ 1 | 1 | 1 | 3,430 ~ 3 | 4 | 6 – 5 = 1 |
| 0349 | 0,546 ~ 1 | 1 | 1 | 2,922 ~ 3 | 3 | 6 – 4 = 2 |
| 0027 | 1,349 ~ 1 | 1 | 1 | 1,778 ~ 2 | 2 | 6 – 3 = 3 |
| 0081 | 0,645 ~ 1 | 1 | 1 | 1,524 ~ 2 | 2 | 6 – 3 = 3 |
| 0168 | 0,964 ~ 1 | 1 | 1 | 1,524 ~ 2 | 2 | 6 – 3 = 3 |
| 0224 | 0,899 ~ 1 | 1 | 1 | 1,651 ~ 2 | 2 | 6 – 3 = 3 |
| 0370 | 0,771 ~ 1 | 1 | 1 | 1,651 ~ 2 | 2 | 6 – 3 = 3 |
| 0117 | 1,188 ~ 1 | 1 | 1 | 1,397 ~ 1 | 2 | 6 – 3 = 3 |
| 0280 | 0,771 ~ 1 | 1 | 1 | 0,635 ~ 1 | 1 | 6 – 2 = 4 |
| 0419 | 0,610 ~ 1 | 1 | 1 | 0,889 ~ 1 | 1 | 6 – 2 = 4 |
| 0306 | 1,285 ~ 1 | 1 | 1 | 0,254 ~ 0 | 0 | 6 – 1 = 5 |
| 0524 | 0,546 ~ 1 | 0 | 1 | 0,127 ~ 0 | 0 | 6 – 1 = 5 |
| 0135 | 0,385 ~ 0 | 0 | 0 | 0,889 ~ 1 | 1 | 6 – 1 = 5 |
| 0241 | 0,482 ~ 0 | 0 | 0 | 0 | 0 | 6 |
| 0395 | 0,289 ~ 0 | 0 | 0 | 0,381 ~ 0 | 1 | 6 – 1 = 5 |
| 0478 | 0,096 ~ 0 | 0 | 0 | 0,381 ~ 0 | 0 | 6 |
| *Összesen* | *17* | *15* | *15* | *43* | *47* | *64* |

A pótcímek kiválasztása lényegében ugyanúgy történik, mint az előző HKF-minta esetén, azaz a program figyelembe veszi a kiválasztott elsődleges címnek a háztartási rétegen túlmenő további jellemzőit (háztartásfő aktivitása, korcsoportja, iskolai végzettsége, a háztartás taglétszáma) és pótcímként egy vagy két ugyanilyen típusú háztartást választ ki, de nem feltétlenül ugyanabból a körzetből, hanem az adott összeíróhoz tartozó összes körzet címanyagából.

## AZ ADATOK FELDOLGOZÁSÁNÁL ALKALMAZANDÓ ELSŐDLEGES SÚLYOK SZÁMÍTÁSA

Olyan mintavételi adatok esetén, ahol az egyes minta elemek kiválasztási valószínűsége nem egyenlő, az ebből eredő potenciális torzítás ellensúlyozására az adatok feldolgozásánál mindig súlyozást kell alkalmazni. Az új HKF-minta kiválasztása során az egyes településnagyság-kategóriákból, illetve Budapest kerületeiből, valamint a *2*., illetve *3*. háztartásrétegből nem egyforma valószínűséggel kerültek be a lakások a mintába. Ezen felül, a szükséges kerekítések miatt, a tényleges kiválasztási arányok egy-egy kategórián belül megyék között is különbözhettek bizonyos mértékig. Az *i*-edik nagyságkategóriában és a *j*-edik megyében a $p_{ij}$ kiválasztási arányt a *megvalósult* mintában lévő $l_{ij}$ *l*akásszámnak és a $L_{ij}$ teljes körű lakásszámnak a hányadosa mutatja, azaz

$$p_{ij} = \frac{l_{ij}}{L_{ij}} \quad (i = 1,\ldots,8) \quad \text{és} \quad (j = 1,\ldots,20)$$

Megjegyzendő, hogy $j = 1$, azaz Budapest esetében egy harmadik, *k* index alkalmazása is indokolt, ami a kerületenkénti kiválasztási arányokat mutatja.

A minta kiválasztása során a megyénként (fővárosi kerületenként) és településnagyság-kategóriánként különböző kiválasztási arányokon kívül háztartási rétegenként is differenciáltuk a kiválasztási valószínűségeket.

Ha $L_{ij}^{(1)}$, illetve $l_{ij}^{(1)}$ jelöli az *i*-edik nem önreprezentáló településnagyság-kategóriában ($i = 1,\ldots,7$) a *j*-edik megye ($j = 2,\ldots,20$) *1*. háztartásrétegbe tartozó lakások (háztartások) számát az alapsokaságban, illetve a megvalósult mintában, akkor az e rétegre vonatkozó tényleges kiválasztási valószínűség

$$p_{ij}^{(1)} = \frac{l_{ij}^{(1)}}{L_{ij}^{(1)}}, \quad \text{a } 2\text{. rétegre pedig} \quad p_{ij}^{(2)} = \frac{l_{ij} - l_{ij}^{(1)}}{L_{ij} - L_{ij}^{(1)}}.$$

Az önreprezentáló települések esetén a három háztartási rétegre vonatkozó tényleges kiválasztási valószínűségek

$$p_{8j}^{(1)} = \frac{l_{8j}^{(1)}}{L_{8j}^{(1)}}, \quad p_{8j}^{(3)} = \frac{l_{8j}^{(3)}}{L_{8j}^{(3)}} \quad \text{és} \quad p_{8j}^{(2)} = \frac{l_{8j} - l_{8j}^{(1)} - l_{8j}^{(3)}}{L_{8j} - L_{8j}^{(1)} - L_{8j}^{(3)}}.$$

Érdemes itt is utalni arra, hogy a fenti tényleges kiválasztási valószínűségek reciprokaiként számított elsődleges súlyok természetesen közelítő értékek, de az így adódó torzítás elhanyagolható a nagy arányú nemválaszolásból eredő torzítás mellett.

Ahhoz, hogy az $s_{ij}^{(h)} = \frac{1}{p_{ij}^{(h)}} = \frac{L_{ij}^{(h)}}{l_{ij}^{(h)}}$ súlyokat alkalmazni lehessen, a népszámlálás továbbvezetett lakásállományából először meg kellett határozni a $L_{ij}^{(h)}$ lakásszámokat ($i = 1,\ldots,8$, $j = 1,\ldots,20$, $h = 1,2,3$). A nem önreprezentáló települések esetén előfordulhat, hogy egy adott nagyságkategóriában az alapsokaságban van település, a mintában azonban nincs. Ilyen esetben az alapsokaságnál össze kellett vonni két szomszédos nagyságkategóriát, és $L_{ij}$ ekkor az összevont kategóriák együttes lakásszámát jelöli. Az $l_{ij}^{(h)}$ mennyiségek természetesen a megvalósult mintából adódnak.

A megyékre és településnagyság-kategóriákon belüli háztartásrétegekre vonatkozó lakás-, illetve háztartásszám-adatok közül az *1*. réteg – azok a háztartások, ahol a háztartásfő 60 éves és idősebb – meghatározása természetesen a kor továbbvezetése alapján történhet, a *3*. réteg – azok a háztartások, ahol a háztartásfő 60 évesnél fiatalabb és felsőfokú végzettségű – esetén viszont az iskolai végzettségre vonatkozó adat a 2001. február 1-jei állapotot tükrözi, mivel a népszámlálás óta eltelt időszakban feltehetőleg nem következett be számottevő változás a fiatalabb felsőfokú végzettségű háztartásfők arányában, illetve a változások egy részét a kor továbbvezetése automatikusan követi. A későbbiekben azonban e feltevés egyre kevésbé lesz tartható, ezért $L_{8j}^{(3)}$ becsléséhez majd a kétévenkénti (először 2004 őszén végrehajtásra kerülő) címbejárás adatai alapján kapott dinamikát kell felhasználni, azzal korrigálni a népszámlálás időpontjára vonatkozó, illetve a kor továbbvezetésével adódó $L_{8j}^{(3)}$ értékeket.

Különösen a *negyedéves* HKF-adatok feldolgozásánál előfordulhat, hogy egyes, más rétegekétől eltérő kiválasztási arányú rétegekben (például egyes budapesti kerületekben) olyan kevés a felvételben közreműködő háztartások száma, hogy ezek adatait csak más, hasonló rétegek (például hasonló jellegű kerületek) adataival összevontan lehet súlyozni. (Kevésnek tekintjük egy adott rétegben a megvalósult minta elemszámát, ha az kevesebb tíznél.)

Bár azáltal, hogy az elsődleges súlyok számításához a *megvalósult* minta adatait használjuk, részben ellensúlyozzuk a nemválaszolásból eredő lehetséges torzításokat, ez nem teszi feleslegessé az eddig is alkalmazott *kalibrálási* számításokat. Célszerű, ha az adatok feldolgozása során továbbra is rendszeresen sor kerül kalibrálásra. Az, hogy a kalibráláshoz milyen külső, a HKF-mintától független adatok kerüljenek felhasználásra, részben az elérhető külső adatforrásoktól, részben a HKF végrehajtását irányító szakemberek megítélésétől függ. A továbbvezetett nemenkénti kormegoszlás lehet ilyen külső adatforrás, de ez a HKF esetén kevésbé hatékony. Lassan változó mutatók esetén bizonyos ideig lehet a népszámlálás adatait használni.

## ROTÁCIÓS ELJÁRÁS

Az új HKF-minta kiválasztása során a háztartási rétegeknél alkalmazott eltérő kiválasztási arány következtében az évenkénti rotációs eljárás valamivel bonyolultabb lesz,

mint a régi mintánál volt. Mivel 2003-ban nem volt címbejárás, így az egyes háztartási rétegek létszáma 2004-re lényegében csak a korosodás következtében változhatott. Az *1*. háztartási rétegbe azok a háztartások fognak 2004-ben tartozni, ahol a háztartásfő 1944-ben vagy korábban született, s ennek megfelelően változik a *2*., illetve *3*. háztartásréteg definíciója is. Ezen felül, esetenként azért is történhetett változás a rétegek létszámában, mert pótcímigénybevétele esetén, ha a címen lakó háztartásréteg besorolása a népszámlálás óta megváltozott, de egyébként késznek mutatkozott a felvételben való közreműködésre, nem kellett ragaszkodni az eredeti réteghez. A közreműködő háztartások a tényleges helyzetnek megfelelő rétegkódot kaptak.

Az alapelv a rotációnál továbbra is az, hogy az elsődleges címek száma a rotáció után minden körzetben 6 legyen.

*Nem önreprezentáló települések*

Legyen egy adott településen az elsődleges címek száma 2003-ban az *1*. háztartásrétegből $n_1(03)$ a *2*. háztartásrétegből pedig $n_2(03)$.

A *megvalósult* minta elemszámai legyenek $n_1'(03)$ és $n_2'(03)$

$$n_1'(03)+n_2'(03)=n'(03)\le n=6k\ ,$$

ahol $k$ az adott településről a mintába választott körzetek száma.

*1. lépés*: a rétegek létszámának esetleges változásait figyelembe véve kiszámolandók a címkiválasztási program alapján az $n_1(04)$ és $n_2(04)$ rétegelemszámok és ezek elosztása a körzetekre, azaz $n_{1i}(04)$ és $n_{2i}(04)$. Mivel a nem válaszoló háztartások pótlása nem szükségszerűen ugyanabból a körzetből, hanem egy adott összeíróhoz tartozó körzetek összességéből történt, egy körzetben a felvételben közreműködő háztartások száma kisebb is, nagyobb is lehet 6-nál. Tekintsük a következő eseteket.

*a*) Az *i*-edik körzetben $n_{1i}'(03)+n_{2i}'(03)\le 4$ ;

a körzetből nem kell háztartást kirotálni és annyi új háztartást kell kiválasztani a 2003. évi minta kiválasztásához leírt eljárással, hogy az elsődleges címek száma pontosan 6 legyen. Az egyes rétegekbe kiválasztandó címek számának meghatározásánál figyelembe kell venni az $n_{1i}(04)-n_{1i}'(03)$, illetve az $n_{2i}(04)-n_{2i}'(03)$ különbségeket, mivel a kijelölt címek és a megvalósult felvételek aránya eltérő lehet a két rétegben.

*b*) Az *i*-edik körzetben $n_{1i}'(03)+n_{2i}'(03)=5$ ;

a körzetből egy háztartás rotálandó ki abból a rétegből, ahol nagyobb volt 2003-ban a közreműködési készség. Kiválasztandó két új háztartás a 2003. évi minta kiválasztásához leírt eljárásnak megfelelően.

*c*) Az *i*-edik körzetben $n_{1i}'(03)+n_{2i}'(03)=6$ ;

a körzetből két háztartás rotálandó ki és két új háztartás választandó ki a két háztartásréteg körzeten belüli arányának megfelelően, azaz olyan módon, hogy

$$\frac{n_{2i}(03)}{n_{1i}(03)}=\frac{n_{2i}(04)}{n_{1i}(04)} \qquad /1/$$

legyen. Ez alól csak az jelentene kivételt, ha 2004-ben a 60 évesek száma az adott településen számottevően eltérne a 60 évesek számától 2003-ban, de ennek kicsi a valószínűsége.

*d*) Az *i*-edik körzetben $n'_{1i}(03)+n'_{2i}(03)\geq 7$ ;

a körzetből három vagy több háztartás rotálandó ki oly módon, hogy rotálás után négy háztartás maradjon a mintában és két új háztartást kell kiválasztani úgy, hogy /1/ teljesüljön.

*2. lépés*: ellenőrzendő hogy a fenti módon végrehajtott rotáció eredményeképp az adott településen az *1*., illetve *2*. háztartásrétegből kiválasztott elsődleges címek száma valóban $n_1(04)$, illetve $n_2(04)$. Ha eltérés mutatkozik, akkor vagy a kirotálandó vagy az újonnan kiválasztandó háztartások közül kell a réteg hovatartozást olyan irányban változtatni, hogy az elsődleges címek száma az *1*. háztartásrétegben pontosan $n_1(04)$ legyen, a *2*. rétegben pedig pontosan $n_2(04)$.

*Önreprezentáló települések*

Jelölje $n_1(03)$ egy adott önreprezentáló településen az *1*. háztartásrétegből a 2003. évi mintába kiválasztott elsődleges címek számát, $n_2(03)$ ugyanezt a *2*. háztartásrétegből, $n_3(03)$ pedig a *3*. rétegből. A megvalósult minta elemszámai az egyes háztartásrétegekből $n_1'(03)$, $n_2'(03)$ és $n_3'(03)$.

$$n'_1(03)+n'_2(03)+n'_3(03)=n'(03)\leq n=6k\ .$$

*1. lépés*: a rétegek létszámának (például az öregedésből fakadó) változását figyelembe véve kiszámolandók a program alapján $n_1(04)$, $n_2(04)$ és $n_3(04)$ és ezek elosztása a körzetekre, azaz $n_{1i}(04)$, $n_{2i}(04)$ és $n_{3i}(04)$, ahol $i = 1,\ldots,k$. Tekintsük az alábbi eseteket!

*a*) Az *i*-edik körzetben $n'_{1i}(03)+n'_{2i}(03)+n'_{3i}(03)\leq 4$ ;

a körzetből nem kell háztartást kirotálni, két (vagy több) új háztartás választandó a mintába 2003. évi minta kiválasztásához leírt eljárással oly módon, hogy az elsődleges címek száma pontosan hat legyen. Az egyes rétegekbe kiválasztásra kerülő címek számának meghatározásánál figyelembe kell venni az $n_{1i}(04) - n_{1i}'(03)$, az $n_{2i}(04) - n_{2i}'(03)$ és az $n_{3i}(04) - n_{3i}'(03)$ különbségeket, mivel a válaszadási arány eltérő lehet az egyes rétegekben.

*b*) Az *i*-edik körzetben $n'_{1i}(03)+n'_{2i}(03)+n'_{3i}(03)=5$ ;

a körzetből egy háztartás rotálandó ki, mégpedig abból a rétegből, ahol a legnagyobb volt 2003-ban a közreműködési készség. Kiválasztandó két új háztartás a 2003. évi mintakiválasztásához leírt eljárásnak megfelelően.

*c*) Az *i*-edik körzetben $n'_{1i}(03)+n'_{2i}(03)+n'_{3i}(03)=6$ ;

a körzetből két háztartás rotálandó ki és két új háztartás választandó ki a háztartásrétegek körzeten belüli arányának megfelelően, azaz oly módon, hogy

$$\frac{n_{2i}(03)}{n_{1i}(03)}\cong\frac{n_{2i}(04)}{n_{1i}(04)} \quad \text{és} \quad \frac{n_{3i}(03)}{n_{1i}(03)}\cong\frac{n_{3i}(04)}{n_{1i}(04)} \qquad /2/$$

legyen. Ha 2004-ben az adott település kiválasztott körzeteiben a 60 évesek száma számottevően eltérne a 60 évesek 2003. évi számától – ennek kicsi a valószínűsége – akkor persze a /2/ összefüggéseknek nem kell fennállniuk.

*d*) az *i*-edik körzetben $n'_{1i}(03)+n'_{2i}(03)+n'_{3i}(03)\geq 7$;

a körzetből három vagy több háztartás rotálandó ki oly módon, hogy rotálás után négy 2003-ban közreműködő háztartás maradjon a mintában és két új háztartást kell kiválasztani úgy, hogy /2/ teljesüljön.

*2. lépés*: itt is ellenőrizni kell, hogy a fenti módon végrehajtott rotáció eredményeképp, az adott településen az *1*., *2*., illetve *3*. háztartásrétegből kiválasztott elsődleges címek száma valóban $n_1(04)$, $n_2(04)$, illetve $n_3(04)$. Ha eltérés mutatkozik valamelyik rétegnél, akkor vagy a kirotálandó vagy az újonnan kiválasztandó háztartások közül kell a réteghovatartozást olyan irányban változtatni, hogy az elsődleges címek száma mindhárom rétegben pontosan az előzetesen megadott $n_1(04)$, $n_2(04)$, illetve $n_3(04)$ legyen.

## TAPASZTALATOK AZ ÚJ MINTÁRÓL 2003-BAN

Az 2003. évi új mintában csak olyan címeket választottak ki, amelyek a 2001. évi népszámláláskor *lakott* lakások voltak. Ez azért is volt követelmény, hogy a háztartásrétegekre vonatkozó eltérő kiválasztási arányt meg lehessen valósítani. Ennek ellenére a felvétel során felkeresett 20 361 elsődleges és pótcím 5,2 százalékát (több mint ezer címet) üres lakásként regisztráltak az összeírók. Egyébként az üres lakásként jelölt lakások aránya megyénként, illetve régiónként viszonylag szűk határok között mozgott. Legalacsonyabb (4,1%) volt Észak-Magyarországon, legmagasabb (6,7%) pedig Nyugat-Dunántúlon. Természetesen a népszámlálás és a 2003. évi felvétel között eltelt átlagosan két és fél év alatt sok lakás válhatott üressé, de felvételi hibák is közrejátszhattak az üres lakások elég magas arányában. Mindenesetre, amikor a közreműködési, illetve meghiúsulási arányokat vizsgáljuk, a felkeresett lakott lakások számához kell viszonytanunk, hiszen az üres lakások nem tartoznak a HKF alapsokaságához.

2003-ban 19 285 lakott lakást kellett az összeíróknak felkeresniük ahhoz, hogy a 11 646 elsődleges címen és a pótcímeken lakó háztartásokat rávegyék a felvételben való közreműködésre. Végül azonban csak 8335 háztartás vállalta, hogy adatokat szolgáltat bevételeiről és kiadásairól. Ez országosan 43,2 százalékos közreműködési arányt jelent, lényegesen kisebbet, mint az előző években. A közreműködési készség csökkenéséhez alapvetően két tényező járult hozzá.

*1*. 2003-tól a KSH még az eddigi nagyon szerény összeggel sem honorálja a háztartásoknak azt a munkát és időráfordítást, amit a bevételek és kiadások naponkénti részletes – mennyiségi adatokat is tartalmazó – feljegyzése igénybe vesz egy hónapon keresztül. Bár az eddig adott nettó összeg sem jelentett a háztartásoknak számottevő bevételi forrást, mégis munkájuk megbecsülésének értékelhették a kapott szerény összeget. Az a tény, hogy a hivatal, pénzügyi megszorítások miatt, 2003-tól még ezzel a szerény összeggel sem méltányolja fontos információforrást nyújtó munkájukat, nyilván számos háztartás esetén meggyengítette a közreműködési készséget.

*2*. A háztartásrétegenként eltérő kiválasztási arány szintén a közreműködési arány csökkenését hozta magával. Az eddigi tapasztalatok szerint a mintában ugyanis az arányosnál kisebb számban szerepelnek a közreműködésre készségesebb idősebb háztartások, viszont nagyobb arányban – legalábbis az önreprezentáló településeken – az általában kevésbé készséges felsőfokú végzettségű személyek háztartásai. Ezáltal a minta összetétele

ugyan jobban közelíti a valóságos arányokat, a közreműködő háztartások száma azonban kisebb lett, mint volt az előző években.

A közel 57 százaléknyi meghiúsulás nagyobb része (31%) megtagadás, de elég jelentős (26%) az egyéb okból történt meghiúsulás is. Ez többnyire azért következett be, mert bár a kiválasztott cím lakott lakás volt, az összeírónak nem sikerült kapcsolatot teremtenie az ott lakó háztartással, vagy nem volt olyan személy, aki a naplóvezetés feladatát el tudta volna látni.

Ha területi bontásban (régiók, illetve megyék szerint) vizsgáljuk a 2003. évi közreműködési és meghiúsulási arányokat, elég jelentős különbségeket találhatunk. Az eddigi tapasztalatokkal egyezően a fővárosban a legkisebb a háztartások készsége a felvételben való közreműködésre, az ilyen háztartások aránya kisebb 30 százaléknál. A megtagadási arány viszont itt a leggyakoribb, ez 2003-ban 38 százalék felett volt. A felvétel szempontjából legsikeresebbnek az Észak-Magyarországhoz, illetve Dél-Alföldhöz tartozó megyék tekinthetők, a közreműködési arány mindkét régióban bőven 50 százalék felett volt, megtagadás ugyanakkor a felkeresett lakott lakások mindössze 20, illetve 24 százalékában fordult elő. Az alábbi, 4. tábla régiónként mutatja a 2003. évi HKF-felvételnél tapasztalt közreműködési és meghiúsulási arányokat. Budapest speciális helyzete indokolja, hogy Közép-Magyarország esetén a főváros és Pest megye adatait külön is bemutassuk.

4. tábla

*A 2003. évi HKF-minta néhány jellemzője*

| Régió | Felkeresett lakott lakások száma | Közreműködési | Megtagadási | Egyéb meghiúsulási |
|---|---|---|---|---|
| | | arány (százalék) | | |
| Budapest | 5 421 | 28,4 | 38,3 | 33,3 |
| Pest megye | 1 522 | 45,4 | 35,2 | 19,4 |
| Közép-Magyarország | 6 943 | 32,1 | 37,6 | 30,3 |
| Közép-Dunántúl | 1 796 | 45,4 | 28,9 | 25,7 |
| Nyugat-Dunántúl | 1 660 | 48,4 | 27,0 | 24,6 |
| Dél-Dunántúl | 1 699 | 43,8 | 31,3 | 24,9 |
| Észak-Magyarország | 2 088 | 55,6 | 20,4 | 24,0 |
| Észak-Alföld | 2 576 | 47,5 | 30,0 | 22,5 |
| Dél-Alföld | 2 523 | 53,5 | 24,0 | 22,5 |
| Ország összesen | 19 285 | 43,2 | 30,7 | 26,1 |

Korábban is köztudott volt, hogy a területi különbségek mellett a háztartások *foglalkozási, demográfiai, iskolázottsági jellemzői* nagymértékben befolyásolják készségüket a HKF-ben való közreműködésre, hiszen ez volt pl. fő indoka annak, hogy az új mintánál a kiválasztási arányt differenciáltuk a háztartási rétegek szerint. A vállalkozók háztartásai például az átlagosnál számottevően nagyobb arányban tagadják meg a felvételben való közreműködést, de e tekintetben is jelentős területi különbségek vannak. Érdekes módon nem is a fővárosban, hanem Pest megyében a legmagasabb a közreműködést elutasító vállalkozó háztartások aránya, míg Budapesten „csak” 40 százalék körül van az arányuk. Ugyanakkor Észak-Magyarország megyéiben a vállalkozók háztartásai közül csak min-

den negyedik tagadta meg a közreműködést, a régión belül is Borsod-Abaúj-Zemplén megyében volt a legkisebb ez az arány, kevesebb, mint 19 százalék. A megtagadók aránya 32 százalék volt azon háztartások körében, ahol a háztartásfő alkalmazásban álló, az inaktívak háztartásainál pedig 27 százalék. Ez utóbbi két rétegen belül is Észak-Magyarország megyéiben a legalacsonyabb a megtagadók aránya a legmagasabb pedig Budapesten. Megjegyzendő ugyanakkor, hogy a fővárosban alig van különbség a két réteg között a megtagadási arányban, mindkettő 37–38 százalék körüli.

*Korcsoportonként* vizsgálva a megtagadási arányt, ez – egy kivétellel – a régiókban is, a megyékben is a középkorú háztartásfők körében a legmagasabb, és országosan a fiatal (30 évesnél fiatalabb) háztartásfők esetében a legalacsonyabb. Budapesten és Közép-Dunántúlon a háztartásfő kora alig befolyásolja a megtagadási arányt, több régióban pedig – különösen Dél Dunántúlon és Pest megyében – az idősek körében lényegesen alacsonyabb a felvételben közreműködni nem hajlandók aránya, mint a fiatalok háztartásaiban.

*Az iskolai végzettség* lényegében csak a felsőfokú végzettségű háztartásfők esetében befolyásolja negatívan a közreműködési készséget, az alap- és középfokú végzettségű háztartásfők között nincs jelentős különbség a megtagadási arányban.

Érdekes módon, a háztartás nagysága másképpen befolyásolja a közreműködési, mint a megtagadási arányt. A közreműködési készség a nagy létszámú (öt- és többtagú) háztartásokban a legnagyobb (országosan közel 50 százalék), ez a régiók és a megyék túlnyomó többségében is így van. Általában az 1-2 tagú háztartásokban legkisebb a közreműködési arány. Ugyanakkor a felvételben való közreműködés határozott elutasítása országosan és a régiók túlnyomó többségében is a kis létszámú háztartásokban a legritkább. Ez azzal magyarázható, hogy a kicsi (főleg az egytagú) háztartásokban fordul elő leggyakrabban az ún. egyéb meghiúsulás, ha a kérdőbiztos többszöri kísérlet során sem találja otthon a lakó(ka)t. A háztartás taglétszáma alapján képzett mindhárom csoporton belül Észak-Magyarországon és a Dél-Alföldön a legalacsonyabb a megtagadás előfordulása, a nagy létszámú háztartásokban csak 17 és 22 százalék.

*

A tanulmányban csupán ízelítőt próbáltam adni azokból a tényezőkből, amelyek a HKF esetén befolyásolják (pontosabban 2003-ban befolyásolták) a háztartások magatartását a felvételben való közreműködés során. A jelenség természetesen mélyebben, további szempontok szerint is vizsgálható lenne. A minta struktúrája nem változik 2004-ben és a következő években, mindazonáltal remélhető, hogy a kérdőbiztosok nagyobb gyakorlottsága és a háztartások meggyőzésére fordított nagyobb erőfeszítése révén 2004-ben és a következő években a mintába került háztartások nagyobb arányban fogják vállalni a közreműködést a HKF-felvételben.

IRODALOM

BENE L. – ÉLTETŐ Ö. [1972]: Az általános célú háztartási minta kialakítása. *Statisztikai Szemle.* 50. évf. 10. sz. 979–992. old.

ÉLTETŐ Ö. – MESZÉNA GY. – ZIERMANN M. [1982]:*Sztochasztikus módszerek és modellek.* Közgazdasági és Jogi Könyvkiadó. Budapest.

ÉLTETŐ Ö. [1987]: Az ELAR-minta és az 1984. évi mikrocenzus mintájának kiválasztási eljárása. *Statisztikai Módszertani Füzetek 24.* Központi Statisztikai Hivatal. Budapest.

ÉLTETŐ Ö. – MIHÁLYFFY L. [2002]: Household Surveys in Hungary. *Statistics in Transition.* 5. évf. 4. sz. 521–540. old.

SUMMARY

The paper presents the theoretical considerations and the from several aspects new methods applied in the course of selecting the new sample of the Household Budget Survey (HBS). The new sample based on the data of the 2001 Census was introduced in 2003. Moreover, some results of the HBS in 2003 in view of the participation and the non-response of the sampled households are also discussed.

# A KISSZERVEZETEK INTEGRÁLT REPREZENTATÍV ÉVKÖZI MEGFIGYELÉSE A 2000-ES ÉVEKBEN

DR. TELEGDI LÁSZLÓ

A szerző ismerteti a kisszervezetek integrált reprezentatív évközi megfigyelését a 2000-es években. A dolgozat foglalkozik a megfigyelés jellemzőivel, a rétegzéssel és a minta kiválasztásával. Tárgyalja az adatgyűjtést és a felhasznált becslési módszereket.

TÁRGYSZÓ: Reprezentatív megfigyelés. Mintavétel. Becslés. Kisszervezet.

A magyar gazdaság egyik fő jellemzője, hogy a különböző nemzetgazdasági ágakban sok kis gazdasági szervezet van, tevékenységük a termelésben számos területen ugyan nem meghatározó, de szinte mindenhol jelentős. A gazdaságstatisztikai megfigyelési rendszer is tükrözi ezt: a gazdaságstatisztikai megfigyelések ezekre a szervezetekre is kiterjednek, a gazdaság fejlődését kifejező statisztikai információk magukban foglalják ezek adatait is. Ily módon a gazdaságstatisztikai megfigyelések nagyszámú, jórészt kis gazdasági szervezetet (vállalkozást, költségvetési és társadalombiztosítási szervezetet, valamint nonprofit szervezetet) ölelnek fel. Ezek nagy többsége vállalkozás: egy részük jogi személyiségű, egy részük anélküli. A jogi személyiség nélküli vállalkozások között vannak jogi személyiség nélküli gazdasági társaságok, jogi személyiség nélküli egyéb vállalkozások és egyéni vállalkozások.

A kis- és középvállalkozásokról, fejlődésük támogatásáról szóló törvény meghatározása szerint kisvállalkozásnak minősül az olyan vállalkozás,

*1*. amelynek összes foglalkoztatotti létszáma 50 főnél kevesebb,

*2*. amelynek éves nettó árbevétele legfeljebb 700 millió forint, vagy mérlegfőösszege legfeljebb 500 millió forint, továbbá

*3*. amelyben az állam, az önkormányzat vagy a nem kisvállalkozásnak minősülő vállalkozások tulajdoni részesedése – tőke vagy szavazati jog alapján – külön-külön és együttesen sem haladja meg a 25 százalékot.

(Az olyan kisvállalkozás, amelynek összes foglalkoztatotti létszáma tíz főnél kevesebb, mikrovállalkozásnak minősül.) A nemzetközi statisztikai gyakorlatban ugyanakkor árbevételtől, mérlegfőösszegtől és tulajdoni részesedéstől függetlenül az 50 főnél kisebb létszámú gazdasági szervezetek tartoznak a kisszervezetek közé (az öt főnél kisebb létszámú kisszervezeteket mikroszervezeteknek is nevezzük).

A gazdaságstatisztikai megfigyelések többnyire ugyan részben teljes körűek, részben azonban mintavételen, mintakiválasztáson alapulnak, reprezentatívak. (A reprezentatív

jelzőt itt tágabb értelemben, a mintavételen alapuló kifejezés szinonimájaként használom, *Telegdi* [2001]-ben foglalkozom a szűkebb értelemben vett reprezentativitással.) A reprezentatív megfigyelések már jó ideje a gazdaságstatisztika lényeges elemei.

A reprezentatív megfigyeléseknek négy összetevője van. Az első maga a *mintavétel*, azaz amikor a minta elemeit kiválasztjuk a sokaságból. A reprezentatív megfigyelés második része az *adatgyűjtés*, amikor összegyűjtjük a mintaelemekre vonatkozó adatokat (a teljes körű megfigyelésekhez hasonlóan). A harmadik összetevő a *becslés*, amelynek során következtetéseket vonunk le a mintából mint részből a sokaságra mint egészre. A reprezentatív megfigyelés negyedik, utolsó összetevője a *becslés helyességének vizsgálata*. Ennek célja egyrészt a becslés helyességének (pontosságának és megbízhatóságának) jellemzése – vagyis a hibaszámítás – és figyelemmel kísérése, másrészt a hibát előidéző legfontosabb tényezők feltárása, lehetőség szerinti kiküszöbölése, illetve hatásuk csökkentése és ezáltal a becslés helyességének fokozása a becslés módosításával.

A gazdaságstatisztikai megfigyelések között gyakoriságuk szerint vannak

– ismétlődő (az év folyamán havonta vagy negyedévente ismétlődő, folyamatos, panel jellegű *évközi* és évente, többévente vagy szabálytalanul ismétlődő), valamint
– eseti (egy alkalommal végzett)

megfigyelések. Jelen dolgozat a kisszervezetek reprezentatív évközi megfigyelését ismerteti a 2000-es években (előzményét illetően lásd *Éltető–Marton–Mihályffy–Telegdi* [1997], *Telegdi* [2001]).

## 1. A MEGFIGYELÉS JELLEMZŐI

A kisszervezetek reprezentatív évközi ipar- és építőipar-statisztikai megfigyelését a Központi Statisztikai Hivatal (KSH) 2000-ben a Gazdasági tevékenységek egységes ágazati osztályozási rendszere, 1998 (TEÁOR '98) szerint meghatározott *C* Bányászat, *D* Feldolgozóipar és *E* Villamosenergia-, gáz-, gőz-, vízellátás iparágakban, illetve az *F* Építőipar gazdasági ágban végezte. Az adatgyűjtésre havi gyakorisággal került sor. A kisszervezeteknek az ún. általános modul részeként történő reprezentatív évközi megfigyelése 2000-ben az *A–K* és *M–O* gazdasági ágakra terjedt ki. Ezen belül a munkaügyi adatok megfigyelése havi, a kisszervezetek teljesítményeinek és beruházásainak megfigyelése negyedéves gyakoriságú volt. A reprezentatív megfigyelés csak a vállalkozásokat ölelte fel (a kijelölt költségvetési és társadalombiztosítási, valamint nonprofit szervezetek munkaügyi adatait és beruházásait teljes körűen megfigyeltük). A teljesítmények reprezentatív megfigyelése nem terjedt ki a *J* Pénzügyi tevékenység gazdasági ágra. A megfigyeléshez tartozó, csak az adott hónapra, illetve negyedévre vonatkozó kérdéseket tartalmazó kérdőívek a következők voltak (zárójelben nyilvántartási számuk az Országos Statisztikai Adatgyűjtési Programban).

– Havi iparstatisztikai adatok, Egyszerűsített jelentés (1043),
– Havi építőipar-statisztikai adatok (1025),
– Havi munkaügyi adatok (1109),
– Negyedéves teljesítmény-adatok, Egyszerűsített jelentés (1762),
– Negyedéves teljesítmény-adatok, Egyszerűsített, kiegészítő jelentés (1768),
– Negyedéves beruházás-statisztikai adatok (1014).

2001-től kezdődően a KSH egységesen az Évközi integrált adatgyűjtések részeként végzi a kisszervezetek reprezentatív gazdaságstatisztikai megfigyelését, amely továbbra is – 2003 óta a Gazdasági tevékenységek egységes ágazati osztályozási rendszere, 2003 (TEÁOR '03) szerint meghatározott – *A–D, F–K, M–O* és – 2003-ig – *E* gazdasági ágakba sorolt vállalkozásokra terjed ki. A megfigyelés részben havi, részben negyedéves gyakoriságú. A megfigyeléshez tartozó, továbbra is csak az adott hónapra, illetve negyedévre vonatkozó kérdéseket tartalmazó kérdőívek a következők.

– Havi (2004-ben integrált) egyszerűsített gazdaságstatisztikai jelentés, Ipar (1043),

– 2001-ben Havi gazdaságstatisztikai jelentés, Építőipar (1025), 2002 óta Havi (2004-ben integrált) egyszerűsített gazdaságstatisztikai jelentés, Építőipar (1938),

– Havi (2004-ben integrált) gazdaságstatisztikai jelentés, Mezőgazdasági (2003 óta Mezőgazdasági, kereskedelmi) és szolgáltatási ágazatok (1872),

– 2001-ben Negyedéves egyszerűsített gazdaságstatisztikai jelentés, Ipar és építőipar (1875), valamint Negyedéves gazdaságstatisztikai jelentés, Pénzügyi vállalkozások (1014), 2002 óta Negyedéves (2004-ben integrált) gazdaságstatisztikai jelentés, Ipari, építőipari és pénzügyi vállalkozások (1874),

– Negyedéves (2004-ben integrált) egyszerűsített gazdaságstatisztikai jelentés, Mezőgazdasági (2003 óta Mezőgazdasági, kereskedelmi) és szolgáltatási ágazatok (1878).

A felvételek során a kisszervezetekről havonta megfigyelt *munkaügyi* ismérvek a következők (voltak):

– a (főállású) teljes munkaidőben alkalmazásban álló fizikai és szellemi foglalkozásúak átlagos állományi létszáma, munkaviszonyból származó összes keresete és egyéb munkajövedelme, az egyéb foglalkoztatottak átlagos állományi létszáma, a szervezet tevékenységében összesen résztvevők átlagos állományi létszáma (a havi átlagban 60 munkaóránál rövidebb munkaidőben foglalkoztatottak nélkül), munkaviszonyból származó összes keresete és egyéb munkajövedelme, a teljes munkaidőben foglalkoztatott fizikaiak teljesített munkaórái és az összes teljesített túlóra, továbbá

– 2001-ig a szervezet tevékenységében összesen résztvevők munkaviszonyból származó összes keresetének és egyéb munkajövedelmének az a része, amelyet a kiegészítő fizetés, a nem havi rendszerességű pótlékok, a prémiumok, a jutalmak, a 13. és további havi fizetés, a felmentési időre fizetett átlagkereset és a nem rendszeres munkajövedelmek összege kitesz,

– 2002-ig az összes teljesített munkaóra,

– 2003-ig a főállású, nem teljes munkaidőben alkalmazásban álló fizikai és szellemi foglalkozásúak, valamint a további munkaviszonyban állók átlagos állományi létszáma, munkaviszonyból származó összes keresete és egyéb munkajövedelme, ezenkívül az egyéb, állományba nem tartozó munkavállalók munkaviszonyból származó összes keresete és egyéb munkajövedelme,

– 2002 óta a (főállású) teljes munkaidőben alkalmazásban álló fizikai és szellemi foglalkozásúak, továbbá a szervezet tevékenységében összesen részt vevők munkaviszonyból származó összes keresetének az a része, amelyet a prémium és jutalom, valamint a 13. és további havi fizetés kitesz, a szervezet tevékenységében összesen részt vevők munkaviszonyból származó összes keresetének az a része, amelyet a le nem dolgozott munkaidőre fizetett munkadíjak kitesznek, a szervezet tevékenységében összesen résztvevők egyéb munkajövedelmének nem havi rendszerességű része és a teljes munkaidőben foglalkoztatott szellemiek teljesített munkaórái,

– 2002-től 2003-ig a főállású, nem teljes munkaidőben alkalmazásban álló fizikai és szellemi foglalkozásúak, a további munkaviszonyban állók és az egyéb, állományba nem tartozó munkavállalók munkaviszonyból származó összes keresetének az a része, amelyet a prémium és jutalom, valamint a 13. és további havi fizetés kitesz,

– 2003-ban a nem teljes munkaidőben foglalkoztatottak és a további munkaviszonyban állók teljesített munkaórái,

– 2004-ben a nem teljes munkaidőben alkalmazásban álló, havi átlagban legalább 60 munkaórát teljesítő fizikai és szellemi foglalkozásúak, továbbá a munkaszerződés szerint havi átlagban 60 munkaóránál rövidebb munkaidőben foglalkoztatottak átlagos állományi létszáma, munkaviszonyból származó összes ke-

resete, egyéb munkajövedelme és a munkaviszonyból származó összes keresetének az a része, amelyet a prémium és jutalom, valamint a 13. és további havi fizetés kitesz, az egyéb, állományba nem tartozó, havi átlagban legalább 60 munkaórát teljesítő munkavállalók munkaviszonyból származó összes keresete, egyéb munkajövedelme és a munkaviszonyból származó összes keresetének az a része, amelyet a prémium és jutalom, valamint a 13. és további havi fizetés kitesz, a szervezet tevékenységében összesen résztvevők átlagos állományi létszáma (a havi átlagban 60 munkaóránál rövidebb munkaidőben foglalkoztatottakkal együtt), valamint a nem teljes munkaidőben alkalmazásban álló, de munkaszerződés szerint legalább 60 munkaóra teljesítésére kötelezettek és a munkaszerződés szerint havi átlagban 60 munkaóránál rövidebb munkaidőben foglalkoztatottak teljesített munkaórái.

A felvételek során a kisszervezetekről havonta megfigyelt további ismérvek az *iparban* a következők (voltak):

– az ipari és az iparon kívüli tevékenység belföldi, export- és összes értékesítésének (nettó) árbevétele, az (összes) értékesítés (nettó) árbevétele, az eladott áruk beszerzési értéke és az alvállalkozói teljesítmények, illetve – 2001 óta – a közvetített szolgáltatások értéke, továbbá

– 2001-től kezdődően az ipari tevékenységhez kötődő fogyasztási és jövedéki adók, valamint az összes termelési érték,

– 2003 óta a (fizetett) bérmunkadíj (az igénybe vett szolgáltatásból), valamint a saját termelésű készletek állománya a hó elején és végén,

– 2004-ben a termelő és importőr árbevételben realizált fogyasztási és jövedéki adó az értékesítés árbevételéből.

A saját termelésű készletek hó végi és hó eleji állományának különbsége a készletváltozás. Ennek és az ipari tevékenység összes értékesítésének összege az ipari termelési érték; 2003 előtt feltételeztük, hogy ez megegyezik az ipari tevékenység összes értékesítésével.

A felvételek során a kisszervezetekről havonta megfigyelt további ismérvek az *építőiparban* a következők (voltak):

– az építőiparban az épületeken és az egyéb építményeken végzett építőipari tevékenység (nettó) árbevétele, az építőipari (összes) és az építőiparon kívüli tevékenység, valamint az (összes) értékesítés (nettó) árbevétele, az eladott áruk beszerzési értéke, ezen kívül – külön-külön az épületekre és az egyéb építményekre – az alvállalkozói teljesítmények, illetve – 2001 óta – a közvetített szolgáltatások értéke, a saját előállítású eszközök aktivált értéke, a saját termelésű készletek állománya a hó elején és végén, a saját építési-szerelési tevékenység összesen, illetve – 2001 óta – az építőipari tevékenység termelési értéke, valamint  csak a 45.1 és a 45.2 alágazatokban  a szerződésállomány a tárgyhó végén és a tárgyhóban kötött új szerződések, továbbá

– 2001-től kezdődően a közvetített szolgáltatások értéke az építőiparon kívüli tevékenységen és összesen, a saját előállítású eszközök aktivált értéke a gépek és egyéb tárgyi eszközök vonatkozásában és összesen, a saját termelésű készletek állománya az építőiparon kívüli tevékenységén és összesen a hó elején és végén, az összes termelési érték, az építőipari tevékenység termelési értéke összesen, ezen kívül  csak a 45.1 és a 45.2 alágazatokban  a szerződésállomány a tárgyhó végén és a tárgyhóban kötött új szerződések összesen;

– 2003 óta a (fizetett) bérmunkadíj (az igénybe vett szolgáltatásból) az épületek és az egyéb építmények építésén, az építőiparon kívüli tevékenységen, valamint összesen;

– 2004-ben a termelő és importőr árbevételben realizált fogyasztási és jövedéki adó az értékesítés árbevételéből, ezen kívül a visszavont (visszamondott) szerződések a tárgyhóban kötött új szerződésekből az épületek és az egyéb építmények építésére, valamint összesen.

A felvételek során a nem a *J* Pénzügyi tevékenység gazdasági ágba sorolt kisszervezetek *teljesítményeiről* negyedévente megfigyelt ismérvek a következők (voltak):

– az iparban és az építőiparban

– 2000-ben a saját termelésű készletek állománya a negyedév elején és végén,

– 2000-től 2001-ig a vásárolt anyagok, valamint áruk és – 2001-ben – szolgáltatások készletének állománya a negyedév elején és végén,
– 2000-től 2001-ig és 2003 óta az összes vásárolt készlet állománya a negyedév elején és végén,

– a többi gazdasági ágban az értékesítés (nettó) árbevétele, az eladott áruk beszerzési értéke, az alvállalkozói teljesítmények, illetve – 2001 óta – a közvetített szolgáltatások értéke, a vásárolt anyagok, valamint áruk és – 2001 óta – szolgáltatások készletének, továbbá az összes vásárolt készletnek az állománya a negyedév elején és végén, ezen kívül
– 2000-ben és 2003 óta a saját termelésű készletek állománya a negyedév elején és végén,
– 2001-től kezdődően az összes termelési érték,
– 2003 óta a (fizetett) bérmunkadíj (az igénybe vett szolgáltatásból),
– 2004-ben a termelő és importőr árbevételben realizált fogyasztási és jövedéki adó az értékesítés árbevételéből.

A felvételek során a kisszervezetek *beruházásairól* negyedévente megfigyelt ismérvek az alábbiak (voltak):

– az új tárgyi eszközök beszerzésének, saját vállalkozásban való létesítésének, a meglevő eszközök bővítésének és felújításának teljesítményértéke az épületekre és egyéb építményekre, a belföldi gépekre és berendezésekre (járművek nélkül), a belföldi járművekre, az import gépekre és berendezésekre (járművek nélkül), az import járművekre, az ültetvényekre és erdőkre, a tenyész- és igásállatokra, a földre, telekre és más nem termelt tárgyi eszközökre, valamint összesen, továbbá
– a használt tárgyi eszközök beszerzésének teljesítményértéke az épületekre és egyéb építményekre, a belföldi gépekre és berendezésekre (járművek nélkül), a belföldi járművekre, az ültetvényekre és erdőkre, a tenyész- és igásállatokra, a földre, telekre és más nem termelt tárgyi eszközökre, valamint összesen (a kétféle teljesítményérték összege a beruházások teljesítményértéke).

Az ismérvek között tehát egyaránt vannak tagismérvek (nem más ismérvek összegeként előállított ismérvek) és összegismérvek (más, tagismérvek összegeként előállított ismérvek). A legfontosabb ismérvek az ipari termelési érték, a saját építési-szerelési tevékenység, illetve az építőipari tevékenység termelési értéke, a szervezet tevékenységében összesen részt vevők átlagos állományi létszáma, az értékesítés (nettó) árbevétele, a saját termelésű készletek állománya a negyedév elején és végén, valamint a kétféle beruházás teljesítményértéke (röviden ipari termelés, építőipari termelés, létszám, értékesítés, készletek és beruházás).

A megfigyelés célja az egyes ágazatokhoz tartozó gazdasági (ipari, építőipari) szervezetek ipari termelésének, építőipari termelésének, létszámának, értékesítésének, készleteinek és beruházásának mérhető (kiszámítható mintavételi hibájú) és elfogadható pontosságú becslése, valamint a becslés szakágazati és megyei bontása.

A reprezentatív megfigyelés célsokasága a tárgyidőszakban működő gazdasági (ipari, építőipari) kisszervezetek közül a megfelelő gazdasági ágakba tartozó, 5–49 fő közötti létszámú kisszervezetek összessége. Súlyuk az ipari és az építőipari termelés, a létszám (csak az 5 főnél nem kisebb létszámú szervezeteket figyelembe véve) és az értékesítés esetén 9, 34, 30, illetve 28 százalék. A tárgyidőszakban működő szervezetek közül az 50 főnél nem kisebb és – 2004-ben – az *E* gazdasági ágba tartozó, 5–49 fő közötti létszámú szervezeteket teljes körűen figyeljük meg. Súlyuk az ipari és az építőipari termelés, a létszám (csak az 5 főnél nem kisebb létszámú szervezeteket figyelembe véve) és az értékesítés esetén 87, 35, 70, illetve 55 százalék. Az 5 főnél kisebb létszámú mikroszervezetekre – súlyuk az ipari és az építőipari termelés, valamint az értékesítés esetén 4, 31, illetve 17 százalék – vonatkozó sokasági értékösszegeket a 6. fejezetben leírtak szerint becsüljük.

A megfigyelés keretét, a mintavételi keretet a Gazdasági szervezetek regisztere (GSZR) biztosítja. A megfigyelési egységek (amelyek egyben a mintavételi egységek is) azok az ebben szereplő működő kisszervezetek, amelyek a célsokaságnak megfelelő besorolásúak. A mintavételi keret nagyságát, vagyis a megfigyelési egységek  a GSZR-ben szereplő megfigyelt kisszervezetek  számát negyedévente (a negyedévek utolsó hónapjában) az 1. tábla mutatja.

1. tábla

*A mintavételi kerethez tartozó kisszervezetek száma 2000. I 2004. I. negyedévben*

| Negyedév | Ipar | Építőipar | Pénzügy | Egyéb | *Összesen* |
|---|---|---|---|---|---|
| 2000. I. | 11 616 | 6 275 | 397 | 31 396 | *49 684* |
| II. | 11 489 | 6 270 | 384 | 31 250 | *49 393* |
| III. | 11 647 | 6 429 | 394 | 31 748 | *50 218* |
| IV. | 11 651 | 6 483 | 391 | 31 762 | *50 287* |
| 2001. I. | 12 094 | 6 668 | 375 | 34 004 | *53 141* |
| II. | 12 166 | 6 668 | 378 | 34 119 | *53 331* |
| III. | 12 175 | 6 746 | 382 | 34 270 | *53 573* |
| IV. | 12 202 | 6 881 | 377 | 34 450 | *53 910* |
| 2002. I. | 12 547 | 7 135 | 393 | 35 914 | *55 989* |
| II. | 12 570 | 7 226 | 389 | 36 098 | *56 283* |
| III. | 12 608 | 7 341 | 390 | 36 323 | *56 662* |
| IV. | 12 716 | 7 534 | 398 | 36 883 | *57 531* |
| 2003. I. | 12 827 | 7 492 | 381 | 36 695 | *57 395* |
| II. | 12 822 | 7 651 | 381 | 36 859 | *57 713* |
| III. | 12 854 | 7 803 | 387 | 37 203 | *58 247* |
| IV. | 12 903 | 7 933 | 391 | 37 627 | *58 854* |
| 2004. I. | 13 141 | 8 290 | 386 | 40 336 | *62 153* |

Annak a nem elhanyagolható változásnak, amely a kisszervezetek számában 2001, 2002 és 2004 első negyedévében az előző negyedévhez képest bekövetkezett, regisztrációs okai vannak. (Az *E* gazdasági ágba tartozó kisszervezetek megfigyelése 2004-ben teljes körű, így ez a 156 ipari kisszervezet ebben az évben nem tartozik a reprezentatív megfigyelés mintavételi keretéhez.)

A reprezentatív adatgyűjtés értelemszerűen azokra a felvétel idején létező és működő kisszervezetekre terjed ki, amelyek a mintavétel során a mintavételi keretből kiválasztásra kerültek. A beérkező adatokat – a hiányzókat a 3–4. fejezetekben leírtak szerint (részben) pótolva – teljeskörűsítjük, majd összevonjuk a teljes körűen megfigyelt és a nem megfigyelt gazdasági szervezetek megfigyelt, illetve becsült adataival.

## 2. A MINTA KIVÁLASZTÁSA

A kisszervezetek reprezentatív évközi megfigyeléséhez a mintakiválasztást rétegzett mintavétellel hajtjuk végre. Ennek során az alábbi rétegeket képezzük.

*1*. Az ágazatonkénti becslés és a szórás csökkentése céljából az ágazati osztályozás alapján megkülönböztetjük

– az egyes ágazatokat,

– az építőipar egyetlen ágazatában (45) az egyes alágazatokat, ezen túlmenően 2001-től a 45.2 alágazatban a 45.21, a 45.25 és a többi szakágazatot,

– a *H* Szálláshely-szolgáltatás, vendéglátás gazdasági ág egyetlen ágazatában (55) az 55.1–55.2 és 55.3–55.5 alágazatcsoportokat,

továbbá külön rétegként kezeltük, illetve kezeljük

– 2000-ben az 50 és 52 ágazaton belül az 50.2, illetve 52.7 alágazatot,
– 2002 óta a 14, 22, 28, 29, 36, 37, 50 és 51 ágazaton, az 55.3–55.5 alágazatcsoporton, valamint a 60, 63 és 80 ágazaton belül a 14.21, 22.22, 28.11, 29.24, 36.14, 37.10, 50.10, 51.70, 55.30, 60.24, 63.40, illetve a 80.42 szakágazatot.

*2*. A szórás csökkentése céljából nagyság szerint megkülönböztetjük a kisszervezeteket. Ennek során felhasználjuk azt a nemzetközi gyakorlatot, amely a gazdasági szervezeteket létszámuk alapján létszám-kategóriákba sorolja. A 40, 30 és 22 létszám-kategóriájú – 20–49, 10–19 és 5–9 fő közötti létszámú – kisszervezeteket megkülönböztetjük.

*3*. Abból a célból, hogy a nemválaszolás tekintetében egységesebb rétegeink legyenek, területi szempontból megkülönböztetjük a budapesti és a vidéki kisszervezeteket. (Ennek egyébként bizonyos mértékű szóráscsökkentő hatása is van.)

A kisszervezeteket ágazati besorolásuk, létszám-kategóriájuk és területi besorolásuk szerint az előzőekkel összhangban megkülönböztetve így

| | |
|---|---|
| 2000-ben | 30 × 3 × 2 ipari + 5 × 3 × 2 építőipari + 30 × 3 × 2 egyéb, |
| 2001-ben | 30 × 3 × 2 ipari + 7 × 3 × 2 építőipari + 28 × 3 × 2 egyéb, |
| 2002–03-ban | 36 × 3 × 2 ipari + 7 × 3 × 2 építőipari + 34 × 3 × 2 egyéb, |
| 2004-ben | 34 × 3 × 2 ipari + 7 × 3 × 2 építőipari + 34 × 3 × 2 egyéb, |

összesen tehát 2001-ig 390, 2002–03-ban 462, 2004-ben (amikor az *E* gazdasági ágba tartozó kisszervezeteket teljes körűen megfigyeljük) pedig 450 réteget képeztünk, illetve képezünk.

A rétegzett reprezentatív megfigyelések egyik leglényegesebb pontja a rétegenkénti mintanagyság meghatározása. Túl nagy méretű minta esetén az adatok csak túl nagy ráfordítással gyűjthetők össze, túl kis méretű minta esetén viszont a kívánt paraméterek nem becsülhetők elég jól. A kisszervezetek reprezentatív évközi megfigyelése során az egész minta nagyságának meghatározása mellett – nem ezután, hanem ezzel egyidejűleg, interaktívan – a minta elosztását is elvégezzük. Ez úgy történik, hogy a becslés helyességére, nevezetesen pontosságára és megbízhatóságára tett különböző feltételek mellett kiszámítjuk a különböző rétegenkénti mintanagyságokat, és ezek közül azokat választjuk, amelyek növelése már nem javítja számottevően a becslést.

A pontosságot a következőképpen értelmezzük. Tekintsük az összes – teljes körűen vagy reprezentatívan megfigyelt – gazdasági szervezetre és a reprezentatívan megfigyelt gazdasági szervezetekre vonatkozó relatív hibahatárt, amelyeket $v^{(m)}$-mel, illetve $v$-vel jelölünk. Jelöljük $w^{(t)}$-vel és $w^{(r)}$-rel a teljes körűen, illetve reprezentatívan megfigyelt gazdasági szervezeteknek a mintakiválasztáskor a tárgyidőszakra előzetesen várt súlyát. Abból, hogy a teljes körűen megfigyelt gazdasági szervezeteket mintavételi hiba nélkül meg tudjuk határozni, következik, hogy ha a reprezentatívan megfigyelt gazdasági szervezetek súlya $k$-ad része az összes megfigyelt gazdasági szervezet súlyának, akkor az előbbiek relatív hibahatára $k$-szor nagyobb lehet az utóbbiaknál.

A megbízhatóságot úgy értelmezzük, hogy előírjuk az említett valószínűség értékét (az igazi sokasági értékösszeget ezzel a valószínűséggel lefedő konfidenciaintervallum számításához használt valószínűségi szintet), és ezzel közvetve meghatározzuk egy standard normális eloszlású valószínűségi változó *u* értékét (ezen *v* és *u* értékek esetén az igazi sokasági értékösszeg a rögzített valószínűséggel a konfidenciaintervallumba esik).

Legyen $v_j$ és $w_j$ a *j*-edik réteg relatív hibahatára, illetve súlya, akkor a relatív hibahatár definíciójából és a mintavételnek az egyes rétegeken belül egymástól függetlenül történő végrehajtásából következik, hogy

$$\sum_j (w_j v_j)^2 = v^2.$$

A $v_j$ értékeket úgy határozzuk meg, hogy nagyobb súlyhoz alacsonyabb relatív hibahatár tartozzon, de a fordított arányosságnál kisebb mértékben: $a^2$-szer nagyobb súlyhoz *a*-szor kisebb relatív hibahatár (négyszer akkora súlyhoz például feleakkora relatív hibahatár). Egyszerű számolással adódik, hogy ekkor az egyes rétegek relatív hibahatára a

$$v_j = \frac{v}{\sqrt{w_j}}$$

összefüggés segítségével határozható meg. (Megjegyzem, hogy ha a $v_j$ értékeket úgy akarnánk meghatározni, hogy *a*-szor nagyobb súlyhoz tartozzon *a*-szor kisebb relatív hibahatár, akkor – *J*-vel jelölve a rétegek számát – az egyes rétegek relatív hibahatára a

$$v_j = \frac{v}{w_j \sqrt{J}}$$

összefüggés segítségével lenne meghatározható, míg ha azt akarnánk, hogy valamennyi réteg relatív hibahatára – a súlytól függetlenül – ugyanaz a $v_1$ érték legyen, akkor ezt az értéket a

$$v_1 = \frac{v}{\sqrt{\sum_j w_j^2}}$$

képlettel számíthatnánk ki. Előbbi – a Neyman-féle optimalizálással megegyezően – az egész sokaságra, utóbbi pedig az egyes, egyformán fontosnak vett rétegekre vonatkozó hiba szempontjából optimális mintaeloszlást eredményez. A választott megoldás a kettő közötti középút.)

A rétegenkénti mintanagyságot a *Telegdi* [1993]-ban leírt módon határozzuk meg (ehhez szükségünk van az egyes rétegek tényleges, illetve előzetesen becsült nagyságára és a relatív szórás előzetesen becsült értékére is). A mintaelemszám meghatározásánál fi-

gyelembe vesszük a KSH területi (megyei) igazgatóságainak teherbíró képességét is (különös tekintettel a Budapesti és Pest megyei Igazgatóságra), és úgy állapítjuk meg a mintanagyságot, hogy az ne idézzen elő jelentős mértékű nem válaszolást (egy 1500 elemű minta 300 nemválaszolóval jobb, mint egy 3000 elemű minta 1000 nemválaszolóval).

A minta nagyságát és az ez alapján meghatározott kiválasztási arányt negyedévente (a negyedévek utolsó hónapjában) a 2. tábla mutatja.

2. tábla

*A mintaelemszám és a kiválasztási arány 2000. I–2004. I. negyedévben*

| Negyedév | Ipar | | Építőipar | | Pénzügy | | Egyéb | | *Összesen* | |
|---|---|---|---|---|---|---|---|---|---|---|
| | elemszám | arány (százalék) | elemszám | arány (százalék) | elemszám | arány (százalék) | elemszám | arány (százalék) | *elemszám* | *arány (százalék)* |
| 2000. I. | 1574 | 13,6 | 1266 | 20,2 | 158 | 39,8 | 6383 | 20,3 | *9381* | *18,9* |
| II. | 1537 | 13,4 | 1233 | 19,7 | 148 | 38,5 | 6214 | 19,9 | *9132* | *18,5* |
| III. | 1517 | 13,0 | 1210 | 18,8 | 147 | 37,3 | 6121 | 19,3 | *8995* | *17,9* |
| IV. | 1506 | 12,9 | 1195 | 18,4 | 143 | 36,6 | 6046 | 19,0 | *8890* | *17,7* |
| 2001. I. | 1544 | 12,8 | 1170 | 17,5 | 162 | 43,2 | 6318 | 18,6 | *9194* | *17,3* |
| II. | 1519 | 12,5 | 1126 | 16,9 | 160 | 42,3 | 6170 | 18,1 | *8975* | *16,8* |
| III. | 1496 | 12,3 | 1119 | 16,6 | 159 | 41,6 | 6126 | 17,9 | *8900* | *16,6* |
| IV. | 1475 | 12,1 | 1101 | 16,0 | 156 | 41,4 | 6065 | 17,6 | *8797* | *16,3* |
| 2002. I. | 1719 | 13,7 | 1133 | 15,9 | 154 | 39,2 | 6318 | 17,6 | *9324* | *16,7* |
| II. | 1695 | 13,5 | 1121 | 15,5 | 152 | 39,1 | 6240 | 17,3 | *9208* | *16,4* |
| III. | 1683 | 13,3 | 1105 | 15,1 | 152 | 39,0 | 6183 | 17,0 | *9123* | *16,1* |
| IV. | 1662 | 13,1 | 1092 | 14,5 | 151 | 37,9 | 6133 | 16,6 | *9038* | *15,7* |
| 2003. I. | 1814 | 14,1 | 1170 | 15,6 | 155 | 40,7 | 6266 | 17,1 | *9405* | *16,4* |
| II. | 1776 | 13,9 | 1150 | 15,0 | 154 | 40,4 | 6174 | 16,8 | *9254* | *16,0* |
| III. | 1748 | 13,6 | 1132 | 14,5 | 153 | 39,5 | 6116 | 16,4 | *9149* | *15,7* |
| IV. | 1729 | 13,4 | 1104 | 13,9 | 151 | 38,6 | 6073 | 16,1 | *9057* | *15,4* |
| 2004. I. | 1743 | 13,3 | 1204 | 14,5 | 157 | 40,7 | 6384 | 15,8 | *9488* | *15,3* |

Annak, hogy a pénzügyi kisszervezetek körében ilyen nagy a kiválasztási arány, kis számuk és nagy súlyuk az oka.

A reprezentatív évközi megfigyelés minden évben az előző év végén kiválasztott mintából történik. A mintavétel során csak azokból a gazdasági szervezetekből választunk, amelyek a GSZR január 1-jére érvényes állapota szerint működők. Ezt a megszorítást, amely a mintavételi keret leszűkítését jelenti, azért tesszük, mert tapasztalataink szerint a létező, de nem működő (felszámolás vagy csődeljárás alatt levő stb.) gazdasági szervezetek teljesítménye elhanyagolható, ugyanakkor körükben a válaszolási arány – érthetően – nagyon rossz.

A kisszervezetek reprezentatív évközi gazdaságstatisztikai megfigyeléseinek sikerességéhez elengedhetetlen a kiválasztott minta karbantartása. Ennek fontos mozzanata a mintaelemek kicserélése bizonyos idő után, más szóval a minta rotációja. Egy-egy reprezentatív megfigyelés esetén ugyanis alapvető kérdés a következő: mennyire megalapozott az a feltételezés, hogy a sokaságot jellemző valamilyen mennyiségnek az igazi értéke közel van a minta alapján becsült értékhez? Bár kicsi a valószínűsége, de előfordulhat, hogy a minta rosszul tükrözi a sokaságot. A rotáció alkalmazását – az adatszolgáltatói terhek csökkentése mellett – általában az teszi indokolttá, hogy védekezzünk ez ellen.

Azok a szempontok (adatszolgáltatási hajlandóság, a mintaelemek lemorzsolódásának mértéke stb.), amelyek figyelembevételével egy minta rotációja kialakításra kerül, erősen függenek az adott felvétel sajátosságaitól. A gazdaságstatisztikai megfigyeléseket az jellemzi, hogy nem mindig könnyű a megfigyelési egységeket megtalálni és bevonni az adatszolgáltatásba. Az Igazgatóságok munkájának megkönnyítése céljából, az adatgyűjtés eredményessége érdekében ezért célunk, hogy a rotáció ne legyen túl nagy.

Mindezeket figyelembe véve az egyes rétegekre a kisszervezetek mintájának kiválasztása a következőképpen történik. A mintavételi kerethez és az adott réteghez tartozó kisszervezetek mindegyikéhez előállítunk egy $v_i$ véletlen számot. E célból vesszük a GSZR-ben a kisszervezethez tartozó, 0 és 1 között egyenletes eloszlású $u_i$ véletlen számot. 2001-ig $v_i$ megegyezett $u_i$-val, 2002 óta viszont

$$v_i = 1 - u_i.$$

Abból a célból, hogy előnyben részesítsük elsősorban azokat a kisszervezeteket, amelyek 2000 esetében 3, 2001-től kezdődően 3–5 évvel korábban nem voltak mintaelemek (ez a rotációt biztosítja), másodsorban azokat, amelyek már az előző éves mintának is elemei voltak (ez a válaszadási arány kedvezőbb alakulását segíti elő), az e kisszervezetekhez tartozó véletlen számokat az első esetben 2-vel, a második esetben 1-gyel növeltük, illetve növeljük, majd a kisszervezeteket az így módosított véletlen számok nagysága szerint csökkenő sorba rendezzük. (Ebben a sorban az adott réteghez tartozó kisszervezetek tehát az alábbi rendben követik egymást: *1*. azok az előző éves mintaelemek, amelyek 3, illetve 3–5 évvel korábban még nem voltak mintaelemek; *2*. azok a kisszervezetek, amelyek nem tartoztak a mintához sem 3, illetve 3–5 évvel korábban, sem az előző évben; *3*. azok az előző éves mintaelemek, amelyek 3, illetve 3–5 évvel korábban is a minta elemei voltak; *4*. azok a kisszervezetek, amelyek 3, illetve 3–5 évvel korábban a mintához tartoztak, de az előző éves mintának már nem voltak elemei. Az $u_i$ 1-ből történő kivonása egyenértékű azzal, mintha e módosítás nélkül a véletlen számokat nagyság szerint növekvő sorba rendeznénk. Ez ahhoz járul hozzá, hogy a mintából kikerült kisszervezetek három év elteltével is csak kevés esélylyel kerüljenek oda újra be.) Az ily módon véletlen sorba rendezett kisszervezetek közül az elsők kerülnek (megfelelő számban) a mintába.

A megfigyelés során az új (az induló minta meghatározását lezáró időpont után alakult, illetve nyilvántartásba került) kisszervezetek folyamatosan bekerülnek a mintavételi keretbe, de belőlük mintát nem választunk ki.

## 3. AZ ADATGYŰJTÉS

A mintához tartozó kisszervezetek az egész évre szólóan postán kapják meg a jelentéseket, melyeket két példányban kell kitölteniük. A két példányból az egyik náluk marad, a másikat postán kell visszaküldeniük a KSH illetékes területi igazgatóságának. A tárgyidőszakban létező és működő mintaelemek állományát az ún. expediálási listán küldjük meg az igazgatóságoknak. Nemválaszolás esetén, ennek alapján kell telefonon vagy postai úton az adatszolgáltatást sürgetniük.

A kitöltött (esetleg nemleges) jelentést visszajuttató kisszervezetek számát és az ennek alapján meghatározott válaszadási arányt negyedévente (a negyedévek utolsó hónapjában) a 3. tábla mutatja.

3. tábla

*A jelentést visszajuttató kisszervezetek száma és a válaszadási arány 2000. I–2004. I. negyedévben*

| Negyedév | Ipar | | Építőipar | | Pénzügy | | Egyéb | | *Összesen* | |
|---|---|---|---|---|---|---|---|---|---|---|
| | szervezet-szám | arány (százalék) | szervezet-szám | arány (százalék) | szervezet-szám | arány (százalék) | szervezet-szám | arány (százalék) | *szervezet-szám* | *arány (százalék)* |
| 2000. I. | 1358 | 86,3 | 1144 | 90,4 | 145 | 91,8 | 5574 | 87,3 | *8221* | *87,6* |
| II. | 1348 | 87,7 | 1136 | 92,1 | 139 | 93,9 | 5643 | 90,8 | *8266* | *90,5* |
| III. | 1340 | 88,3 | 1120 | 92,6 | 138 | 93,9 | 5568 | 91,0 | *8166* | *90,8* |
| IV. | 1331 | 88,4 | 1058 | 88,5 | 133 | 93,0 | 5520 | 91,3 | *8042* | *90,5* |
| 2001. I. | 1272 | 82,4 | 984 | 84,1 | 152 | 93,8 | 5236 | 82,9 | *7644* | *83,1* |
| II. | 1276 | 84,0 | 980 | 87,0 | 148 | 92,5 | 5097 | 82,6 | *7501* | *83,6* |
| III. | 1253 | 83,8 | 973 | 87,0 | 150 | 94,3 | 5222 | 85,2 | *7598* | *85,4* |
| IV. | 1234 | 83,7 | 963 | 87,5 | 149 | 95,5 | 5206 | 85,8 | *7552* | *85,8* |
| 2002. I. | 1501 | 87,3 | 994 | 87,7 | 143 | 92,9 | 5552 | 87,9 | *8190* | *87,8* |
| II. | 1483 | 87,5 | 980 | 87,4 | 143 | 94,1 | 5511 | 88,3 | *8117* | *88,2* |
| III. | 1474 | 87,6 | 968 | 87,6 | 141 | 92,8 | 5457 | 88,3 | *8040* | *88,1* |
| IV. | 1458 | 87,7 | 958 | 87,7 | 139 | 92,1 | 5369 | 87,5 | *7924* | *87,7* |
| 2003. I. | 1620 | 89,3 | 1011 | 86,4 | 150 | 96,8 | 5576 | 89,0 | *8357* | *88,9* |
| II. | 1598 | 90,0 | 988 | 85,9 | 149 | 96,8 | 5506 | 89,2 | *8241* | *89,1* |
| III. | 1569 | 89,8 | 974 | 86,0 | 146 | 95,4 | 5397 | 88,2 | *8086* | *88,4* |
| IV. | 1543 | 89,2 | 950 | 86,1 | 144 | 95,4 | 5234 | 86,2 | *7871* | *86,9* |
| 2004. I. | 1528 | 87,7 | 990 | 82,2 | 148 | 94,3 | 5418 | 84,9 | *8084* | *85,2* |

A 3. tábla mutatja a válaszadás 2001-ben bekövetkezett romlását. Valószínűleg azért csökkent a válaszadási arány, mert a mintavétel során leszűkítettük azoknak a kisszervezeteknek a körét, amelyeket előnyben részesítettünk, és ezáltal nagyobb rotációt hajtottunk végre.

A jelentést visszajuttató kisszervezetek adataiból tudunk következtetéseket levonni a kisszervezetek évközi gazdálkodását jellemző ismérvekre, vagyis az e kisszervezetek mintájában megvalósuló megfigyeléseket általánosítjuk, teljeskörűsítjük.

## 4. A NEMVÁLASZOLÁS KEZELÉSE

Az adatgyűjtés eredményességét kedvezőtlenül befolyásolhatja a nemválaszolás. Ennek, vagyis a nem teljesített adatszolgáltatásnak okairól a KSH érkeztető rendszere nyújt információt az ún. MV19 kóddal. Ennek értékei, az érkeztető kódok az alábbiak.

000 Még nem tisztázott ok
1 Gazdasági szervezethez tartozó okok
101 A szervezet jogutód nélkül megszűnt
102 Felszámolás vagy végelszámolás alatt levő, nem működő szervezet
103 Csődeljárás alatt levő, nem működő szervezet
104 Még nem működő szervezet
105 Működését egyéb okból szüneteltető szervezet
107 A szervezet elköltözött, címe ismeretlen
108 Nem létező cím
111 Az ágazati besorolás helytelen
112 A létszám-kategóriába sorolás helytelen
113 A szervezet nem az adott megyében folytatja a tevékenységét

115 A szervezet jogutóddal megszűnt
116 Működő, felszámolás vagy végelszámolás alatt álló szervezet
117 Működő, csődeljárás alatt álló szervezet
118 A szervezet nem folytat rendszeres gazdasági tevékenységet
2 A szervezet tevékenységéhez kapcsolódó okok
201 A szervezetnek nincs az adatgyűjtésre vonatkozó tevékenysége
202 A szervezetnek az adatgyűjtésre vonatkozó tevékenysége megszűnt
203 Az adott időszakban a szervezetnek nincs az adatgyűjtésre vonatkozó tevékenysége
205 Egyéb ok miatt nemleges lenne a jelentés
8 Szubjektív tényezők
801 Az adatszolgáltatást megtagadta
802 A jelentést késve fogja küldeni
803 A szervezettel a kapcsolatfelvétel nem sikerült
999 Beérkezett adatjelentés

A nem teljesített adatszolgáltatás okairól összefoglalóan a következők mondhatók. A nemválaszolásnak a korábbi évekhez képest kevesebb, átlagosan 15–30 százaléka adódik a GSZR részben szükségszerű hibáiból (a gazdasági szervezethez tartozó és a szervezet tevékenységéhez kapcsolódó okok). Ezekben az esetekben a nem válaszolás a gazdasági szervezet szempontjából voltaképpen jogos: vagy meg sem kapja a jelentést, vagy tulajdonképpen indokolatlanul kapja meg, és nem tud válaszolni. A nem mindig egyértelműen elválasztható „még nem tisztázott ok” és „a szervezettel a kapcsolatfelvétel nem sikerült” esetek vannak a legtöbben (átlagosan 35–70 százalék). A korábbi évekhez képest több az olyan kisszervezet, amely nyíltan megtagadja az adatszolgáltatást (átlagosan 15–30 százalék). A kisszervezetek átlagosan 5–15 százaléka a jelentést késve küldi, vagyis csak a feldolgozás időpontjában tekinthető nem válaszolónak (a nem válaszolással részletesebben foglalkozik *Telegdi* [1999]).

A nemválaszoló kisszervezetek közül egyedileg 0-val pótoljuk, más szóval imputáljuk azok hiányzó adatait, amelyekről az feltételezhető, hogy nemleges jelentést küldtek volna be. (Azt, hogy melyek ezek a kisszervezetek, az érkeztető rendszer által szolgáltatott MV19 kód és teljes körű – részben külső forrásból származó – adatokkal való utólagos összehasonlításból származó tapasztalataink alapján állapítjuk meg.) A többi nem válaszoló kisszervezet hiányzó adatait 2001-ig egyedileg nem pótoltuk. (Mint azonban a következő fejezetben látni fogjuk, az alkalmazott becslés egyenértékű volt azzal, mintha ezeket a hiányzó adatokat az átlaggal pótoltuk volna – de nem egyedileg, hanem összességükben.) 2002 óta ezeknek a nem válaszoló kisszervezeteknek is pótoljuk a hiányzó adatait. A pótlás módját a következő fejezetben ismertetjük.

## 5. A BECSLÉS

A becslés folyamán a mintába kiválasztott és jelentést visszajuttató kisszervezetek adataiból kell következtetéseket levonnunk az összes kisszervezet évközi megfigyelésének ismérveiről, vagyis az ebben a mintában elvégzett megfigyeléseket kell teljeskörűsítenünk. Egyes rétegeket két részre bontunk. Ennek során a réteg kiugró (kiugróan nagy) értékekkel, más szóval outlierekkel rendelkező kisszervezeteit elkülönítjük. Ezáltal a reprezentatívan megfigyelt és így teljeskörűsítésre kerülő, közönséges rétegek mellett teljes körűen megfigyelt és így teljeskörűsítésre nem kerülő, kiemelt rétegeket is képezünk. 2002 óta a kiemelt kisszervezetek egy része a GSZR szerinti árbevétel adatuk

alapján minden hónapban, illetve negyedévben kiemelt (ha válaszolt; ellenkező esetben nem tekintjük kiemeltnek). A kiemelt kisszervezetek másik részét havonta, illetve negyedévente jelöljük ki megfigyelési adataik alapján.

Mind az állandóan, mind a nem állandóan kiemelt kisszervezeteket az alábbi módon határozzuk meg (részletesebben lásd *Csereháti* [2004]). Ismérvenként a különböző rétegek összehasonlíthatóvá tétele céljából a kisszervezetek adatait – az ismérv rétegátlagát levonva és a különbséget az ismérv rétegbeli szórásával osztva – standardizáljuk, majd a standardizált értéket a réteg mintanagyságának függvényében módosítjuk (ezt a módosítást az teszi szükségessé, hogy kevesebb adathoz képest nagyobb valószínűséggel fordul elő nagy érték). A szóban forgó ismérvre vonatkozó módosított standardizált értékeket nagyság szerint csökkenő sorba rendezzük. Az ily módon sorba rendezett értékek közül a matematikai és tapasztalati megfontolások alapján megállapított küszöbnél nagyobb értékűeket tekintjük outliernek. Azokat a kisszervezeteket emeljük ki, amelyeknek legalább egy ismérvre vonatkozó adata outlier.

A kiemelt kisszervezetek meghatározása után – 2002 óta – pótoljuk azoknak a nem válaszoló kisszervezeteknek a hiányzó adatait, amelyek adatait nem pótoltuk 0-val. Ezeket a hiányzó adatokat a Budapest és a vidék összevonásával, a kiemelt kisszervezetek elhagyásával kiszámított megfelelő rétegátlagokkal pótoljuk.

A feldolgozás folyamán több, a megfigyelés ismérveire vonatkozó mennyiség paraméterbecslését is elvégezzük egyrészt az egyes reprezentatívan megfigyelt rétegekre, másrészt összevonva ezeket, valamint a kiemelt és az egyéb teljes körűen megfigyelt (ilyen létszám-kategóriákba sorolt) rétegeket. A becslést a mintakiválasztás rétegeire végezzük, majd előbb a teljeskörűsített és a megfelelő kiemelt réteget, azután Budapestet és a vidéket, ezt követően az egyes létszám-kategóriákat, végül az ágazati egységeket vonjuk össze.

A kisszervezetek összességének jellemzése érdekében becsüljük az egyes ismérvek sokasági értékösszegét. Közülük a tagismérvek értékösszegének becslését az elemi adatokból közvetlenül végezzük. Egy-egy reprezentatívan megfigyelt rétegen belül a következőképpen járunk el. Meghatározzuk azt a $q_j$ értékarányt, amely az összes *j*-edik rétegbeli gazdasági szervezet $N_j$ számának (a réteg nagyságának), valamint az ezek közül a mintába kiválasztott és válaszoló, nemlegesnek vagy – 2002 óta – nem nemlegesnek pótolt gazdasági szervezetek $n_j$ számának a hányadosa. A számításoknál egy $k_j$ korrekciós tényezőt alkalmazunk, melyet a minta reprezentativitása alapján határozunk meg. $k_j$ értéke attól függően tér el 1-től, hogy a GSZR-ből teljes körűen rendelkezésünkre álló korábbi éves árbevételnek a mintára vonatkozó átlaga mennyire különbözik a megfelelő sokasági átlagtól. Az egyes rétegeken belül a $Y_j$ sokasági értékösszeget úgy becsüljük, hogy a mintaelemekre vonatkozó $y_j$ értékösszeget megszorozzuk a korrekciós tényező és az értékarány szorzatával:

$$Y_j = k_j q_j y_j .$$

A korrekciós tényezőtől eltekintve tehát egyszerű felszorzással becslünk. Korábban (lásd *Telegdi* [2001]) alkalmaztunk a tárgyidőszaknál 2-vel korábbi év (későbbi nem áll

rendelkezésre) árbevételét mint segédinformációt felhasználó hányadosbecslést, ez azonban – éppen az időbeli távolság miatt – nem bizonyult jobbnak az egyszerű felszorzásnál. (Igaz, a korrekciós tényező valamit „visszahoz" a hányadosbecslésből.)

Mind az egyes rétegekre, mind ezekre együttesen meghatározzuk az előző évi hasonló és az előző havi, illetve negyedéves sokasági értékösszeghez mint bázishoz viszonyított százalékos növekedést, vagyis a

$$100\frac{Y_j}{Y_{0j}}, \qquad \text{illetve} \qquad 100\frac{Y}{Y_0},$$

(százalékos) értékindexeket.

Mind az összeg-, mind a tagismérvekre a (tárgyidőszaki) sokasági értékösszegből 2001-ig a

$$\overline{Y_j} = \frac{Y_j}{k_j N_j}, \qquad \overline{Y} = \frac{Y}{\sum_j k_j N_j}.$$

2002 óta a szokásos ($k_j$ nélküli) képletek segítségével becsültük, illetve becsüljük a sokasági átlagot az egyes rétegekre és rétegekre együttesen.

Mivel az egyes rétegekre 2002-től kezdődően

$$\overline{Y}_j = \frac{Y_j}{N_j} = \frac{k_j y_j}{n_j} = k_j \overline{y_j},$$

azért a $\overline{Y_j}$ sokasági átlag azóta csak akkor egyezik meg a $\overline{y_j}$ mintaátlaggal, ha $k_j$ értéke 1.

A reprezentatívan megfigyelt rétegekre 2001-ig a szokásos, 2002 óta a

$$\sigma_j^2 = \frac{1}{n_j - 1}\sum_{i=1}^{n_j}(k_j y_{ji} - \overline{Y_j})^2 = \frac{1}{n_j - 1}(k_j^2 \sum_{i=1}^{n_j} y_{ji}^2 - n_j \overline{Y_j}^2)$$

képlet segítségével meghatároztuk, illetve meghatározzuk a korrigált tapasztalati szórásnégyzetet, ahol $y_{ji}$ a *j*-edik réteg mintába kiválasztott *i*-edik válaszoló, nemlegesnek vagy – 2002 óta – nem nemlegesnek pótolt kisszervezetének (tárgyhavi, illetve -negyedéves) ismérvértéke, és az egyes rétegek

$$C_j = \frac{\sigma_j}{\overline{Y_j}}$$

relatív szórását.

Egy-egy reprezentatívan megfigyelt rétegen belül az egyes ismérvek sokasági értékösszegének $\sigma_{Y_j}$ szórását, a mintavételi vagy standard hibát 2001-ig a

$$\sigma_{Y_j} = \frac{k_j N_j \sigma_j}{\sqrt{n_j}} \sqrt{1 - \frac{n_j}{k_j N_j}},$$

2002 óta a szokásos képlet segítségével becsültük, illetve becsüljük. (Rétegekre együttesen a $\sigma_Y$ standard hiba az egyes rétegek melletti standard hibák négyzetösszegének négyzetgyöke.)

Mind az egyes reprezentatívan megfigyelt rétegekre, mind rétegekre együttesen a szokásos módon meghatározzuk a *relatív hibát*.

Az egyes ismérvek értékösszegének becslése köré a standard hiba segítségével konfidenciaintervallum jelölhető ki. Nevezetesen a 2. fejezetben említett megbízhatósági követelménynek megfelelően (amikor is a valószínűségi szint 0,95) meghatározható az a

$$\Delta_j = 1{,}96\sigma_{Y_j}, \qquad \text{illetve} \qquad \Delta = 1{,}96\sigma_Y$$

*abszolút hibahatár*, amelyre 0,95 valószínűséggel a

$$(Y_j - \Delta_j, Y_j + \Delta_j), \qquad \text{illetve} \qquad (Y - \Delta, Y + \Delta)$$

(abszolút) konfidenciaintervallum közrefogja az „igazi" sokasági értékösszeget.

Az abszolút hibahatárból a szokásos módon meghatározzuk az értékösszeg relatív hibahatárát (amely egyben a relatív konfidenciaintervallum sugara).

Mind az egyes reprezentatívan megfigyelt rétegekre, mind a rétegekre együttesen a

$$\sigma_{Y_j}^{(i)} = f_j \frac{\sigma_{Y_j}}{Y_{0j}}, \qquad \text{illetve a} \qquad \sigma_Y^{(i)} = \frac{\sqrt{\sum_j [Y_{0j}\sigma_{Y_j}^{(i)}]^2}}{Y_0}$$

képlet segítségével meghatározzuk az értékindexek standard hibáját. Az első képletben szereplő $f_j$ szorzót a következőképpen számítjuk ki (lásd például *Éltető–Mészéna–Ziermann* [1982], 63–65. oldal). Egy-egy reprezentatívan megfigyelt rétegen belül csak azokat a válaszoló, nemlegesnek vagy – 2002 óta – nem nemlegesnek pótolt kisszervezeteket tekintjük, amelyek a bázisidőszakban is ugyanehhez a réteghez tartoztak, és válaszoló vagy nemlegesnek pótolt mintaelemek voltak. Jelölje ezek számát $m_j$, tárgy- és bázisidőszaki ismérvértéküket $z_{ji}$, illetve $x_{ji}$ (az $m_j$ számú $z_{ji}$ érték az $n_j > m_j$ számú $y_{ji}$ érték közül kerül ki), összegüket pedig $z_j$ és $x_j$, akkor a megfelelő átlagok

$$\overline{z_j} = \frac{z_j}{m_j} \qquad \text{és} \qquad \overline{x_j} = \frac{x_j}{m_j},$$

a szórások

$$\sigma_{1j} = \sqrt{\frac{1}{m_j - 1}\left[ (\sum_{i=1}^{m_j} z_{ji}^2) - m_j (\overline{z_j})^2 \right]} \qquad \text{és} \qquad \sigma_{0j} = \sqrt{\frac{1}{m_j - 1}\left[ (\sum_{i=1}^{m_j} x_{ji}^2) - m_j (\overline{x_j})^2 \right]},$$

a relatív szórások

$$C_{1j} = \frac{\sigma_{1j}}{\overline{z_j}} \qquad \text{és} \qquad C_{0j} = \frac{\sigma_{0j}}{\overline{x_j}},$$

a tárgy- és bázisidőszaki ismérvérték korrelációs együtthatója pedig

$$r_j = \frac{\frac{1}{m_j - 1}\left[ (\sum_{i=1}^{m_j} z_{ji} x_{ji}) - m_j \overline{z_j} \cdot \overline{x_j} \right]}{\sigma_{1j} \sigma_{0j}}.$$

Az $f_j$ szorzót az

$$f_j = \frac{\sqrt{C_{1j}^2 + C_{0j}^2 - 2 r_j C_{1j} C_{0j}}}{C_{1j}}$$

képlet segítségével határozzuk meg. A képletből következik, hogy a relatív szórás időbeli állandósága, vagyis $C_{1j} = C_{0j}$ esetén

$$v_j^{(i)} = v_j \sqrt{2(1 - r_j)}.$$

Innen látszik, hogy $r_j = 0{,}5$ esetén $v_j^{(i)} = v_j$, $r_j > 0{,}5$ esetén viszont $v_j^{(i)} < v_j$ (az $r_j = 1$ szélsőséges esetben $v_j^{(i)} = 0$). Az egyes teljes körűen megfigyelt rétegekre értelemszerűen $v_j^{(i)} = 0$.

## 6. A SOKASÁGI ÉRTÉKÖSSZEGEK SZAKÁGAZATI ÉS MEGYEI BONTÁSÚ BECSLÉSE

A relatív szórás nagysága és a kisszervezetek száma nem teszi lehetővé, hogy a teljeskörűsítés során a szakágazatokat – néhány kivételtől eltekintve – külön rétegként kezeljük. Ennek ellenére – mivel igény van rá – becsüljük az egyes szakágazatok különböző ismérveinek sokasági értékösszegét: az előző alfejezetben leírtak szerint teljeskörűsített értékeket létszám-kategóriánként, külön Budapestre és külön a vidékre

mindegyik ágazati réteg esetén megbontjuk a szakágazatok között, mégpedig a minta értékösszegét a beérkezett adatok, a sokasági értékösszeg fennmaradó részét a szakágazatok elemszáma és súlya alapján.

Az egyes szakágazatok becsléséhez létszám-kategóriánként, külön Budapestre és külön a vidékre szükség van a GSZR-ből az egyes szakágazatokba tartozó kisszervezetek $M_j$ számára, valamint a kisszervezetek árbevételére és foglalkoztatottjainak létszámára. Ez utóbbiakból meghatározzuk az egyes szakágazatokba tartozó gazdasági szervezetek $X_j^{(i)}$ sokasági értékösszegeit ($i$ = 1: árbevétel, $i$ = 2: létszám). Ezekből az adatokból kiszámítjuk az

$$\overline{X_j^{(i)}} = \frac{X_j^{(i)}}{M_j}$$

átlagokat.

Az évközi feldolgozás során az egyes, szakágazatnál bővebb ágazati rétegeken és létszám-kategóriákon belül külön Budapestre és külön a vidékre a következőképpen járunk el. Jelölje $Y^{(i)}$ az adott létszám-kategória és ágazati réteg Budapestre vagy a vidékre vonatkozó $i$-edik típusú teljeskörűsített értékösszegét ($i$ = 1: nem munkaügyi adatok, $i$ = 2: munkaügyi adatok), $y_j^{(i)}$ és

$$y^{(i)} = \sum_j y_j^{(i)}$$

a mintaelemekre vonatkozó megfelelő értékösszegeket az ágazati réteghez tartozó szakágazatokban, illetve az egész ágazatban, $N_j$ a megfelelő sokaságokhoz tartozó, $n_j$ pedig a mintába kiválasztott és válaszoló vagy pótolt gazdasági szervezetek számát. Képezzük az

$$s_j^{(i)} = \frac{(N_j - n_j)\overline{X_j^{(i)}}}{\sum_j (N_j - n_j)\overline{X_j^{(i)}}}$$

súlyszámokat. A mintaelemekre vonatkozó értékösszegekkel csökkentett $\left[Y^{(i)} - y^{(i)}\right]$ teljeskörűsített értékösszegeket e súlyszámok alapján megbontjuk a szakágazatok között:

$$Y^{(i)} - y^{(i)} = \sum_j W_j^{(i)},$$

ahol

$$W_j^{(i)} = s_j^{(i)}\left[Y^{(i)} - y^{(i)}\right].$$

Az egyes szakágazatok sokasági értékösszegét a

$$Y_j^{(i)} = y_j^{(i)} + W_j^{(i)}$$

képlet segítségével becsüljük.

Az ágazati rétegre vonatkozó $\sigma_{Y_j}$ standard hibából – egy-egy rétegen belül a szakágazatokra vonatkozóan homoszkedaszticitást, vagyis egyenlő szórást feltételezve – a

$$\sigma_{Y_{jl}} = \sqrt{t_{jl}}\,\sigma_{Y_j}$$

képlettel számítjuk ki az egyes szakágazatokra vonatkozó standard hibákat, ahol

$$t_{jl} = \frac{N_{jl}^2 / n_{jl}}{\sum_l (N_{jl}^2 / n_{jl})}.$$

Az értékindex ágazati rétegre vonatkozó $\sigma_{Y_j}^{(i)}$ standard hibájából – ugyancsak homoszkedaszticitást feltételezve – a

$$\sigma_{Y_{jl}}^{(i)} = \sqrt{t_{jl}}\,\sigma_{Y_j}^{(i)}$$

képlettel számítjuk ki az értékindexnek az egyes szakágazatokra vonatkozó standard hibáját.

A relatív szórás nagysága és a kisszervezetek száma azt sem teszi lehetővé, hogy a teljeskörűsítés folyamán a vidéken belül a megyéket külön rétegként kezeljük. Mivel azonban egyre nagyobb igény van az egyes megyék sokasági értékösszegeinek ismeretére, szükséges a megyei értékösszegek becslése: a vidékre vonatkozó, a fentiek szerint meghatározott szakágazati értékösszegeket megbontjuk a megyék között. Ennek során az egyes szakágazatokon és létszám-kategóriákon belül hasonlóan járunk el, mint az ágazati értékösszegek szakágazatok közti megbontása esetén. Az egyetlen különbség, hogy a súlyszámok képzésénél csak a megfelelő sokaságokhoz tartozó, valamint a mintába kiválasztott és válaszoló vagy pótolt kisszervezetek számát vesszük figyelembe, a GSZR adataiból számított átlagokat nem.

## 7. A NEM MEGFIGYELT MIKROSZERVEZETEK SOKASÁGI ÉRTÉKÖSSZEGÉNEK BECSLÉSE

A közvetlenül nem megfigyelt (5 főnél kisebb létszámú) ipari mikroszervezetek ipari termelési értékének sokasági értékösszegét az alábbi eljárás segítségével becsüljük. A GSZR-ben rendelkezésünkre álló utolsó év adatai alapján mindhárom iparágban (*C*, *D*, *E*) meghatározzuk egyrészt az (5 főnél kisebb létszámú) mikroszervezetek, másrészt a 22, 30 és 40 létszám-kategóriájú kisszervezetek átlagos árbevételének $c_j$ hányadosát ($j$ = 1, 2, 3).

Feltételezzük, hogy a három iparágban egyrészt a mikroszervezetek, másrészt a 22, 30 és 40 létszám-kategóriájú kisszervezetek átlagos ipari termelésének a mindenkori tárgyhónapra vonatkozó hányadosa a tárgyév folyamán állandó és megegyezik $c_j$ értékével.

Az adott, *i*-edik tárgyhónapban a mikroszervezetek $Y_{0-4}^{(i)}$ ipari termelését az egész iparra vonatkozóan úgy számítjuk ki, hogy az egyes iparágakban a 22, 30 és 40 létszám-kategóriájú szervezetek $\overline{Y_{22,j}^{(i)}}$, $\overline{Y_{30,j}^{(i)}}$ és $\overline{Y_{40,j}^{(i)}}$ átlagos ipari termelését megszorozzuk 0,5-del, 0,3-del, illetve 0,2-del (a súlyszámok a mikroszervezetek, valamint a 22, 30 és 40 létszám-kategóriájú szervezetek becsült korrelációjával arányosak), a szorzatokat összeadjuk, az összeget megszorozzuk a megfelelő $c_j$ hányadossal és a mikroszervezetek $N_{0-4,j}^{(i)}$ aktuális számával, majd összegezzük az egyes iparágakhoz tartozó értékeket:

$$Y_{0-4}^{(i)} = \sum_{j=1}^{3} c_j N_{0-4,j}^{(i)} (0{,}5\overline{Y_{22,j}^{(i)}} + 0{,}3\overline{Y_{30,j}^{(i)}} + 0{,}2\overline{Y_{40,j}^{(i)}}).$$

Az előbbi módszerben meghatározó szerepet játszik a 22, 30 és 40 létszám-kategóriájú kisszervezetek, valamint a mikroszervezetek átlagos ipari termelésének hányadosa. Nyilvánvalóan a megfelelő átlagos ipari termelések között akkor van nem formális, oksági kapcsolat, ha az ipari termelés átlaga a ténylegesen működő szervezetekre vonatkozik. Az ilyen mikroszervezetek számát pontosan ugyan nem ismerjük, de a GSZR szerint működő, 22 létszám-kategóriájú és legfeljebb 4 fős szervezetek számából, valamint a ténylegesen működő, 22 létszám-kategóriájú szervezetek számából becsüljük oly módon, hogy feltételezzük, a kétféle szervezetszám aránya a 22 létszám-kategóriában és a mikroszervezetek körében megegyezik.

A közvetlenül nem megfigyelt építőipari mikroszervezetek építőipari termelésének értékét az egész építőiparra vonatkozóan korábban a következőképpen becsültük. Abból a feltételezésből kiindulva, hogy éves szinten az építőipari termelés egy főre jutó értéke mind az 1–4, mind a 0 fős szervezetek körében ugyanakkora, mint a 22 létszám-kategóriában, a korábbi építőipar-statisztikai, munkaügyi és áfa adatok alapján meghatároztuk az 1–4 fős építőipari szervezeteknek az év során állandónak tekintett $\overline{M_{1-4}}$ átlagos létszámát, a 0 fős alkalmazotti létszámú építőipari szervezetek $\overline{M_0}$ átlagos létszámát pedig 1-nek vettük. Az adott *i*-edik tárgyhónapban a meg nem figyelt *l* létszámcsoport (*l* = 1–4, 0) építőipari termelésének $Y_l^{(i)}$ értékösszegét az

$$Y_l^{(i)} = N_l^{(i)} \overline{M_l} \frac{Y_{22}^{(i)}}{N_{22}^{(i)} \overline{M_{22}^{(i)}}} = \frac{\overline{M_l}}{\overline{M_{22}^{(i)}}} N_l^{(i)} \overline{Y_{22}^{(i)}}$$

képlettel határoztuk meg, ahol $N_l^{(i)}$ a tárgyhónapban működő, *l* létszámcsoportú építőipari szervezetek száma, $Y_{22}^{(i)}$ a 22 létszám-kategóriájú építőipari szervezetek építőipari termelésének tárgyhavi értékösszege, $N_{22}^{(i)}$ a tárgyhónapban működő, 22 létszám-

kategóriájú építőipari szervezetek száma, $\overline{M_{22}^{(i)}}$ a tárgyhónapban működő, 22 létszám-kategóriájú építőipari szervezetek átlagos létszáma, $\overline{Y_{22}^{(i)}}$ pedig a 22 létszám-kategória építőipari termelésének egy szervezetre jutó átlagos tárgyhavi értéke.

A korábbi évek tapasztalatai szerint a megfigyelt létszám-kategóriákban az építőipari szervezetek átlagos létszáma és építőipari termelésük értéke jelentős szezonális ingadozást mutat. Ezért módosítottuk a becslést. Mivel hasonló szezonalitás valószínűsíthető a mikroszervezetek körében is, ezért ha már az $\overline{M_l^{(i)}}$ tárgyhavi átlagos létszámokkal nem tudunk számolni a fenti képletben, az $\overline{M_{22}^{(i)}}$ tárgyhavi átlagos létszámok használatától is eltekintünk, és a különböző hónapokban a valóságot minden bizonnyal jobban közelítő, az év során állandó

$$\frac{\overline{M_l}}{\overline{M_{22}}}$$

hányadossal számolunk. Ez lényegében az iparban használt módszer azzal a különbséggel, hogy feltesszük, az építőipari termelés értékének aránya az összes értékesítésen belül nem függ lényegesen a létszámtól. A becslést oly módon is módosítottuk, hogy a számolást alágazatonként végezzük.

## 8. HAVI ÉS NEGYEDÉVES MUNKAÜGYI ISMÉRVEK KÉPZÉSE ÉS TELJESKÖRŰSÍTÉSE

Az 1. fejezetben ismertetett munkaügyi alapismérvekből havi és negyedéves tömegérték- és fajlagos ismérveket képzünk, amelyeknek elvégezzük a teljeskörűsítését. Az egyes kisszervezetekre havonta a következő tömegérték-ismérveket képezzük:

– a teljes munkaidősek létszáma, kereset- és munkajövedelem-tömege, valamint – 2002-től kezdődően – óratömege,

– a nem teljes munkaidősek és az alkalmazásban állók létszáma,

– a teljes munkaidősek, nem teljes munkaidősek és további munkaviszonyban állók egyenértékes létszáma, kereset- és munkajövedelem-tömege, valamint óratömege.

A havonta képzett tömegérték-ismérvekből rétegenként negyedévente képezzük ugyanezeket az ismérveket oly módon, hogy az egyes rétegekre összeadjuk a megfelelő teljeskörűsített havi értékeket. Ugyancsak rétegenként mind havonta, mind negyedévente fajlagos ismérveket is képzünk. A teljeskörűsített tömegérték-ismérvekből az egyes rétegekre havonta az alábbi, egy főre vonatkozó fajlagos ismérveket képezzük:

– a teljes munkaidősek havi átlagkeresete, munkajövedelme és – 2002 óta – óra/fő értéke, – a teljes munkaidősek, nem teljes munkaidősek és további munkaviszonyban állók havi munkajövedelme és óra/fő értéke.

A teljeskörűsített tömegérték-ismérvekből az egyes rétegekre havonta az alábbi, egy órára vonatkozó fajlagos ismérveket képezzük:

– a teljes munkaidősek, nem teljes munkaidősek és további munkaviszonyban állók órás átlagkeresete és munkajövedelme.

A negyedéves tömegérték-ismérvekből rétegenként ugyanezeket az egy főre, illetve egy órára vonatkozó fajlagos ismérveket képezzük.

Mivel a nem válaszoló kisszervezetek közül havonta 0-val pótoljuk azoknak a kisszervezeteknek az alapismérvekre vonatkozó minden adatát, amelyekről az feltételezhető, hogy nemleges jelentést küldtek volna be, értelemszerűen ezeknek a kisszervezeteknek a havi tömegérték-ismérvekre vonatkozó minden adata is 0 lesz. 2002 óta havonta pótoljuk azoknak a nem válaszoló kisszervezeteknek az alapismérvekre vonatkozó hiányzó adatait, amelyek adatait nem pótoltuk 0-val. Ezen adatokból az egyes kisszervezetekre havonta képezzük a tömegérték-ismérveket.

A teljeskörűsítés során becsüljük a havi és negyedéves tömegérték- és fajlagos ismérvek sokasági értékét. A havi tömegérték-ismérvek sokasági értékösszegét úgy becsüljük, hogy összeadjuk a megfelelő alapismérvekre vonatkozó sokasági értékösszegek becslését. A negyedéves tömegérték-ismérvek sokasági értékösszegét úgy becsüljük, hogy összeadjuk a megfelelő havi tömegérték-ismérvekre vonatkozó sokasági értékösszegek becslését.

Mind a havi, mind a negyedéves tömegérték-ismérvek esetén rétegekre együttesen a *Y* sokasági értékösszeget az egyes rétegbecslések összegével becsüljük, továbbá az alapismérvekhez hasonlóan becsüljük a sokasági átlagot az egyes rétegekre és rétegekre együttesen.

A fajlagos ismérvek sokasági értékét úgy becsüljük, hogy képezzük a megfelelő tömegérték-ismérvek sokasági értékösszegére vonatkozó becslések hányadosát.

A havi és negyedéves tömegérték- és fajlagos ismérvek esetén mind az egyes rétegekre, mind rétegekre együttesen meghatározzuk az előző évi hasonló és a (tárgyévi) előző havi, illetve negyedéves sokasági értékösszeghez mint bázishoz viszonyított értékindexeket.

A reprezentatívan megfigyelt rétegekre a havi tömegérték-ismérvek esetén az alapismérvekhez hasonlóan, a negyedéves tömegérték-ismérvek esetén a

$$\sigma_j^2 = \sum_{k=1}^{3} \sigma_{jk}^2 + 2 \sum_{k=1}^{2} \sum_{l=k+1}^{3} co_{jkl}$$

képlet segítségével meghatározzuk az ismérv $\sigma_j^2$ szórásnégyzetét, ahol $\sigma_{jk}^2$ a negyedéves tömegérték-ismérvhez tartozó *k*-adik havi tömegérték-ismérv szórásnégyzete, $co_{jkl}$ a negyedéves tömegérték-ismérvhez tartozó *k*-adik és *l*-edik havi tömegérték-ismérv 2001-ig a

$$co_{jkl} = \frac{1}{m_{jkl} - 1} \left[ \left( \sum_{i=1}^{m_{jkl}} y_{jik} y_{jil} \right) - m_{jkl} \overline{Y_{jk}} \cdot \overline{Y_{jl}} \right]$$

2002 óta a

$$co_{jkl} = \frac{1}{m_{jkl} - 1} \left[ \left( k_j^2 \sum_{i=1}^{m_{jkl}} y_{jik} y_{jil} \right) - m_{jkl} \overline{Y_{jk}} \cdot \overline{Y_{jl}} \right]$$

képlet segítségével meghatározott kovarianciája, $m_{jkl}$ a *j*-edik réteg mintába kiválasztott kisszervezetei közül azoknak a száma, amelyek mind a *k*-adik, mind az *l*-edik hónapban válaszolók, nemlegesnek vagy – 2002 óta – nem nemlegesnek pótoltak voltak, $y_{jik}$ és $y_{jil}$ az *i*-edik ilyen kisszervezet *k*-adik, illetve *l*-edik havi adata, $\overline{Y_{jk}}$ és $\overline{Y_{jl}}$ pedig a *k*-adik, illetve *l*-edik havi sokasági átlag. Mind a havi, mind a negyedéves tömegérték-ismérvekre az alapismérvekhez hasonlóan meghatározzuk a relatív szórást és becsüljük a standard hibát.

Mind a havi, mind a negyedéves tömegérték-ismérvek esetén rétegekre együttesen a standard hiba a rétegenkénti standard hibák négyzetösszegének négyzetgyöke.

Mind az egyes reprezentatívan megfigyelt rétegekre, mind rétegekre együttesen a

$$\sigma_{Y_j} = f_j \frac{\sigma_{Z_j}}{X_j}, \qquad \text{illetve a} \quad \sigma_Y = \frac{\sqrt{\sum_j (X_j \sigma_{Y_j})^2}}{X}$$

képlet segítségével meghatározzuk a havi fajlagos ismérvek standard hibáját, ahol $\sigma_{Z_j}$ a fajlagos ismérv számlálójában szereplő (havi) tömegérték-ismérv standard hibája, $X_j$ és $X$ a fajlagos ismérv nevezőjében szereplő (havi) tömegérték-ismérv sokasági értékösszege a *j*-edik rétegben, illetve a rétegek együttesében. Az első képletben szereplő $f_j$ szorzót az 5. fejezet végén leírtakhoz hasonlóan számítjuk ki.

Mind az egyes reprezentatívan megfigyelt rétegekre, mind rétegekre együttesen a fenti képletek segítségével közelítőleg meghatározzuk a negyedéves fajlagos ismérvek standard hibáját, ahol $\sigma_{Z_j}$ a fajlagos ismérv számlálójában szereplő (negyedéves) tömegérték-ismérv standard hibája, $X_j$ és $X$ a fajlagos ismérv nevezőjében szereplő (negyedéves) tömegérték-ismérv sokasági értékösszege a *j*-edik rétegben, illetve a rétegek együttesében. Az $f_j$ szorzót az

$$f_j = \sqrt[3]{\prod_{k=1}^{3} f_{jk}}$$

képlet segítségével határozzuk meg, ahol $f_{jk}$ a negyedéves fajlagos ismérvhez tartozó *k*-adik havi fajlagos ismérv szorzója.

A havi és negyedéves tömegérték- és fajlagos ismérvek esetén mind az egyes reprezentatívan megfigyelt rétegekre, mind rétegekre együttesen meghatározzuk a relatív hibát.

IRODALOM

CSEREHÁTI Z. [2004]: Outlierek meghatározása és kezelése gazdaságstatisztikai felvételekben. *Statisztikai Szemle*. 82. évf. 8. sz. 728–746. old.

ÉLTETŐ Ö. – MESZÉNA Gy. – ZIERMANN M. [1982]: Sztochasztikus módszerek és modellek. Közgazdasági és Jogi Könyvkiadó. Budapest.

ÉLTETŐ, Ö. – MARTON, Á. – MIHÁLYFFY, L. – TELEGDI, L. [1997]: Sampling surveys in Hungary. *Statistics in Transition*. 3. évf. 2. sz. 267–279. old.

TELEGDI L. [1993]: Az ipari és építőipari kisszervezetek reprezentatív megfigyelése. *Statisztikai Szemle*. 71. évf. 3. sz. 226–244. old.

TELEGDI L. [1999]: A nemválaszolás megelőzése és kezelése a gazdaságstatisztikában. I–II. *Gazdaság és Statisztika*. 11. (50.) évf. 4. sz. 43–64. old. és 5. sz. 28–56. old.

TELEGDI L. [2001]: Az ipari és építőipari kisszervezetek reprezentatív megfigyelésének egy évtizede. Központi Statisztikai Hivatal. Budapest.

## SUMMARY

The author reviews the sub-annual sampling survey of small enterprises in Hungary in the recent decade. The paper deals with general characteristics of the survey, stratification and selection of the sample. Data collection and methods of estimation are discussed as well.

# AZ ADATFELFEDÉS ELLENI VÉDELEM STATISZTIKAI ESZKÖZEI

ERDEI VIRÁG – HORVÁTH ROLAND

A tanulmány az adatfelfedés elleni védelem statisztikai eszközeit mutatja be, az adatvédelem problémáinak tárgyalása mellett és azok összefüggésében. Ismerteti az adatvédelem európai és magyar jogi alapjait, a tájékoztatási formák bővülését is. Az eszközök, módszerek tárgyalásakor sor kerül a táblázatos- és mikroadatokban lévő adatfelfedési kockázat, majd a táblázatos adatokra vonatkozó védelmi eszközök és a mikroadat védelem különböző módjainak ismertetésére, gyakorlati példákon keresztül.

TÁRGYSZÓ: Adatvédelem. Adatfelfedés elleni statisztikai eszközök

Talán nem szerénység azt állítani, hogy lassan immár 15 éves demokráciánkban az adatvédelem szó mindenkinek ismerősen cseng. Rádió- és tévéműsorok állandó szereplője az adatvédelmi biztos, gyakran újságok vezető híre az adatvédelemmel kapcsolatos valamely aktuális téma. Az utca embere természetesen azt látja a kifejezés mögött, hogy a korábbi mindent tudó állammal szemben napjainkban már inkább a semmit nem tudó állam áll. Érdekvédő szervezetek és jogászok hada áll szemben az állammal, illetve minden egyéb magán illetve hivatalos szervvel, amennyiben az szeretne valami nem jogában állót megtudni rólunk, hiszen személyes adataink védettek, mi rendelkezünk felőlük, és jogi felhatalmazás híján nehezen tudható meg tőlünk bármi is.

A magyarországi demokrácia érésének folyamán a személyes adatok védelme volt az első, amit mindenki megismert, de az évek során az egyre tudatosabban viselkedő az állampolgárok annak is tudatában kezdenek lenni, hogy a közérdekű adatok nyilvánosságához is joguk van, valamint általában az információhoz. Állampolgárként ugyanolyan vehemenciával igényelhetünk információkat, mint amilyen mértékben ragaszkodunk személyes adataink védelméhez. Az Európai Unióhoz történő csatlakozás nyomán Európa és a világ kitágul számunkra. Nő az információigényünk, egyre jobban tisztában vagyunk a jogainkkal és a lehetőségeinkkel. Számítani lehet arra, hogy a csatlakozás hatására az emberek egyre jobban felmérik a lehetőségeiket, és élni is fognak velük, például egyre több információt fognak igényelni.

Az adatgyűjtők, így a statisztikai hivatalok is, hatalmas adatvagyonnal rendelkeznek, mégis egyes becslések szerint ennek csak 30–40 százaléka hasznosul, kerül nyilvánosságra. Ennek egyik fő oka az adatvédelem. Az informatika óriási térnyerése következté-

ben megnőtt az adatfelfedés lehetősége. A statisztikai hivatalnak meg kell felelnie a törvényi adatvédelmi kötelezettségeknek, s ezt annál is inkább meg kell tennie, mivel egy esetleges adatfelfedés nagyban aláásná az adatszolgáltatói bizalmat, és így a statisztikai tevékenységet. A kötelező adatvédelemmel szemben azonban az információszabadság állampolgári joga áll.

Tanulmányunk célja, hogy bemutassuk az adatfelfedés elleni védelem statisztikai eszközeit. Ezen eszközök birtokában válhat lehetővé a minél szélesebb körű biztonságos adatközlés, az adatfelfedés egyidejű elkerülésével.

## AZ ADATFELFEDÉS ELLENI VÉDELEM KÖRNYEZETE

A statisztikai és egyéb adatgyűjtések célja az adatok elemzés, feldolgozás utáni nyilvánosságra hozatala. Ez az adatfelvételek végső és legérzékenyebb pontja. A magyar statisztikai törvény kimondja, hogy a statisztikai módszerekkel felvett, feldolgozott, tárolt és elemzett adatok az államhatalmi és a közigazgatási szervek, valamint a társadalom szervezetei és tagjai tájékoztatását szolgálják.

### *Tájékoztatási kötelezettség, szélesedő lehetőségek*

Az állami, központi költségvetésből finanszírozott statisztikai szerveknek a törvényben rögzítetten túl erkölcsi kötelessége is az adatok legteljesebb mértékű közzététele, hiszen az adatokat mi, állampolgárok térítés nélkül szolgáltatjuk, és az állam statisztikákat felhasználó tevékenysége, munkája is a mi érdekünkben történik. A magyar statisztikai törvény szerint a hivatalos statisztikai szolgálathoz tartozó szervek által végrehajtott adatgyűjtések eredményei – az adatvédelemre vonatkozó szabályok betartása mellett – nyilvánosak.[1]

A nyilvánosságra hozatal, a tájékoztatás „kiadványokból és más adathordozókon lévő adatállományokból történő közlésekből áll".[2] A papír alapú tájékoztatás magában foglalja a különböző kiadványokat, évkönyveket, tájékoztatókat, brosúrákat stb., ám napjainkban a gyors és nagy információigény miatt egyre inkább tért nyer az egyéb adatközlés. Ilyenek az internetes adatközlés, a CD-k, és az egyéb nem papír alapú adathordozók, de akár a telefonon keresztül történő adatszolgáltatás is.

Az adatközléseknél különbséget tehetünk aszerint is, hogy azok egy konkrét „legyártott" adatot, táblázatot tartalmaznak, vagy a felhasználó, adatkérő közreműködésével egy állományból egyedi beállítás alapján lekérhető adatokat, táblázatokat. Az adatközlés egy harmadik típusa a mikroadat-állomány közzététele, amely rekordsorosan tartalmazhat egy adatfelvételt vagy annak egy részét.

A tájékoztatás kötelezettségét, annak alapelveit az uniós statisztikai jogszabályok is részletesen rögzítik. A tájékoztatási tevékenységgel kapcsolatban megfigyelhető az a tendencia, hogy egyre nyitottabbá válnak a statisztikai szervezetek, egyre több adat kerül nyilvánosságra. Egyre több formában válik lehetségessé a tájékoztatás, egy-egy adat, szám közlése mellett egyre részletesebb összesítések, táblázatok jelennek meg, és akár teljes adatállományok is hozzáférhetővé válnak. Ennek konkrét bizonyítéka, hogy euró-

[1] 1993. évi XLVI. Törvény a statisztikáról 17.§ (1)
[2] 1993. évi XLVI. Törvény a statisztikáról 23.§ (2)

pai uniós szinten jogilag is megnyílt a lehetőség a kutatók, tudományos élet képviselői előtt, hogy bizalmas, egyedi adatokhoz férjenek hozzá. (Az Európai Unió 1997-ben született statisztikai törvénye már megfogalmazta ezt a lehetőséget (17. cikk), 2002-ben azonban rendelet is született, amely részletezi azt.)

A 831/2002/EK rendelet a bizalmas adatokhoz való tudományos célú hozzáférésről[3] lehetővé teszi a közösségi hatóság (az Európai unió statisztikai hivatala, más néven Eurostat) hivatali helyiségeiben a bizalmas adatokhoz való hozzáférést, és anonimizált mikroadatok kibocsátását is. Egyetemek, felsőoktatási intézmények, tudományos kutatással foglalkozó szervezetek, intézmények, hivatalok, számára nyitott ez a lehetőség (részletesen lásd 831/2002/EK rendelet 3. cikk). A rendelet az adatokhoz való hozzáférés módjáról, engedélyezéséről szól, annak érdekében, hogy pontosan tudható legyen – az adatok bizalmas volta miatt –, hogy az adathozzáférés folyamán ki mikor jut hozzá valamihez és mi alapján, mit tehet, mik a kötelezettségei stb.

Adatvédelmi intézkedések természetesen itt is vannak, a mikroadatok kiadásakor eltávolítják a közvetlen azonosítókat, és a rendelkezésre álló legjobb eljárás alkalmazásával minimálisra csökkentik az érintett statisztikai egységek közvetett azonosításának veszélyét. Az Eurostat hivatali helyiségeiben (a gyakorlatban kutatószoba) engedélyezhető hozzáférés pedig mindig csak hivatalos személy felügyelete mellett történhet, és a kutatás eredményeit – mielőtt kikerülnek az intézményből – ellenőrzik, biztosítva, hogy azok nem tartalmaznak bizalmas adatokat.

Nevesítve a hozzáférés négy felmérésből, illetve statisztikai adatforrásból lehetséges: a közösségi háztartási panelből, a munkaerő-felmérésből, a közösségi innovációs felmérésből és a szakmai továbbképzési felmérésből. (Az adatszolgáltató nemzeti statisztikai hivatalok megtagadhatják az adataikhoz történő hozzáférést, de engedélyezhetik is a felsoroltaktól eltérő bizalmas adatokhoz való hozzáférést.)

A rekordsoros adatokhoz történő hozzáférés nagyon nagy nyitottságot jelent az adatgazda statisztikai hivataloktól a tájékoztatásban, ezért is követeli meg a legszigorúbb adatvédelmet.

## DEFINÍCIÓK

A cikkben tárgyalt statisztikai információk bizonyos jogszabályokon alapulnak, így azokra támaszkodunk mi is. Az előbbiekben használtuk a *bizalmas adatok* kifejezést. A következőkben ismertetjük a *személyes adat*, a *bizalmas adat* és az *azonosíthatóság* fogalmát, amelyek témánk szempontjából meghatározóak. A magyar adatvédelmi, statisztikai törvények, így a fogalmak is számos európai jogszabály és ajánlás alapján születtek. Először az európai uniós megfogalmazások lényegét ismertetjük, majd röviden a hazairól szólunk.

*Személyes adat:* A személyhez kapcsolódó adat definíciója alapvetően fontos, hiszen a statisztikai felmérések nagy része emberekre vonatkozik. A fogalom igen jól körülírható. „Személyes adat bármely, azonosított vagy azonosítható természetes személyre (’adatalany’) vonatkozó információ; a személy különösen akkor tekinthető azonosíthatónak, ha őt – közvetlenül vagy közvetve – azonosítószám vagy egy vagy több fizikai, fizi-

[3] A Bizottság 831/2002/EK Rendelete (2002. május 17.) a bizalmas adatokhoz való tudományos célú hozzáférés tekintetében a közösségi statisztikáról szóló 322/97/EK tanácsi rendelet végrehajtásáról.

ológiai, mentális, gazdasági, kulturális vagy szociális azonosságára jellemző tényező alapján azonosítani lehet."[4]

*Bizalmas adat:* Ez ugyancsak kulcsfontosságú fogalom, hiszen ez alapján definiálhatjuk majd a statisztikai titkosságot. A bizalmas adat a személyes adatnál bővebb kategória. A személyes adaton túl egyéb adatok is beletartoznak, pl. a gazdasági szervezetek adatai. A bizalmas adat lényeges tulajdonsága, hogy az a megfigyelési egységekre – személyekre, cégekre stb. – vonatkozó adat, információ. Az „adatok bizalmasnak tekintendők, amennyiben segítségükkel a statisztikai egységek akár közvetlenül, akár közvetve azonosíthatók és így egyedi információt fednek fel."[5]

A bizalmas – egyes szövegekben védettnek nevezett – adat alapján a *statisztikai titkosság* magának a tevékenységnek, az egyes statisztikai egységekkel kapcsolatos adatoknak a védelme.

A bizalmas adat tehát megköveteli, hogy ne lehessen sem közvetlenül, sem közvetve azonosítani a vonatkoztatási, statisztikai tárggyal. (A bizalmas adatokhoz való tudományos célú hozzáférésről szóló rendeletben bizalmas adatok alatt már csak a közvetett azonosíthatóságot értik, hiszen a statisztikai munkában a közvetlen azonosítást a feldolgozási folyamat elején lehetetlenné teszik, illetve az idősoros elemzéseknél külön kezelik az azonosítókat.)

A nemzetközi joganyagok fogalmai egységesek a tekintetben, hogy megkövetelik a közvetlen azonosítók leválasztását, illetve, hogy az egyértelmű azonosíthatóságot és a lehetséges azonosíthatóságot is a fogalom részévé teszik. Az igazán lényegi információt azonban azok a meghatározások adják, amelyek magáról a kikövetkeztethetőségről, azonosíthatóságról szólnak.

*Azonosíthatóság:* A közvetlen azonosíthatóság egyértelműen definiálható az egyedi azonosítók leválasztásával (személyeknél: név, lakcím; gazdasági szervezeteknél: név, telephely vagy azonosítószám).

A közvetett azonosíthatóságról vagy felfedésről már csak durva körülhatárolás lehetséges:

– „A statisztikai egység azonosíthatóságának megállapításakor figyelembe kell venni mindazokat az eszközöket, amelyeket egy harmadik fél ésszerűen *(reasonably)* igénybe vehet az említett statisztikai egység azonosításához."[6] (A harmadik fél úgy értendő, hogy az első két fél az adatszolgáltató és a statisztikai hivatal, hiszen ők jogosultak az adatot ismerni.)

– „A személy nem tekinthető azonosíthatónak, ha az azonosítása ésszerűtlenül hosszú időt és munkabefektetést igényel."[7]

Magyarországon két alaptörvény szabályozza a kérdéskört, a statisztikai törvény (1993. évi XLVI. Törvény), valamint az adatvédelmi törvény (1992. évi LXIII. Törvény), hivatalos nevén Törvény a személyes adatok védelméről és a közérdekű adatok nyilvánosságáról.

Az adatvédelmi törvény határozza meg a személyes adatot.

[4] Az Európai Parlament és a Tanács 95/46/EC Irányelve az egyénnek a személyes adatok feldolgozásával kapcsolatos védelméről és ezeknek az adatoknak a szabad áramlásáról 2. cikk (a.)

[5] A Tanács 1997. február 17-i. 322/97. (EK) számú rendelete a közösségi statisztikákról 13. cikk (1)

[6] A Tanács 1997. február 17-i. 322/97. (EK) számú rendelete a közösségi statisztikákról V. fejezet 13. cikk

[7] A tagállamok minisztereinek bizottsága által 1997. szeptember 30.-án elfogadott 97/18 sz. ajánlás a statisztikai célból gyűjtött és feldolgozott személyes adatok védelméről Fogalmak 1. bekezdés

*Személyes adat:* bármely meghatározott (azonosított vagy azonosítható) természetes személlyel kapcsolatba hozható adat, az adatból levonható, az érintettre vonatkozó következtetés. A személyes adat az adatkezelés során mindaddig megőrzi e minőségét, amíg kapcsolata az érintettel helyreállítható. A személy különösen akkor tekinthető azonosíthatónak, ha őt – közvetlenül vagy közvetve – név, azonosító jel, illetőleg egy vagy több, fizikai, fiziológiai, mentális, gazdasági, kulturális vagy szociális azonosságára jellemző tényező alapján azonosítani lehet.[8]

A statisztika és a statisztikai törvény azonban a nemzetközi gyakorlat alapján védi a többi adattípust is, például a gazdasági szervezetek adatait. Ennek érdekében bevezeti az egyedi adat fogalmát és azt védi.

*Egyedi adat:* a statisztikai célt szolgáló, a természetes és a jogi személy, valamint a jogi személyiséggel nem rendelkező adatszolgáltatóval kapcsolatba hozható adat.[9] Egyedi adat tehát az, ami a nemzetközi joganyagokban bizalmas vagy védett adat. (A jelenleg folyó uniós jogszabályok fordításában elképzelhető, hogy a bizalmas adatok helyett egyedi adat szerepel majd.) Egyedi adat csak statisztikai célra használható.

*Azonosíthatóság:* A hazai gyakorlat, jog is elsődleges védelmi kritériumként az egyedi azonosítók leválasztását követeli meg. Az azonosítók leválasztása a közvetlen azonosítás megakadályozását szolgálja: „A természetes személy személyére vonatkozó adatgyűjtésnél az érintett nevét és a lakcímét (személyazonosító adat) – kivéve azt, amelynek adathordozóját a levéltári anyag védelmére vonatkozó jogszabály értelmében levéltári őrizetbe kell adni – a statisztikai feldolgozás befejezésekor, az adatok teljességének és összefüggésének ellenőrzését követően, de legkésőbb a tárgyidőszakot követő egy éven belül kell törölni, adatátadás esetén ezt megelőzően is."[10]

(„Az egy évnél hosszabb időszakra vonatkozó idősoros vizsgálatok esetében az adatállományt belső azonosítóval kell ellátni, amelyből az érintett személyazonossága nem állapítható meg. Az érintett személyazonosító adatait az adatállománytól elkülönítetten kell kezelni."[11])

A gazdálkodó szervezet akkor tekinthető anonimnak, ha elnevezése és telephelye nincs feltüntetve (*Statisztikai igazgatás* [2000]).

Egyetlen kritérium van a statisztikai törvény végrehajtási rendeletében, amely a közvetett azonosítást kívánja megakadályozni. Azt mondja a szabály, hogy összesítve sem lehet nyilvánosságra hozni olyan adatot, amelynél az adatszolgáltatók száma háromnál kevesebb.[12]

A jogszabályok definíciói után szeretnénk tisztázni egy, a gyakorlatban elterjedt félreértést. Az adatvédelem során gyakori, hogy megkülönböztetik a jogi védelmet a technikai védelemtől, mondván, hogy amikor például egy szerződést ír alá valaki egy adathozzáférésről, akkor az jogi védelem, míg amikor beavatkozást végzünk egy táblázaton, vagy adatbázison, akkor az technikai. A valóságban ez a két dolog nem különíthető így el, hanem egyik a másikon alapul. A jogszabályok megfogalmazzák a kereteket, fogalmakat, teendőket, s ennek alapján készülnek a gyakorlatban technikák, módszerek azok megvalósítására.

[8] 1992. évi LXIII. Törvény a személyes adatok védelméről és a közérdekű adatok nyilvánosságáról 2.§ 1.
[9] 1993. évi XLVI. Törvény a statisztikáról 17.§ (2)
[10] 1993. évi XLVI. Törvény a statisztikáról 19.§ (1)
[11] 1993. évi XLVI. Törvény a statisztikáról 19.§ (2)
[12] 1993. évi XLVI. Törvény végrehajtásáról szóló 170/1993. (XII. 3.) Kormány rendelet 19.§

## AZ ADATKÖZLÉS PROBLEMATIKÁJA

Az adatközlés egyik, nagy problémát jelentő kérdése a *közvetett azonosíthatóság*, azaz az adatfelfedés lehetősége. Maga az *adatvédelem* az a technika vagy módszer, amely alkalmazásával minimálisra csökkenthető a statisztikai egységek azonosításának veszélye.

Az adatközlés során az adatvédelmet készítők maguk döntik el, hogy a jogszabályban megfogalmazott „nagy időbefektetés során lehetővé válható kikövetkeztethetőség" mikor válhat lehetségessé. A közvetett felfedés elleni védekezés bonyolult, komoly munkát igényel, hiszen egy külső, harmadik fél technikai és tudásbeli hátterével szemben kell eszközöket találni. A külső fél, a lehetséges adatfelfedő jó- és rosszindulatú is lehet, különféle motivációkkal és eszközökkel. Az adatközlés számos publikációs formában ölthet testet, a papír alapútól az internetes közlésen át, és ezek eltérő védelmi technikákat, stratégiákat igényelnek.

Az adatfelfedés teljes mértékű megakadályozása által tökéletesen lehetetlenné válna az adatközlés, az adatokhoz való hozzájutás. Az egyre biztonságosabb adatközlés, az egyre nagyobb védelem mindig együtt jár azzal, hogy egyre több és több adatot kell elzárni a felhasználók elöl, és végül az elrejtett információknak köszönhetően használhatatlanná válhatnak adatbázisok.

A cél és egyben a legnagyobb kihívás a felfedés elleni védekezésben az, hogy megtaláljuk azt az optimális arányt az elrejtett, védett és a tájékoztatás révén közölt adatok közt, amivel már biztonságosnak tekinthetőek az adatok, és a felhasználók is hozzájuthatnak a megfelelő részletezettségű információkhoz. Ehhez ismernünk kell, hogy milyen kockázat rejlik a különféle adatközlésekben, és kik lehetnek a felhasználók *(Eurostat* [1999]).

(Azonosításon, azonosíthatóságon azt értjük, hogy egy anonim információhoz valamilyen módon hozzárendelhető, hozzákapcsolható egy egyedi azonosító (azonosítószám vagy kulcs). E mellett az adatfelfedés azt jelenti, hogy egy személyre vagy egy intézményre vonatkozóan új, plusz információ birtokába jutunk az azonosítás által. A két kategória tehát egymásból következik, hiszen plusz információ birtokába akkor jutunk, ha azonosítjuk a személyt. Tanulmányunkban mi e két kategóriával, s a kialakítandó védelemmel együtt foglalkozunk.)

### *Felfedési lehetőségek és kockázatok*

Az informatika nagyfokú elterjedtségének és technikai fejlődésének következtében a közölt adatok analizálásával, kombinálásával olyan új információ birtokába juthat egy külső, harmadik személy, amelyet az adatközlőnek nem állt szándékában közölni. Az adatok felfedése, kikövetkeztethetősége az adatok egyedisége, bizalmassága miatt kockázatossá válhat.

A területi szintű tájékoztatásban kiemelten jelentkezik a probléma: a terület nagysága, az alacsony lélekszám, vagy az adattartalom miatt válik nem közölhetővé az adat. Például:

– Ritka foglalkozások közlése (például: a budapesti agglomeráció egyik kis településén élő operaénekesnő közölt adatai név nélkül egyértelmű felfedést jelentenek).

– Egy átlagos foglalkozású (például bolti eladó) ember is azonosíthatóvá válik, ha csak egy emberről van szó a területen.

– Ugyancsak védendőek bizonyos egyedi, ritka családi vagy egyéb körülmények kis területre vonatkozó adatközlésben (például: 8 gyermekes család; magas jövedelmű személy).

Gazdasági szervezetek adatközlésénél számos probléma merülhet fel. A legkiemelkedőbb a dominancia problémája, vagy a monopol pozíciójú szervezetek, cégek adatai. Azonos jellemzőkkel rendelkező, azonos adatszolgáltatói csoportba tartozó, azonos terméket gyártó, azonos szolgáltatást nyújtó gazdálkodó szervezetek adatai statisztikai összesítés formájában bármikor közölhetők, ám amint valamelyik szervezet egyik mutatója kiugró, domináns értékkel bír (például legmagasabb foglalkoztatotti szám, legnagyobb bevétel, előállított egyedi termék stb.), akkor érzékennyé válik az adat.

Közérdekű és védendő adat együttes közlése során is felmerülhetnek adatvédelmi aggályok. (A Központi Statisztikai Hivatal pontosan felsorolja a közérdekű adatok körét.[13]) Közérdekű adat például a központi vagy helyi önkormányzati költségvetésből finanszírozott bölcsődei ellátásra vonatkozóan az ellátók száma, az ellátottak száma, a forgalom, a befogadóképesség és az ellátottak által fizetett hozzájárulás összesen. Amennyiben egy településen három bölcsőde működik, amelyből kettő állami és egy magán, akkor nagyon megfontoltnak kell lenni a felsorolt adattípusok együttes közlésekor, hiszen az egy magán bölcsőde adata így felfedhetővé, azaz nyilvánossá válik.

A mintavételes felvételek védelmét gyakran feleslegesnek tartják, holott a tájékoztatás módjától függően bizonyos esetekben védendő adattá válnak:

– Ha a megszerzett adatokból, tehát a mintából becslünk egy tulajdonságot egy legalább három fős sokaságra, akkor ezek az adatok nem lesznek védendőek, még akkor sem, ha becsült (és egyben a tájékoztatott) adatok egybeesnek valamely mintaelemmel.

– Abban az esetben viszont, ha a mintaelemeket „nyersen", mikroadat formájában szeretnénk közreadni, védelemmel kell ellátni őket. Gondoljuk csak el, hogy a szomszédunk elmeséli, hogy egy kérdezőbiztos a napi időbeosztásáról és tevékenységéről érdeklődött. Amennyiben birtokában vagyunk egy-két alapinformációnak szomszédunkról, az illető könnyen beazonosíthatóvá válik az adatbázis segítségével.

Előfordulhat, hogy az adatszolgáltatók magas száma ellenére is védenünk kell a cellát, például a kategóriák alacsony száma miatt. Olyan kérdésnél, amire igennel illetve nemmel lehet válaszolni, vagy kevés számú válaszlehetőség van – különösen, ha az adott kérdés valamely kényes, különleges dologra kérdez rá (például betegség, vallási hovatartozás, politikai vélemény) –, fokozottan figyelnünk kell. Ha ugyanis minden válaszadó azonosan, mondjuk igennel válaszol, akkor, amennyiben a többi, nem érzékeny kérdésre adott válaszból felfedünk valakit, akkor arról az egyedről olyan plusz, és érzékeny információ birtokába jutunk, aminek nem kellene tudtunkra jutnia: például, hogy milyen betegsége van, droghasználó-e vagy sem, milyen vallási közösség tagja stb. (Ez a fajta adatvédelmi probléma egyébként meglehetősen ritkán merül fel a nagy esetszámok miatt.)

Problémát okozhat az ugyanazon kiadványban vagy ugyanazon adatbázison alapuló különböző adatközlésekben a különböző táblák összeolvasásából azonosítható adatszolgáltató. Annak megoldása, hogy a keresztinformációkból ne váljon kikövetkeztethetővé az adatszolgáltató, a nagyon pontosan megtervezett és nyomon követett adatközléseken múlik.

## FELHASZNÁLÓK ÉS MOTIVÁCIÓK

Az adatok felhasználói az igényelt adatok és az igénylés módja szerint jól definiálható csoportokra oszlanak. Fontos ismernünk a különböző felhasználókat, hogy tudjuk, kitől, mikor és miért várható az adatfelfedés, és milyen következményekkel számolhatunk.

[13] IV/1997. (SK 3.) KSH szabályzat a statisztikáról szóló jogszabályokból adódó feladatok végrehajtásáról X.

A felhasználók alábbi osztályozását, amelyet az OECD a legjobbnak minősített, a dán statisztikai hivatal állította össze:

*1. „Farmerek" (szerződéses ügyfelek)*: Mindig ugyanarra a statisztikai adatra, szolgáltatásra van szükségük, általában ciklikusan. Az adatok közt nem válogatnak, csak olvassák azokat. Az igényeik kielégítése védett lekérdezés biztosításával, gyors adatátadással (például e-mailen keresztül), speciális információk előjegyzésével, valamint speciális formák (általuk elkészített egyéni táblázatok) alkalmazásával történik. A felhasználóknak ebbe a körébe tartoznak a pénzügyi szektor, a gazdálkodó és egyéb szervezetek képviselői, akik a statisztikán keresztül általában saját szakterületük alakulására kíváncsiak.

*2. „Turisták" (alkalmi böngészők)*: Ezek a felhasználók általános statisztikai adatokat igényelnek különböző területekről, témákról. Az adatok, dokumentumok könnyű, gyors elérésében érdekeltek. Mindenképpen szükséges számukra köznapi fogalmak és nyelvezet alkalmazásával magyarázatot fűzni a számokhoz. Ők azok, aki csak „felnéznek" az Internetre, és ők teszik ki a felhasználók mintegy 15 százalékát. Közéjük tartoznak a köznapi emberek, a diákok, a sajtó mindenre kíváncsi munkatársai.

*3. „Bányászok" (szakértők)*: Mélyre ásnak az adatokban, igénylik a minél részletesebb metszeteket. Vizsgálataikat sok és részletes információval, bizonyítékkal szeretnék alátámasztani. Egyedi adatokra is szükségük van, amelyekhez külön szerződés keretében hozzá is juthatnak. Elsősorban a kutatók és a tervezők tartoznak közéjük.

Az adatfelfedés szempontjából a farmerek felől érkező támadás a legvalószínűtlenebb. A *Bányászok* csoportba sorolható egyre több és több részletes adatkérőt viszont már potenciális támadónak kell tekintenünk. A legnagyobb – akár jó- vagy rosszindulatú – támadás a turisták, alkalmi böngészők csoportja felől érkezhet. Ugyanis a diákok, egyetemisták akár véletlenül is felfedhetnek bizalmas adatot, a média embere pedig szándékosan is kereshet bizalmas, védendő adatot.

Az adatfelfedést motiválhatják társadalmi, gazdasági, pszichológiai, politikai tényezők. Az adatvédelmi stratégia kialakításához fontos ezen tényezők ismerete.

Az azonosítási kísérletek elkövetőit két nagy csoportba sorolhatjuk a szerint, hogy elsősorban információszerzés céljából követik-e el a támadást, vagy pedig az adatgyűjtések, a statisztikai hivatal lejáratása, ilyen módon a közbizalom rontása a céljuk.

A támadás eszközei is széles skálán mozognak, a statisztikai és informatikai ismeretek alkalmazásától a nagy fokú számítógépes támogatásig. Ehhez járulnak a meglévő, mindenki számára elérhető információk, a köztudomású tények, a nyilvános adatbázisok, és az egyéb ismeretek egy témáról.

Természetesen minden egyes statisztikai felvétel esetén maghatározhatók támadási okok, motivációs célok, attól függően, hogy milyen érzékeny vagy érzékenynek vélhető kérdések szerepelnek a kérdőívben.

Egy angol tanulmány – elsősorban a népszámlálási adatok feltörése kapcsán – a következő lehetséges forgatókönyveket állapítja meg az adatfelfedési motivációkra, célokra nézve (*Elliot* [1996]):

*1*. Adatbázis gazdagítása népszámlálási adatokkal;

*2*. Adatbázis adatainak összevetése és megerősítése (lopott) népszámlálási adatokkal;

*3*. Egy jó újságíró sztori megírása annak bizonyítására, hogy adataink mennyire nem védettek;
*4*. Az állami adatgyűjtéseket és a mindenkori kormányzat hitelét rontó támadások;
*5*. Személyazonosító (szám vagy kulcs) ellopása;
*6*. Gazdasági versenytárs adatainak azonosítása.

(A példa az angol népszámlálásra vonatkozik.)

Az adatfelfedés hatása természetesen függ a behatoló céljától, a kísérlet sikerétől vagy kudarcától. Legyen azonban szó akárcsak egy adatszolgáltató személy azonosításáról, vagy egy gazdasági versenytárs adatainak megszerzéséről, az adatbiztonság mindenképpen sérül. Egy adatfelfedési kísérlet megtörténtének nyilvánosságra hozatala mindenképpen rontja a közhangulatot – akár sikeres volt a kísérlet, akár nem –, hiszen maga a tény, hogy egy behatolás (azonosítás) véghezvihető, önmagában rontja a statisztikai szolgálat hitelét. Ennek eredménye végső soron az lehet, hogy megszűnik maga a minőségi adatszolgáltatás.

## ADATVÉDELMI TECHNIKÁK, TÁBLÁZATOS ÉS MIKROADAT VÉDELEM

A természetes jóindulatú állampolgári adatigény és az ezzel szemben jelentkező rosszindulatú támadások, lejáratások miatt az adatközlőknek szükségük van egy olyan stratégiára, amellyel biztonsággal közölhetnek adatokat, és minimálisra csökkentik a támadható felületet.

A magyar jogszabály által nevesített szabály, azaz a minimum 3 elem egy cellában jó és szükséges védelmi szabály, de nem minden esetben nyújt elégséges védelmet a felfedés ellen. A világ országaiban számos jól bevált technika létezik az adatfelfedés megakadályozására. Ezek alapjait mutatjuk be a következőkben.

Az adatfelfedhetőség szempontjából alapvetően kétféle tájékoztatási formát különböztetünk meg. Az első, hagyományosnak mondható forma az, amikor táblázatos formában, azaz bizonyos dimenziók (tulajdonságok) kereszthivatkozásaiban hozzuk nyilvánosságra az adatokat. A másik esetben az adatokat rekordsorosan tesszük közzé. Ez utóbbit nevezzük mikroadatoknak. Ennek megfelelően beszélünk *táblázatos*-, illetve *mikro*-adatvédelemről. A kétféle védelem alapjait tekintve hasonlít egymásra, mégis különböző technikák alkalmazását igénylik.

### *Kockázat*

Az adatvédelem kialakításának első lépéseként felmérjük és megbecsüljük az adatközlésben felmerülő felfedési kockázatot. Ehhez ismernünk kell az adott statisztikai felvétel összes jellemzőjét: a sokaságot, az elemszámot, az esetleges mintát, a mintakiválasztás módját, a változókat, az adatokat, a főbb megoszlásokat stb. Ennek ismeretében tudunk dönteni valamelyik védelmi technika mellett, és határozhatjuk meg az adatvédelmi stratégiát.

Adatfelfedés több tényező együttes jelenléte esetén történhet meg, így a felfedési kockázat becslésére is több mód kínálkozik.

*Táblázatos adatokban rejlő kockázat*

A felfedés kockázata a táblázatos adatoknál párhuzamban áll a cella érzékenységének kritériumával. Az egyes cellák érzékenységének megléte és mennyisége határozza meg a kockázat mértékét. A gyakorlatban négy alapvető módszer terjedt el arra, hogy egy celláról kiderítsük, szükséges-e védeni vagy sem (*Carlson* [2002], *Merola* [2003]).

*Jelölések:*

$n$ – az adott cellába tartozó adatszolgáltatók száma,
$z_1 \geq z_2 \geq ... \geq z_n \geq 0$ – az adatszolgáltatóktól származó adatok nemnövekvő rendszere,
$T$ – a cella értéke, azaz $T = \sum_{j=1}^{n} z_j$ .

Ehhez kapcsolódóan definiálunk még három értéket:

$$t_m = \sum_{j=1}^{m} z_j, \qquad r_m = \sum_{j=m+1}^{n} z_j, \qquad R_{l,m} = \sum_{j=m+1}^{m+l} z_j ,$$

ahol $1 \leq m \leq n$.

*Küszöb szabály*: Ha az adatszolgáltatók száma egy meghatározott $M$ küszöbértéknél ($M \geq 1$) kevesebb, akkor a cella érzékeny. A cella biztonságosnak tekinthető, ha $n > M$.

*Dominancia szabály*: Érzékenynek tekinthető a cella, ha az értékét adó $z$-k közül $m$ db legnagyobb összegének a $T$-hez viszonyított aránya meghalad egy $k$ értéket (azaz dominánsak a cellában), ahol $0 < k < 1$. Az $m$ és a $k$ változtatásával alakíthatjuk a rendszerünk biztonságát: nagy $m$-mel és kicsi $k$-val nagy biztonság érhető el.

Választott $m$ és $k$ mellett a cella biztonságosnak tekinthető, ha

$$\frac{t_m}{T} < k .$$

Ez a szabály tulajdonképpen azt méri, hogy a legnagyobb elem vagy elemek mekkora arányban szerepelnek a teljes összegben. Ha egy elem 99 százalékát adja a cellaértéknek, akkor ezt nyílván nem szabad közölni, mivel nagyon kis hibával lehet következtetni erre az értékre. A nemzetközi gyakorlatban a két legnagyobb elem 80-85 százalékos részesedésénél már veszélyesnek tekintik a cellát. A paraméterekre nézve ez azt jelenti hogy: $m$=2, $k$=0,8 – 0,85.

*p-szabály*: Ez a szabály közvetlenül vizsgálja az egyes adatszolgáltatók adatainak a részvételét a teljes értékösszegben és feltételezi, hogy a támadó személy a cellát alkotó válaszadók közül kerül ki ($z_i$). Függetlenül az $n$ nagyságától, $T - z_i$ –t tekinthetjük úgy is, hogy egy becslés minden egyes $z_h$-ra ($1 \leq h \leq n$), azaz $\hat{z}_h = T - z_i$. A felfedési kockázat mértéke ennek a becslésnek a relatív hibája: $(\hat{z}_h - z_h)/ z_h$. Minél kisebb a $z_h$, annál roszszabb ez a becslés. Nyílván a $T$-hez legközelebb álló értékek ($z_1$, $z_2$) adják a legjobb becslést, és ebben az esetben a legvalószínűbb is az adatfelfedés. Tehát ezt alapul véve kell megállapítani a kockázatot:($h$=1, $i$=2): $(T - z_1 - z_2)/ z_1$. A szabály megköveteli, hogy ez a

relatív hiba nagyobb legyen, mint egy előre megadott $p > 0$ érték (*Cox* [1981]). Így biztonságosnak tekinthető egy cella, ha

$$\frac{r_2}{z_1} > p\,.$$

*pq-szabály*: Ez tulajdonképpen a p-szabály általánosítása, ahol a $p$-t alulról korlátozzuk egy $q \geq 0$ számmal. Tehát a $0 \leq q < p$ figyelembevételével biztonságosnak mondható a cella, ha

$$\frac{r_2}{z_1} > \frac{q}{p}\,.$$

*A mikroadatokban rejlő kockázat*

Számos módszer kínálkozik adataink ellenőrzésére. A szakirodalom (*Skinner–Elliot* [2002], *Carlson, M.* [2002]) alaposan tárgyalja ezeket a számítási módokat, ezek közül a legáltalánosabbat részletezzük. A módszer alapjául az egyednek a sokaságban való előfordulási gyakorisága szolgál. Első lépésként megvizsgáljuk minden egyes egyed gyakoriságát a sokaságban, kiszámítjuk az egyes egyedek, rekordok kockázatát, majd ez után kiszámítható a teljes adatstruktúra kockázata. Fontos betartani ezt a két lépcsős számítást, mivel az egyes rekordokban rejlő esetleges alacsony kockázat nem jelenti automatikusan a teljes adatstruktúra biztonságosságát. Ennek az az oka, hogy a rekordszintű kockázati valószínűségek összeadódnak.

*Példa*: $U$ jelöli a teljes (véges) sokaságot, $X$ az azonosító változók lehetséges kombinációinak összességét, $J$ pedig a kombinációk számát. Ekkor az $X$ például olyan elemekből fog állni, hogy „Férfi–50éves–Fogorvos”. Minden egyes ilyen (rész)sokaságnak meg kell határozni a gyakoriságát, azaz hogy hány egyed tartozik ebbe a tulajdonságkörbe. $F_j$ a $j$-edik sokaság gyakorisága. $I$ az indikátor függvényt jelöli.

$$F_j = \sum_{i \in U} I(X_i = j), \qquad j = 1,\ldots,J$$

Ebből már látható, hogy milyen gyakoriságúak az egyes sokaságok. Fontos tudni azt is, hogy az egyes gyakoriságokból hány darab van, mivel ha kétszer több egyelemű sokaság van, akkor kétszer nagyobb a felfedés kockázata is. Ha

$$N_r = \sum_{j=1}^{J} I(F_j = r), \qquad r = 1,2,\ldots\,,$$

ami a gyakoriságok gyakoriságát jelöli, akkor ez alapján fel tudjuk írni a felfedési kockázatot:

$$P = \frac{N_1}{N} = \frac{\sum_j I(F_j = 1)}{N}.$$

$N$ a sokaság méretét jelöli. $N_1$ került a számlálóba, mivel a felfedés legnagyobb kockázata az egy elemű sokaságokban ($N_1$) rejlik. Ha $N_1$=0 akkor a következő legkisebb nemnulla részsokaság gyakoriságát kell tekintenünk. A $P$ meghatározza, milyen valószínűséggel fedhetőek fel az adataink. A rendszerünktől megkövetelt biztonságától függ, hogy mikor tekintjük ezt elfogadhatónak és mikor nem. Az alacsony elemszámú sokaságok megszűntetésével természetszerűen csökkenthető a $P$ értéke. Ennek módjáról a következőkben mutatunk be módszereket.

## TÁBLÁZATOS ADATVÉDELEM

A tájékoztatás szempontjából kétféle táblázatot különböztetünk meg. Az egyik a gyakorisági (frequency), másik a értékösszeg (magnitude) tábla. A gyakorisági tábla tartalmazza az adatszolgáltatók számát, az értékösszeg tábla pedig az ezen adatszolgáltatók által szolgáltatott adatok összességét. A védendő adatok feltérképezéséhez minden egyes tájékoztatásra kerülő táblához el kell készíteni annak gyakorisági tábláját is. A következő példa ezt szemlélteti.

1. tábla

*Értékösszeg tábla*
*Árbevétel (millió forint)*

| | Ipar | Mező-gazdaság | … | Összesen |
|---|---|---|---|---|
| *1*. város | 124 | 0 | … | … |
| *2*. város | 236 | 377 | … | … |
| *Összesen* | *360* | *377* | … | … |

*Gyakorisági tábla*
*Vállalatok száma*

| | Ipar | Mező-gazdaság | … | Összesen |
|---|---|---|---|---|
| *1*. város | | 0 | … | … |
| *2*. város | 6 | 1 | … | … |
| *Összesen* | *7* | *1* | … | … |

Természetesen a két tábla megegyezik abban az esetben, ha a tájékoztatásra kerülő adataink pont az adatszolgáltatók számát jelöli.

Ha a két táblázat nem egyezik meg, és csak az értékösszeg táblát jelentetjük meg, akkor az adatokból nem derül ki, hogy mely cellák rejtenek mindössze 1 vagy 2 adatszolgáltatót. Ez is jelent önmagában egy minimális védelmet, de a védelem kialakításánál fel kell tételeznünk, hogy ezen információ megszerzéséhez nem kell különösebb detektív képességgel rendelkezni, hiszen például a gazdasági életben a cégek tudják, hány hozzájuk hasonló van a piacon.

A példákban a sötéttel jelzett cellák értékei jelentik azokat az adatokat, amelyeket nem közölhetünk. Az egyszerűség kedvéért az értékösszeg- és gyakorisági tábla adatai egyértelműen megfeleltethetők egymásnak, és az érzékenység kritériuma az 1 vagy 2 adatszolgáltató ténye. A cél tehát az, hogy „megszüntessük” ezeket a cellákat.

### *Aggregálás*

A módszer lényege, hogy oszlopok illetve sorok összevonásával cellákat egyesítünk, növelve ezzel az egy cellában lévő adatszolgáltatók számát (*Eurostat* [1996]).

Az összevonás alapja a következő két ismérv:

– *minőségi ismérv*: Két hasonló, vagy minimális számú hasonló dimenzióértékeket vonunk össze;
– *mennyiségi ismérv*: A skálázás alapjául vett mennyiségértékeknek állapítunk meg új határokat.

Példák a módszer szemléltetésére:

*a*) *A tábla méretének kicsinyítése* (minőségi ismérv)

2. tábla

*Eredeti tábla*

| | Kék szemű | Zöld szemű | Barna szemű | Albínó (piros) szemű | *Összesen* |
|---|---|---|---|---|---|
| Férfi | 12 | 10 | **2** | 6 | *30* |
| Nő | 24 | **2** | 6 | 8 | *40* |
| *Összesen* | *36* | *12* | *8* | *14* | *70* |

Az „Zöld szemű” és „Barna szemű” oszlopokban kis értékű cellákat találhatunk. Összevonjuk őket, feltételezve azt, hogy ezek a dimenzióértékek egy meghatározott szempont szerint összetartozónak tekinthetők. Egy érzékeny cellát tartalmazó oszlopot természetesen összevonhatunk olyan oszloppal is, amelyben nem szerepel érzékeny adat.

3. tábla

*Védett tábla*

| | Kék szemű | Zöld és barna szemű | Albínó (piros) szemű | *Összesen* |
|---|---|---|---|---|
| Férfi | 12 | 12 | 6 | *30* |
| Nő | 24 | 8 | 8 | *40* |
| *Összesen* | *36* | *20* | *14* | *70* |

*b*) *A karakterisztika újrakódolása* (mennyiségi ismérv)

4. tábla

*Eredeti tábla*

| Kor | <12 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | >20 | *Összesen* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Férfi | 23 | 3 | 3 | 7 | 7 | 3 | 4 | 4 | 7 | 4 | 15 | *80* |
| Nő | **2** | **2** | **1** | **1** | **1** | **2** | **2** | **2** | **1** | **1** | 5 | *20* |
| *Összesen* | *25* | *5* | *4* | *8* | *8* | *5* | *6* | *6* | *8* | *5* | *20* | *100* |

Ennél a módszernél egy meghatározott skálázási tulajdonság alapján új intervallumokat állapítunk meg.

5. tábla

*Védett tábla*

| Kor | <13 | 13-15 | 16-19 | 20 vagy < | *Összesen* |
|---|---|---|---|---|---|
| Férfi | 26 | 20 | 19 | 15 | *80* |
| Nő | 4 | 5 | 6 | 5 | *20* |
| *Összesen* | *30* | *25* | *25* | *20* | *100* |

Ezt az adatvédelmi technikát gyakran alkalmazzák a statisztikai munkában. (Papír alapú kiadványokban előfordul, hogy nem pusztán a védelem miatt használják, hanem a táblázatok kisebb mérete miatt az összevont kategóriák áttekinthetőbbek. Az összevonás során például minden egyes korév helyett öt évenként összevont korcsoportok jelennek meg.)

*Az aggregálás előnye:*

– az adatok nem torzulnak, azaz a táblázat adatai a valóságot tükrözik, ami a a *legfontosabb* szempont felhasználók számára;
– könnyen megvalósítható;
– az adatbázisban léteznek olyan ismérvek (régió-megye-város stb.), amelyek a hierarchikus felépítés miatt közvetlen alapjául szolgálhatnak az eljárásnak.

*Az aggregálás hátránya:*

– az összevonások során a háttérbe kerülhetnek részletes tulajdonságok, vagyis *információveszteséggel* kell számolni. A fenti példában minden egyes korév helyett például csupán négy összevont korcsoport kategória jelenik meg.

Igen szemléletes példáját láthatjuk itt az adatvédelmi mérlegelésnek. Dönteni kell, hogy megengedhető-e az összes korév megjelenése, vagy 5, esetleg 10 éves korcsoportokat kell közölni, netán csak 3-4 korcsoport kategória jelenhet csak meg

*Cellaelnyomás*

Amennyiben adatok nagyfokú részletességgel történő közlése a cél, az egyik megoldás a cellák elnyomásának módszere.

A cellaelnyomás lényege, hogy az érzékenynek ítélt cellák tartalmát egyszerűen kitöröljük, ezt nevezzük *elsődleges elnyomásnak*. Mivel a kitörölt adat sorában lévő többi adatból, illetve az „összesen" mezőből ezt követően is egyértelműen meghatározható lenne a cella értéke, ezért a biztonság növelése érdekében további cellákat kell „elnyomni", ezt nevezzük *másodlagos elnyomásnak*. Különféle algoritmusok léteznek annak meghatározására, hogy mely cellákat kell még járulékosan kitörölni a védelem biztosításához (*Hundepool* [1999]).

Kétféle cellaelnyomás létezik:

– a cella tartalmának teljes kitörlése; illetve
– olyan intervallum megadása a cellában, amelybe a cella értéke belelesik.

Példa a módszer szemléltetésére:

6. tábla

*Eredeti tábla*

| | Barna szemű | Kék szemű | *Összesen* |
|---|---|---|---|
| Fekete hajú | Védendő cella | 3 | *7* |
| Barna hajú | 2 | 1 | *3* |
| Szőke | 3 | 3 | *6* |
| *Összesen* | *9* | *7* | *16* |

7. tábla

*Elsődlegesen és másodlagosan elnyomott cellák*

| | Barna szemű | Kék szemű | *Összesen* |
|---|---|---|---|
| Fekete hajú | X | X | *7* |
| Barna hajú | X | X | *3* |
| Szőke | 3 | 3 | *6* |
| *Összesen* | *9* | *7* | *16* |

| | Barna szemű | Kék szemű | *Összesen* |
|---|---|---|---|
| Fekete hajú | 3-6 | 1-4 | *7* |
| Barna hajú | 0-3 | 0-3 | *3* |
| Szőke | 3 | 3 | *6* |
| *Összesen* | *9* | *7* | *16* |

Amennyiben nem kívánjuk teljesen elrejteni a számokat, egyetlen tartományt adunk meg az elsődlegesen és másodlagosan elnyomott cellákra.

*A cellaelnyomás előnye:*

– a látható adatokat részletes felosztásban kapjuk meg, ami több információt jelent;
– a látható adatok a valóságot tükrözik, vagyis nem torzítottak;
– léteznek szoftverek, melyek optimalizálják a másodlagos cellaelnyomást.

*A cellaelnyomás hátránya:*

– a másodlagos cellaelnyomásokkal olyan cellák is rejtve maradnak, melyek egyébként közölhetőek lennének. Egy érzékeny cellához további 2-3 cellára kell alkalmazni a másodlagos cellaelnyomást, ennek következtében jelentősen megritkulhat a táblázat;

– bonyolult és hosszadalmas algoritmus végrehajtása szükséges ahhoz, hogy meghatározzuk azt a minimális számú törlendő cellát, amellyel a védelem még fennáll.

*Kerekítés*

Ennél az adatfelfedés elleni módszernél nem követeljük meg a cellaadatoktól, hogy pontosan tükrözzék a valóságot. A felfedési valószínűséget úgy is lehet csökkenteni, ha

nem szolgáltatunk a felhasználónak pontos értékeket, hanem az összes cella értékét – beleértve az „összesen" cellákat is – kerekítjük egy hozzá közel eső szintre.

A módszer legnagyobb előnye hihetetlenül egyszerű megvalósításában rejlik, a felhasználók körében mégsem arat osztatlan sikert, mivel meglehetősen bizalmatlanul kezelik ezeket az adatokkal. A felhasználóknak azonban nem szabad elfeledkezniük arról, hogy az általunk „elrejtett" értékek is hordozhatnak magukban hibákat (például mintavételi hibát).

*A kerekítési folyamat:*

Elsődlegesen megválasztunk egy *b* értékét, amit az egészszámú kerekítés *alapjának* nevezünk. $N_{ij}$ jelöli az *i*-edik sor *j*-edik oszlopának cellaértékét.

Egy cellában lévő érték kerekítésének lépései (*Eurostat* [1996]):

*1*. Meghatározzuk azt a legnagyobb *h*-t (*szorzóérték*) amelyre teljesül, hogy $N_{ij} \geq h*b$

*2*. Így adódik a $r_{ij}$ maradék: $N_{ij} = h*b + r_{ij}$, ahol $0 \leq r_{ij} < b$.

*3*. A maradékot kerekítjük, 0-ra vagy *b*-re. Így adódik a cella új értéke ($N_{ij}$'): ha $r_{ij}=0$ vagy *b*, akkor nyilvánvalóan: $N_{ij} = N_{ij}'$.

A különböző megoldási módozatok sajátosságai a *b* értékének megválasztásában rejlenek.

8. tábla

*Eredeti tábla*

| | Kék szemű | Zöld szemű | Barna szemű | *Összesen* |
|---|---|---|---|---|
| Fekete hajú | 1 | 4 | 0 | *5* |
| Barna hajú | 15 | 10 | 10 | *35* |
| Vörös hajú | 2 | 10 | 8 | *20* |
| Szőke | 2 | 6 | 15 | *23* |
| *Összesen* | *20* | *30* | *33* | *83* |

*a*) *Rögzített kerekítés*

A *b* értéket rögzítjük, és az előbb leírtak alapján alkalmazzuk a kerekítést.

A tábla minden egyes cellájára elvégezzük a kerekítést (példánkban legyen *b*=5) és akkor a táblázat az alábbi módon alakul.

9. tábla

*Rögzített kerekítéssel védett tábla*

| | Kék szemű | Zöld szemű | Barna szemű | *Összesen* |
|---|---|---|---|---|
| Fekete hajú | 0 | 5 | 0 | *5* |
| Barna hajú | 15 | 10 | 10 | *35* |
| Vörös hajú | 0 | 10 | 10 | *20* |
| Szőke | 0 | 5 | 15 | *25* |
| *Összesen* | *20* | *30* | *35* | *85* |

A módszer előnye, hogy egyszerűen kiszámolható, és minimalizálja a tényleges értéktől való eltérést.

*b*) *Véletlen kerekítés*

A *b* értéke itt is rögzített, viszont a kerekítés már nem a hagyományos módon történik. Az $N_{ij}$ értéket *p* valószínűséggel kerekítjük lefelé, és 1–*p* valószínűséggel kerekítjük felfelé. Ez a következőt jelenti:

Ha *b*=5, akkor a kerekítés valószínűségei a maradék függvényében a következőképpen alakulnak:

| $N_{ij}$ b-vel való osztásának a maradéka | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| 0-ra való kerekítés valószínűségei, *p*= | 1 | 4/5 | 3/5 | 2/5 | 1/5 | 0 |
| 1-re való kerekítés valószínűségei, 1–*p*= | 0 | 1/5 | 2/5 | 3/5 | 4/5 | 1 |

A *p*-t tehát egyenletesen kell megválasztani, a maradék és a *b* hányadosaként. Ily módon a valószínűséggel történő kerekítés biztosítja a módszer torzítatlanságát, azaz $E(N_{ij}')=N_{ij}$ (*Eurostat* [1996]). (Ha például a maradék 1, akkor $E(N_{ij}')=4/5(N_{ij}-1)+1/5(N_{ij}+4)= N_{ij}$)

A módszer sem biztosítja, hogy az oszlop és sorösszegek kiadják az egyes elemek összegét, mivel minden egyes elemre (beleértve az összegeket is) külön-külön végzzük el a kerekítéseket, és nem vesszük figyelembe az elem és az összegértékek viszonyát.

10. tábla

*Véletlen kerekítéssel védett tábla (b=5)*

| | Kék szemű | Zöld szemű | Barna szemű | *Összesen* |
|---|---|---|---|---|
| Fekete hajú | 0 | 0 | 0 | *5* |
| Barna hajú | 15 | 10 | 10 | *35* |
| Vörös hajú | 0 | 10 | 10 | *20* |
| Szőke | 0 | 10 | 15 | *20* |
| *Összesen* | *20* | *30* | *35* | *85* |

*c*) *Ellenőrzött kerekítés*

A kerekítésnek ez a fajtája annyiban különbözik a véletlen kerekítéstől, hogy járulékos ellenőrzéssel megpróbálunk eleget tenni az additivitásnak is, vagy annak, hogy a sorok és oszlopok kiadják az „összesen” mező értékeit. Ennek megvalósítására a leggyakrabban a Cox & Ernst algoritmust használják, melynek során a fel-le kerekítéseket úgy határozzák meg, hogy az kiadja a sor illetve oszlopösszegeket. (*Fischetti– Salazar-González* [1998], *Eurostat* [1996], *Ernst* [1989]).

11. tábla

*Ellenőrzött kerekítéssel védett tábla*

| | Kék szemű | Zöld szemű | Barna szemű | *Összesen* |
|---|---|---|---|---|
| Fekete hajú | 0 | 5 | 0 | *5* |
| Barna hajú | 15 | 10 | 10 | *35* |
| Vörös hajú | 0 | 10 | 10 | *20* |
| Szőke | 5 | 5 | 15 | *25* |
| *Összesen* | *20* | *30* | *35* | *85* |

*Dimenziókorlátozás*

Ez a védelmi módszer csak az elektronikus tájékoztatási formánál alkalmazható. Egyes tájékoztatási rendszereknél olyan formában érhetőek el az adatok, hogy a felhasználó által kiválasztott tulajdonságoknak (dimenziók) megfelelő adatokat kapja meg az adatigénylő táblázatos formában. Az adatkérő a kiválasztott tulajdonságok növelésével egyre részletesebb adatokhoz jut, és egyben növeli az azonosíthatóságot és ezzel együtt a felfedési kockázatot is. Ilyen esetben célravezető védelmi megoldás, hogy maximalizáljuk a választható tulajdonságok számát (legyen ez a szám $n$). Ezt az értéket úgy kell megválasztani, hogy a tulajdonságokból bármely $n$ darabot választva sem juthassunk olyan információhoz, ami védendőnek tekinthető.

A módszer egyszerű és könnyen megvalósítható. A probléma csak az, hogy sok információ maradhat rejtve, ha az adatstruktúra egy részében kevés tulajdonság választása esetén is sok védendő adatot kapunk, és emiatt kicsire kell választanunk az $n$-t.

Ebből kifolyólag a gyakorlatban ezt a módszert csak „elővédelemnek” szokták alkalmazni, olyan formában, hogy egy alkalmas $n$ választásával levágják az adathalmaz peremét (mivel itt a legvalószínűbbek az egyedi adatok), a továbbiakban felmerülő eseteket pedig lokális védelemmel látják el.

A dimenziókorlátozás klasszikus módszertanának vannak változatai, amelyekkel átfogóbb védelmet alakíthatunk ki:

*Szelektív dimenziókorlátozás*: Meg kell vizsgálni, hogy mi az a maximális $n$, amely mellett nem érhető el védendő cella. Az $n$-t 3-4-nél kevesebbre nincs értelme választani, még akkor sem, ha a vizsgálatok azt bizonyítják, hogy kevesebbnél kellene meghúzni a határt, mivel az elérhető adatok aránya vészesen lecsökken. A gyakorlatban megfigyelhető, hogy sokszor csak egy-két dimenziópárosítás választásával érhetőek el védendő cellák. Ezeknek a párosításoknak a letiltásával növelhetjük az $n$ értékét.

*Differenciált dimenziókorlátozás*: A lekérdezett dimenziók számához különböző részletességű adatbázist párosítunk. A választott dimenziók számának növelésével csökkentjük a megjelenítendő adatok részletességét. A felhasználó természetesen egy dimenzió választása esetén kapja a legbővebb adatbázist.

## MIKROADAT-VÉDELEM

Mikroadatok alatt az egy statisztikai egységről birtokunkban lévő legrészletesebb adatokat értjük. Ezek az adatok a gyakorlatban rekordsoros állományokban vannak eltárolva**,** az *egy sor egy adatszolgáltató* elv alapján.

A jogszabályok alapján anonimizált mikroadatok olyan egyedi statisztikai adatok, amelyeket annak érdekében módosítottak, hogy a mindenkori legjobb eljárással összhangban minimálisra csökkenjen az érintett statisztikai egységek azonosításának veszélye.

Ilyen adatállományok teljes vagy részleges publikálása is csak az megfelelő anonimizálás után tehető meg. A jogi részben megismertek alapján az egyedi azonosítók esetében nincs mérlegelési jogkörünk azok megtartására, egyszerűen *ki kell* törölni őket. A további vizsgálataink tárgyát a fennmaradó oszlopok képezik.

12. tábla

*Személyek adataiból álló mikroadatbázis*

| Név | Lakhely | Születési hely | Születési idő | Foglalkozás | Vallás | … |
|---|---|---|---|---|---|---|
| Kala Pál | Iszapszentmotoros | Iszapszentmotoros | 1881.01.02. | Tűzkő árus | – | … |
| Hó Virág | Tápiórettentő | Tápiórettentő | 2031.02.12. | Tűzoltó | Szombatista | … |
| … | … | … | … | … | … | … |

13. tábla

*Gazdasági szervezetek adataiból álló mikroadatbázis*

| Cégnév | Telephely | Alapítás dátuma | Tevékenység | Alaptőke (millió forint) | … |
|---|---|---|---|---|---|
| Gépolaj Rt. | Markotabödöge | 1844.06.12. | Szállítmányozás | 234 | … |
| Sikattyu Kft. | Nagybajom | 2021.12.22. | Költöztetés | 133 | … |
| … | … | … | … | … | … |

Látható, hogy a mikroadatok esetében sokkal szorosabb kapcsolat van az *azonosítás* és a *felfedés* között, mint a táblázatos adatoknál. Ezért beszélünk itt anonimizálásról, nem pedig felfedésről: a speciális rekordsorban szereplő adatok miatt az azonosítás itt önmagában felfedést is jelent. Tehát itt nem az érzékeny adatok elrejtésén van a hangsúly, hanem a rekordnak az egyénhez való társításának megakadályozásán. Ahhoz, hogy kicsi legyen a kockázat, csökkenteni kell a legkisebb gyakoriságú részsokaságok számát.

Az anonimizálás itt is tartalmaz bizonyos fokú információveszteséget, de a védelmi technikáknál éppen az a célunk, hogy megtaláljuk azokat az adatokat amelyek elrejtésével a legkevesebb az információveszteség, és közben az anonimitásnak is eleget teszünk.

### *Csonkolás*

Ez a technika a legnyilvánvalóbb és egyben első helyen alkalmazott. Csonkolásnál egy teljes oszlopot kitörlünk az adatbázisból. Ezt a módszert alkalmazzuk akkor is, amikor az egyedi azonosítókat leválasztjuk az adatbázisról.

14. tábla

*Védelem kialakítása csonkolással*

| Születési hely | Szül.idő | Foglalkozás | Vallás | … |
|---|---|---|---|---|
| XXX | 1881.01.02. | Tűzkő árus | – | … |
| XXX | 2031.02.12. | Tűzoltó | Szombatista | … |
| XXX | … | … | … | … |

A fő probléma annak eldöntésében rejlik, hogy mely oszlop kitörlésével érhetjük el a kellő anonimitást. Ennek eldöntésére meg kell vizsgálnunk, hogy vannak-e olyan tu-

lajdonság- kombinációk (például: „Születési hely” és „Születési idő”) amelyek egyedivé teszik az egyes vagy akár az összes rekordokat. Ezek azok az oszlopok potenciális jelöltjei a csonkolásnak. A csonkolási technika alkalmazása egy ciklikus folyamat. Minden egyes lépésnél csakis egyetlen oszlopot szabad kitörölni, és ezután újra meg kell vizsgálni, mely rekordok maradtak még továbbra is kritikusak a felfedés szempontjából.

A csonkolási technika adatbázisok védelménél igen durva beavatkozásnak számít, hiszen hatására dimenziók tűnnek el. Mivel a tájékoztatás célja, hogy minél több információt biztosítsunk a felhasználóknak, így a mikroadatbázis egészénél a csonkolás mellett gyakran más módszereket is alkalmaznak.

*Cellaelnyomás*

A cellaelnyomás során egyes tulajdonságok „vészesen kevés számú” előfordulásait kell kitörölni. A tulajdonságok vészesen kevés előfordulásai során arra kell gondolni, hogy az adatbázis információi egyediek, így közvetlenül beazonosítható az adatszolgáltató. A minőségileg egyedi és a mennyiségileg kevés vagy kiugróan sok elemszám teszi kritikussá, azonosíthatóvá a rekordot, és így az adatszolgáltatót is.

*Példa*: Ha Iszapszentmotoroson csak egy tűzkőárus van, akkor ez egyértelmű azonosítást, közvetett adatfelfedést tesz lehetővé. Tehát itt a lakhely és a foglalkozás kombinációja kritikus a védelem szempontjából. Bármelyik cella rejtetté tétele megoldja a problémát. A döntés a védelmet kialakító egyéntől függ, illetve attól, hogy lakhelynek vagy a foglalkozásnak a kombinációja fontos-e a többi adatéval összevetve.

15. tábla

*Védelem kialakítása cellaelnyomással*

| Születési hely | Születési idő | Foglalkozás | Vallás | … |
|---|---|---|---|---|
| Iszapszentmotoros | 1881.01.02. | XXX | – | … |
| Tápiórettentő | 2031.02.12. | Tűzoltó | Szombatista | … |
| … | … | … | … | … |

Ez a módszer enyhébb, mintha csonkolással eltávolítottuk volna az egész foglalkozási vagy vallási oszlopot, de tény, hogy ez is adatvesztéssel jár.

*Átkódolás*

Az adatszolgáltatók kilétével kapcsolatos bizonytalanság kialakítható úgy is, ha nem az általunk ismert legpontosabb adatot írjuk a cellába. Ez nem az jelenti, hogy a cellák nem valós értékeket tartalmaznak, hanem csak annyit, hogy egy bővebb tartományba helyezzük át a tulajdonságot, tulajdonságokat.

*Példa*: Ha egy városban csak egy balettcipő-készítő van, akkor érdemes összevonni a cipőkészítővel. Ennek megfelelően a cipőkészítőt és a balettcipő-készítőt át kell írni „Cipő- és balettcipő-készítő”-re. A következő két példa talán még szemléletesebbé teszi az elvet.

16. tábla

*Védelem kialakítása átkódolással*

| Születési hely | Születési idő | Foglalkozás | Vallás | … |
|---|---|---|---|---|
| Iszapszentmotoros | 1881.01.02. | Tűzoltó-Tűzkőárus | – | … |
| Tápiórettentő | 2031.02.12. | Tűzoltó-Tűzkőárus | Szombatista | … |
| … | … | … | … | … |

| Alapítás dátuma | Tevékenység | Alaptőke (millió forint) | … |
|---|---|---|---|
| 1844.06.12. | Szállítmányozás-Költöztetés | 234 | … |
| 2021.12.22. | Szállítmányozás-Költöztetés | 133 | … |
| … | … | … | … |

## *Kerekítés*

A táblázatos adatok védelménél bemutatott kerekítés, teljesen azonos módon alkalmazható számértékű mikroadatoknál is. A *b* érték növelésével növekszik a bizonytalanság a tényleges értékre, viszont ezzel arányosan sajnos növekszik az adat használhatatlansága is. A legnagyobb feladat az optimális *b* választásában rejlik, amely mindig függ a kerekítendő szám nagyságrendjétől, illetve attól, hogy az egyes értékek milyen tartományban mozognak. Nyílván lényegesen eltérő *b*-t kell választani milliós, illetve ezres nagyságrendű értékeknél. Ügyelni kell arra is, hogy ne forduljon elő az, hogy a kerekítést követően olyan számot kapjunk, amelyet az egyébként jellemzett tulajdonság fel sem vehet.

17. tábla

*Védelem kialakítása kerekítéssel (b=50 millió forint)*

| Alapítás dátuma | Tevékenység | Alaptőke (millió forint) | … |
|---|---|---|---|
| 1844.06.12. | Szállítmányozás | 250 | … |
| 2021.12.22. | Költöztetés | 150 | … |
| … | … | … | … |

## *Összekeverés*

A felfedési kockázatot azzal is tudjuk csökkenteni, hogy az egyes emberekhez tartozó érzékeny adatokat véletlenszerűen összekeverjük. Így az egyes emberek adatain nincs mit felfedni, viszont a sokaságok egészére nézve nem változott semmi.

18. tábla

*Védelem kialakítása összekeveréssel*

| Születési idő | Lakhely | Foglalkozás | Vallás | … |
|---|---|---|---|---|
| … | … | … | … | … |
| 1881.01.02. | Iszapszentmotoros | Tűzkő árus | – | |
| 2031.02.12. | Tápiórettentő | Tűzoltó | Szombatista | … |
| … | … | … | … | … |

*

Tanulmányunkban bemutattuk az adatfelfedés elleni védelem környezetét, valamint a a védelem különböző eszközeit. Mi ezen eszközök „tiszta" bemutatására törekedtünk, és olyan példákat hoztunk, amelyek megfelelően reprezentálták az eszközök bemutatását. A gyakorlati statisztikai munkában azonban komoly háttérmunkát jelent annak eldöntése, hogy mely módszer vagy módszerek alkalmazása a legcélravezetőbb egy adott statisztikai felvétel közlésekor; a megvalósítás pedig összehangolt statisztikai-matematikai-informatikai eszköztárat igényel.

Egyes módszerek információ-veszteséggel vagy információ-torzulással járnak. Mint korábban jeleztük, az adatvédelmi technikák alkalmazása egy finom mérleghez hasonlít, ahol azt kalkuláljuk, hogy a tájékoztatás formájának célja, módja milyen védelmi technikát igényel, azaz mit nyerünk az egyik oldalon és mit veszítünk a másikon, s hogyan kerül a kettő egyensúlyba.

A tájékoztatás során a felhasználót mindig tájékoztatják az adatvédelemről. Ez történhet úgy, hogy egy kerekítésnél megadják a kerekítés mértékét, lábjegyzetben vagy mellékletben jelezik, hogy milyen eljárással módosították az adatokat, mikroadat-állomány közlése során pedig csatolnak az állományhoz egy leírást az adatvédelem módjáról és hatásáról. Egyes országok nem közlik pontosan az alkalmazott adatvédelmi technikát, csak a pontosság és megbízhatóság mértékét, illetve, hogy az adatok milyen korlátok és feltételek között alkalmazhatóak.

A védelmi eljárás alkalmazásának tényéről azonban mindig történik tájékoztatás. Ez a nemzetközi gyakorlat is, hiszen ez az etikai lépés biztosítja a kölcsönös bizalmat, és a korrekt statisztikai munkát az adatközlő és a felhasználó oldalán is.

A célhoz, a minél szélesebb körű tájékoztatáshoz kezünkben vannak tehát az eszközök, amelyek alkalmazásával elégedett lehet mind a tájékoztató statisztikai intézmény, mind az adatigénylő felhasználó.

## IRODALOM


BÁNSZEGI K. [1997]: Felfedést akadályozó módszerek a statisztikai tájékoztatásban. *Statisztikai Szemle*. 75. évf. 12. sz. 1039–1046. old.

BÁNSZEGI K. – LAKATOS M. [1994]: Információszabadság – adatvédelem – statisztika (III.). *Statisztikai Szemle*. 72. évf. 10. sz. 761–777. old.

CARLSON, M. [2002]: Assessing microdata disclosure risk using the poisson inverse gaussian distribution. Stockholm. Kézirat. (http://www.matstat.umu.se/banocoss/papers/carlson.pdf)

COX, L. H. [1981]: Linear sensitivity measure in statistical disclosure control. *Journal of Statistical Planning and Inference*. 5. évf. 2. sz. 153–164. p.

DUNCAN, G. T. – KELLER-MCNULTY, S. A. – STOKES, S. L. [2001]: *Disclosure risk vs. data utility: The R-U confidentiality map*. Kézirat. (http://www.niss.org/technicalreports/tr121.pdf)

ELLIOT, M. [1996]: Attacks on census confidentiality using the sample of anonymised records: an analysis. 3rd International Seminar on statistical confidentiality. Bled 1996.

ERDEI V. – SÁNTA J. [2000]: *A statisztikai adatok védelmének nemzetközi szabályozása, módszertani kérdései.* Népszámlálások az ezredfordulón 3. *(Tanulmányok)* Központi Statisztikai Hivatal. Budapest.

ERNST, L. R. [1989]: Further application on linear programming to sampling problems. Kézirat. (http://www.census.gov/srd/papers/pdf/rr89-05.pdf)

*Eurostat* [1996]: *Manual on disclosure control methods*. Luxemburg.

*Eurostat* [1999]: Statistical data confidentiality.

FAGAN, J. T. – GREENBERG, B. V. – HEMMING, B. [1988]: *Controlled rounding of three dimensional tables.* Kézirat. (http://www.census.gov/srd/papers/pdf/rr88-02.pdf)

FISCHETTI, M. – SALAZAR-GONZÁLEZ, J. J. – Caprar, A. [1998]: *Computational experience with the controlled rounding problem in statistical disclosure control*. Padova. Kézirat. (http://neon.vb.cbs.nl/casc/ISIBerlin/Salazar.pdf)

FISCHETTI, M. – SALAZAR-GONZÁLEZ, J. J. [1998]: Experiments with controlled rounding for statistical disclosure control in tabular data with linear constraints. *Journal of Official Statistics*. 4. évf. 4. sz.

HUNDEPOOL, A. [1999]: Statistical disclosre limitation in practice. Kézirat. (http://europa.eu.int/en/comm/eurostat/research/conferences/etk-99/papers/hundepool.pdf)

LAKATOS M. [1994]: Információszabadság – adatvédelem – statisztika (I.). *Statisztikai Szemle*. 72. évf. 7. sz. 547–559. old.
MEROLA, G. [2003]: *Generalized risk measure for tabular data*. Roma. Kézirat. (http://neon.vb.cbs.nl/casc/ISIBerlin/merola.pdf)
SKINNER, C.J. – ELLIOT, M. J. [2002]: *A measure of disclosure risk for microdata*. Kézirat. (http://www.ccsr.ac.uk/publications/occasion/occ23.pdf)
*Statistical Journal of the United Nations ECE* [2001]. Data confidentilaity. 285–407. old.
*Statisztikai igazgatás* [2000]. (Közigazgatási szakvizsga tankönyv). Budapest.

## SUMMARY

This study is about the statistical tools on data confidentiality and the background of the confidentiality issue. It gives an overview on the European and Hungarian legislation, the different and new forms of dissemination, and the confidentiality problems. It shows the different risks on micro and tabular data dissemination and the possible statistical tools of protection with examples.

# A SZEZONÁLIS KIIGAZÍTÁS HARMONIZÁCIÓJA A KÖZPONTI STATISZTIKAI HIVATALBAN

BAUER PÉTER – FÖLDESI ERIKA

A Központi Statisztikai Hivatalban (KSH) 2004. áprilisában lépett életbe a szezonális kiigazítás egységes gyakorlatáról szóló szabályzat, amely a KSH összes kiigazításra kerülő idősorára vonatkozik. A szabályzat kialakításának története 2001-ig nyúlik vissza, ekkor merült fel először a KSH-n belüli egységes szezonális kiigazítási politika kialakításának igénye. 2002-től kezdődően az új rendszer szerint történik a kiigazítás, 2003 folyamán került sor az első év tapasztalatainak szakértői és felhasználói fórumokon történő megvitatására és elfogadására, majd 2004 elején szabályzatba foglaltuk a gyakorlatot.

Cikkünkben röviden kitérünk arra, milyen megfontolások álltak a szezonális kiigazítás gyakorlatának egységesítése mögött, hogyan zajlott a módszertan megváltoztatásának és az új gyakorlat bevezetésének folyamata. Részletesen bemutatjuk a jelenleg alkalmazott módszert és a kialakított gyakorlatot, majd röviden szólunk az eddigi problémákról és tapasztalatokról.

TÁRGYSZÓ: Szezonális kiigazítás. Trend. Munkanap-hatás. TRAMO/SEATS.

A gazdasági életben egyre nagyobb jelentősége van a *minőség*nek, és ez a statisztikai adatok előállításával kapcsolatban is így van. Korábban a statisztikai adatok minőségét az adatok pontosságával jellemezték, de ma már az egész adatelőállítási folyamatot, a statisztikai hivatalok szervezetét is értékeljük az adatminőséggel összefüggésben.

Ahhoz, hogy a KSH meg tudjon felelni az Európai Statisztikai Rendszer (ESR) által meghatározott minőségfogalomnak és az ezen alapuló, az Európai Unió statisztikai hivatala, az Eurostat által megfogalmazott minőségi kritériumoknak[1], minden részterületen, így a szezonális kiigazítás gyakorlatában is biztosítani kell az előbb említett tényezők megfelelő színvonalát.

Az Eurostat által közreadott „Quality in the European Statistical System – The way forward" című kiadvány tartalmazza a minőség kérdésével foglalkozó Szakértői Csoport (Leadership Expert Group – LEG) zárójelentését és a minőséggel kapcsolatos ajánlásait. Ennek keretében javasolják, hogy a Nemzeti Statisztikai Hivataloknak leggyakoribb eljárásaikra ki kell fejleszteniük a Jelenlegi Legjobb Módszereket (Current Best Methods),

[1] Relevancia, pontosság, időszerűség és időbeli pontosság, hozzáférhetőség és átláthatóság, összehasonlíthatóság, koherencia, teljesség (*Lyberg* [2003]). Ezek a kritériumok kiegészíthetőek az adatgyűjtési költségekkel és adatszolgáltatói terhekkel, amelyek ugyan nem tartoznak szorosan a minőség definíciójához, de a minőség értékelésekor szükséges ezeket is figyelembe venni. A minőség statisztikában betöltött szerepéről lásd *Szép Katalin* és *Vígh Judit* cikkét.

erről készíteniük kell egy kézikönyvet, és folyamatosan felül kell vizsgálni a módszereket. A már összegyűjtött legjobb módszereket pedig terjeszteni kell az Európai Statisztikai Rendszeren belül.

A szezonális kiigazítás KSH-n belüli egységes gyakorlatának kialakításánál figyelembe vettük ezeket az ajánlásokat, és összegyűjtöttük az egyes főosztályok által alkalmazott eljárásokat, továbbá megvizsgáltuk az Európai Unió tagállamainak gyakorlatát, valamint a nemzetközi ajánlásokat, különös tekintettel az Eurostat ajánlásaira.

Az Eurostat által javasolt „Proposal for Quality Report for Seasonal Adjustment" [2001] ajánlásai alapján a kiigazítás minőségét az igazolja, ha egyrészt a szezonális kiigazítás eredményei magyarázhatóak az eredeti adatsorra hatással lévő, ismert társadalmi-gazdasági folyamatokkal; másrészt, ha indokolható a munkanap- és a húsvét-hatás, valamint a kiugró értékek jelenléte.

Az új gyakorlat kialakítása során elkészültek a kiigazításra kerülő fontosabb idősorok belső dokumentációi. Az egyes dokumentumok tartalmazták az idősorok főbb jellemzőit (hossz, bázisév, gyakoriság, publikálás), az aktuális kiigazítás eredményeit (paraméterek, kiugró értékek, munkanap- és húsvét-hatás) és ezek gazdasági értelmezését. Ezek a dokumentumok biztosítják az elvégzett munka folyamatos nyomon követhetőségét, a hozzáférhetőséget és az átláthatóságot.

## SZEZONÁLIS KIIGAZÍTÁS – AZ EGYSÉGES GYAKORLAT KIALAKÍTÁSA

A szezonális kiigazítás módszertana az egyes statisztikai területek módszertanától nagyrészt független eljárás, bár a kiigazítás végeredményét nagyban befolyásolja a szakstatisztikusok véleménye a kiugró értékek (outlierek) magyarázhatóságáról, illetve a munkanap- és húsvét-hatás meglétéről. Mivel a szezonális kiigazítás speciális matematikai statisztikai ismereteket igényel, és az Eurostat előírásai a fő irányelvek tekintetében minden területen azonosak, ezért lehetővé vált az a megoldás, hogy ne valamelyik szakfőosztály koordinálja a szezonális kiigazítást, hanem az erre a területre specializálódott munkatársak végezzék el ezt a feladatot. Több statisztikai hivatal gyakorlatához hasonlóan a módszertannal foglalkozó részleg (a KSH-ban ez a Mintavételi és módszertani osztály[2]) feladata lett a KSH egységes módszer alapján történő szezonális kiigazításának koordinálása, a kapcsolódó tudományos eredmények, a hazai és nemzetközi gyakorlat és ajánlások nyomon követése, illetve kutatások végzése.

### *A szezonális kiigazítás alapfogalmai*

Gazdasági vagy társadalmi folyamatok vizsgálatakor gyakran a folyamatok időbeli alakulását leíró idősorokat használnak. Az idősorok viselkedését nagymértékben befolyásolhatják olyan tényezők, amelyek különböző évek azonos időszakaiban (például hónap vagy negyedév) azonos irányban és közel azonos mértékben hatnak az idősor alakulására. Ilyen tényezők lehetnek az időjárás, különféle adminisztratív hatások (például iskolai tanév kezdése és befejezése) vagy kulturális-vallási hagyományok (rögzített dátumú ünnepek, például a karácsony). Ezeket a tényezőket együttesen szezonális hatásnak nevez-

[2] 2004. áprilisig Statisztikai mintavételi és módszertani osztály néven önálló osztályként, jelenleg a Statisztikai kutatási és oktatási főosztály részeként működik.

zük. Az elemzők gyakran a folyamatok olyan jellemzőire kíváncsiak, amelyeket a nagymértékű szezonális hatás elfed. Ilyen jellemzők lehetnek például a növekedés, csökkenés, fordulópont és a más folyamatokkal való kapcsolat. A szezonális hatás kiszűrésével kapott, szezonálisan kiigazított idősor alkalmasabb lehet a fenti jellemzők vizsgálatára, mint az eredeti idősor. A szezonális hatás kiszűrését szezonális kiigazításnak nevezzük. A szezonális hatás egyszerű kiszűrésére használt eredeti idősorra vett „időszak / előző év azonos időszaka" mutatóval szemben előny, hogy a szezonálisan kiigazított idősor esetén tetszőleges időszakok értékei hasonlíthatók össze, így például „időszak / előző időszak" mutatókat lehet számolni, és kevesebb időszak adatai alapján, azaz hamarabb azonosítható például a vizsgált folyamat egy növekedési szakasza vagy egy fordulópont.

A szezonális kiigazítás során az idősort komponensekre bontjuk. Ezek a komponensek a trend[3], a szezonális komponens és az irreguláris komponens. A trendben jelenik meg az idősor alapirányzatának hosszú távú változása, a szezonális komponens a fentiekben részletezett szezonális hatás számszerűsített értéke, az irreguláris komponens pedig a maradék, amely a véletlenszerű hatásokat tartalmazza. Megkülönböztethetünk additív vagy multiplikatív összekapcsolódású modellt, attól függően, hogy a komponensek összegének vagy szorzatának kell eredményül adnia az eredeti idősort. A két modell között az dönt, hogy a szezonális ingadozás mértéke állandó vagy pedig az idősor szintjével arányos. A szezonálisan kiigazított idősort a trend és az irreguláris tényező összegeként vagy szorzataként kapjuk.

A statisztikai hivatalok rendszerint a kiigazítatlan adatok mellett szezonálisan kiigazított adatokat, illetve azokból képzett mutatókat is publikálnak. A trend publikálása általában grafikon formájában történik. Megfigyelhető az a – nem biztos, hogy pozitív – nemzetközi tendencia, hogy a kiigazítatlan adatokkal szemben egyre inkább a szezonálisan kiigazított adatok kerülnek előtérbe.

*Két szezonális kiigazító módszer – az X12-ARIMA és a TRAMO/SEATS*

Jelenleg a két legelterjedtebb szezonális kiigazító módszer az X12-ARIMA és a TRAMO/SEATS. A következőkben röviden áttekintjük a két módszert. A TRAMO/SEATS speciális beállításairól, a kiugró értékekről és a munkanap-hatásról bővebben a KSH gyakorlatánál írunk.

Az X12-ARIMA a U.S. Census Bureau által, az X11-ARIMA-ból továbbfejlesztett, mozgóátlagokon alapuló módszer. Míg a komponensekre bontáshoz ugyanazt a módszert alkalmazza, mint az X11-ARIMA (azaz mozgóátlagolású rögzített szűrőkkel állítja elő a szezonálisan kiigazított idősort és a trendet), addig a kiugró értékek és a munkanap-hatás kezelésére az úgynevezett RegARIMA-eljárást használja. Ez egy olyan regresszió, amelynek a zajtényezője ARIMA-folyamat. A módszer az illesztett ARIMA-modellt használja az idősor előre- és hátrafelé történő meghosszabbítására is; segítségével az idősor elején és végén ugyanazzal a mozgóátlaggal számolhatók a szezonálisan kiigazított adatok, mint az idősor közepén. Az X12-ARIMA a kiigazítás után számos, a kiigazítás minőségét jellemző mutatót ad eredményül. Ebben a tekintetben egyértelműen felülmúlja riválisát, a TRAMO/SEATS-et, aminél jóval kevesebb ilyen mutatót kapunk. Az X12-ARIMA további előnye, hogy az alkalmazott módszerek egyszerűbbek, könnyebben

[3] Pontosabban *trend-ciklus*, mert az egy évnél hosszabb periódusú, ciklikus hatásokat is tartalmazza.

megérthetőek, mint a TRAMO/SEATS-nél, amely erősen támaszkodik az idősorelemzés matematikai elméletére.

A TRAMO/SEATS-módszert a Spanyol Nemzeti Bankban fejlesztette ki Gómez és Maravall. A TRAMO/SEATS egy ARIMA-modell alapú szezonális kiigazítási módszer. Ez azt jelenti, hogy az eljárás egy ARIMA-modellt[4] illeszt az idősorra, majd a komponensekre (trend, szezonális komponens, irreguláris komponens) szintén ARIMA-modelleket határoz meg, amelyeket az eredeti idősorra illesztett ARIMA-modellből vezet le. Az eredeti idősort a komponensekre kapott ARIMA-modellek alapján bontja fel[5].

A TRAMO/SEATS két együttműködő részből áll: a TRAMO-ból és a SEATS-ből. A TRAMO a komponensek multiplikatív vagy additív összekapcsolódásáról képes automatikusan dönteni, és multiplikatív összekapcsolódás esetén az eredeti idősor logaritmusát veszi, amelyet a továbbiakban additív összekapcsolódású modellként kezel. A TRAMO végzi az idősor előzetes korrekcióját is, nevezetesen a kiugró értékek szűrését, a hiányzó megfigyelések interpolációját, valamint a munkanap- és húsvét-hatás kezelését. Ezen kívül egyéb, a felhasználó által meghatározott regresszorokat is képes kezelni. Hasonlóan az X12-ARIMA-hoz, a korrekciókat olyan regresszióval számolja, ahol a zajtényező ARIMA-modellt követ, és ez a zajtényező lesz a felsorolt hatásoktól megtisztított idősor. Az ARIMA-modell azonosítását és a hozzátartozó paraméterek becslését a TRAMO automatikusan végrehajtja, de lehetőség van a modell és a paraméterek kézi beállítására is.

A SEATS átveszi a TRAMO által átadott, már megszűrt idősort és az illesztett ARIMA-modellt, majd a korábban leírtaknak megfelelően komponensekre bontja. Végső lépésként a TRAMO által kiszűrt hatások visszakerülnek a meghatározott komponensekbe, így az eredményül kapott komponensek együttesen kiadják az eredeti idősort.

A kiigazítás után a kiigazítás minőségére vonatkozó diagnosztikákat kapunk. Ezek a TRAMO-modell illesztésének jóságára vonatkozó statisztikák. A legfontosabb diagnosztikák a következők: a reziduumokra, illetve azok négyzetére vonatkozó Ljung-Box és Box-Pierce statisztikák, a normalitásra, a csúcsosságra és a ferdeségre vonatkozó tesztek eredményei, valamint a kiugró értékek számának aránya az idősor hosszához képest.

*A szezonális kiigazítás módszertanának egységesítése a KSH-ban*

2002 előtt a KSH-ban csak néhány területen végeztek szezonális kiigazítást, többnyire az X11 és X11-ARIMA-módszerrel.

2000–2001 táján megnőtt az igény egy új módszer bevezetésére és jóval több szezonálisan kiigazított adat publikálására. Két oldalról nehezedett nyomás a KSH-ra: egyrészt a hazai felhasználók (különösen a Magyar Nemzeti Bank) fogalmaztak meg kritikát a KSH szezonális igazításával kapcsolatosan, másrészt az Európai Unió harmonizációs igénye is indokolta egy új módszer bevezetését.

Az Eurostat már jóval a csatlakozás előtt igényelte a leendő tagországok adatait, nem csak alapadat formájában, hanem munkanappal kiigazítva, illetve szezonálisan és munkanappal kiigazítva is. Annak érdekében, hogy az egyes országok adatai összehasonlíthatóak, valamint aggregátumok képzésére alkalmasak legyenek, egységes elvárásokat fo-

[4] Pontosabban multiplikatív szezonális ARIMA (SARIMA)-modellekről van szó.

[5] Valójában a komponensek spektrumának meghatározásáról van szó, és az idősor komponensekre bontásához csak a spektrumokra van szükség. A komponensek ARIMA-modelljeinek a spektrumokból történő kiszámítása lehetséges, de nem szükséges.

galmazott meg a jelenlegi és leendő tagországok számára (*Seasonal*... [1998]). Két kiigazítási módszert ajánlott: az X12-ARIMA- és a TRAMO/SEATS-módszert. A könnyebb használhatóság kedvéért kifejlesztetett egy kezelőfelületet (Demetra), amely mindenki számára ingyenesen hozzáférhető, és tartalmazza mindkét módszert.

2001 második felében a KSH Statisztikai mintavételi és módszertani osztályának munkatársaiból és az egyes szakstatisztikák szakértőiből megalakult „A szezonális kiigazítás harmonizációja" munkacsoport. Munkája során a csoport egy olyan, elemzésekre épülő, megalapozott javaslatot kívánt tenni a szezonális kiigazítás új rendszerére vonatkozóan, amely mind a hivatal, mind a partnerintézmények és a felhasználók számára elfogadható. A vizsgálat során a cél egy olyan egységes rendszer kialakítása volt, amely megfelelő minőséggel alkalmazható valamennyi, a KSH-ban kiigazításra kerülő idősorra, ugyanakkor 2002 elejétől bevezetésre kerülhet. További szempont volt, hogy a rendszer legyen összhangban az Eurostat ajánlásaival. A munkacsoport ezen szempontok figyelembe vételével vizsgálta meg az Eurostat által ajánlott két módszert: az X12-ARIMA- és a TRAMO/SEATS-eljárásokat.

A Demetra segítségével, 176 idősoron végzett vizsgálatok, tesztelések során a munkacsoport a következő 11 szempont alapján értékelte a módszereket:

– megfelel-e a nemzetközi ajánlásoknak;
– tudományosan elfogadott, korszerű módszerről van-e szó;
– rövid idősorok megfelelő kezelése;
– kezeli-e magyar ünnepnapokat;
– stabil eredményeket ad-e az idősorok végén;
– a szezonalitás szűrésének hatásossága;
– az eredmények statisztikai diagnosztikája;
– az automatikus futtatás lehetősége;
– hogyan kezel nagyszámú idősort;
– input- és output-fájlok könnyű kezelhetősége;
– felhasználóbarát kezelőfelület.

A próbaszámítások elvégzése és kiértékelése után a szakértői értekezlet a TRAMO/SEATS-módszert – amelyet egyébként az MNB is ajánlott – találta jobbnak az X12-vel szemben. Így a KSH a 2002. évtől kezdődően új, egységes módszertant vezetett be, amely összhangban van a hazai és az Eurostat elvárásokkal, és teljesíti a kiigazítással szemben támasztott követelményeket. A módszertan alkalmazásához a KSH a Demetra szoftver mellett döntött.

A 2002. február 5-i Gazdaságstatisztikai Felhasználói Fórum elfogadta az új gyakorlatot, majd a 2002. február 18-i Elnöki Értekezlet jóváhagyta a módszertani váltást. Ezután a KSH a 2002. március 14-i sajtóközleményben tájékoztatta a közvéleményt a Hivatal szezonális kiigazítási módszertanában történt változásról.

A tapasztalatok alapján 2003 szeptemberében belső szakértői értekezlet keretében megtárgyaltuk a felmerült problémákat és azok megoldásait. Vezetői Kollégiumi Előterjesztési Javaslatban rögzítettük az alapelveket és a kialakult gyakorlatot, amelynek alapján a jövőben is szeretnénk folytatni a szezonális kiigazítást.

A *Statisztikai Szemle* 2003. júliusi számában jelentek meg *Friss Péternek* a szezonális kiigazítás alkalmazásával kapcsolatos kérdései, amelyekre a 2003. szeptemberi számban kíséreltünk meg választ adni (*Friss* [2003], *Bauer-Földesi* [2003]).

A kérdések alapvetően a módszertan gyakorlati megvalósításának a tájékoztatással kapcsolatos problémáira vonatkoztak. Ezek közül az egyik leginkább gondot okozó probléma a szezonálisan kiigazított adatok folyamatos, visszamenőleges változása. Ez Friss Péter szerint a felhasználók számára nehezen elfogadható, hiszen minden egyes publikálásnál más adatokkal találkozhatnak a korábbi időszakokra vonatkozóan, ugyanakkor ez a módszer lényegéből alapvetően következik. (A revízióról részletesebben is írunk a következőkben.)

Szintén problémaként merült fel a szezonálisan kiigazított adat és a trend együttes publikálása, illetve a trend közlése az utolsó időszakig. Itt alapvetően a kiugró értékek befolyásolhatják a szezonálisan kiigazított és a trend adatok eltérő viselkedését, illetve okozhatják a trend utolsó adatainak bizonytalanságát. (A trend publikálásának lehetséges módjairól a későbbiekben szót ejtünk.)

2003. december 9-én az MTA Statisztikai Bizottságának Módszertani albizottsága is napirendre tűzte a szezonális kiigazítás témáját, ahol az elméleti előadás után a résztvevők megismerkedtek az MNB és a KSH szezonális kiigazítási gyakorlatával[6]. A vita során a KSH gyakorlatával szemben nem merült fel kritika.

Szintén decemberben (11-én) a Gazdaságstatisztikai felhasználói fórumon is bemutattuk a KSH szezonális kiigazítási gyakorlatát. Mindkét fórum megerősített minket abban, hogy a jövőben nagyobb hangsúlyt kell fektetni az új módszer bemutatására, oktatására, a felhasználókkal való alaposabb megismertetésre és az eredmények részletes magyarázására.

A 2004. elején elfogadott szabályzatot a KSH-n belül mindenki számára hozzáférhetővé tettük, és annak érdekében, hogy a szezonális kiigazítással foglalkozó szakértők és a KSH szezonálisan kiigazított adatait a felhasználók is megismerhessék, az új gyakorlatról 2004 folyamán a KSH egy kiadványt jelentet meg.

## A KSH JELENLEG ÉRVÉNYES ÚJ GYAKORLATA

A szezonális kiigazításhoz a TRAMO/SEATS módszert alkalmazzuk a Demetra program felhasználásával. A TRAMO/SEATS havi, negyedéves vagy féléves gyakoriságú idősorokat tud kezelni (a KSH-ban jelenleg csak havi és negyedéves idősorokat igazítunk szezonálisan).

A KSH-ban általában bázisévhez viszonyított indexeket igazítunk, de előfordul abszolút számokból álló idősor és adott év adott hónapjához viszonyított indexek igazítása is (ez utóbbira példa a fogyasztóiár-index idősora, amelynél 1994. december = 100%). Nem igazítunk „előző év azonos időszaka = 100%” típusú indexeket vagy „bázisév azonos időszaka = 100%” típusú indexeket, mert ezeknél az eredeti idősorhoz képest információt veszítenénk, és megváltozna az idősor dinamikája is.

A szezonális kiigazításhoz szükséges, hogy az igazítandó idősor havi gyakoriság esetén legalább 36 megfigyelésből álljon, azaz legalább 3 év hosszú legyen, negyedéves gyakoriság esetén pedig legalább 16 megfigyelés, azaz 4 éves hosszúság szükséges. Ezek természetesen a minimum értékek, ahhoz, hogy a program diagnosztikákat számoljon, ennél hosszabb idősorokra lehet szükség az idősorra illesztett ARIMA modelltől függő-

[6] Az elméleti bevezetőt *Sugár András* (a BKÁE adjunktusa), a korreferátumokat az MNB részéről *Montvai Beáta*, a KSH részéről *Bauer Péter* tartotta.

en. Rövid idősorok szezonális kiigazítása nem ad megbízható, jó minőségű eredményt, ezért törekedni kell arra, hogy minél hosszabb idősorok álljanak rendelkezésre. Önmagában azonban a hosszú idősor nem garantálja a jó eredményt: az idősornak homogénnek is kell lennie, azaz előállítása időben változatlan módszertan szerint történjen, és hasonló viselkedésűnek kell lennie az idősor teljes hosszában. Kisebb változások megengedhetőek, például a szezonális tényező lassú változását, azaz a mozgó szezonalitást, valamint a kiugró értékként kezelhető változásokat a módszer figyelembe veszi. Ha azonban az idősor viselkedése, elsősorban a szezonalitás lényegesen megváltozik, és egységes modell már nem illeszthető rá teljes hosszúságában, akkor szükségessé válik az idősor csonkítása. Nyilvánvaló, hogy ez a lehetőség elsősorban hosszú idősorok esetében adott, a jelenlegi magyar idősorok még csonkítás nélkül is rövidnek tekinthetők.

Szezonálisan kiigazított adatokat és trendet a statisztikai hivatalok, így a KSH is ugyanazon idősorra ismételten közölnek, ahogy az idősor kiegészül az újabb és újabb időszakok megfigyeléseivel. Nyilvánvaló, hogy a több megfigyelésből álló idősor több információt hordoz a sor komponensekre bontásához, emiatt a szezonálisan kiigazított adatok és a trend is visszamenőlegesen változik a korábbi eredményekhez képest, vagyis revízióra kerül sor. Ennek mérséklésére, illetve ütemezésére van lehetőség, mégpedig az idősorra illesztett ARIMA-modell és paramétereinek korlátozott újrabecslésével. A következő lehetőségek adottak:

– Az ARIMA-modell és paramétereinek újrabecslése minden egyes alkalommal, amikor új időszakra vonatkozó megfigyeléssel egészül ki az idősor. Ez a megközelítés teljes mértékben felhasználja az összes rendelkezésre álló információt, és így a lehető legjobb eredményt adja, viszont nagy mértékű és minden egyes publikációnál jelentkező revízióval jár.

– Az ARIMA-modell rögzítése, de a paraméterek újrabecslése minden egyes új megfigyelésnél. Az ARIMA-modell frissítésére évente egyszer kerül sor. Ilyenkor a rendelkezésre álló információkat korlátozottan használjuk fel, de még így is viszonylag jó eredményt kapunk. A revízió kisebb mértékű, mint ha a modellt is újrabecsülnénk minden új megfigyelésnél. Nagyobb revízió csak az évenkénti modell-frissítésnél következik be.

– Az ARIMA-modell és a paraméterek rögzítése, évente egyszeri újrabecsléssel. Ilyenkor még korlátozottabb a rendelkezésre álló információk felhasználása, de még ez is elég jó eredményt ad. A revízió viszonylag kis mértékű, nagyobb revízió következik be az évenkénti modell- és paraméter-frissítésnél.

– A szezonális kiigazító módszer évenkénti egyszeri használata, amikor a következő évi szezonális tényezőt előre jelezzük, a már ismert időszakokra pedig újrabecsüljük. Az év során, az új megfigyelések adatainak beérkezésekor az előre jelzett szezonális tényezővel számolva állítjuk elő a szezonálisan kiigazított adatokat. Év közben egyáltalán nem használjuk fel az új megfigyelésekből a szezonalitásra vonatkozó információt. Revízió csak a szezonális kiigazító módszer évi egyszeri használatakor jelentkezik, de akkor nagymértékben.

A KSH számára fontos a lehetőség szerinti kis mértékű revízió, de az eredmények jó minősége is, ezért a harmadik lehetőséget, tehát az évi egyszeri modell- és paraméter-frissítést választottuk. A modell és a paraméterek rögzítése rendszerint az év utolsó adatainak beérkezésekor történik, az új modell és paraméterek szerinti kiigazítás eredménye pedig a következő év első adatával jelenik meg. A modellnek és a paramétereknek a tervezettől eltérő, év közbeni újrarögzítésére akkor kerülhet sor, ha alapadatokat érintő revízió történik azokra a megfigyelésekre, amelyeket felhasználtunk a korábbi modell- és paraméter-becslésnél. Újrarögzítés történhet akkor is, ha a kiigazító program diagnosztikái alapján az újabb időszakokra vonatkozó megfigyelésekkel kiegészült idősor lényegesen más viselkedést mutat, mint ami a rögzített modellnek és paramétereknek megfelel.

A KSH-ban munkamegosztást alakítottunk ki a szezonális kiigazítás elvégzésére. A szakfőosztályok végzik saját idősoraik évközi kiigazítását, ők rendelkeznek a szükséges szakértői információkkal és felelősek az eredmények publikálásáért. A Mintavételi és módszertani osztályon történik az évenkénti (és az esetleges évközi) modell- és paraméter-rögzítés, az évközi igazítás eredményeinek matematikai statisztikai ellenőrzése, ezen kívül itt történik a szezonális kiigazítás módszertani koordinációja és a témakörben jelentkező tudományos eredmények követése is.

*Néhány módszertani kérdés és azok lehetséges megoldásai*

A szezonális kiigazítás során felmerülő egyik probléma az úgynevezett munkanap-hatás kezelése. A munkanap-hatás az a jelenség, amikor az egy időszakra vonatkozó megfigyelés értékét befolyásolja az időszak munkanapjainak száma. A problémát az okozza, hogy a munkanapok száma nemcsak időszakról időszakra változhat, hanem még a különböző évek azonos időszakaiban is különbözhet, így közönséges szezonális hatásként nem kezelhető. A program a munkanap-hatást regressziós változókkal jellemzi, és a TRAMO rész szűri ki az idősorból. Attól függően használhatunk 1 vagy 6 regresszort, hogy a munkanap-hatásnál csak a munkanapokat és a szabadnapokat (szombat, vasárnap és ünnepnap) különböztetjük meg, vagy a hét minden egyes napjának különböző hatást tulajdonítunk. Az 1 és a 6 regresszoros esetben is a változók értékei úgy lettek kialakítva, hogy az egy hétre vonatkozó hatás nulla legyen[7]. Ennek következtében a munkanappal való korrekció nem valamiféle átlagos munkanap számhoz történő igazítást jelent.

Lehetőség van a szökőéveknél jelentkező februári szökőnap hatásának figyelembe vételére. Ez még egy regresszort jelent[8].

A munkanap-hatás kezelésekor figyelembe vesszük a magyar ünnepnapokat, ami a Demetrában beépített lehetőség a TRAMO/SEATS-módszer használata esetén.

A húsvét hatásának kiemelt kezelésére bizonyos idősoroknál (például kiskereskedelmi forgalom) szükség lehet, mert a húsvét mozgó ünnep: lehet márciusban vagy áprilisban, és hatása nagymértékű lehet a húsvétot megelőző egy hetes időszakban is. A húsvétot tehát egyszer figyelembe vesszük a munkanap-hatás kezelésekor mint munkaszüneti napot (1 regresszornál hétvégi napnak, 6 regresszornál vasárnapnak számít), másrészt a megelőző $d$ nap hatását figyelembe véve alakítunk ki egy külön regressziós változót a húsvét-hatás kezelésére[9]. A figyelembe vett napok száma, azaz $d$ értéke állítható, a KSH-ban a $d = 6$ alapbeállítást használjuk.

A TRAMO képes automatikusan tesztelni, hogy az idősornál jelentkezik-e munkanap- vagy húsvét-hatás, azonban a hatások meglétével kapcsolatban célszerű figyelembe

[7] Az 1 regresszoros esetben a regressziós változó értéke minden egyes időszakra az időszak munkanapjainak száma mínusz a szabadnapok száma szorozva ötketteddel. A 6 regresszoros esetben a regressziós változók: a hétfők száma mínusz vasárnapok száma, keddek száma mínusz vasárnapok száma, …, szombatok száma mínusz vasárnapok száma, ahol a vasárnapokhoz számítanak az esetleges ünnepnapok is.

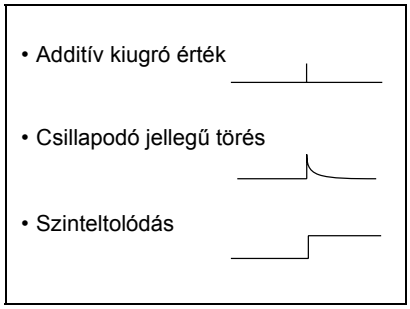
[8] Értéke februárok kivételével nulla, szökőévben lévő (29 napos) februárnál 0,75, nem szökőévben lévő februárok esetén –0,25.

[9] Havi gyakoriságú idősort feltételezve: ennek március és április hónapok kivételével az összes hónapra nulla az értéke. Márciusra a változó értéke $p_M - m_M$, ahol $p_M$ a $d$ napon belül azon napok aránya, amelyek márciusra esnek, $m_M$ pedig a sok éven át vett átlaga az ilyen $p_M$ értékeknek. Az áprilisi hónapokhoz rendelt érték $p_A - m_A$, ahol $p_A$ és $m_A$ hasonlóképpen van definiálva. Jó közelítéssel $m_M = m_A = 1/2$. Ily módon a március és április hatása együttesen nulla, hiszen $p_A = 1 - p_M$.

venni a rendelkezésre álló szakértői információkat is. Ez különösen fontos a rövid idősorok esetében, ahol a statisztikai tesztek nem elég megbízhatóak. Arra is figyelni kell, hogy a regressziós együtthatók előjele értelmezhető legyen, azaz, ha az idősornál feltételezhető, hogy a több munkanap növeli az időszak értékét, akkor a munkanapok regressziós együtthatója pozitív előjelű legyen. Általános szabály, hogy ha az idősor rövid, akkor nem célszerű sok regresszor használata.

A TRAMO képes automatikusan kezelni a kiugró értékeket is: ezeknek mind helyét, mind típusát automatikusan detektálja, majd valamennyi talált kiugró értéknek megfeleltet egy-egy regressziós változót, és számszerűsíti a kiugró értékek hatását, azaz meghatározza a regressziós változók együtthatóját[10]. Három fajta kiugró értéket veszünk figyelembe: additív kiugró értéket, csillapodó jellegű törést és szinteltolódást. Az additív kiugró értéknél csak egyetlen megfigyelés értéke tér el jelentősen; a csillapodó jellegű töréstnél egy megfigyelés értéke kiugró, majd a következő megfigyeléseknél fokozatosan (exponenciálisan) csökken a kiugrás mértéke[11]; a szinteltolódásnál pedig egy időszaktól kezdve az összes megfigyelés értéke ugyanannyival ugrik ki a korábbi értékekhez képest, azaz az idősor szintje tartósan megváltozik.

*1. ábra. A kiugró értékek típusai*




Annak érdekében, hogy a kiigazítás eredményeként kapott komponensek kiadják (a kapcsolódás típusának megfelelően összegezve vagy összeszorozva) az eredeti idősort, a TRAMO által kiszűrt hatások a SEATS által végzett dekompozíció után visszakerülnek a megfelelő komponensbe: a munkanap- és húsvét-hatás a szezonális komponensbe, az additív kiugró érték és a csillapodó törés az irreguláris komponensbe, a szinteltolódás pedig a trendbe kerül. A kiugró értékekre vonatkozó választást az indokolja, hogy a szinteltolódás hosszú távon hat, míg a másik két típusú kiugró érték rövid távon. Mivel a szezonálisan kiigazított idősor a trendből és az irreguláris komponensből tevődik össze, ezért a szezonálisan kiigazított sor az összes kiugró értéket tartalmazza.

A kiigazítás során figyelembe vesszük az esetleges szakértői információkat arról, hogy egy adott időszakban lehetett-e kiugró érték, illetve a TRAMO által automatikusan talált kiugró értékeknek mi lehet a közgazdasági, társadalmi, időjárási vagy egyéb külső tényezőkben keresendő magyarázata. Különösen fontos a szakértői információ az idősorok végén megjelenő kiugró értékeknél, mert ezek típusa matematikai szempontból soká-

[10] Itt lényegesen egyszerűsítettük az eljárás menetét, a pontos leírás megtalálható: *Gómez* és *Maravall* [1998].

[11] A TRAMO/SEATS-nél 0,7 hatványaival csökken a kiugrás hatása. Ezen a beállításon lehet változtatni, de a KSH gyakorlatában erre még nem volt példa.

ig bizonytalan lehet, a típus későbbi módosulása pedig nagymértékű revízióhoz vezethet. Mint a munkanap-hatásnál, itt is elmondható, hogy a túl sok regresszor használatát lehetőleg kerülni kell, emiatt az egy idősorhoz rendelt kiugró értékek számát a Demetrában 5 százalékban maximálják az idősor összes megfigyeléseinek számához képest. Amennyiben több kiugró érték kezelésére van igény, amire rövid, különösen negyedéves idősoroknál lehet szükség, az 5 százalékos érték átállítható.

A KSH-ban gyakori, hogy aggregált idősorokat és azok alágazatait (komponenseit) is ki kell igazítani szezonálisan. A szezonális kiigazítási módszereket nem ismerő felhasználó azt várná, hogy a szezonálisan kiigazított adatokra is teljesül az, hogy az alágazatok együttese kiadja az aggregátumot, azaz abszolút számok esetén az alágazatok összege megegyezik az aggregátummal, illetve indexek esetén az alágazatok súlyozott átlaga egyezik meg az aggregátummal. Ez sajnos automatikusan nem feltétlenül teljesül sem a TRAMO/SEATS módszernél, sem számos más szezonális kiigazító módszernél, így az X12-ARIMA módszernél sem. A probléma kezelésére kézenfekvő a következő lehetőségek egyikét választani:

– Direkt igazítás. Ekkor az alágazatokat és az aggregátumot külön-külön igazítjuk. Az eredményül kapott szezonálisan kiigazított idősorokra nem teljesül az aggregációs feltétel, viszont mind az alágazatokra, mind az aggregátumra a legjobb minőségű eredményt kapjuk.

– Indirekt igazítás. Ilyenkor az alágazatokat igazítjuk ki, és a kiigazított idősorok aggregátumát fogadjuk el az aggregátum kiigazítottjának. Ilyenkor az aggregációs feltétel definíció szerint teljesül, viszont az aggregátum kiigazításának minősége romlik.

– Szétosztásos módszer. A direkt igazítás egy változata, amikor az aggregátum kiigazítottja és az alágazatok kiigazítottjainak aggregációjával kapott sorok közötti eltérést szétosztjuk az alágazatok között. Ezzel az alágazatok kiigazításának minőségét némileg rontjuk, viszont az aggregációs feltétel teljesül és az aggregátum szezonális kiigazítása is jó minőségű marad.

E három lehetőség közül a KSH-ban rendszerint az elsőt, azaz a direkt igazítást választjuk, tekintve, hogy ez adja a legjobb eredményt mind az alágazatokra, mind az aggregátumra. Amennyiben előírás vagy határozott igény van az aggregációs feltétel teljesülésére, akkor a szétosztásos módszert alkalmazzuk, mivel az aggregátum minősége fontosabb, mint az alágazatoké.

Szintén probléma jelentkezik az időbeli aggregációnál: a szezonálisan kiigazított adatok éves összege nem feltétlenül adja ki az eredeti, kiigazítatlan adatok éves összegét. Amennyiben igény van arra, hogy ez a két összeg megegyezzen, akkor a különbség szétosztható az év szezonálisan kiigazított adatai között, de ezzel rontjuk a kiigazítás minőségét, éppen ezért ezt a megoldást lehetőség szerint nem alkalmazzuk.

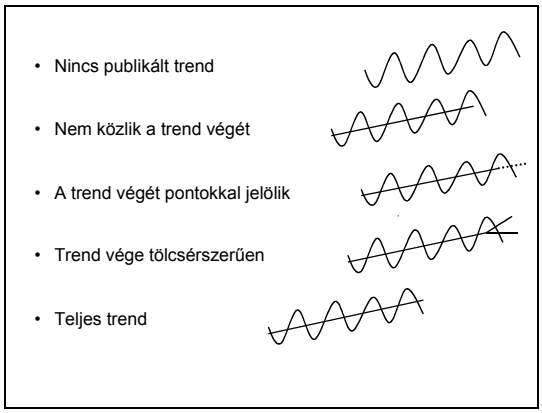
Gyakori, hogy a KSH a szezonálisan kiigazított adatok mellett (vagy helyett) – elsősorban az Eurostat igényeit kielégítve – úgynevezett munkanappal kiigazított adatokat is előállít. Ezek az idősorok úgy állnak elő, hogy az eredeti idősorokból csak a munkanap- és húsvét-hatást szűrjük ki, a szezonális hatást nem. Amennyiben az idősornál jelentkezik munkanap- vagy húsvét-hatás, akkor a munkanappal kiigazított adatok alkalmasabbak lehetnek az „időszak / előző év azonos időszaka” típusú összehasonlításra, mint az eredeti idősor adatai. A munkanappal való kiigazítás és a szezonális kiigazítás között összhang van, mert a szezonális kiigazításnál becsült munkanap- és húsvét-hatást szűrjük ki a csak munkanappal való kiigazításnál is. A munkanappal kiigazított idősoroknál is felmerül az alágazatok aggregációs problémája és az időbeli aggregáció problémája, a választott megoldások itt is ugyanazok.

Az Eurostat rövidtávú mutatókra (STS) vonatkozó legfrissebb ajánlása alapján a munkanappal kiigazított idősoroknál biztosítjuk, hogy a bázisév indexeinek átlaga a kiigazítatlan és a munkanappal kiigazított adatokra megegyezzen, azaz 100 százalék legyen. Ezt a munkanappal kiigazított idősor átskálázásával oldjuk meg, ami azt jelenti, hogy a munkanappal kiigazított idősor bázisévi átlagával leosztjuk a kiigazított sor minden egyes értékét. Ez az átskálázás nem változtatja az idősor dinamikáját, azaz az egyes időszakok értékeinek egymáshoz képesti aránya nem változik, csupán a bázisévhez viszonyított indexek értékei változnak meg az átskálázás során. Hasonló előírás a szezonálisan kiigazított adatokra nincs, ezért azokat nem skálázzuk át.

*A trend publikációjának kérdései*

Publikációs szempontból problémát jelent, hogy a trend utolsó néhány értéke meglehetősen bizonytalan, az újabb megfigyelésekkel kiegészült idősorra végrehajtott kiigazítás során nagymértékben változhat visszamenőlegesen. Vannak, akik fontosnak tartják, hogy ez a bizonytalanság valamilyen módon a trend grafikonján is jelezve legyen, mások azon az állásponton vannak, hogy mivel a trend utolsó néhány adata szezonális kiigazítási módszertől függetlenül mindenképpen bizonytalan, ezért ezt felesleges külön jelezni. Éppen ezért a trend publikációjára többféle gyakorlat létezik, amelyek alkalmazása nemhogy statisztikai hivatalonként, de gyakran egy hivatalon belül témakörönként is változik.

*2. ábra. A trend publikálásának lehetőségei*




– Az egyik lehetőség az, hogy egyáltalán nem publikálunk trendet. Ez az eljárás nem igazán szerencsés, véleményünk szerint a trend publikációjára szükség van, mert a kevésbé hozzáértő olvasó számára többet mond, mint a szezonálisan kiigazított adatsor.

– Lehetséges, hogy a trend végét nem publikáljuk. Igaz, hogy így kiiktatjuk a trend végének bizonytalanságát, de a trend információ tartalmát is csökkentjük, mert az a tény, hogy a trend végének adatai változhatnak, nem jelenti, hogy ezek az adatok nem hordoznak információt, és a trend végének változásából (az ún. csapkodó farok) is lehet következtetéseket levonni. Különösen hasznos lehet, ha a trendet a szezonálisan kiigazított adatokkal együtt vizsgáljuk. Itt érdemes megjegyezni, hogy ha a trendet egyszerűen az eredeti adatsorra illesztett mozgóátlaggal számoljuk, akkor a trend végének kiszámítására nincs is lehetőség (hacsak nem készítünk előrejelzést az eredeti idősorra).

– Elképzelhető, hogy a trend végének néhány adatára nem hosszabbítjuk meg a trendvonalat, hanem csak ponttal jelöljük értéküket. Ezzel a megoldással nem vesztünk információt, miközben a trend végének bizonytalanságát is jeleztük.

– Lehetséges olyan megközelítés is, amikor a trend végét tölcsérszerű ábrával rajzoljuk, azaz megadjuk azt a tartományt, amelyben a trend nagy valószínűséggel haladni fog.

– Végül pedig a legegyszerűbb (a trend elhagyásán kívül), hogy a trendet teljes hosszúságában külön jelölések nélkül közöljük.

A KSH-n belül a trend publikációjának kérdésében nem született konszenzus, ezért a főosztályok saját hatáskörben, a Tájékoztatási főosztállyal egyeztetve döntenek a trend publikációjának módjáról. Jelenleg a trend teljes hosszúságú publikációját, a trend végének elhagyását és a trend közlésének elhagyását is alkalmazzák a KSH-ban.

*A KSH szezonálisan kiigazított idősorai*

Jelenleg a KSH több területen is végez szezonális kiigazítást, a kiigazított adatoknak azonban csak egy része jelenik meg hazai publikációkban, a többi adatsor az Eurostat részére készül, ahol részben publikálják, részben nemzetközi aggregátumok képzésére használják fel a magyar adatokat.

A felhasználók számára talán legfontosabb adatsorok a fogyasztói árindex és a maginfláció szezonálisan kiigazított adatai, illetve a változatlan áras GDP adatai, amelyeket mind a termelési oldal szerinti, mind a felhasználási oldal szerinti bontásban is közöl a KSH, sőt 2004-től már a folyóáras felhasználási adatokat is igazítjuk, de ezek eredményei hazai publikációkban még nem jelennek meg.

Legnagyobb számban azonban ipari idősorokat igazítunk, amelyeknél a termelést, az export-, a belföldi-, és az összes értékesítést a TEÁOR szerinti két betűs bontásban közöljük. Szintén az iparstatisztikához tartozik az építőipar és alágazatai, illetve a beruházás és az energiafelhasználás.

Ezeken kívül a munkaügy-statisztika néhány adatsorát (foglalkoztatottak, munkanélküliek, átlagkereset, munkaerő-költség, teljesített munkaóra alakulása), a kiskereskedelem adatsorait, a szolgáltatási árbevétel adatait, és a külkereskedelem adatait is igazítjuk különböző mélységű bontásban.

*3. ábra. Szezonálisan kiigazított idősorok az egyes főosztályokon (2004. június)*

| Főosztály | Idősorok száma | Témakör | Gyakoriság |
|---|---|---|---|
| Életszínvonal- és emberi erőforrás-statisztikai | 5 | Foglalkoztatottság, Munkanélküliség | Negyedéves |
| | 16 | Átlagkereset | Havi |
| | 59 | Teljesített munkaóra | Havi |
| | 27 | Munkaerő költség | Negyedéves |
| Fogyasztás és felhalmozás statisztikai | 18 | GDP felhasználási oldal | Havi |
| | 9 | Fogyasztói árindex, Maginfláció | Havi |
| Iparstatisztikai | 228 | Ipar | Havi |
| | 10 | Építőipar | Havi |
| | 1 | Beruházás | Negyedéves |
| | 1 | Energiafelhasználás | Havi |
| Külkereskedelem-statisztikai | 14 | Kivitel, Behozatal | Havi |
| Nemzeti számlák | 20 | GDP termelési oldal | Havi |
| Szolgáltatás-statisztikai | 32 | Kiskereskedelem | Havi |
| | 10 | Szolgáltatások árbevétele | Negyedéves |
| *Összesen* | *450 idősor* | | |

*Néhány megoldatlan probléma*

A KSH-ban folyó szezonális kiigazítással jelenleg a legnagyobb probléma és a kiigazítás minőségét leginkább negatívan befolyásoló tényező a rendelkezésre álló homogén idősorok rövidsége. Néhány kivételtől eltekintve a legtöbb idősor az 1998-as évvel kezdődik. Ennek az az oka, hogy a 2000-es bázisra való áttéréskor módszertani problémák miatt a legtöbb esetben az idősorok visszaszámolása az 1998-at megelőző időszakra nem történt meg. Az idősorok rövidsége az egyik oka annak, hogy az évenkénti modell- és paraméter-frissítéskor nagy mértékű revízió történik a szezonálisan kiigazított adatokban. A kiugró értékek helye és típusa gyakran változik, a munkanap-hatás meglétére vonatkozó statisztikai tesztek sem elég megbízhatóak, emiatt előfordul, hogy egyik évben még szignifikáns a munkanap-hatás, a következő évi paraméter-rögzítéskor pedig már nem, vagy fordítva.

További problémát jelent – mint azt már a *Statisztikai Szemle* korábbi számában is jeleztük (*Bauer–Földesi* [2003], 826. old.) –, hogy nem vagy csak részben ismerjük a szezonálisan kiigazított adatok felhasználóit. A 2004 év eleji Gazdaságstatisztikai felhasználói fórum meghívottai részére készítettünk egy kérdőívet, hogy megtudjuk kik, és hogyan használják adatainkat. Megkérdeztük azt is, hogy a jelenlegi publikálási gyakorlat megfelelő-e számukra. Sajnos az alacsony visszaérkezési arány miatt messzemenő következtetések ezekből nem szűrhetőek le.

Ugyanakkor az is gondot okoz, hogy legnagyobb „felhasználónk", az Eurostat egyre több szezonálisan kiigazított adatot vár a KSH-tól, így a kiigazításra kerülő idősorok száma folyamatosan növekszik. Ezért egyre jobban kihasználjuk a programban rejlő automatizálási lehetőségeket. Sajnos az idősorok egyedi kezelésére csak nagy jelentőségű idősoroknál illetve különleges esetekben van lehetőség.

*

2002-től tehát a KSH a TRAMO/SEATS-módszert alkalmazza a szezonálisan kiigazított adatok előállításához. Az évi egyszeri paraméterrögzítés során a magyar ünnepnapok hatását is figyelembe vesszük, és szükség esetén a szökőév és a húsvét hatását is számszerűsítjük. A szakfőosztályok és a módszertani osztály közötti munkamegosztás biztosítja, hogy mind a matematikai statisztikai, mind a közgazdasági, szakmai szempontokat érvényesíteni tudjuk a kiigazítás során.

Ugyanakkor a fentiekben leírtakból is látható, hogy van még jó néhány nyitott kérdés a szezonális kiigazítás módszertanában illetve az azon alapuló gyakorlatban. Ezek megoldását leginkább a felhasználói igények és a nemzetközi előírások befolyásolják. Mivel ezen igényeket és előírásokat a szezonális kiigazítás során mindenkor szeretnénk figyelembe venni, ezért a jelen cikkben ismertetett szabályzatot nem tekintjük véglegesnek és megváltoztathatatlannak.

FORRÁS- ÉS IRODALOMJEGYZÉK

BAUER P.–FÖLDESI E. [2003]: Észrevételek az idősorelemzési módszerek alkalmazásával kapcsolatos kérdésekhez. *Statisztikai Szemle*. 81. évf. 9. sz. 826–831. old.

*Demetra program* http://forum.europa.eu.int/Public/irc/dsis/eurosam/library?l=/&vm=detailed&sb=Title. (A Demetra felhasználói kézikönyve és a szezonális kiigazítás módszertanával kapcsolatos anyagok is letölthetőek innen.)

FRISS P. [2003]: Kérdések az idősor-elemzési módszerek alkalmazásáról. *Statisztikai Szemle*. 81. évf. 7. sz. 588–595. old.
GÓMEZ, V. – MARAVALL, A. [1997]: *Programs TRAMO and SEATS, instructions for the user*. Working Paper 97001, Dirección General de Análisis y P.P. Ministerio de Economía y Hacienda. Madrid.
GÓMEZ, V. – MARAVALL, A. [2000]: *Automatic modeling methods for univariate series,* Chapter 7 In: *A course in time series analysis*. J. Wiley and Sons.
LYBERG, L. [2003]: *Definitions and measurements of survey quality,* May, 28.
http://www.gallup-europe.be/presentation/MEE%20Lars%20Lyberg%20Gallup.pdf
MARAVALL, A. [1999]: *Unobserved components in economic time series,* Chapter 1 In: *Handbook of applied econometrics*, Macroeconomics. Blackwell Publishing.
*Proposal for quality report for seasonal adjustment* [2001]. Assessment of quality in statistics – fourth meeting. Luxembourg.
*Quality in the European statistical system – The way forward* [2002]. Luxembourg. Office for Official Publications of the European Communities. http://europa.eu.int
CHAMBER, R. – SKINNER, CH. [2003]: *Statistical tools for improving quality of data*, Amrads course material. Rome.
Seasonal Adjustment Policy – Some Eurostat Proposals [1998]. SAM98 Seminar. Bucharest
SUGÁR A. [1999]: Szezonális kiigazítási eljárások (II.), *Statisztikai Szemle.* 77. évf. 10–11. sz. 816–832. old.
*TRAMO/SEATS program.* http://www.bde.es/servicio/software/econome.htm. *(Itt számos Maravall cikk is letölthető.)*

## SUMMARY

The regulation of unified seasonal adjustment procedure was introduced in the Hungarian Central Statistical Office (HCSO) in April 2004. Its development started in 2001 as a practice recommended for all departments dealing with seasonal adjustment within the HCSO. From the first quarter of 2002 the seasonal adjustment has been going on with the new system.

In 2003 the experiences of the previous (first) year was discussed both with experts and users, and on that basis the regulation was launched in early 2004.

In the study the authors write shortly about the reason of the standardisation of seasonal adjustment practice, how the system was changed and the new practice was introduced. The recently applied method and the problems arisen are discussed too.

# AZ ADATFELFEDÉS ELLENI VÉDELEM STATISZTIKAI ESZKÖZEI

ERDEI VIRÁG – HORVÁTH ROLAND

A tanulmány az adatfelfedés elleni védelem statisztikai eszközeit mutatja be, az adatvédelem problémáinak tárgyalása mellett és azok összefüggésében. Ismerteti az adatvédelem európai és magyar jogi alapjait, a tájékoztatási formák bővülését is. Az eszközök, módszerek tárgyalásakor sor kerül a táblázatos- és mikroadatokban lévő adatfelfedési kockázat, majd a táblázatos adatokra vonatkozó védelmi eszközök és a mikroadat védelem különböző módjainak ismertetésére, gyakorlati példákon keresztül.

TÁRGYSZÓ: Adatvédelem. Adatfelfedés elleni statisztikai eszközök

Talán nem szerénység azt állítani, hogy lassan immár 15 éves demokráciánkban az adatvédelem szó mindenkinek ismerősen cseng. Rádió- és tévéműsorok állandó szereplője az adatvédelmi biztos, gyakran újságok vezető híre az adatvédelemmel kapcsolatos valamely aktuális téma. Az utca embere természetesen azt látja a kifejezés mögött, hogy a korábbi mindent tudó állammal szemben napjainkban már inkább a semmit nem tudó állam áll. Érdekvédő szervezetek és jogászok hada áll szemben az állammal, illetve minden egyéb magán illetve hivatalos szervvel, amennyiben az szeretne valami nem jogában állót megtudni rólunk, hiszen személyes adataink védettek, mi rendelkezünk felőlük, és jogi felhatalmazás híján nehezen tudható meg tőlünk bármi is.

A magyarországi demokrácia érésének folyamán a személyes adatok védelme volt az első, amit mindenki megismert, de az évek során az egyre tudatosabban viselkedő az állampolgárok annak is tudatában kezdenek lenni, hogy a közérdekű adatok nyilvánosságához is joguk van, valamint általában az információhoz. Állampolgárként ugyanolyan vehemenciával igényelhetünk információkat, mint amilyen mértékben ragaszkodunk személyes adataink védelméhez. Az Európai Unióhoz történő csatlakozás nyomán Európa és a világ kitágul számunkra. Nő az információigényünk, egyre jobban tisztában vagyunk a jogainkkal és a lehetőségeinkkel. Számítani lehet arra, hogy a csatlakozás hatására az emberek egyre jobban felmérik a lehetőségeiket, és élni is fognak velük, például egyre több információt fognak igényelni.

Az adatgyűjtők, így a statisztikai hivatalok is, hatalmas adatvagyonnal rendelkeznek, mégis egyes becslések szerint ennek csak 30–40 százaléka hasznosul, kerül nyilvánosságra. Ennek egyik fő oka az adatvédelem. Az informatika óriási térnyerése következté-

ben megnőtt az adatfelfedés lehetősége. A statisztikai hivatalnak meg kell felelnie a törvényi adatvédelmi kötelezettségeknek, s ezt annál is inkább meg kell tennie, mivel egy esetleges adatfelfedés nagyban aláásná az adatszolgáltatói bizalmat, és így a statisztikai tevékenységet. A kötelező adatvédelemmel szemben azonban az információszabadság állampolgári joga áll.

Tanulmányunk célja, hogy bemutassuk az adatfelfedés elleni védelem statisztikai eszközeit. Ezen eszközök birtokában válhat lehetővé a minél szélesebb körű biztonságos adatközlés, az adatfelfedés egyidejű elkerülésével.

## AZ ADATFELFEDÉS ELLENI VÉDELEM KÖRNYEZETE

A statisztikai és egyéb adatgyűjtések célja az adatok elemzés, feldolgozás utáni nyilvánosságra hozatala. Ez az adatfelvételek végső és legérzékenyebb pontja. A magyar statisztikai törvény kimondja, hogy a statisztikai módszerekkel felvett, feldolgozott, tárolt és elemzett adatok az államhatalmi és a közigazgatási szervek, valamint a társadalom szervezetei és tagjai tájékoztatását szolgálják.

### *Tájékoztatási kötelezettség, szélesedő lehetőségek*

Az állami, központi költségvetésből finanszírozott statisztikai szerveknek a törvényben rögzítetten túl erkölcsi kötelessége is az adatok legteljesebb mértékű közzététele, hiszen az adatokat mi, állampolgárok térítés nélkül szolgáltatjuk, és az állam statisztikákat felhasználó tevékenysége, munkája is a mi érdekünkben történik. A magyar statisztikai törvény szerint a hivatalos statisztikai szolgálathoz tartozó szervek által végrehajtott adatgyűjtések eredményei – az adatvédelemre vonatkozó szabályok betartása mellett – nyilvánosak.[1]

A nyilvánosságra hozatal, a tájékoztatás „kiadványokból és más adathordozókon lévő adatállományokból történő közlésekből áll".[2] A papír alapú tájékoztatás magában foglalja a különböző kiadványokat, évkönyveket, tájékoztatókat, brosúrákat stb., ám napjainkban a gyors és nagy információigény miatt egyre inkább tért nyer az egyéb adatközlés. Ilyenek az internetes adatközlés, a CD-k, és az egyéb nem papír alapú adathordozók, de akár a telefonon keresztül történő adatszolgáltatás is.

Az adatközléseknél különbséget tehetünk aszerint is, hogy azok egy konkrét „legyártott" adatot, táblázatot tartalmaznak, vagy a felhasználó, adatkérő közreműködésével egy állományból egyedi beállítás alapján lekérhető adatokat, táblázatokat. Az adatközlés egy harmadik típusa a mikroadat-állomány közzététele, amely rekordsorosan tartalmazhat egy adatfelvételt vagy annak egy részét.

A tájékoztatás kötelezettségét, annak alapelveit az uniós statisztikai jogszabályok is részletesen rögzítik. A tájékoztatási tevékenységgel kapcsolatban megfigyelhető az a tendencia, hogy egyre nyitottabbá válnak a statisztikai szervezetek, egyre több adat kerül nyilvánosságra. Egyre több formában válik lehetségessé a tájékoztatás, egy-egy adat, szám közlése mellett egyre részletesebb összesítések, táblázatok jelennek meg, és akár teljes adatállományok is hozzáférhetővé válnak. Ennek konkrét bizonyítéka, hogy euró-

[1] 1993. évi XLVI. Törvény a statisztikáról 17.§ (1)
[2] 1993. évi XLVI. Törvény a statisztikáról 23.§ (2)

pai uniós szinten jogilag is megnyílt a lehetőség a kutatók, tudományos élet képviselői előtt, hogy bizalmas, egyedi adatokhoz férjenek hozzá. (Az Európai Unió 1997-ben született statisztikai törvénye már megfogalmazta ezt a lehetőséget (17. cikk), 2002-ben azonban rendelet is született, amely részletezi azt.)

A 831/2002/EK rendelet a bizalmas adatokhoz való tudományos célú hozzáférésről[3] lehetővé teszi a közösségi hatóság (az Európai unió statisztikai hivatala, más néven Eurostat) hivatali helyiségeiben a bizalmas adatokhoz való hozzáférést, és anonimizált mikroadatok kibocsátását is. Egyetemek, felsőoktatási intézmények, tudományos kutatással foglalkozó szervezetek, intézmények, hivatalok, számára nyitott ez a lehetőség (részletesen lásd 831/2002/EK rendelet 3. cikk). A rendelet az adatokhoz való hozzáférés módjáról, engedélyezéséről szól, annak érdekében, hogy pontosan tudható legyen – az adatok bizalmas volta miatt –, hogy az adathozzáférés folyamán ki mikor jut hozzá valamihez és mi alapján, mit tehet, mik a kötelezettségei stb.

Adatvédelmi intézkedések természetesen itt is vannak, a mikroadatok kiadásakor eltávolítják a közvetlen azonosítókat, és a rendelkezésre álló legjobb eljárás alkalmazásával minimálisra csökkentik az érintett statisztikai egységek közvetett azonosításának veszélyét. Az Eurostat hivatali helyiségeiben (a gyakorlatban kutatószoba) engedélyezhető hozzáférés pedig mindig csak hivatalos személy felügyelete mellett történhet, és a kutatás eredményeit – mielőtt kikerülnek az intézményből – ellenőrzik, biztosítva, hogy azok nem tartalmaznak bizalmas adatokat.

Nevesítve a hozzáférés négy felmérésből, illetve statisztikai adatforrásból lehetséges: a közösségi háztartási panelből, a munkaerő-felmérésből, a közösségi innovációs felmérésből és a szakmai továbbképzési felmérésből. (Az adatszolgáltató nemzeti statisztikai hivatalok megtagadhatják az adataikhoz történő hozzáférést, de engedélyezhetik is a felsoroltaktól eltérő bizalmas adatokhoz való hozzáférést.)

A rekordsoros adatokhoz történő hozzáférés nagyon nagy nyitottságot jelent az adatgazda statisztikai hivataloktól a tájékoztatásban, ezért is követeli meg a legszigorúbb adatvédelmet.

## DEFINÍCIÓK

A cikkben tárgyalt statisztikai információk bizonyos jogszabályokon alapulnak, így azokra támaszkodunk mi is. Az előbbiekben használtuk a *bizalmas adatok* kifejezést. A következőkben ismertetjük a *személyes adat*, a *bizalmas adat* és az *azonosíthatóság* fogalmát, amelyek témánk szempontjából meghatározóak. A magyar adatvédelmi, statisztikai törvények, így a fogalmak is számos európai jogszabály és ajánlás alapján születtek. Először az európai uniós megfogalmazások lényegét ismertetjük, majd röviden a hazairól szólunk.

*Személyes adat:* A személyhez kapcsolódó adat definíciója alapvetően fontos, hiszen a statisztikai felmérések nagy része emberekre vonatkozik. A fogalom igen jól körülírható. „Személyes adat bármely, azonosított vagy azonosítható természetes személyre (’adatalany’) vonatkozó információ; a személy különösen akkor tekinthető azonosíthatónak, ha őt – közvetlenül vagy közvetve – azonosítószám vagy egy vagy több fizikai, fizi-

[3] A Bizottság 831/2002/EK Rendelete (2002. május 17.) a bizalmas adatokhoz való tudományos célú hozzáférés tekintetében a közösségi statisztikáról szóló 322/97/EK tanácsi rendelet végrehajtásáról.

ológiai, mentális, gazdasági, kulturális vagy szociális azonosságára jellemző tényező alapján azonosítani lehet."[4]

*Bizalmas adat:* Ez ugyancsak kulcsfontosságú fogalom, hiszen ez alapján definiálhatjuk majd a statisztikai titkosságot. A bizalmas adat a személyes adatnál bővebb kategória. A személyes adaton túl egyéb adatok is beletartoznak, pl. a gazdasági szervezetek adatai. A bizalmas adat lényeges tulajdonsága, hogy az a megfigyelési egységekre – személyekre, cégekre stb. – vonatkozó adat, információ. Az „adatok bizalmasnak tekintendők, amennyiben segítségükkel a statisztikai egységek akár közvetlenül, akár közvetve azonosíthatók és így egyedi információt fednek fel."[5]

A bizalmas – egyes szövegekben védettnek nevezett – adat alapján a *statisztikai titkosság* magának a tevékenységnek, az egyes statisztikai egységekkel kapcsolatos adatoknak a védelme.

A bizalmas adat tehát megköveteli, hogy ne lehessen sem közvetlenül, sem közvetve azonosítani a vonatkoztatási, statisztikai tárggyal. (A bizalmas adatokhoz való tudományos célú hozzáférésről szóló rendeletben bizalmas adatok alatt már csak a közvetett azonosíthatóságot értik, hiszen a statisztikai munkában a közvetlen azonosítást a feldolgozási folyamat elején lehetetlenné teszik, illetve az idősoros elemzéseknél külön kezelik az azonosítókat.)

A nemzetközi joganyagok fogalmai egységesek a tekintetben, hogy megkövetelik a közvetlen azonosítók leválasztását, illetve, hogy az egyértelmű azonosíthatóságot és a lehetséges azonosíthatóságot is a fogalom részévé teszik. Az igazán lényegi információt azonban azok a meghatározások adják, amelyek magáról a kikövetkeztethetőségről, azonosíthatóságról szólnak.

*Azonosíthatóság:* A közvetlen azonosíthatóság egyértelműen definiálható az egyedi azonosítók leválasztásával (személyeknél: név, lakcím; gazdasági szervezeteknél: név, telephely vagy azonosítószám).

A közvetett azonosíthatóságról vagy felfedésről már csak durva körülhatárolás lehetséges:

– „A statisztikai egység azonosíthatóságának megállapításakor figyelembe kell venni mindazokat az eszközöket, amelyeket egy harmadik fél ésszerűen *(reasonably)* igénybe vehet az említett statisztikai egység azonosításához."[6] (A harmadik fél úgy értendő, hogy az első két fél az adatszolgáltató és a statisztikai hivatal, hiszen ők jogosultak az adatot ismerni.)

– „A személy nem tekinthető azonosíthatónak, ha az azonosítása ésszerűtlenül hosszú időt és munkabefektetést igényel."[7]

Magyarországon két alaptörvény szabályozza a kérdéskört, a statisztikai törvény (1993. évi XLVI. Törvény), valamint az adatvédelmi törvény (1992. évi LXIII. Törvény), hivatalos nevén Törvény a személyes adatok védelméről és a közérdekű adatok nyilvánosságáról.

Az adatvédelmi törvény határozza meg a személyes adatot.

[4] Az Európai Parlament és a Tanács 95/46/EC Irányelve az egyénnek a személyes adatok feldolgozásával kapcsolatos védelméről és ezeknek az adatoknak a szabad áramlásáról 2. cikk (a.)

[5] A Tanács 1997. február 17-i. 322/97. (EK) számú rendelete a közösségi statisztikákról 13. cikk (1)

[6] A Tanács 1997. február 17-i. 322/97. (EK) számú rendelete a közösségi statisztikákról V. fejezet 13. cikk

[7] A tagállamok minisztereinek bizottsága által 1997. szeptember 30.-án elfogadott 97/18 sz. ajánlás a statisztikai célból gyűjtött és feldolgozott személyes adatok védelméről Fogalmak 1. bekezdés

*Személyes adat:* bármely meghatározott (azonosított vagy azonosítható) természetes személlyel kapcsolatba hozható adat, az adatból levonható, az érintettre vonatkozó következtetés. A személyes adat az adatkezelés során mindaddig megőrzi e minőségét, amíg kapcsolata az érintettel helyreállítható. A személy különösen akkor tekinthető azonosíthatónak, ha őt – közvetlenül vagy közvetve – név, azonosító jel, illetőleg egy vagy több, fizikai, fiziológiai, mentális, gazdasági, kulturális vagy szociális azonosságára jellemző tényező alapján azonosítani lehet.[8]

A statisztika és a statisztikai törvény azonban a nemzetközi gyakorlat alapján védi a többi adattípust is, például a gazdasági szervezetek adatait. Ennek érdekében bevezeti az egyedi adat fogalmát és azt védi.

*Egyedi adat:* a statisztikai célt szolgáló, a természetes és a jogi személy, valamint a jogi személyiséggel nem rendelkező adatszolgáltatóval kapcsolatba hozható adat.[9] Egyedi adat tehát az, ami a nemzetközi joganyagokban bizalmas vagy védett adat. (A jelenleg folyó uniós jogszabályok fordításában elképzelhető, hogy a bizalmas adatok helyett egyedi adat szerepel majd.) Egyedi adat csak statisztikai célra használható.

*Azonosíthatóság:* A hazai gyakorlat, jog is elsődleges védelmi kritériumként az egyedi azonosítók leválasztását követeli meg. Az azonosítók leválasztása a közvetlen azonosítás megakadályozását szolgálja: „A természetes személy személyére vonatkozó adatgyűjtésnél az érintett nevét és a lakcímét (személyazonosító adat) – kivéve azt, amelynek adathordozóját a levéltári anyag védelmére vonatkozó jogszabály értelmében levéltári őrizetbe kell adni – a statisztikai feldolgozás befejezésekor, az adatok teljességének és összefüggésének ellenőrzését követően, de legkésőbb a tárgyidőszakot követő egy éven belül kell törölni, adatátadás esetén ezt megelőzően is."[10]

(„Az egy évnél hosszabb időszakra vonatkozó idősoros vizsgálatok esetében az adatállományt belső azonosítóval kell ellátni, amelyből az érintett személyazonossága nem állapítható meg. Az érintett személyazonosító adatait az adatállománytól elkülönítetten kell kezelni."[11])

A gazdálkodó szervezet akkor tekinthető anonimnak, ha elnevezése és telephelye nincs feltüntetve (*Statisztikai igazgatás* [2000]).

Egyetlen kritérium van a statisztikai törvény végrehajtási rendeletében, amely a közvetett azonosítást kívánja megakadályozni. Azt mondja a szabály, hogy összesítve sem lehet nyilvánosságra hozni olyan adatot, amelynél az adatszolgáltatók száma háromnál kevesebb.[12]

A jogszabályok definíciói után szeretnénk tisztázni egy, a gyakorlatban elterjedt félreértést. Az adatvédelem során gyakori, hogy megkülönböztetik a jogi védelmet a technikai védelemtől, mondván, hogy amikor például egy szerződést ír alá valaki egy adathozzáférésről, akkor az jogi védelem, míg amikor beavatkozást végzünk egy táblázaton, vagy adatbázison, akkor az technikai. A valóságban ez a két dolog nem különíthető így el, hanem egyik a másikon alapul. A jogszabályok megfogalmazzák a kereteket, fogalmakat, teendőket, s ennek alapján készülnek a gyakorlatban technikák, módszerek azok megvalósítására.

[8] 1992. évi LXIII. Törvény a személyes adatok védelméről és a közérdekű adatok nyilvánosságáról 2.§ 1.
[9] 1993. évi XLVI. Törvény a statisztikáról 17.§ (2)
[10] 1993. évi XLVI. Törvény a statisztikáról 19.§ (1)
[11] 1993. évi XLVI. Törvény a statisztikáról 19.§ (2)
[12] 1993. évi XLVI. Törvény végrehajtásáról szóló 170/1993. (XII. 3.) Kormány rendelet 19.§

## AZ ADATKÖZLÉS PROBLEMATIKÁJA

Az adatközlés egyik, nagy problémát jelentő kérdése a *közvetett azonosíthatóság*, azaz az adatfelfedés lehetősége. Maga az *adatvédelem* az a technika vagy módszer, amely alkalmazásával minimálisra csökkenthető a statisztikai egységek azonosításának veszélye.

Az adatközlés során az adatvédelmet készítők maguk döntik el, hogy a jogszabályban megfogalmazott „nagy időbefektetés során lehetővé válható kikövetkeztethetőség" mikor válhat lehetségessé. A közvetett felfedés elleni védekezés bonyolult, komoly munkát igényel, hiszen egy külső, harmadik fél technikai és tudásbeli hátterével szemben kell eszközöket találni. A külső fél, a lehetséges adatfelfedő jó- és rosszindulatú is lehet, különféle motivációkkal és eszközökkel. Az adatközlés számos publikációs formában ölthet testet, a papír alapútól az internetes közlésen át, és ezek eltérő védelmi technikákat, stratégiákat igényelnek.

Az adatfelfedés teljes mértékű megakadályozása által tökéletesen lehetetlenné válna az adatközlés, az adatokhoz való hozzájutás. Az egyre biztonságosabb adatközlés, az egyre nagyobb védelem mindig együtt jár azzal, hogy egyre több és több adatot kell elzárni a felhasználók elöl, és végül az elrejtett információknak köszönhetően használhatatlanná válhatnak adatbázisok.

A cél és egyben a legnagyobb kihívás a felfedés elleni védekezésben az, hogy megtaláljuk azt az optimális arányt az elrejtett, védett és a tájékoztatás révén közölt adatok közt, amivel már biztonságosnak tekinthetőek az adatok, és a felhasználók is hozzájuthatnak a megfelelő részletezettségű információkhoz. Ehhez ismernünk kell, hogy milyen kockázat rejlik a különféle adatközlésekben, és kik lehetnek a felhasználók *(Eurostat* [1999]).

(Azonosításon, azonosíthatóságon azt értjük, hogy egy anonim információhoz valamilyen módon hozzárendelhető, hozzákapcsolható egy egyedi azonosító (azonosítószám vagy kulcs). E mellett az adatfelfedés azt jelenti, hogy egy személyre vagy egy intézményre vonatkozóan új, plusz információ birtokába jutunk az azonosítás által. A két kategória tehát egymásból következik, hiszen plusz információ birtokába akkor jutunk, ha azonosítjuk a személyt. Tanulmányunkban mi e két kategóriával, s a kialakítandó védelemmel együtt foglalkozunk.)

### *Felfedési lehetőségek és kockázatok*

Az informatika nagyfokú elterjedtségének és technikai fejlődésének következtében a közölt adatok analizálásával, kombinálásával olyan új információ birtokába juthat egy külső, harmadik személy, amelyet az adatközlőnek nem állt szándékában közölni. Az adatok felfedése, kikövetkeztethetősége az adatok egyedisége, bizalmassága miatt kockázatossá válhat.

A területi szintű tájékoztatásban kiemelten jelentkezik a probléma: a terület nagysága, az alacsony lélekszám, vagy az adattartalom miatt válik nem közölhetővé az adat. Például:

– Ritka foglalkozások közlése (például: a budapesti agglomeráció egyik kis településén élő operaénekesnő közölt adatai név nélkül egyértelmű felfedést jelentenek).

– Egy átlagos foglalkozású (például bolti eladó) ember is azonosíthatóvá válik, ha csak egy emberről van szó a területen.

– Ugyancsak védendőek bizonyos egyedi, ritka családi vagy egyéb körülmények kis területre vonatkozó adatközlésben (például: 8 gyermekes család; magas jövedelmű személy).

Gazdasági szervezetek adatközlésénél számos probléma merülhet fel. A legkiemelkedőbb a dominancia problémája, vagy a monopol pozíciójú szervezetek, cégek adatai. Azonos jellemzőkkel rendelkező, azonos adatszolgáltatói csoportba tartozó, azonos terméket gyártó, azonos szolgáltatást nyújtó gazdálkodó szervezetek adatai statisztikai összesítés formájában bármikor közölhetők, ám amint valamelyik szervezet egyik mutatója kiugró, domináns értékkel bír (például legmagasabb foglalkoztatotti szám, legnagyobb bevétel, előállított egyedi termék stb.), akkor érzékennyé válik az adat.

Közérdekű és védendő adat együttes közlése során is felmerülhetnek adatvédelmi aggályok. (A Központi Statisztikai Hivatal pontosan felsorolja a közérdekű adatok körét.[13]) Közérdekű adat például a központi vagy helyi önkormányzati költségvetésből finanszírozott bölcsődei ellátásra vonatkozóan az ellátók száma, az ellátottak száma, a forgalom, a befogadóképesség és az ellátottak által fizetett hozzájárulás összesen. Amennyiben egy településen három bölcsőde működik, amelyből kettő állami és egy magán, akkor nagyon megfontoltnak kell lenni a felsorolt adattípusok együttes közlésekor, hiszen az egy magán bölcsőde adata így felfedhetővé, azaz nyilvánossá válik.

A mintavételes felvételek védelmét gyakran feleslegesnek tartják, holott a tájékoztatás módjától függően bizonyos esetekben védendő adattá válnak:

– Ha a megszerzett adatokból, tehát a mintából becslünk egy tulajdonságot egy legalább három fős sokaságra, akkor ezek az adatok nem lesznek védendőek, még akkor sem, ha becsült (és egyben a tájékoztatott) adatok egybeesnek valamely mintaelemmel.

– Abban az esetben viszont, ha a mintaelemeket „nyersen", mikroadat formájában szeretnénk közreadni, védelemmel kell ellátni őket. Gondoljuk csak el, hogy a szomszédunk elmeséli, hogy egy kérdezőbiztos a napi időbeosztásáról és tevékenységéről érdeklődött. Amennyiben birtokában vagyunk egy-két alapinformációnak szomszédunkról, az illető könnyen beazonosíthatóvá válik az adatbázis segítségével.

Előfordulhat, hogy az adatszolgáltatók magas száma ellenére is védenünk kell a cellát, például a kategóriák alacsony száma miatt. Olyan kérdésnél, amire igennel illetve nemmel lehet válaszolni, vagy kevés számú válaszlehetőség van – különösen, ha az adott kérdés valamely kényes, különleges dologra kérdez rá (például betegség, vallási hovatartozás, politikai vélemény) –, fokozottan figyelnünk kell. Ha ugyanis minden válaszadó azonosan, mondjuk igennel válaszol, akkor, amennyiben a többi, nem érzékeny kérdésre adott válaszból felfedünk valakit, akkor arról az egyedről olyan plusz, és érzékeny információ birtokába jutunk, aminek nem kellene tudtunkra jutnia: például, hogy milyen betegsége van, droghasználó-e vagy sem, milyen vallási közösség tagja stb. (Ez a fajta adatvédelmi probléma egyébként meglehetősen ritkán merül fel a nagy esetszámok miatt.)

Problémát okozhat az ugyanazon kiadványban vagy ugyanazon adatbázison alapuló különböző adatközlésekben a különböző táblák összeolvasásából azonosítható adatszolgáltató. Annak megoldása, hogy a keresztinformációkból ne váljon kikövetkeztethetővé az adatszolgáltató, a nagyon pontosan megtervezett és nyomon követett adatközléseken múlik.

## FELHASZNÁLÓK ÉS MOTIVÁCIÓK

Az adatok felhasználói az igényelt adatok és az igénylés módja szerint jól definiálható csoportokra oszlanak. Fontos ismernünk a különböző felhasználókat, hogy tudjuk, kitől, mikor és miért várható az adatfelfedés, és milyen következményekkel számolhatunk.

[13] IV/1997. (SK 3.) KSH szabályzat a statisztikáról szóló jogszabályokból adódó feladatok végrehajtásáról X.

A felhasználók alábbi osztályozását, amelyet az OECD a legjobbnak minősített, a dán statisztikai hivatal állította össze:

*1. „Farmerek" (szerződéses ügyfelek)*: Mindig ugyanarra a statisztikai adatra, szolgáltatásra van szükségük, általában ciklikusan. Az adatok közt nem válogatnak, csak olvassák azokat. Az igényeik kielégítése védett lekérdezés biztosításával, gyors adatátadással (például e-mailen keresztül), speciális információk előjegyzésével, valamint speciális formák (általuk elkészített egyéni táblázatok) alkalmazásával történik. A felhasználóknak ebbe a körébe tartoznak a pénzügyi szektor, a gazdálkodó és egyéb szervezetek képviselői, akik a statisztikán keresztül általában saját szakterületük alakulására kíváncsiak.

*2. „Turisták" (alkalmi böngészők)*: Ezek a felhasználók általános statisztikai adatokat igényelnek különböző területekről, témákról. Az adatok, dokumentumok könnyű, gyors elérésében érdekeltek. Mindenképpen szükséges számukra köznapi fogalmak és nyelvezet alkalmazásával magyarázatot fűzni a számokhoz. Ők azok, aki csak „felnéznek" az Internetre, és ők teszik ki a felhasználók mintegy 15 százalékát. Közéjük tartoznak a köznapi emberek, a diákok, a sajtó mindenre kíváncsi munkatársai.

*3. „Bányászok" (szakértők)*: Mélyre ásnak az adatokban, igénylik a minél részletesebb metszeteket. Vizsgálataikat sok és részletes információval, bizonyítékkal szeretnék alátámasztani. Egyedi adatokra is szükségük van, amelyekhez külön szerződés keretében hozzá is juthatnak. Elsősorban a kutatók és a tervezők tartoznak közéjük.

Az adatfelfedés szempontjából a farmerek felől érkező támadás a legvalószínűtlenebb. A *Bányászok* csoportba sorolható egyre több és több részletes adatkérőt viszont már potenciális támadónak kell tekintenünk. A legnagyobb – akár jó- vagy rosszindulatú – támadás a turisták, alkalmi böngészők csoportja felől érkezhet. Ugyanis a diákok, egyetemisták akár véletlenül is felfedhetnek bizalmas adatot, a média embere pedig szándékosan is kereshet bizalmas, védendő adatot.

Az adatfelfedést motiválhatják társadalmi, gazdasági, pszichológiai, politikai tényezők. Az adatvédelmi stratégia kialakításához fontos ezen tényezők ismerete.

Az azonosítási kísérletek elkövetőit két nagy csoportba sorolhatjuk a szerint, hogy elsősorban információszerzés céljából követik-e el a támadást, vagy pedig az adatgyűjtések, a statisztikai hivatal lejáratása, ilyen módon a közbizalom rontása a céljuk.

A támadás eszközei is széles skálán mozognak, a statisztikai és informatikai ismeretek alkalmazásától a nagy fokú számítógépes támogatásig. Ehhez járulnak a meglévő, mindenki számára elérhető információk, a köztudomású tények, a nyilvános adatbázisok, és az egyéb ismeretek egy témáról.

Természetesen minden egyes statisztikai felvétel esetén maghatározhatók támadási okok, motivációs célok, attól függően, hogy milyen érzékeny vagy érzékenynek vélhető kérdések szerepelnek a kérdőívben.

Egy angol tanulmány – elsősorban a népszámlálási adatok feltörése kapcsán – a következő lehetséges forgatókönyveket állapítja meg az adatfelfedési motivációkra, célokra nézve (*Elliot* [1996]):

*1*. Adatbázis gazdagítása népszámlálási adatokkal;

*2*. Adatbázis adatainak összevetése és megerősítése (lopott) népszámlálási adatokkal;

*3*. Egy jó újságíró sztori megírása annak bizonyítására, hogy adataink mennyire nem védettek;
*4*. Az állami adatgyűjtéseket és a mindenkori kormányzat hitelét rontó támadások;
*5*. Személyazonosító (szám vagy kulcs) ellopása;
*6*. Gazdasági versenytárs adatainak azonosítása.

(A példa az angol népszámlálásra vonatkozik.)

Az adatfelfedés hatása természetesen függ a behatoló céljától, a kísérlet sikerétől vagy kudarcától. Legyen azonban szó akárcsak egy adatszolgáltató személy azonosításáról, vagy egy gazdasági versenytárs adatainak megszerzéséről, az adatbiztonság mindenképpen sérül. Egy adatfelfedési kísérlet megtörténtének nyilvánosságra hozatala mindenképpen rontja a közhangulatot – akár sikeres volt a kísérlet, akár nem –, hiszen maga a tény, hogy egy behatolás (azonosítás) véghezvihető, önmagában rontja a statisztikai szolgálat hitelét. Ennek eredménye végső soron az lehet, hogy megszűnik maga a minőségi adatszolgáltatás.

## ADATVÉDELMI TECHNIKÁK, TÁBLÁZATOS ÉS MIKROADAT VÉDELEM

A természetes jóindulatú állampolgári adatigény és az ezzel szemben jelentkező rosszindulatú támadások, lejáratások miatt az adatközlőknek szükségük van egy olyan stratégiára, amellyel biztonsággal közölhetnek adatokat, és minimálisra csökkentik a támadható felületet.

A magyar jogszabály által nevesített szabály, azaz a minimum 3 elem egy cellában jó és szükséges védelmi szabály, de nem minden esetben nyújt elégséges védelmet a felfedés ellen. A világ országaiban számos jól bevált technika létezik az adatfelfedés megakadályozására. Ezek alapjait mutatjuk be a következőkben.

Az adatfelfedhetőség szempontjából alapvetően kétféle tájékoztatási formát különböztetünk meg. Az első, hagyományosnak mondható forma az, amikor táblázatos formában, azaz bizonyos dimenziók (tulajdonságok) kereszthivatkozásaiban hozzuk nyilvánosságra az adatokat. A másik esetben az adatokat rekordsorosan tesszük közzé. Ez utóbbit nevezzük mikroadatoknak. Ennek megfelelően beszélünk *táblázatos*-, illetve *mikro*-adatvédelemről. A kétféle védelem alapjait tekintve hasonlít egymásra, mégis különböző technikák alkalmazását igénylik.

### *Kockázat*

Az adatvédelem kialakításának első lépéseként felmérjük és megbecsüljük az adatközlésben felmerülő felfedési kockázatot. Ehhez ismernünk kell az adott statisztikai felvétel összes jellemzőjét: a sokaságot, az elemszámot, az esetleges mintát, a mintakiválasztás módját, a változókat, az adatokat, a főbb megoszlásokat stb. Ennek ismeretében tudunk dönteni valamelyik védelmi technika mellett, és határozhatjuk meg az adatvédelmi stratégiát.

Adatfelfedés több tényező együttes jelenléte esetén történhet meg, így a felfedési kockázat becslésére is több mód kínálkozik.

*Táblázatos adatokban rejlő kockázat*

A felfedés kockázata a táblázatos adatoknál párhuzamban áll a cella érzékenységének kritériumával. Az egyes cellák érzékenységének megléte és mennyisége határozza meg a kockázat mértékét. A gyakorlatban négy alapvető módszer terjedt el arra, hogy egy celláról kiderítsük, szükséges-e védeni vagy sem (*Carlson* [2002], *Merola* [2003]).

*Jelölések:*

$n$ – az adott cellába tartozó adatszolgáltatók száma,
$z_1 \geq z_2 \geq ... \geq z_n \geq 0$ – az adatszolgáltatóktól származó adatok nemnövekvő rendszere,
$T$ – a cella értéke, azaz $T = \sum_{j=1}^{n} z_j$ .

Ehhez kapcsolódóan definiálunk még három értéket:

$$t_m = \sum_{j=1}^{m} z_j, \qquad r_m = \sum_{j=m+1}^{n} z_j, \qquad R_{l,m} = \sum_{j=m+1}^{m+l} z_j ,$$

ahol $1 \leq m \leq n$.

*Küszöb szabály*: Ha az adatszolgáltatók száma egy meghatározott $M$ küszöbértéknél ($M \geq 1$) kevesebb, akkor a cella érzékeny. A cella biztonságosnak tekinthető, ha $n > M$.

*Dominancia szabály*: Érzékenynek tekinthető a cella, ha az értékét adó $z$-k közül $m$ db legnagyobb összegének a $T$-hez viszonyított aránya meghalad egy $k$ értéket (azaz dominánsak a cellában), ahol $0 < k < 1$. Az $m$ és a $k$ változtatásával alakíthatjuk a rendszerünk biztonságát: nagy $m$-mel és kicsi $k$-val nagy biztonság érhető el.

Választott $m$ és $k$ mellett a cella biztonságosnak tekinthető, ha

$$\frac{t_m}{T} < k .$$

Ez a szabály tulajdonképpen azt méri, hogy a legnagyobb elem vagy elemek mekkora arányban szerepelnek a teljes összegben. Ha egy elem 99 százalékát adja a cellaértéknek, akkor ezt nyílván nem szabad közölni, mivel nagyon kis hibával lehet következtetni erre az értékre. A nemzetközi gyakorlatban a két legnagyobb elem 80-85 százalékos részesedésénél már veszélyesnek tekintik a cellát. A paraméterekre nézve ez azt jelenti hogy: $m$=2, $k$=0,8 – 0,85.

*p-szabály*: Ez a szabály közvetlenül vizsgálja az egyes adatszolgáltatók adatainak a részvételét a teljes értékösszegben és feltételezi, hogy a támadó személy a cellát alkotó válaszadók közül kerül ki ($z_i$). Függetlenül az $n$ nagyságától, $T - z_i$ –t tekinthetjük úgy is, hogy egy becslés minden egyes $z_h$-ra ($1 \leq h \leq n$), azaz $\hat{z}_h = T - z_i$. A felfedési kockázat mértéke ennek a becslésnek a relatív hibája: $(\hat{z}_h - z_h)/ z_h$. Minél kisebb a $z_h$, annál roszszabb ez a becslés. Nyílván a $T$-hez legközelebb álló értékek ($z_1$, $z_2$) adják a legjobb becslést, és ebben az esetben a legvalószínűbb is az adatfelfedés. Tehát ezt alapul véve kell megállapítani a kockázatot:($h$=1, $i$=2): $(T - z_1 - z_2)/ z_1$. A szabály megköveteli, hogy ez a

relatív hiba nagyobb legyen, mint egy előre megadott $p > 0$ érték (*Cox* [1981]). Így biztonságosnak tekinthető egy cella, ha

$$\frac{r_2}{z_1} > p\,.$$

*pq-szabály*: Ez tulajdonképpen a p-szabály általánosítása, ahol a $p$-t alulról korlátozzuk egy $q \geq 0$ számmal. Tehát a $0 \leq q < p$ figyelembevételével biztonságosnak mondható a cella, ha

$$\frac{r_2}{z_1} > \frac{q}{p}\,.$$

*A mikroadatokban rejlő kockázat*

Számos módszer kínálkozik adataink ellenőrzésére. A szakirodalom (*Skinner–Elliot* [2002], *Carlson, M.* [2002]) alaposan tárgyalja ezeket a számítási módokat, ezek közül a legáltalánosabbat részletezzük. A módszer alapjául az egyednek a sokaságban való előfordulási gyakorisága szolgál. Első lépésként megvizsgáljuk minden egyes egyed gyakoriságát a sokaságban, kiszámítjuk az egyes egyedek, rekordok kockázatát, majd ez után kiszámítható a teljes adatstruktúra kockázata. Fontos betartani ezt a két lépcsős számítást, mivel az egyes rekordokban rejlő esetleges alacsony kockázat nem jelenti automatikusan a teljes adatstruktúra biztonságosságát. Ennek az az oka, hogy a rekordszintű kockázati valószínűségek összeadódnak.

*Példa*: $U$ jelöli a teljes (véges) sokaságot, $X$ az azonosító változók lehetséges kombinációinak összességét, $J$ pedig a kombinációk számát. Ekkor az $X$ például olyan elemekből fog állni, hogy „Férfi–50éves–Fogorvos”. Minden egyes ilyen (rész)sokaságnak meg kell határozni a gyakoriságát, azaz hogy hány egyed tartozik ebbe a tulajdonságkörbe. $F_j$ a $j$-edik sokaság gyakorisága. $I$ az indikátor függvényt jelöli.

$$F_j = \sum_{i \in U} I(X_i = j), \qquad j = 1,\ldots,J$$

Ebből már látható, hogy milyen gyakoriságúak az egyes sokaságok. Fontos tudni azt is, hogy az egyes gyakoriságokból hány darab van, mivel ha kétszer több egyelemű sokaság van, akkor kétszer nagyobb a felfedés kockázata is. Ha

$$N_r = \sum_{j=1}^{J} I(F_j = r), \qquad r = 1,2,\ldots\,,$$

ami a gyakoriságok gyakoriságát jelöli, akkor ez alapján fel tudjuk írni a felfedési kockázatot:

$$P = \frac{N_1}{N} = \frac{\sum_j I(F_j = 1)}{N}.$$

$N$ a sokaság méretét jelöli. $N_1$ került a számlálóba, mivel a felfedés legnagyobb kockázata az egy elemű sokaságokban ($N_1$) rejlik. Ha $N_1$=0 akkor a következő legkisebb nemnulla részsokaság gyakoriságát kell tekintenünk. A $P$ meghatározza, milyen valószínűséggel fedhetőek fel az adataink. A rendszerünktől megkövetelt biztonságától függ, hogy mikor tekintjük ezt elfogadhatónak és mikor nem. Az alacsony elemszámú sokaságok megszűntetésével természetszerűen csökkenthető a $P$ értéke. Ennek módjáról a következőkben mutatunk be módszereket.

## TÁBLÁZATOS ADATVÉDELEM

A tájékoztatás szempontjából kétféle táblázatot különböztetünk meg. Az egyik a gyakorisági (frequency), másik a értékösszeg (magnitude) tábla. A gyakorisági tábla tartalmazza az adatszolgáltatók számát, az értékösszeg tábla pedig az ezen adatszolgáltatók által szolgáltatott adatok összességét. A védendő adatok feltérképezéséhez minden egyes tájékoztatásra kerülő táblához el kell készíteni annak gyakorisági tábláját is. A következő példa ezt szemlélteti.

1. tábla

*Értékösszeg tábla*
*Árbevétel (millió forint)*

| | Ipar | Mező-gazdaság | … | Összesen |
|---|---|---|---|---|
| *1*. város | 124 | 0 | … | … |
| *2*. város | 236 | 377 | … | … |
| *Összesen* | *360* | *377* | … | … |

*Gyakorisági tábla*
*Vállalatok száma*

| | Ipar | Mező-gazdaság | … | Összesen |
|---|---|---|---|---|
| *1*. város | | 0 | … | … |
| *2*. város | 6 | 1 | … | … |
| *Összesen* | *7* | *1* | … | … |

Természetesen a két tábla megegyezik abban az esetben, ha a tájékoztatásra kerülő adataink pont az adatszolgáltatók számát jelöli.

Ha a két táblázat nem egyezik meg, és csak az értékösszeg táblát jelentetjük meg, akkor az adatokból nem derül ki, hogy mely cellák rejtenek mindössze 1 vagy 2 adatszolgáltatót. Ez is jelent önmagában egy minimális védelmet, de a védelem kialakításánál fel kell tételeznünk, hogy ezen információ megszerzéséhez nem kell különösebb detektív képességgel rendelkezni, hiszen például a gazdasági életben a cégek tudják, hány hozzájuk hasonló van a piacon.

A példákban a sötéttel jelzett cellák értékei jelentik azokat az adatokat, amelyeket nem közölhetünk. Az egyszerűség kedvéért az értékösszeg- és gyakorisági tábla adatai egyértelműen megfeleltethetők egymásnak, és az érzékenység kritériuma az 1 vagy 2 adatszolgáltató ténye. A cél tehát az, hogy „megszüntessük” ezeket a cellákat.

### *Aggregálás*

A módszer lényege, hogy oszlopok illetve sorok összevonásával cellákat egyesítünk, növelve ezzel az egy cellában lévő adatszolgáltatók számát (*Eurostat* [1996]).

Az összevonás alapja a következő két ismérv:

– *minőségi ismérv*: Két hasonló, vagy minimális számú hasonló dimenzióértékeket vonunk össze;
– *mennyiségi ismérv*: A skálázás alapjául vett mennyiségértékeknek állapítunk meg új határokat.

Példák a módszer szemléltetésére:

*a*) *A tábla méretének kicsinyítése* (minőségi ismérv)

2. tábla

*Eredeti tábla*

| | Kék szemű | Zöld szemű | Barna szemű | Albínó (piros) szemű | *Összesen* |
|---|---|---|---|---|---|
| Férfi | 12 | 10 | **2** | 6 | *30* |
| Nő | 24 | **2** | 6 | 8 | *40* |
| *Összesen* | *36* | *12* | *8* | *14* | *70* |

Az „Zöld szemű” és „Barna szemű” oszlopokban kis értékű cellákat találhatunk. Összevonjuk őket, feltételezve azt, hogy ezek a dimenzióértékek egy meghatározott szempont szerint összetartozónak tekinthetők. Egy érzékeny cellát tartalmazó oszlopot természetesen összevonhatunk olyan oszloppal is, amelyben nem szerepel érzékeny adat.

3. tábla

*Védett tábla*

| | Kék szemű | Zöld és barna szemű | Albínó (piros) szemű | *Összesen* |
|---|---|---|---|---|
| Férfi | 12 | 12 | 6 | *30* |
| Nő | 24 | 8 | 8 | *40* |
| *Összesen* | *36* | *20* | *14* | *70* |

*b*) *A karakterisztika újrakódolása* (mennyiségi ismérv)

4. tábla

*Eredeti tábla*

| Kor | <12 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | >20 | *Összesen* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Férfi | 23 | 3 | 3 | 7 | 7 | 3 | 4 | 4 | 7 | 4 | 15 | *80* |
| Nő | **2** | **2** | **1** | **1** | **1** | **2** | **2** | **2** | **1** | **1** | 5 | *20* |
| *Összesen* | *25* | *5* | *4* | *8* | *8* | *5* | *6* | *6* | *8* | *5* | *20* | *100* |

Ennél a módszernél egy meghatározott skálázási tulajdonság alapján új intervallumokat állapítunk meg.

5. tábla

*Védett tábla*

| Kor | <13 | 13-15 | 16-19 | 20 vagy < | *Összesen* |
|---|---|---|---|---|---|
| Férfi | 26 | 20 | 19 | 15 | *80* |
| Nő | 4 | 5 | 6 | 5 | *20* |
| *Összesen* | *30* | *25* | *25* | *20* | *100* |

Ezt az adatvédelmi technikát gyakran alkalmazzák a statisztikai munkában. (Papír alapú kiadványokban előfordul, hogy nem pusztán a védelem miatt használják, hanem a táblázatok kisebb mérete miatt az összevont kategóriák áttekinthetőbbek. Az összevonás során például minden egyes korév helyett öt évenként összevont korcsoportok jelennek meg.)

*Az aggregálás előnye:*

– az adatok nem torzulnak, azaz a táblázat adatai a valóságot tükrözik, ami a a *legfontosabb* szempont felhasználók számára;
– könnyen megvalósítható;
– az adatbázisban léteznek olyan ismérvek (régió-megye-város stb.), amelyek a hierarchikus felépítés miatt közvetlen alapjául szolgálhatnak az eljárásnak.

*Az aggregálás hátránya:*

– az összevonások során a háttérbe kerülhetnek részletes tulajdonságok, vagyis *információveszteséggel* kell számolni. A fenti példában minden egyes korév helyett például csupán négy összevont korcsoport kategória jelenik meg.

Igen szemléletes példáját láthatjuk itt az adatvédelmi mérlegelésnek. Dönteni kell, hogy megengedhető-e az összes korév megjelenése, vagy 5, esetleg 10 éves korcsoportokat kell közölni, netán csak 3-4 korcsoport kategória jelenhet csak meg

*Cellaelnyomás*

Amennyiben adatok nagyfokú részletességgel történő közlése a cél, az egyik megoldás a cellák elnyomásának módszere.

A cellaelnyomás lényege, hogy az érzékenynek ítélt cellák tartalmát egyszerűen kitöröljük, ezt nevezzük *elsődleges elnyomásnak*. Mivel a kitörölt adat sorában lévő többi adatból, illetve az „összesen" mezőből ezt követően is egyértelműen meghatározható lenne a cella értéke, ezért a biztonság növelése érdekében további cellákat kell „elnyomni", ezt nevezzük *másodlagos elnyomásnak*. Különféle algoritmusok léteznek annak meghatározására, hogy mely cellákat kell még járulékosan kitörölni a védelem biztosításához (*Hundepool* [1999]).

Kétféle cellaelnyomás létezik:

– a cella tartalmának teljes kitörlése; illetve
– olyan intervallum megadása a cellában, amelybe a cella értéke belelesik.

Példa a módszer szemléltetésére:

6. tábla

*Eredeti tábla*

| | Barna szemű | Kék szemű | *Összesen* |
|---|---|---|---|
| Fekete hajú | Védendő cella | 3 | *7* |
| Barna hajú | 2 | 1 | *3* |
| Szőke | 3 | 3 | *6* |
| *Összesen* | *9* | *7* | *16* |

7. tábla

*Elsődlegesen és másodlagosan elnyomott cellák*

| | Barna szemű | Kék szemű | *Összesen* |
|---|---|---|---|
| Fekete hajú | X | X | *7* |
| Barna hajú | X | X | *3* |
| Szőke | 3 | 3 | *6* |
| *Összesen* | *9* | *7* | *16* |

| | Barna szemű | Kék szemű | *Összesen* |
|---|---|---|---|
| Fekete hajú | 3-6 | 1-4 | *7* |
| Barna hajú | 0-3 | 0-3 | *3* |
| Szőke | 3 | 3 | *6* |
| *Összesen* | *9* | *7* | *16* |

Amennyiben nem kívánjuk teljesen elrejteni a számokat, egyetlen tartományt adunk meg az elsődlegesen és másodlagosan elnyomott cellákra.

*A cellaelnyomás előnye:*

– a látható adatokat részletes felosztásban kapjuk meg, ami több információt jelent;
– a látható adatok a valóságot tükrözik, vagyis nem torzítottak;
– léteznek szoftverek, melyek optimalizálják a másodlagos cellaelnyomást.

*A cellaelnyomás hátránya:*

– a másodlagos cellaelnyomásokkal olyan cellák is rejtve maradnak, melyek egyébként közölhetőek lennének. Egy érzékeny cellához további 2-3 cellára kell alkalmazni a másodlagos cellaelnyomást, ennek következtében jelentősen megritkulhat a táblázat;

– bonyolult és hosszadalmas algoritmus végrehajtása szükséges ahhoz, hogy meghatározzuk azt a minimális számú törlendő cellát, amellyel a védelem még fennáll.

*Kerekítés*

Ennél az adatfelfedés elleni módszernél nem követeljük meg a cellaadatoktól, hogy pontosan tükrözzék a valóságot. A felfedési valószínűséget úgy is lehet csökkenteni, ha

nem szolgáltatunk a felhasználónak pontos értékeket, hanem az összes cella értékét – beleértve az „összesen" cellákat is – kerekítjük egy hozzá közel eső szintre.

A módszer legnagyobb előnye hihetetlenül egyszerű megvalósításában rejlik, a felhasználók körében mégsem arat osztatlan sikert, mivel meglehetősen bizalmatlanul kezelik ezeket az adatokkal. A felhasználóknak azonban nem szabad elfeledkezniük arról, hogy az általunk „elrejtett" értékek is hordozhatnak magukban hibákat (például mintavételi hibát).

*A kerekítési folyamat:*

Elsődlegesen megválasztunk egy *b* értékét, amit az egészszámú kerekítés *alapjának* nevezünk. $N_{ij}$ jelöli az *i*-edik sor *j*-edik oszlopának cellaértékét.

Egy cellában lévő érték kerekítésének lépései (*Eurostat* [1996]):

*1*. Meghatározzuk azt a legnagyobb *h*-t (*szorzóérték*) amelyre teljesül, hogy $N_{ij} \geq h*b$

*2*. Így adódik a $r_{ij}$ maradék: $N_{ij} = h*b + r_{ij}$, ahol $0 \leq r_{ij} < b$.

*3*. A maradékot kerekítjük, 0-ra vagy *b*-re. Így adódik a cella új értéke ($N_{ij}$'): ha $r_{ij}=0$ vagy *b*, akkor nyilvánvalóan: $N_{ij} = N_{ij}'$.

A különböző megoldási módozatok sajátosságai a *b* értékének megválasztásában rejlenek.

8. tábla

*Eredeti tábla*

| | Kék szemű | Zöld szemű | Barna szemű | *Összesen* |
|---|---|---|---|---|
| Fekete hajú | 1 | 4 | 0 | *5* |
| Barna hajú | 15 | 10 | 10 | *35* |
| Vörös hajú | 2 | 10 | 8 | *20* |
| Szőke | 2 | 6 | 15 | *23* |
| *Összesen* | *20* | *30* | *33* | *83* |

*a*) *Rögzített kerekítés*

A *b* értéket rögzítjük, és az előbb leírtak alapján alkalmazzuk a kerekítést.

A tábla minden egyes cellájára elvégezzük a kerekítést (példánkban legyen *b*=5) és akkor a táblázat az alábbi módon alakul.

9. tábla

*Rögzített kerekítéssel védett tábla*

| | Kék szemű | Zöld szemű | Barna szemű | *Összesen* |
|---|---|---|---|---|
| Fekete hajú | 0 | 5 | 0 | *5* |
| Barna hajú | 15 | 10 | 10 | *35* |
| Vörös hajú | 0 | 10 | 10 | *20* |
| Szőke | 0 | 5 | 15 | *25* |
| *Összesen* | *20* | *30* | *35* | *85* |

A módszer előnye, hogy egyszerűen kiszámolható, és minimalizálja a tényleges értéktől való eltérést.

*b*) *Véletlen kerekítés*

A *b* értéke itt is rögzített, viszont a kerekítés már nem a hagyományos módon történik. Az $N_{ij}$ értéket *p* valószínűséggel kerekítjük lefelé, és 1–*p* valószínűséggel kerekítjük felfelé. Ez a következőt jelenti:

Ha *b*=5, akkor a kerekítés valószínűségei a maradék függvényében a következőképpen alakulnak:

| $N_{ij}$ b-vel való osztásának a maradéka | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| 0-ra való kerekítés valószínűségei, *p*= | 1 | 4/5 | 3/5 | 2/5 | 1/5 | 0 |
| 1-re való kerekítés valószínűségei, 1–*p*= | 0 | 1/5 | 2/5 | 3/5 | 4/5 | 1 |

A *p*-t tehát egyenletesen kell megválasztani, a maradék és a *b* hányadosaként. Ily módon a valószínűséggel történő kerekítés biztosítja a módszer torzítatlanságát, azaz $E(N_{ij}')=N_{ij}$ (*Eurostat* [1996]). (Ha például a maradék 1, akkor $E(N_{ij}')=4/5(N_{ij}-1)+1/5(N_{ij}+4)= N_{ij}$)

A módszer sem biztosítja, hogy az oszlop és sorösszegek kiadják az egyes elemek összegét, mivel minden egyes elemre (beleértve az összegeket is) külön-külön végzzük el a kerekítéseket, és nem vesszük figyelembe az elem és az összegértékek viszonyát.

10. tábla

*Véletlen kerekítéssel védett tábla (b=5)*

| | Kék szemű | Zöld szemű | Barna szemű | *Összesen* |
|---|---|---|---|---|
| Fekete hajú | 0 | 0 | 0 | *5* |
| Barna hajú | 15 | 10 | 10 | *35* |
| Vörös hajú | 0 | 10 | 10 | *20* |
| Szőke | 0 | 10 | 15 | *20* |
| *Összesen* | *20* | *30* | *35* | *85* |

*c*) *Ellenőrzött kerekítés*

A kerekítésnek ez a fajtája annyiban különbözik a véletlen kerekítéstől, hogy járulékos ellenőrzéssel megpróbálunk eleget tenni az additivitásnak is, vagy annak, hogy a sorok és oszlopok kiadják az „összesen” mező értékeit. Ennek megvalósítására a leggyakrabban a Cox & Ernst algoritmust használják, melynek során a fel-le kerekítéseket úgy határozzák meg, hogy az kiadja a sor illetve oszlopösszegeket. (*Fischetti– Salazar-González* [1998], *Eurostat* [1996], *Ernst* [1989]).

11. tábla

*Ellenőrzött kerekítéssel védett tábla*

| | Kék szemű | Zöld szemű | Barna szemű | *Összesen* |
|---|---|---|---|---|
| Fekete hajú | 0 | 5 | 0 | *5* |
| Barna hajú | 15 | 10 | 10 | *35* |
| Vörös hajú | 0 | 10 | 10 | *20* |
| Szőke | 5 | 5 | 15 | *25* |
| *Összesen* | *20* | *30* | *35* | *85* |

*Dimenziókorlátozás*

Ez a védelmi módszer csak az elektronikus tájékoztatási formánál alkalmazható. Egyes tájékoztatási rendszereknél olyan formában érhetőek el az adatok, hogy a felhasználó által kiválasztott tulajdonságoknak (dimenziók) megfelelő adatokat kapja meg az adatigénylő táblázatos formában. Az adatkérő a kiválasztott tulajdonságok növelésével egyre részletesebb adatokhoz jut, és egyben növeli az azonosíthatóságot és ezzel együtt a felfedési kockázatot is. Ilyen esetben célravezető védelmi megoldás, hogy maximalizáljuk a választható tulajdonságok számát (legyen ez a szám $n$). Ezt az értéket úgy kell megválasztani, hogy a tulajdonságokból bármely $n$ darabot választva sem juthassunk olyan információhoz, ami védendőnek tekinthető.

A módszer egyszerű és könnyen megvalósítható. A probléma csak az, hogy sok információ maradhat rejtve, ha az adatstruktúra egy részében kevés tulajdonság választása esetén is sok védendő adatot kapunk, és emiatt kicsire kell választanunk az $n$-t.

Ebből kifolyólag a gyakorlatban ezt a módszert csak „elővédelemnek” szokták alkalmazni, olyan formában, hogy egy alkalmas $n$ választásával levágják az adathalmaz peremét (mivel itt a legvalószínűbbek az egyedi adatok), a továbbiakban felmerülő eseteket pedig lokális védelemmel látják el.

A dimenziókorlátozás klasszikus módszertanának vannak változatai, amelyekkel átfogóbb védelmet alakíthatunk ki:

*Szelektív dimenziókorlátozás*: Meg kell vizsgálni, hogy mi az a maximális $n$, amely mellett nem érhető el védendő cella. Az $n$-t 3-4-nél kevesebbre nincs értelme választani, még akkor sem, ha a vizsgálatok azt bizonyítják, hogy kevesebbnél kellene meghúzni a határt, mivel az elérhető adatok aránya vészesen lecsökken. A gyakorlatban megfigyelhető, hogy sokszor csak egy-két dimenziópárosítás választásával érhetőek el védendő cellák. Ezeknek a párosításoknak a letiltásával növelhetjük az $n$ értékét.

*Differenciált dimenziókorlátozás*: A lekérdezett dimenziók számához különböző részletességű adatbázist párosítunk. A választott dimenziók számának növelésével csökkentjük a megjelenítendő adatok részletességét. A felhasználó természetesen egy dimenzió választása esetén kapja a legbővebb adatbázist.

## MIKROADAT-VÉDELEM

Mikroadatok alatt az egy statisztikai egységről birtokunkban lévő legrészletesebb adatokat értjük. Ezek az adatok a gyakorlatban rekordsoros állományokban vannak eltárolva**,** az *egy sor egy adatszolgáltató* elv alapján.

A jogszabályok alapján anonimizált mikroadatok olyan egyedi statisztikai adatok, amelyeket annak érdekében módosítottak, hogy a mindenkori legjobb eljárással összhangban minimálisra csökkenjen az érintett statisztikai egységek azonosításának veszélye.

Ilyen adatállományok teljes vagy részleges publikálása is csak az megfelelő anonimizálás után tehető meg. A jogi részben megismertek alapján az egyedi azonosítók esetében nincs mérlegelési jogkörünk azok megtartására, egyszerűen *ki kell* törölni őket. A további vizsgálataink tárgyát a fennmaradó oszlopok képezik.

12. tábla

*Személyek adataiból álló mikroadatbázis*

| Név | Lakhely | Születési hely | Születési idő | Foglalkozás | Vallás | … |
|---|---|---|---|---|---|---|
| Kala Pál | Iszapszentmotoros | Iszapszentmotoros | 1881.01.02. | Tűzkő árus | – | … |
| Hó Virág | Tápiórettentő | Tápiórettentő | 2031.02.12. | Tűzoltó | Szombatista | … |
| … | … | … | … | … | … | … |

13. tábla

*Gazdasági szervezetek adataiból álló mikroadatbázis*

| Cégnév | Telephely | Alapítás dátuma | Tevékenység | Alaptőke (millió forint) | … |
|---|---|---|---|---|---|
| Gépolaj Rt. | Markotabödöge | 1844.06.12. | Szállítmányozás | 234 | … |
| Sikattyu Kft. | Nagybajom | 2021.12.22. | Költöztetés | 133 | … |
| … | … | … | … | … | … |

Látható, hogy a mikroadatok esetében sokkal szorosabb kapcsolat van az *azonosítás* és a *felfedés* között, mint a táblázatos adatoknál. Ezért beszélünk itt anonimizálásról, nem pedig felfedésről: a speciális rekordsorban szereplő adatok miatt az azonosítás itt önmagában felfedést is jelent. Tehát itt nem az érzékeny adatok elrejtésén van a hangsúly, hanem a rekordnak az egyénhez való társításának megakadályozásán. Ahhoz, hogy kicsi legyen a kockázat, csökkenteni kell a legkisebb gyakoriságú részsokaságok számát.

Az anonimizálás itt is tartalmaz bizonyos fokú információveszteséget, de a védelmi technikáknál éppen az a célunk, hogy megtaláljuk azokat az adatokat amelyek elrejtésével a legkevesebb az információveszteség, és közben az anonimitásnak is eleget teszünk.

### *Csonkolás*

Ez a technika a legnyilvánvalóbb és egyben első helyen alkalmazott. Csonkolásnál egy teljes oszlopot kitörlünk az adatbázisból. Ezt a módszert alkalmazzuk akkor is, amikor az egyedi azonosítókat leválasztjuk az adatbázisról.

14. tábla

*Védelem kialakítása csonkolással*

| Születési hely | Szül.idő | Foglalkozás | Vallás | … |
|---|---|---|---|---|
| XXX | 1881.01.02. | Tűzkő árus | – | … |
| XXX | 2031.02.12. | Tűzoltó | Szombatista | … |
| XXX | … | … | … | … |

A fő probléma annak eldöntésében rejlik, hogy mely oszlop kitörlésével érhetjük el a kellő anonimitást. Ennek eldöntésére meg kell vizsgálnunk, hogy vannak-e olyan tu-

lajdonság- kombinációk (például: „Születési hely” és „Születési idő”) amelyek egyedivé teszik az egyes vagy akár az összes rekordokat. Ezek azok az oszlopok potenciális jelöltjei a csonkolásnak. A csonkolási technika alkalmazása egy ciklikus folyamat. Minden egyes lépésnél csakis egyetlen oszlopot szabad kitörölni, és ezután újra meg kell vizsgálni, mely rekordok maradtak még továbbra is kritikusak a felfedés szempontjából.

A csonkolási technika adatbázisok védelménél igen durva beavatkozásnak számít, hiszen hatására dimenziók tűnnek el. Mivel a tájékoztatás célja, hogy minél több információt biztosítsunk a felhasználóknak, így a mikroadatbázis egészénél a csonkolás mellett gyakran más módszereket is alkalmaznak.

*Cellaelnyomás*

A cellaelnyomás során egyes tulajdonságok „vészesen kevés számú” előfordulásait kell kitörölni. A tulajdonságok vészesen kevés előfordulásai során arra kell gondolni, hogy az adatbázis információi egyediek, így közvetlenül beazonosítható az adatszolgáltató. A minőségileg egyedi és a mennyiségileg kevés vagy kiugróan sok elemszám teszi kritikussá, azonosíthatóvá a rekordot, és így az adatszolgáltatót is.

*Példa*: Ha Iszapszentmotoroson csak egy tűzkőárus van, akkor ez egyértelmű azonosítást, közvetett adatfelfedést tesz lehetővé. Tehát itt a lakhely és a foglalkozás kombinációja kritikus a védelem szempontjából. Bármelyik cella rejtetté tétele megoldja a problémát. A döntés a védelmet kialakító egyéntől függ, illetve attól, hogy lakhelynek vagy a foglalkozásnak a kombinációja fontos-e a többi adatéval összevetve.

15. tábla

*Védelem kialakítása cellaelnyomással*

| Születési hely | Születési idő | Foglalkozás | Vallás | … |
|---|---|---|---|---|
| Iszapszentmotoros | 1881.01.02. | XXX | – | … |
| Tápiórettentő | 2031.02.12. | Tűzoltó | Szombatista | … |
| … | … | … | … | … |

Ez a módszer enyhébb, mintha csonkolással eltávolítottuk volna az egész foglalkozási vagy vallási oszlopot, de tény, hogy ez is adatvesztéssel jár.

*Átkódolás*

Az adatszolgáltatók kilétével kapcsolatos bizonytalanság kialakítható úgy is, ha nem az általunk ismert legpontosabb adatot írjuk a cellába. Ez nem az jelenti, hogy a cellák nem valós értékeket tartalmaznak, hanem csak annyit, hogy egy bővebb tartományba helyezzük át a tulajdonságot, tulajdonságokat.

*Példa*: Ha egy városban csak egy balettcipő-készítő van, akkor érdemes összevonni a cipőkészítővel. Ennek megfelelően a cipőkészítőt és a balettcipő-készítőt át kell írni „Cipő- és balettcipő-készítő”-re. A következő két példa talán még szemléletesebbé teszi az elvet.

16. tábla

*Védelem kialakítása átkódolással*

| Születési hely | Születési idő | Foglalkozás | Vallás | … |
|---|---|---|---|---|
| Iszapszentmotoros | 1881.01.02. | Tűzoltó-Tűzkőárus | – | … |
| Tápiórettentő | 2031.02.12. | Tűzoltó-Tűzkőárus | Szombatista | … |
| … | … | … | … | … |

| Alapítás dátuma | Tevékenység | Alaptőke (millió forint) | … |
|---|---|---|---|
| 1844.06.12. | Szállítmányozás-Költöztetés | 234 | … |
| 2021.12.22. | Szállítmányozás-Költöztetés | 133 | … |
| … | … | … | … |

## *Kerekítés*

A táblázatos adatok védelménél bemutatott kerekítés, teljesen azonos módon alkalmazható számértékű mikroadatoknál is. A *b* érték növelésével növekszik a bizonytalanság a tényleges értékre, viszont ezzel arányosan sajnos növekszik az adat használhatatlansága is. A legnagyobb feladat az optimális *b* választásában rejlik, amely mindig függ a kerekítendő szám nagyságrendjétől, illetve attól, hogy az egyes értékek milyen tartományban mozognak. Nyílván lényegesen eltérő *b*-t kell választani milliós, illetve ezres nagyságrendű értékeknél. Ügyelni kell arra is, hogy ne forduljon elő az, hogy a kerekítést követően olyan számot kapjunk, amelyet az egyébként jellemzett tulajdonság fel sem vehet.

17. tábla

*Védelem kialakítása kerekítéssel (b=50 millió forint)*

| Alapítás dátuma | Tevékenység | Alaptőke (millió forint) | … |
|---|---|---|---|
| 1844.06.12. | Szállítmányozás | 250 | … |
| 2021.12.22. | Költöztetés | 150 | … |
| … | … | … | … |

## *Összekeverés*

A felfedési kockázatot azzal is tudjuk csökkenteni, hogy az egyes emberekhez tartozó érzékeny adatokat véletlenszerűen összekeverjük. Így az egyes emberek adatain nincs mit felfedni, viszont a sokaságok egészére nézve nem változott semmi.

18. tábla

*Védelem kialakítása összekeveréssel*

| Születési idő | Lakhely | Foglalkozás | Vallás | … |
|---|---|---|---|---|
| … | … | … | … | … |
| 1881.01.02. | Iszapszentmotoros | Tűzkő árus | – | |
| 2031.02.12. | Tápiórettentő | Tűzoltó | Szombatista | … |
| … | … | … | … | … |

*

Tanulmányunkban bemutattuk az adatfelfedés elleni védelem környezetét, valamint a a védelem különböző eszközeit. Mi ezen eszközök „tiszta" bemutatására törekedtünk, és olyan példákat hoztunk, amelyek megfelelően reprezentálták az eszközök bemutatását. A gyakorlati statisztikai munkában azonban komoly háttérmunkát jelent annak eldöntése, hogy mely módszer vagy módszerek alkalmazása a legcélravezetőbb egy adott statisztikai felvétel közlésekor; a megvalósítás pedig összehangolt statisztikai-matematikai-informatikai eszköztárat igényel.

Egyes módszerek információ-veszteséggel vagy információ-torzulással járnak. Mint korábban jeleztük, az adatvédelmi technikák alkalmazása egy finom mérleghez hasonlít, ahol azt kalkuláljuk, hogy a tájékoztatás formájának célja, módja milyen védelmi technikát igényel, azaz mit nyerünk az egyik oldalon és mit veszítünk a másikon, s hogyan kerül a kettő egyensúlyba.

A tájékoztatás során a felhasználót mindig tájékoztatják az adatvédelemről. Ez történhet úgy, hogy egy kerekítésnél megadják a kerekítés mértékét, lábjegyzetben vagy mellékletben jelezik, hogy milyen eljárással módosították az adatokat, mikroadat-állomány közlése során pedig csatolnak az állományhoz egy leírást az adatvédelem módjáról és hatásáról. Egyes országok nem közlik pontosan az alkalmazott adatvédelmi technikát, csak a pontosság és megbízhatóság mértékét, illetve, hogy az adatok milyen korlátok és feltételek között alkalmazhatóak.

A védelmi eljárás alkalmazásának tényéről azonban mindig történik tájékoztatás. Ez a nemzetközi gyakorlat is, hiszen ez az etikai lépés biztosítja a kölcsönös bizalmat, és a korrekt statisztikai munkát az adatközlő és a felhasználó oldalán is.

A célhoz, a minél szélesebb körű tájékoztatáshoz kezünkben vannak tehát az eszközök, amelyek alkalmazásával elégedett lehet mind a tájékoztató statisztikai intézmény, mind az adatigénylő felhasználó.

## IRODALOM


BÁNSZEGI K. [1997]: Felfedést akadályozó módszerek a statisztikai tájékoztatásban. *Statisztikai Szemle*. 75. évf. 12. sz. 1039–1046. old.

BÁNSZEGI K. – LAKATOS M. [1994]: Információszabadság – adatvédelem – statisztika (III.). *Statisztikai Szemle*. 72. évf. 10. sz. 761–777. old.

CARLSON, M. [2002]: Assessing microdata disclosure risk using the poisson inverse gaussian distribution. Stockholm. Kézirat. (http://www.matstat.umu.se/banocoss/papers/carlson.pdf)

COX, L. H. [1981]: Linear sensitivity measure in statistical disclosure control. *Journal of Statistical Planning and Inference*. 5. évf. 2. sz. 153–164. p.

DUNCAN, G. T. – KELLER-MCNULTY, S. A. – STOKES, S. L. [2001]: *Disclosure risk vs. data utility: The R-U confidentiality map*. Kézirat. (http://www.niss.org/technicalreports/tr121.pdf)

ELLIOT, M. [1996]: Attacks on census confidentiality using the sample of anonymised records: an analysis. 3rd International Seminar on statistical confidentiality. Bled 1996.

ERDEI V. – SÁNTA J. [2000]: *A statisztikai adatok védelmének nemzetközi szabályozása, módszertani kérdései.* Népszámlálások az ezredfordulón 3. *(Tanulmányok)* Központi Statisztikai Hivatal. Budapest.

ERNST, L. R. [1989]: Further application on linear programming to sampling problems. Kézirat. (http://www.census.gov/srd/papers/pdf/rr89-05.pdf)

*Eurostat* [1996]: *Manual on disclosure control methods*. Luxemburg.

*Eurostat* [1999]: Statistical data confidentiality.

FAGAN, J. T. – GREENBERG, B. V. – HEMMING, B. [1988]: *Controlled rounding of three dimensional tables.* Kézirat. (http://www.census.gov/srd/papers/pdf/rr88-02.pdf)

FISCHETTI, M. – SALAZAR-GONZÁLEZ, J. J. – Caprar, A. [1998]: *Computational experience with the controlled rounding problem in statistical disclosure control*. Padova. Kézirat. (http://neon.vb.cbs.nl/casc/ISIBerlin/Salazar.pdf)

FISCHETTI, M. – SALAZAR-GONZÁLEZ, J. J. [1998]: Experiments with controlled rounding for statistical disclosure control in tabular data with linear constraints. *Journal of Official Statistics*. 4. évf. 4. sz.

HUNDEPOOL, A. [1999]: Statistical disclosre limitation in practice. Kézirat. (http://europa.eu.int/en/comm/eurostat/research/conferences/etk-99/papers/hundepool.pdf)

LAKATOS M. [1994]: Információszabadság – adatvédelem – statisztika (I.). *Statisztikai Szemle*. 72. évf. 7. sz. 547–559. old.
MEROLA, G. [2003]: *Generalized risk measure for tabular data*. Roma. Kézirat. (http://neon.vb.cbs.nl/casc/ISIBerlin/merola.pdf)
SKINNER, C.J. – ELLIOT, M. J. [2002]: *A measure of disclosure risk for microdata*. Kézirat. (http://www.ccsr.ac.uk/publications/occasion/occ23.pdf)
*Statistical Journal of the United Nations ECE* [2001]. Data confidentilaity. 285–407. old.
*Statisztikai igazgatás* [2000]. (Közigazgatási szakvizsga tankönyv). Budapest.

## SUMMARY

This study is about the statistical tools on data confidentiality and the background of the confidentiality issue. It gives an overview on the European and Hungarian legislation, the different and new forms of dissemination, and the confidentiality problems. It shows the different risks on micro and tabular data dissemination and the possible statistical tools of protection with examples.

# AZ OUTLIEREK MEGHATÁROZÁSA ÉS KEZELÉSE GAZDASÁGSTATISZTIKAI FELVÉTELEKBEN

CSEREHÁTI ZOLTÁN

A tanulmány első részében az outlierek fogalmával, különféle helyzetekben való előfordulásukkal foglalkozom. Ezután kitérek arra, hogy miért olyan fontos azonosításuk és kezelésük, milyen hatással lehetnek a becslések pontosságára. Ezt követően egy speciális terület, a regressziós modellek példáján vizsgálom meg, hogy milyen zavart okozhatnak a kiugró értékek, és hogyan lehet ezt orvosolni robusztus módszerekkel. Majd rátérek a gazdaságstatisztikai megfigyelések sajátosságaira, és röviden ismertetem az outlierek kiszűrésére leggyakrabban használt eljárásokat. Bemutatom, hogy milyen robusztus eljárások segíthetnek a gyakran előforduló „elfedési effektus" kiküszöbölésében. Néhány javaslatot teszek arra, hogyan lehet egyszerűbb eljárások ötvözésével újabb, testreszabott módszereket kidolgozni. Szót ejtek az outlier-súlyok alkalmazásának lehetőségéről, mérlegre téve annak előnyeit és hátrányait. Ismertetem az eredeti Grubbs-féle módszert, majd ennek egy továbbfejlesztett, módosított változatát, mely alkalmas arra, hogy egy többrétegű mintából kiszűrje a gyanúsan kiugró értékeket. Ezt követően egy többváltozós adathalmazokra kidolgozott szimulációs módszert mutatok be. Az outlierek kiemelése nemcsak a becslés hibájára van hatással, hanem annak torzítatlanságára is. Erről és egyéb, a kiugró értékek által felvetett problémákról szólok a cikkem utolsó részében.

TÁRGYSZÓ: Gazdaságstatisztika. Lineáris regresszió. Outlier. Grubbs-módszer. Robusztus eljárás.

Az outlierek, azaz a kiugró értékek problémája egyike a statisztika legnehezebben kezelhető kérdéseinek. Nem létezik olyan módszer, mellyel a probléma minden változata megoldható lenne. Valójában sok speciális eljárás létezik, azonban alkalmazhatósági körük többnyire meglehetősen szűk. Vannak szélesebb körben használható módszerek is, ezek azonban kevésbé jó eredményt adnak bizonyos esetekben. Bátran mondhatjuk, hogy ez a probléma elvileg megoldhatatlan, ugyanakkor valamiféle megoldást mégiscsak igényel a gyakorlatban. Igazából „jó módszerek" helyett indokoltabb lenne „kevésbé rossz", illetve „rosszabb" eljárásokról beszélni. A probléma természetéből adódik, hogy nemigen lehet mérni, egy-egy módszer hatásosságát, ezért nehézségeket okoz az összehasonlítás feladata.

A kidolgozott eljárásokat két nagy csoportba oszthatjuk. Vannak, amelyek modell alapúak, azaz bizonyos eloszlást, vagy különféle sokasági jellemzőket feltételeznek, illetve vannak olyanok, melyek robusztusabbak abban az értelemben, hogy kevésbé érzékenyek az eloszlás típusára. Értelemszerűen a modell alapú módszerek szűkebb körben al-

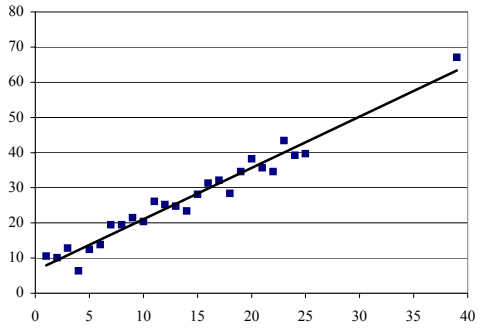
kalmazhatóak, viszont ott jobb eredményt szolgáltatnak. A robusztus eljárások tágabb körben alkalmazhatók, ám többnyire gyengébben teljesítenek.

A statisztikai elemzések kiindulópontja, hogy rendelkezésünkre áll valamilyen adathalmaz. Ezzel dolgozunk a továbbiakban: különféle statisztikai függvényeket, elemzéseket, próbákat alkalmazunk az adatokra. Az viszont, hogy ezeknek az elemzéseknek a végén milyen eredményre jutunk, nagy mértékben függ a kiindulási adathalmaz tulajdonságaitól. Ezek az adatok bizonyos értelemben a véletlen eredményei. Például bizonytalan kimenetelű kísérleti eredményekből vagy véletlen mintavételből, különféle mérésekből származnak.

## AZ OUTLIEREK FOGALMA

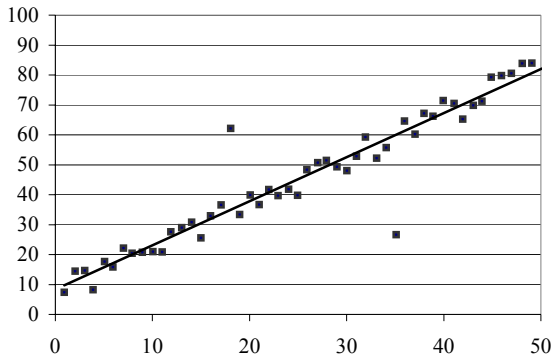
Előfordulhat, hogy adataink között vannak olyan értékek, melyek nem tűnnek hihetőnek, mintha „kilógnának” a többi szám közül. Amennyiben ez a gyanúnk alaposnak bizonyul, szükség szerint el kell távolítani vagy legalábbis más módon kell kezelni az ilyen értékeket, ha nem akarjuk, hogy a későbbi elemzések eredményeit eltorzítsák. Az ilyen kiugró értékeket nevezzük az angol nyelvű szakirodalomban elterjedt kifejezés szerint outliereknek. Általában a túl nagy vagy a túl kicsi értékeket szoktuk outliereknek hívni, de ettől némely esetben eltérünk. Ha a sokaság elemei csak pozitív értékeket vehetnek fel, és a kicsi értékeknek csekély a jelentőségük, akkor csak a kiugróan nagy értékek érdekesek, ezért ezekre szűkítjük le az outlier fogalmát. Előfordulhat azonban, hogy olyan értékeket is outlierként azonosítunk, amelyek nem tartoznak a legnagyobbak közé. Az itt következő példák egy lineáris regressziós modell illesztésekor adódhatnak nemcsak a változó értékei, hanem a pontoknak a regressziós egyenestől való távolsága alapján is indokolt kiugró értékekről beszélnünk. Az 1. ábra olyan esetet mutat be, ahol van egy olyan eleme a sokaságnak, amelyre mindkét változó értéke jóval nagyobb, mint a sokaság többi elemére, ennek az elemnek a jelenléte mégsem befolyásolja jelentős mértékben a regressziós egyenes helyzetét. Az árindexek esetében például a meglepően kicsi értékek is legalább olyan érdekesek, mint a hihetetlenül nagyok.

*1. ábra. Regressziós egyenes illesztése egy rendellenes érték esetén*



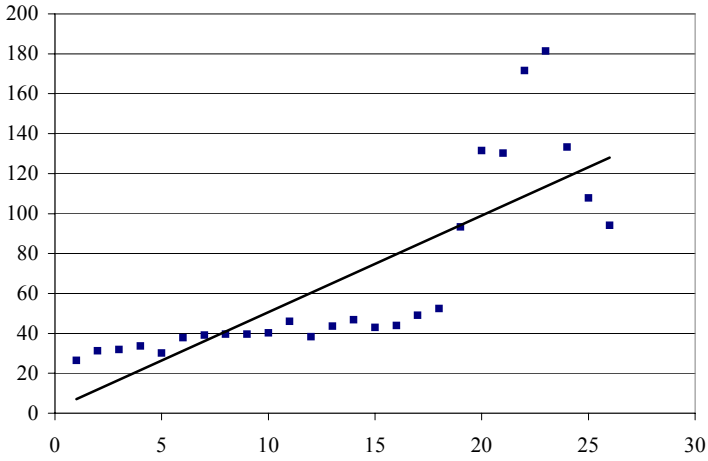

A 2. ábrán két olyan elemet láthatunk, melyek nem illeszkednek bele a lineáris trendbe. Ha csak az egyik lenne jelen, akkor jelentősen eltorzíthatná a regressziós egyenes állását, így viszont kétoldalról kiegyenlítődik a hatásuk.

*2. ábra. Regressziós egyenes illesztése két rendellenes érték esetén*




A 3. ábrán az utolsó hét pont helyzete jelentősen eltér a várhatótól, valószínűleg hibás adatok vannak a jelenség mögött, esetleg a modell nem alkalmazható egy bizonyos határon túl. A kilógó értékek jelenléte jelentősen eltorzította a regressziós egyenes helyzetét, ezért itt a szokásos – legkisebb négyzetek elvén készített – becslés helyett robusztus illesztési technika segíthetne.

*3. ábra. Regressziós egyenes illesztése több rendellenes érték esetén*




Az outlierek fogalmának nincs egységesen elfogadott definíciója a szakirodalomban. Hogy mit tekintünk kiugró értéknek, illetve kevéssé hihető, vagy a modellünkbe nem jól illeszkedő adatnak, az nehezen fejezhető ki egzakt módon. Ezért a továbbiakban tárgyalt módszerek ismertetésekor is képlékenyen kezelem ezt a fogalmat.

Milyen okokból jelenhetnek meg kiugró értékek az adataink között? Az alkalmazott statisztikai munkában többnyire mérési eredményekkel dolgoznak. A fizika, kémia, biológia, szociológia és sok egyéb tudományág különböző területein szükség lehet arra,

hogy mérési eredményekből vonjunk le következtetéseket. Ha valamilyen okból egy mérés hibás (szennyezett volt a kémcső, nem kalibrálták helyesen a feszültségmérőt, nem vették figyelembe a hőmérséklet-ingadozást stb.), akkor az eredményül kapott mérési adat jelentősen eltérhet a valós értéktől. Előfordulhat azonban, hogy ez az eltérés csak akkor válik szembetűnővé, amikor az összes mérési eredményt egybevetve azt látjuk, hogy egy-két adat nem illik bele a képbe. Ekkor azonosítanunk kell ezeket a kiugró értékeket és el kell távolítani őket az adathalmazból.

A véletlen mintavétel esetén is előfordulhat hasonló jellegű hiba például elírás, rögzítési hiba, osztályba sorolási tévedés következtében. Ilyenkor ezeket az értékeket korrigálni kell. Általában azonban másról van szó. Az alapsokaság vagy az abból kiválasztott minta akkor is tartalmazhat kiugró értéket, ha az adott érték mögött valós folyamat rejlik és nincsen semmilyen hiba a háttérben. Ekkor egészen más okból kell foglalkozni a kiugró értékekkel, mert a mintából történő becslés során torzítást okozhatnak. (Ennek a részletesebb taglalására a későbbi fejezetekben térünk ki.) Annak a kiderítése, hogy hibás adatról van-e szó, sokszor nem könnyű feladat, ehhez további külső információk szükségesek.

Bár a továbbiakban a gazdaságstatisztika szemszögéből vizsgálom az egyes módszereket, a lakossági felvételekben előforduló gyanús, kiugró értékek kezelése is fontos feladat, az ismertetett, illetve szakirodalomban fellelhető további módszerek ezekre az adatgyűjtésekre is adaptálhatók.

*A gazdaságstatisztikai megfigyelések sajátosságairól*

A KSH 1991 óta végzi a kisszervezetek reprezentatív megfigyelését. A reprezentatív megfigyelés során kiindulópontunk a minta, amelyből mint részből következtetéseket vonunk le a sokaságra mint egészre. Ezeknek a következtetéseknek, vagyis a minta alapján történő becsléseknek a helyessége jelentős mértékben függ a minta reprezentativitásától. Véletlen minta esetében általában feltételezhető, hogy jól reprezentálja a megfelelő sokaságot. Ez azonban nem mindig van így. Pusztán a véletlen szeszélye folytán is előfordulhatnak bizonyos anomáliák. Ilyen nem várt jelenség lehet, hogy a sokaság valamely része, például a legnagyobb értékekkel rendelkező néhány szervezet túlreprezentált. (A kisebb értékekkel rendelkező szervezetek esetében ez szintén előfordulhat, de ezekből több van, a súlyuk pedig kicsi, így néhány „fölös" mintaelem jelenléte nem zavarhatja nagyon a becslést.) Az ily módon megfigyelt kiugró értékek, az outlierek vizsgálata, azonosítása és kezelése a becslések javításának fontos eszköze minden reprezentatív megfigyelés esetén, így a kisszervezeteknek az éves integrált adatgyűjtés keretében történő reprezentatív megfigyelésénél is.

A tapasztalatok szerint a gazdasági szervezetek termelési adatai közelítőleg negatív exponenciális eloszlást követnek mind teljes sokaságukat, mind egyes rétegeiket tekintve. (Feltéve, hogy egy-egy kérdéses réteg nem túl kicsi.) Ennek az a következménye, hogy a legnagyobb szervezetek adata az átlagos érték többszöröse lehet.

A becslés rétegezett mintavétel alapján történik. Az egyes rétegekre vonatkozó becslésekből számítjuk a teljes sokaságra vonatkozó becsült adatokat. A rétegek képzésénél a következő szempontok játszanak szerepet. Bizonyos ágazatok jelentősége indokolttá teszi, hogy megfelelő becsléssel szolgáljunk az ilyen specifikus területekre. Ez

már önmagában indokolja a rétegzést. Ha azonban ilyen speciális igények nem merülnének fel, azaz csak az országos becslésre koncentrálnánk, akkor is érdemes lenne rétegezni a mintát, mivel kimutatható, hogy mindig javíthatunk a becslés pontosságán, ha sikerül viszonylag homogénebb rétegeket elkülönítenünk, majd kialakítani a rétegeken belüli minta-elemszámokat. Látjuk, hogy kettős oka van a rétegzésnek. Ez a kettős szempontrendszer kell tehát, hogy tükröződjék az outlierek kezelésénél is. Ezért az outliereket az egyes rétegek jellemzőinek figyelembe vételével kell meghatározni és kezelni. A becslés során egy rétegen belül a mintaelemek adatait a mintahányad reciprokával felszorozzuk. Ez azt jelenti, hogy úgy tekintjük, mintha minden mintaelem ugyanannyi hozzá közeli értékű sokasági elemet reprezentálna. Ha tehát egy kiugróan nagy értékkel rendelkező szervezet bekerül a mintába, akkor adatának felszorzásával azt feltételezzük, hogy van még a sokaságban jó néhány hozzá hasonló érték. Tekintettel a negatív exponenciális eloszlásra, ez erősen kétséges, ha valóban egy, a többitől jelentősen eltérő értékről van szó. Ilyen outlier jelenlétekor mindenképpen változtatni kell a becslési módszeren. Elsőként azt kell megvizsgálni, hogy nem hibás-e a kérdéses adat, és ha hibás, ki kell javítani. A továbbiakban végig feltételezzük, hogy a valóságnak megfelelő adatokkal van dolgunk.

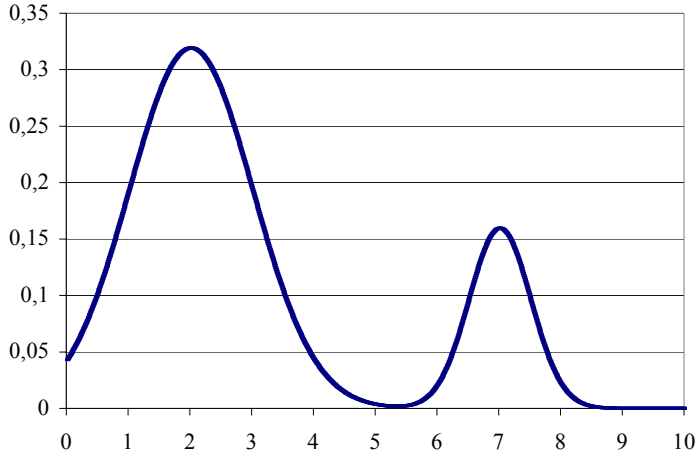
A szakirodalomban leggyakrabban egy egységesen kezelt sokaság az outlier-kezelés tárgya. Számos eljárást dolgoztak ki különböző elméleti, illetve alkalmazott statisztikai tudományágak igényeihez igazodva. Bizonyos módszerek célja a hibás adatok kirostálása, míg másoké az, hogy a feltételezetten helyes adatokból kiszűrje és korrigálja azokat a szélsőséges értékeket, melyek nem kellően reprezentatívak. Az általunk vizsgált megfigyelések adataira ezekből a módszerekből egyik sem alkalmazható közvetlenül. Ennek egyik oka, hogy az adatok számát, azok feltételezett eloszlását is sokszor figyelembe veszik egy-egy módszer kialakításakor, így azok nem használhatók fenntartások nélkül eltérő adatstruktúrák vizsgálatára. Még lényegesebb probléma, hogy esetünkben számos megfigyelési réteggel rendelkezünk. Az egyes rétegek becslésének javításán túlmenően azonban feladatunk elsősorban az, hogy a teljes becslést javítsuk.

Az évközi adatgyűjtésekben a legalább 50 főt foglalkoztató szervezetek, az éves adatgyűjtéseknél pedig a legalább 20 fővel rendelkező vállalkozások megfigyelése teljes körű. Ezeknél is fontos a kiugró értékek azonosítása, de itt az esetleges hibák feltárása a cél, hiszen ebben a körben nincs felszorzás, tehát átsúlyozásra sincs szükség.

*Az outlierek azonosításának és kezelésének fontossága, hatásuk a becslés pontosságára*

Az outlierek azonosítása azért rendkívül fontos, mert egy-két oda nem illő kiugró érték jelentősen befolyásolhatja az egész statisztikai elemzés, becslés pontosságát. Azonosításuk viszont csak úgy történhet meg, ha előre rögzítünk egy megfelelő módszert a kiszűrésükre. Kell tehát már előzetesen is rendelkeznünk valamilyen képpel arról, hogy mit tekintünk normális, elfogadható adatnak és mit kiugró, rendkívüli értéknek. Mihez képest rendkívüli egy érték? Kell, hogy legyen egy előzetes modellünk a vizsgált mutató eloszlásáról, hogy ezt eldönthessük. Tegyük fel például, hogy egy képzeletbeli eloszlás sűrűségfüggvénye olyan, mint amilyet az 4. ábrán láthatunk. Ilyen esetben a minta akkor is „produkálhat” outliereket, ha azok valójában jól illeszkednek a sokaságba.

*4. ábra. Feltételezett kétmóduszú sokaság sűrűségfüggvénye*




Ha 10 elemű mintát veszünk egy olyan sokaságból, amelyre ez az eloszlás jellemző, akkor valószínűleg olyan értékeket kapunk, amelyek közül 8-9 érték 1 és 3 között van, míg 1-2 érték 7 körüli. Ha nem tudnánk, hogy ilyen sajátságos alakú az eloszlásunk sűrűségfüggvénye, akkor azt gondolhatnánk, hogy rendkívüli, kiugró értékekről van szó. Valójában azonban nagyon jól beleilleszkednek abba a képbe, amit az eloszlás jellege mutat.

Vizsgáljuk meg egy nagyon egyszerű példán, hogy miként befolyásolják az outlierek a becslés pontosságát.

*A sokasági értékösszeg becslései*

| A minta sorszáma | Mintaelemek | | A mintaelemek átlaga | Értékösszeg-becslés | Eltérés a valódi értékösszegtől | |
|---|---|---|---|---|---|---|
| 1 | 4 | 6 | 5,0 | 30 | | –170 |
| 2 | 4 | 9 | 6,5 | 39 | | –161 |
| 3 | 4 | 13 | 8,5 | 51 | | –149 |
| 4 | 4 | 18 | 11,0 | 66 | | –134 |
| 5 | 4 | 150 | 77,0 | 462 | 262 | |
| 6 | 6 | 9 | 7,5 | 45 | | –155 |
| 7 | 6 | 13 | 9,5 | 57 | | –143 |
| 8 | 6 | 18 | 12,0 | 72 | | –128 |
| 9 | 6 | 150 | 78,0 | 468 | 268 | |
| 10 | 9 | 13 | 11,0 | 66 | | –134 |
| 11 | 9 | 18 | 13,5 | 81 | | –119 |
| 12 | 9 | 150 | 79,5 | 477 | 277 | |
| 13 | 13 | 18 | 15,5 | 93 | | –107 |
| 14 | 13 | 150 | 81,5 | 489 | 289 | |
| 15 | 18 | 150 | 84,0 | 504 | 304 | |
| *Átlag* | | | | *200* | *280* | *–140* |

Legyen adott egy 6 elemű sokaság, amiből 2 elemű mintát veszünk egyszerű véletlen mintavétellel. A kiválasztható minták száma $\binom{6}{2} = \frac{6 \cdot 5}{2} = 15$. Legyenek a sokaság elemei

sorba rendezve: 4; 6; 9; 13; 18; 150. Azonnal látszik, hogy a legnagyobb érték jóval nagyobb, mint a többi. Tekintsük az összes lehetséges mintát, és adjunk becslést a sokasági értékösszegre (amelynek valódi értéke 200).

Mint az előző oldalon levő táblából láthatjuk, a 15 mintából kapott értékösszegbecslések átlaga megegyezik a sokasági értékösszeggel. Ennek így is kell lennie, hiszen tudjuk, hogy az egyszerű véletlen mintavétel esetén az átlagbecslés torzítatlan. A torzítatlanság viszont nem jelenti azt, hogy minden egyes becslésnek ugyanakkora a hibája. Az alulbecslések átlagosan 140-nel térnek el a valódi értéktől, míg a felülbecslések 280-nal. A konkrét példa kapcsán megfigyelt jelenség általánosabban is érvényes. Minden olyan sokaságnál, melynek kellően ferde az eloszlása (nem feltétlenül kell olyan egyértelműen kiugró értéknek jelen lenni, mint a példánkban), az összes lehetséges mintát tekintve igazak a következők: *1*. a felülbecslések átlagos hibája mindig nagyobb, mint az alulbecsléseké, *2*. kevesebb a felülbecslő minták száma, mint az alulbecslőké.

## AZ OUTLIEREK KIMUTATÁSA

Az outlierek azonosítására gyakran használt eljárások egyik csoportja a következő elven működik. Tekintsük a mintának valamilyen középértékét. Ez lehet a számtani vagy a mértani közép, a medián, esetleg más, ritkábban használt függvény. Ezek után veszünk valamilyen szóródási mutatót. A gyakrabban használtak a mintából számított korrigált szórás, az átlagtól való átlagos abszolút eltérés és a mediántól való abszolút eltérések mediánja. Egy mintaelem szélsőségességének mérőszáma az az érték lesz, mely megadja, hogy az adott mintaelemnek a középértéktől való távolsága hányszorosa a szóródási mutatónak. Ez az adott elemnek a középértéktől való relatív távolsága. Ezt az *i*-edik elemre $d_i$-vel jelöljük. Ha az így számított érték egy adott, előre rögzített korlátot meghalad, akkor a mintaelemet outliernek tekintjük. (Ennek a korlátnak a meghatározására nehéz általános módszert adni. Általában a vizsgált sokaság sajátosságait ismerő tapasztalt szakemberek feladata, hogy a gyakorlat során kialakítsák az erre vonatkozó irányelveket.) A módszerek mind a pozitív, mind a negatív irányú eltérések azonosítására alkalmasak, de esetünkben az adatok eloszlását tekintve nincs értelme a túl kicsi értékeket outliernek tekinteni. Ezért, bár a következő módszerek mind alkalmasak a kétoldali outlier-tesztelésre, ezentúl mindig csak a jobboldali kiugró értékekre koncentrálunk. Amikor a dolgozatom egyes helyein outlier-tesztekről beszélek, akkor ezen nem a statisztikai tesztek hagyományos fogalmára kell gondolni. Csupán azért használom ezt a fogalmat, mert a szakirodalomban sok helyütt elterjedt ez e szóhasználat.

Tekintsünk néhány példát.

*1*. $d_i = \dfrac{|y_i - m|}{s}$, ahol $m = \dfrac{\sum_{j=1}^{n} y_j}{n}$ a mintaátlag, $s = \sqrt{\dfrac{\sum_{j=1}^{n} (y_j - m)^2}{n-1}}$ a korrigált szórás.

*2*. $d_i = \dfrac{|y_i - m|}{\bar{s}}$, ahol $\bar{s} = \dfrac{\sum_{j=1}^{n} |y_j - m|}{n}$ az átlagtól való átlagos abszolút eltérés.

*Az „elfedési effektus" által felvetett problémák kiküszöbölése robusztus eljárások használatával*

Az említett fenti két módszer hátránya, hogy az outlierek jelenléte erősen eltorzíthatja mind a középértéket, mind a szóródási mutatót, és ezen keresztül a középértéktől való relatív távolságot. Ennek az lehet a következménye, hogy miközben nyilvánvaló az outlier jelenléte, a hozzá tartozó $d_i$ érték mégsem jelez jelentős eltérést. Az átlag, illetve a szórás számításakor ugyanis „egybemosódnak" az elemek, azaz nem tűnik ki, hogy lényegében egyetlen kiugró érték jelenléte okozza a nagyobb értékű mutatókat. Ezt nevezzük *elfedési effektusnak*. Hatásosabbak lehetnek az olyan robusztus módszerek, amelyek alkalmazása esetén az eljárások által szolgáltatott értékeket kevésbé torzítja el az outlierek jelenléte. Ilyenre példa a további két eljárás (ezekkel ritkán találkozhatunk a szakirodalomban, pedig éppen olyan esetekben lehetnek hasznosak, amikor egy sokaság döntő többségének a viselkedésére vagyunk kíváncsiak tekintet nélkül arra, hogy esetleg egy-egy „renitens" elem is jelen van).

*3*. $d_i = \dfrac{\left| y_i - \underset{j}{\text{medián}}\left(y_j\right) \right|}{MAD}$, ahol $MAD = \underset{i}{\text{medián}}\left( \left| y_i - \underset{j}{\text{medián}}\left(y_j\right) \right| \right)$.

*4*. $d_i = \dfrac{\left| y_i - \underset{j}{\text{medián}}\left(y_j\right) \right|}{q_{0,75} - q_{0,5}}$, ahol $q_{0,75} - q_{0,5}$ az ún. felső interkvartilis terjedelem: $q_{0,75}$ a harmadik kvartilis, $q_{0,5}$ pedig a második kvartilis, azaz a medián.

A *3*. és *4*. módszer egymáshoz hasonló tulajdonságokkal rendelkezik. A medián és az interkvartilis terjedelem kevéssé érzékeny az outlierek torzító hatására. Ezen túlmenően mindkét eljárás egyszerűen számítható. A *4*. módszer általánosabban használt, mint a *3*., azonban van néhány hátránya. Előfordulhat ugyanis, hogy a felső interkvartilis terjedelem szokatlanul szűk, azaz a medián és a *3*. kvartilis kevéssé térnek el egymástól. Ez lehet a helyzet, ha a medián felett sok hasonló érték található. Ekkor a *4*. teszt által adott $d_i$ érték akkor is nagy lehet, ha $y_i$ nem igazán kiugró érték. Sőt ekkor a felső negyedből számos értéket minősíthet outliernek az eljárás, ami önmagában sem jó, hiszen egy-egy rétegben nem kívánatos egy-két elemnél többet kiemelni. Érdemes eleve csak egy-egy réteg maximális mintaelemére gyanakodni. Másik probléma az, hogy ha a mintaelemek eloszlása nem egyenletes – márpedig nálunk negatív exponenciális eloszlásról van szó –, akkor a $d_i$-kre meghatározandó kritikus érték függ a minta elemszámától. (Nagy minta esetén a középértéktől való nagyobb relatív távolság is tolerálható.) Ez az elemszámtól való függés csak hosszas kísérletezgetéssel korrigálható. A kérdéssel foglalkozó szakirodalomban azonban nem találtam ezzel kapcsolatos vizsgálatokat.

*Egyéb módszerek – egyszerűbb eljárások ötvözése*

Másik lehetséges eljárás az adatok logaritmizálásán alapul. Vegyük tehát a mintaelemek logaritmusát. Rendezzük csökkenő sorrendbe az így kapott értékeket. Jelöljük a

mintaelemek számát *n*-nel. Tekintsük a szomszédos elemek különbségeit. Amennyiben a két legnagyobb elem különbsége meghaladja a többi különbség átlagának egy előre rögzített konstansszorosát, akkor tekintsük a legnagyobb elemet outliernek. Ez a módszer $n \geq 4$ esetén használható jól, főként akkor, ha a logaritmizált elemek közel egyenletes eloszlást követnek. Amennyiben nem ez a helyzet, akkor előnyösebb, ha nem az összes különbséget vesszük alapul a számításnál, hanem csupán a *3*. kvartilisnál nagyobb elemekéit. Ha mindkét eljárás outliernek minősíti a legnagyobb elemet, akkor elfogadhatjuk, hogy ez az elem valóban kiemelendő. A különbségek átlaga helyett lineáris regressziót is végezhetünk, vizsgálva a legnagyobb elemnek a regressziós egyenestől való távolságát. A regressziós egyenes nem alkalmazkodik kellőképpen az adatok eloszlásának jellegzetességeihez. Előfordulhat, hogy a regressziós egyenes közel kerül egy kiugró értékhez (lásd az 1. ábrát). Egyenes helyett más, alkalmasabb regressziós görbét használva javíthatunk a helyzeten, ehhez azonban minden réteg esetén külön előzetes mérlegelés lenne szükséges. A fentebb leírt, a differenciák átlagán alapuló eljárás robusztusabb abban a tekintetben, hogy kevéssé érzékeny az egyedi eloszlás jellemzőinek zavaró hatására.

Ez a módszer számos előnnyel bír. Az adott rétegben tapasztalható „tipikus” növekedési ütemhez viszonyítva határoz meg korlátot az outlier számára, így az eloszlásra vonatkozó minden előzetes feltevés hiányában is jól alkalmazkodik annak jellegéhez. Nem függ a minta elemszámától, így nem kell bonyolult függvényekkel torzítani a módszert, hogy a változó elemszám függvényében állítsuk be a kritikus korlátot. Kevéssé érzékeny olyan anomáliákra, melyek néhány más módszert bizonyos esetekben megbízhatatlanná tesznek (ilyen például a szűk interkvartilis terjedelem). Ezen túlmenően egyszerűen számítható, és az eredmény grafikusan is szemléletesen megjeleníthető.

Előfordulhat, hogy a fent vázolt eljárások nem mutatják ki egyik elemről sem, hogy outlier lenne, de „ránézésre” jól látható egy erősen kiugró érték. További technikai nehézséget jelenthet, hogy két vagy három elemű minta esetén nincs sok értelme outliert keresni. Legfeljebb akkor lehet ez indokolt, ha ugyanezen réteg korábbi havi adataihoz képest is erősen kiugró a nagyobbik mintaelem. Mindezek a problémák indokolják, hogy ne egy egyszerű tesztet alkalmazzunk csupán, hanem próbáljuk meg a különböző módszereket ötvözni valamilyen módon. Erre egy lehetőség például az, ha több eljárást is lefuttatunk, és azok eredményeit figyelembe véve határozzuk meg azt a korlátot, amely felett outlierként azonosítjuk a maximális mintaelemet. A *4*. teszt alapján a következő korlátot adhatjuk meg:

$$k_4 = q_{0,5} + \left(q_{0,75} - q_{0,5}\right) \cdot 10 \cdot \log_2 n \,.$$

Ez többé-kevésbé torz eredményt adhat, ha $n \leq 4$, illetve abban az esetben, ha a felső interkvartilis terjedelem kisebb a vártnál. Ezért érdemes tekinteni egy olyan tesztet is, mely erre nem érzékeny. Legyen ez a következő:

$$k_5 = q_{0,5} \cdot 6 \cdot \log_2 n \,.$$

(Mindkét esetben a 2-es alapú logaritmusfüggvény szolgáltatja az elemszám nagyságrendjének megfelelő kiigazítást.) Ez viszont túlzott egyszerűsége miatt nem tekinthető önmagában hatásos tesztnek. A kettő ötvözésével kapott $k_6 = \sqrt{k_4 \cdot k_5}$, azaz a két korlát

mértani közepe jó jelölt egy általános outlier-tesztre. (A képletekben szereplő 10 és 6 konstansok természetesen tetszőlegesen változtathatók aszerint, mekkora szigorral kívánunk eljárni az outlier-gyanús elemekkel szemben.)

Az így nyert teszt annyiban korrigálható még, hogy megpróbáljuk figyelembe venni azt az egyszerű heurisztikát, hogy ha a legnagyobb elem jóval nagyobb, mint a második, akkor érdemes azt outlierként kezelni, függetlenül attól, hogy a többi elem eloszlása milyen. Egy lehetséges korlát ekkor $k_7 = 6 \cdot y_2$, ahol $y_2$ a második legnagyobb elem. A végső korlát tehát $k_8 = \min(k_6, k_7)$. Így biztosítható, hogy észleljük a kiugró értéket, bármelyik módszer is figyelmeztet erre.

(Az iménti bekezdésekben felvázolt képletekkel azt próbáltam érzékeltetni, hogy miként lehet az adathalmazzal kapcsolatos elvárások heurisztikus, képlékeny világát a matematikai formulák nyelvére lefordítani.)

*A Grubbs-féle módszer*

Az outlierek kiszűrésére általánosan használt eljárás a Grubbs-féle teszt. (*Grubbs*, [1969]) Ez bonyolult, számításigényes eljárás, mely a mintaelemek eloszlására vonatkozó információt is felhasznál, azaz avval az előfeltevéssel él, hogy azok normális eloszlást követnek. A teszt a következőképpen zajlik. Vegyük a minta legnagyobb elemét, ezt jelöljük $y_{\max}$ -szal. Számítsuk ki az 1. teszt képletének megfelelően a következő *Z*-vel jelölt standardizált értéket:

$$Z = \frac{|y_{\max} - m|}{s}.$$

Ez után történik a

$$T = \sqrt{\frac{n \cdot (n-2) \cdot Z^2}{(n-1)^2 - n \cdot Z^2}}$$

érték számítása, ahol *n* a mintaelemszám. Ezt követően kiszámítjuk az (*n*–2) szabadsági fokú Student-eloszlás *T* paraméterhez tartozó értékét. Ezt jelöljük $P_0$-val. Legyen most $P = n \cdot P_0$. Az így számított *P* érték annak a valószínűségét adja meg, hogy egy *n* elemű, normális eloszlásból származó minta legnagyobb eleme az általunk tapasztalt eltérést mutatja a többi elemtől. Ha előre rögzítünk egy *P* értéket, akkor a különböző elemszámok esetére közelítőleg meghatározhatjuk a *Z* mennyiségnek azt a kritikus korlátját, amelyre a fenti számítások a *P* valószínűségi értéket adják eredményül. Jelölje ezt a korlátot $\overline{Z}$. Az ide vonatkozó szakirodalomban közlik a *P*=0,05-höz tartozó $\overline{Z}$ értékek táblázatát a 3-tól 140-ig terjedő minta-elemszámokra. A fönti képleteket használó algoritmus segítségével jó közelítéssel meghatározhatók ezek a kritikus értékek nagyobb *n*-ekre is.

A Grubbs-teszt előnyös tulajdonsága, hogy a Student-eloszlás felhasználásával különböző kritikus értékeket határoz meg különböző mintaelemszámok esetén. Hátránya az,

hogy az outlierek torzító hatására a rendkívül érzékeny *1*. módszert alkalmazza. További hátránya pedig, hogy negatív exponenciális eloszlás esetén nem alkalmazható. Ezen úgy lehet segíteni, hogy nem az eredeti adatokkal, hanem azok logaritmusaival dolgozunk. Ekkor sok esetben már normálishoz közeli eloszlást kapunk. Ne felejtsük el, hogy a negatív exponenciális eloszlás is csak egy alkalmasnak tűnő közelítés, amelytől többékevésbé eltérhet az adott minta, különösen, ha kicsi. A Grubbs-teszt alkalmazása során azt tapasztaltam, hogy 10 alatti mintanagyság esetén (márpedig ez egy-egy réteget tekintve meglehetősen gyakori) erőteljesen jelentkezhet a kiugró értékeknek az *1*. eljárásra gyakorolt torzító hatása, ennek következtében a módszer hajlamos nem felismerni olyan értékeket, melyek egyértelműen outliernek látszanak. Ezen úgy segíthetünk, hogy az átlag és a szórás számításakor a legnagyobb elemet nem vesszük figyelembe.

*Az outlierek szimultán detektálása és kezelése*

Az előzőkben ismertetett módszerek csak egy-egy réteg vizsgálatára használhatók. Az általunk kitűzött cél viszont a teljes sokaságra vonatkozó becslés javítása. Ezért a különböző rétegek adatainak együttes elemzésére van szükség. Ez többféleképpen is megtehető. Szem előtt kell tartanunk azonban néhány alapelvet. Nem szabad túl sok outliert kiemelnünk. Természetes jelenség, hogy bizonyos mértékű alulbecslés, illetve túlbecslés jelentkezik egy-egy rétegben. A rétegek nagy száma miatt ezek jól kiegyenlíthetik egymást. Olyankor érdemes csak beavatkozni a hagyományos becslési módszerbe, ha egy rétegben olyan kiugró érték található, mely nemcsak az adott rétegen belül, hanem más, azonos ágazati, illetve létszám-kategóriába tartozó rétegek összességén belül is jelentősen kimagaslik a többi közül.

*A Grubbs-módszer adaptálása többrétegű minta esetében*

Az outlierek szimultán azonosítására használhatjuk a Grubbs-féle tesztet a következő módon. Először logaritmizáljuk az adatokat. Ezt követően minden rétegben meghatározzuk a legnagyobb elemhez tartozó $Z$ értéket, illetve a megfelelő mintaelemszámhoz tartozó kritikus $\overline{Z}$ korlátot. Ezek után az $R = Z / \overline{Z}$ hányados értékét vizsgáljuk. Az eredeti Grubbs-teszt minden olyan elemet outliernek tekint, melyre $R > 1$. Ezek közül most csak a legnagyobb $R$ értékekkel rendelkező elemeket emeljük ki. Ez a módszer elméletileg megalapozott és első ránézésre használhatónak tűnik, azonban van egy hiányossága. Nem veszi figyelembe az abszolút számok közötti nagyságrendi különbségeket. Nyilvánvaló, hogy nagyobb figyelmet kell szentelnünk azoknak az értékeknek, melyek önmagukban is nagyobbak. Világos, hogy a túlbecslés mértéke nemcsak attól függ, hogy mennyire kiugró egy érték valamely rétegen belül, hanem attól is, hogy a teljes sokaságban mennyire jelentős a súlya. Ennek megfelelően a következő módosítás tűnik ésszerűnek.

Határozzuk meg rétegenként minden logaritmizált elemre a következő értékeket. Legyen az $\ln(y)$ standardizált értéke

$$Z_y = \frac{|\ln(y) - m|}{s}.$$

Ezt követően határozzuk meg az $R_y = Z_y / \overline{Z}$ hányadost, ahol $\overline{Z}$ a megfelelő rétegelemszámhoz tartozó Grubbs-féle korlát. Nevezzük $R_y$-t a továbbiakban módosított standardizált értéknek. Ezek után képezzük a $T_y = \ln(y) \cdot \sqrt{R_y}$ szorzatot. Azok az *y* értékek kerülnek kiemelésre, melyekre a kapott $T_y$ érték a legnagyobb. A $T_y$ nagysága két tényezőtől függ: az *y* érték nagyságrendjétől – ezt fejezi ki az $\ln(y)$ – és a megfelelő rétegen belüli szélsőségességének mértékétől – ennek leírására szolgál a $\sqrt{R_y}$ tényező. Ez a képlet tehát egyszerre veszi figyelembe azt, hogy mennyire kiugró egy érték a saját rétegén belül, és azt, hogy nagyságrendje folytán mekkora hatással van a becslésre. (Ha $R_y$ negatív értékű, akkor nem értelmezhető ez a képlet, de ekkor a megfelelő *y* érték amúgy sem outlier-gyanús, tehát nyugodtan figyelmen kívül hagyhatjuk.)

Javasolható az előzőhöz hasonló alternatív teszt is, mely a konkrét számítások tapasztalatai alapján némely esetben jobb eredményt hozhat. (Különösen akkor, ha sok 0 adat van bizonyos rétegekben.) Az alkalmazandó képlet most $\overline{T_y} = y \cdot R_y^2$. A fő különbség az, hogy most *y* nagyobb súllyal szerepel $R_y$-hoz képest. (A sok 0 jelenléte eltorzíthatja a szórást és az átlagot, így a standardizált és a módosított standardizált értékeket is, ezért jó, ha ilyenkor az eredeti *y* érték erősebben befolyásolja $\overline{T_y}$ értékét.)

A sok 0 érték problémájára egy másik lehetséges megoldás, ha a számításokat úgy is elvégezzük, hogy ezeket az értékeket figyelmen kívül hagyjuk.

Miután a havi reprezentatív megfigyelések több létszámkategóriát érintenek, továbbá így egy évre vonatkozóan is 12 különböző adathalmaz áll rendelkezésre a módszerek tesztelésére, ezért részletes vizsgálataimban a havi megfigyelések adataira összpontosítottam.

A gyakorlatban még néhány további értékelő szempont is hasznosnak bizonyult. Ezért a következő mennyiségeket is kiszámítottam a fönti módon elkészített „toplista” elemeire:

– az adott réteg becslésében a kérdéses elem kiemelése miatt bekövetkezett változás nagyságának abszolút értéke,

– ennek a változásnak a nagysága a becsült érték százalékában.

A felvázolt módszert már élesben is használjuk. A fentebb leírt algoritmust egy SAS-program formájában valósítottam meg. A program minden megfigyelt mutatóra különkülön elvégzi a számításokat, mégpedig nemcsak az egyes rétegek legnagyobb elemeire, hanem az összes szervezet adataira is. Ezután a leírt módon elkészíti a leginkább kiugró értékek listáját minden egyes mutatóra, mellékelve mindazokat az említett mennyiségeket, melyek segítenek eldönteni, hogy mekkora hatással lehet az adott szervezet kiemelése a kérdéses mutató adott rétegbeli becslésére. A kiugró értékek azonosítása előtt alaposan szemügyre vesszük a program által számított értékeket. Figyelembe vesszük továbbá azt is, hogy milyen mutatók alapján tűnik kiugrónak az adott szervezet, az adott réteg korábbi adataival összehasonlítva mennyire meglepők az értékei, és ki volt-e emelve korábban.

## AZ OUTLIEREK KIEMELÉSE, SÚLYOZÁSA

Az egyszerű véletlen mintavételen alapuló hagyományos felszorzásos becslés torzítatlan, azaz az összes lehetséges mintát tekintve az azokból származó értékösszeg-becslések átlaga megegyezik a valódi értékösszeggel. A korábbiakban már bemutattam egy példán, hogy egy erősen ferde eloszlású sokaság esetén egy véletlen mintából származó becslés nagy valószínűséggel kicsit alulbecslő lesz, míg kis valószínűséggel jelentősen túlbecsüli a sokasági értéket. A kismértékű alulbecslés nem feltűnő, azonban a jelentős túlbecslés ténye megsejthető a mintaelemek vizsgálatával. Erre éppen a korábban tárgyalt outlierszűrő algoritmusok használhatók.

*Az outlierek kiemelésének a hatása a becslés torzítatlanságára és hibájára*

Mi történik tehát akkor, amikor egy kiugró értéket azonosítunk és azt kivesszük a felszorzásból? Ezzel nagy valószínűséggel tompítottuk egy jelentős túlbecslés mértékét. Ha a mintavételt sokszor megismételnénk, akkor azt tapasztalnánk, hogy módszerünk segítségével számos túlbecslés mértéke csökkenthető, tehát kisebb lesz a becsléseink szórása. Ez jó, de sajnos azzal jár együtt, hogy torzítottá válik a becslésünk, hiszen egyoldalúan korrigáltuk a becsléseket: csak a felülbecsléseket csökkentettük, az alulbecslések megmaradtak. Átlagban tehát alulbecsüljük a valódi értékösszeget.

*Outlier-súlyok alkalmazása*

Mint az már az eddigiekből is kitűnt, sokszor nehéz éles határvonalat húzni az outlierek és a többi adat között. Felmerülhet az az igény is, hogy valami módon próbáljunk javítani a becslésünkön akkor is, ha nincsenek jelen egyértelműen azonosítható kiugró értékek. Ilyenkor ahelyett, hogy egy egyszerű logikai értéket rendelnénk minden adathoz aszerint, hogy outliernek minősítjük-e vagy sem, finomabban is különbséget tehetünk közöttük úgy, hogy egy olyan értéket rendelünk hozzájuk, mely azt fejezi ki, mennyire tekinthető outliernek az adott szám. Ennek főleg akkor van jelentős szerepe, ha egy becslés során felszorzásra kerülnek az értékek. Míg hagyományosan minden értéket ugyanazzal a számmal szorzunk, az outlierek kiszűrését követően ezt úgy módosítottuk, hogy az ilyen értékek 1-es szorzót kaptak. Ez tovább finomítható úgy, hogy minden egyes mintaértéknek a felszorzási súlyán változtathatunk. Ennek a mértéke pedig attól függ, hogy mennyire tekinthető outliernek az a bizonyos érték.

Ennek a módszernek sok előnye, de számos hátránya is van. Előnye, hogy finomabb különbségtételt tesz lehetővé az adatok között. Segítségével jól számszerűsíthető például egy olyan verbális értékelés, mely azt fejezi ki, hogy bizonyos kétségeim vannak afelől, vajon kiugró értéknek minősítsek-e valamit. További előnye, hogy segítségével elkerülhetők az olyan idősorbeli törések, melyek abból származnak, hogy egy szervezet értékét az egyik időszakban már éppen kiugrónak minősítem, míg az előző időszakban még éppen nem minősült annak.

Hátránya, hogy erősen beleavatkozik a becslés menetébe. Aggályossá válhat a becslés torzítatlansága, továbbá nagyban megnehezítheti a mintavételi hiba számítását. Ezen túlmenően a súlyok előállítása önmagában is hosszadalmas procedúra sok rejtett hibalehetőséggel, nem is beszélve az adatbázis-technikai problémákról.

Most következzen egy módszer az outlier-súlyok képzésének gyakorlati megvalósítására. A továbbiakban feltételezzük, hogy valamilyen – elméleti vagy tapasztalati – megfontolás alapján azt állíthatjuk, hogy a sokaság eloszlása jól közelíthető valamilyen jól ismert eloszlással. Az egyszerűség kedvéért tételezzük fel, hogy ez normális eloszlás. Vizsgáljuk meg a mintánkat. Számítsuk ki a mintaelemek átlagát és szórását. Tekintsük ezután azt a normális eloszlást, amelynek két paramétere: a várható értéke és a szórása rendre megegyezik a mintából számított átlaggal és szórással. Ennek az eloszlásnak jó közelítéssel meg kell egyeznie a sokaságra jellemző eloszlással. Ezek után vegyünk mesterségesen egy „egyenletes" mintát ebből a normális eloszlásból. Ennek a mesterséges mintának az elemszáma egyezzen meg az eredeti minta elemszámával. Az „egyenletesség" a következőt jelenti. Tekintsük a modellként kapott normális eloszlás eloszlásfüggvényét. Ennek az értékkészlet-halmaza a (0,1) nyílt intervallum. Jelöljük az eloszlásfüggvényt $F$-fel, a minta elemszámát $n$-nel, a mesterséges minta elemeit pedig $\overline{m}_i$-vel ($i$=1, 2, ... , $n$). Ekkor legyen $\overline{m}_i = F^{-1}\left(\frac{1}{2 \cdot n} + \frac{i-1}{n}\right)$, ahol $F^{-1}$ az $F$ függvény inverzét jelöli. Itt valójában arról van szó, hogy az értékkészlet halmazban egyenletesen elosztva elhelyezünk $n$ számú pontot, majd ezekhez megkeressük a megfelelő értékeket. Ezzel mintegy biztosítjuk, hogy a mesterséges mintánk a lehető „legszebb" legyen. A becslés ezután egyszerűen úgy történhet, hogy ezzel a mesterséges mintával dolgozunk, ezzel végezzük el a felszorzást. Mindez megfogalmazható a súlytényezők „nyelvén" is. Nevezetesen: rakjuk növekvő sorba az eredeti minta elemeit is. Jelölje az eredeti minta sorrendben $i$-edik elemét $m_i$. Párosítsuk össze az azonos sorszámú elemeket. Ezek után a $w_i$ súlyok a következő módon képezhetők: $w_i = \overline{m}_i / m_i$. Ha ezekkel a tényezőkkel súlyozzuk a mintaelemeket, akkor eredményül ugyanazt kapjuk, mint a fentebb leírtak alapján. A súlyok korrekciós szerepe jól érzékelhető, ha a következőkre gondolunk. A modellbeli normális eloszlás illesztésekor nem várható, hogy minden egyes érték jól illeszkedjen a modellbe. Azok, amelyek eltérnek tőle, annál inkább 1-től eltérő korrekciós súlyt kapnak, minél inkább jelentős az illeszkedési hiba. Ez az érték lehet 1-nél kisebb. Az outliernek minősülő értékek esetében annál kisebb, minél inkább kiugró értékről van szó. Lehet viszont 1-nél nagyobb is. Ilyen módon némileg korrigálható az is, ha az elvárhatónál több kicsi érték kerül bele a mintába.

Az imént ismertetett konkrét módszernek az általánosságban felsoroltakon kívül további hibái is vannak.

*1*. Csak olyan esetben alkalmazható, ha egy egyszerű, jól parametrizálható eloszlással hatékonyan modellezhető a sokaság.

2. Az outlierek jelenléte eltorzíthatja az átlag-, illetve szórásszámítást. Ezen úgy segíthetünk, ha valamilyen módon megpróbáljuk robusztussá tenni ezeknek a számítását. Ez történhet úgy, hogy egyszerűen kihagyjuk az alsó, illetve a felső néhány percentilist a számításokból. Az így előálló mutatók valóban robusztusak lesznek, de így könnyen a másik végletbe eshetünk. Előnyösebb lehet, ha nem hagyjuk ki a számításokból a legkisebb, illetve a legnagyobb elemeket, hanem valamilyen módon olyan elemekkel pótoljuk őket, melyek jobban illeszkednek a többi érték által meghatározott eloszlásba. Ez megtehető például a következő iteratív eljárással. Először elvégezzük a fentebb ismertetett modellillesztést, ezután első lépésben csak az illesztett eloszlástól leginkább eltérő értékeket „súlyozzuk át", majd az így módosított adathalmazra újra elvégezzük a modellillesztést, és így tovább, egészen addig, amíg az iteráció $k$-adik lépésében már egyik adat sem igényel egy előre meghatározott mértékűnél nagyobb átsúlyozást.

*Egy szimulációs eljárás outlierek azonosítására többváltozós adathalmazokban*

A statisztikai munkában gyakran előfordul, hogy egy-egy mintavételi egységtől több adatot gyűjtünk be. Az így előálló adataink egy többváltozós adathalmazt alkotnak, amelyben minden egyes szervezethez az adatoknak egy rendezett sorozata tartozik. Ilyenkor minden egyes változóra külön-külön el kell végezni nemcsak a teljeskörűsítést, hanem az outlierek kiszűrését is. Előfordulhat, hogy egy bizonyos változó esetében kiugrónak talált szervezet egy másik változó esetében nem lóg ki a többi közül. Ekkor döntenünk kell arról, hogy melyik változót tartjuk meghatározó jelentőségűnek és ennek alapján kiemeljük-e a kérdéses adatszolgáltatót mint outliert.

A következőkben egy olyan módszert mutatok be, amellyel megoldható a többváltozós adatsorok kiugró értékeinek azonosítása úgy, hogy egyszerre vesszük figyelembe az összes változó értékét.

Tegyük fel, hogy $n$ számú adatszolgáltatótól $p$ darab változó értékét gyűjtöttük be. Az így kialakult adathalmazt tekinthetjük úgy is, mint $n$ darab pont halmazát a $p$-dimenziós euklideszi térben. Kiválasztjuk azt a pontot, melynek a többitől való átlagos távolsága a legkisebb. Ez a pont olyan helyen lesz, ahol a ponthalmazunk a leginkább sűrűsödik. Ebből a pontból elindítunk egy szimulált „járványt". Kezdetben csak ez a pont fertőzött. A pontrendszer állapota diszkrét időegységenként változik. Minden „óraütésre" a következő történik. Minden olyan pont, amely eddig már megfertőződött, fertőzött is marad. Az olyan pontok, amelyek még nem voltak fertőzöttek, bizonyos valószínűséggel megfertőződhetnek. Annak a valószínűsége, hogy egy fertőzött pont megfertőzzön egy egészségeset, a távolságukkal arányosan csökken. (Hogy a távolság növekedésével milyen arányban csökken ez a valószínűség, egy megfelelő folytonos függvénnyel leírható, mely monoton fogyó, és értékkészlet-halmaza a (0,1) intervallum.) Így előbb vagy utóbb minden pont megfertőződik. Minden pontnál feljegyezzük, hogy mikor érte el a járvány. Ilyen módon egy sztochasztikus függvényt kapunk, melynek értelmezési tartománya a pontok halmaza, értékei pedig időpontok (a megfertőződés ideje). Nyilvánvaló, hogy azok a pontok maradnak legtovább egészségesek, melyek a leginkább izoláltan helyezkednek el. Ezért azok a pontjaink lesznek outlier-gyanúsak, melyeknél az imént leírt sztochasztikus függvény a legnagyobb értékeket veszi fel.

A módszer legfőbb hátránya az, hogy rendkívül számításigényes. Minden pontpár távolságát ki kell számítani, ezenkívül minden egyes időpontban minden pontpárra meg kell vizsgálni, hogy fennáll-e az egyiknek a másik általi megfertőződésének a veszélye és ha igen, akkor egy véletlenszám generálásával, a megfelelő függvény alapján dönteni arról, hogy egészséges maradjon-e.

Másik hátránya az, hogy nehezen lehet számszerűsíteni, mennyire találunk fontosnak egy-egy változót. Ezért nehéz beépíteni ezt a fontos többletinformációt a modellbe. Nyilvánvaló, hogy ha a változók közül egy vagy kettő sokkal fontosabb, mint a többi, akkor ezt az információt érdemes beépíteni a szimulációs modellbe. Ez megtehető például úgy, hogy a kérdéses változó által meghatározott irányban „megnyújtjuk" a terünket. Ezáltal számszerűen is érzékeltetni tudjuk azt, hogy az adott változó értékei közötti eltérés fontosabb számunkra, mint a többi változó esetében.

## A REGRESSZIÓS OUTLIEREKRŐL

Gyakran előfordul, hogy egy sokaság elemeit két változó szemszögéből vizsgáljuk abból a célból, hogy az egyes változók által felvett értékek között valamilyen összefüggést találjunk. A sokaság minden elemére két értékünk van, ezért ezeket kényelmesen ábrázolhatjuk egy síkbeli koordinátarendszerben. Általában valamilyen regressziós modellt próbálunk ráilleszteni a pontjainkra. A legegyszerűbb esetben ez egy egyenes, ekkor tehát lineáris regresszióról beszélünk. Ez nemcsak a leggyakrabban előforduló regressziós függvény, hanem több más (például logaritmikus, exponenciális) regresszió is egyszerűen visszavezethető rá. Ezért a következőkben fordítsuk figyelmünket a lineáris regresszióra.

### *Az outlierek előfordulása a regressziós modelleknél*

Egy-egy outlier jelenléte megzavarhatja a regressziós modellt. Érdekes módon azonban bizonyos típusú kiugró értékekre nem érzékeny a regressziós illesztés. Ilyenre láthattunk példát az 1. ábrán. Máskor olyan pontok jelenléte is megzavarhatja a regressziós görbe illesztését, melyek – legalábbis az egyik változó alapján – nem tűnnek kiugrónak. Mindez indokolja, hogy ez esetben regressziós outlierekről beszéljünk, kiemelve ezzel azt, hogy a legfőbb szempont az outliernek a regressziós illesztésre gyakorolt hatása.

A regressziós outlierek azonosítása két okból fontos lehet. Az egyik a hibás értékek (mérési eredmények, megfigyelések) kiszűrése. Ez az elsődleges célja minden más outlier-tesztnek is. A másik fontos oka az, hogy ezáltal elkülöníthetünk olyan megfigyelési értékeket, melyek nem illeszkednek az általános modellbe, ezért magyarázatukhoz más megközelítésmód, esetleg paradigmaváltás szükséges. (Ilyenre látványos példát szolgáltatnak bizonyos csillagászati mérések, melyeknél éppen a regressziós outlierek hívták fel a figyelmet egy új típusú égitest létezésére.) Egy későbbi fejezetben lesz szó a többváltozós outlierek azonosításáról (itt minden elemhez két érték tartozik), ezért erről itt bővebben nem szólok.

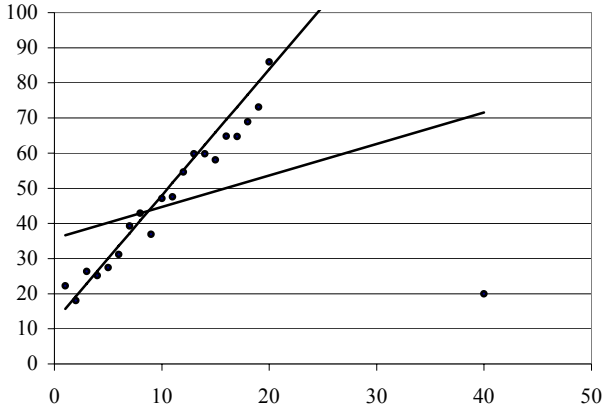
### *A robusztus regressziós illesztés*

Eddig arról volt szó, hogy az outlierek zavaró hatását úgy próbáljuk megszüntetni, hogy azonosítjuk, majd szükség szerint eltávolítjuk őket az adathalmazból. Egy másik lehetőség az, hogy olyan regressziós illesztési technikákat alkalmazunk, amelyek kevéssé érzékenyek kis számú kiugró érték jelenlétére, általában arra, ha az alapadatok egy kisebb része – akár jelentős mértékben – megváltozik. Ezeket nevezzük robusztus eljárásoknak. A következőkben egy példán keresztül fölvázoljuk a hagyományos regressziós technika által szolgáltatott eredményt, majd pedig egy olyan robusztus eljárást, mely alternatívaként javasolható. Mint azt az 5. ábrán láthatjuk, egy outlier megzavarhatja regressziós egyenesünket.

Gondoljuk át, hogyan is történik a regressziós illesztés. Adott $n$ darab pont a síkon: $(x_i, y_i)$, $i = 1, 2, ..., n$. Lényegében arról van szó, hogy minimalizáljuk a következő mennyiséget: $e = \sum_{i=1}^{n} r_i^2$, ahol $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_i$ a keresett regressziós egyenes egyenlete,

$r_i = y_i - \hat{y}_i$ pedig az *i*-edik pont reziduuma. (Valójában azt szeretnénk, ha minden egyes $r_i$ kicsi lenne.) Mivel minimalizálásról van szó, ezért nyugodtan oszthatunk a fenti formulában *n*-nel. Ezek szerint azzal egyenértékű a fenti formula, hogy az eltérés-négyzetek átlagát minimalizáljuk. Tudjuk, hogy az átlagfüggvény nagyon érzékeny egy-egy érték kilengésére, azaz nem robusztus. Ez okozza azt, hogy a regressziós egyenes irányát könnyen „eltéríti" egy-egy outlier. Ezen könnyen tudunk segíteni úgy, hogy az egyszerű számtani átlag helyett egy robusztusabb függvénnyel dolgozunk. Erre jó jelölt a medián. Ha az $e_m = medián\ e_i^2$ mennyiséget minimalizáljuk, akkor egy sokkal robusztusabb regressziós egyeneshez jutunk, mely nem érzékeny néhány pont kilengéseire. Ezt mutatja a 5. ábrán a nagyobb meredekségű egyenes. Egyszerű szemléletes jelentést adhatunk ennek az egyenesnek. Vegyük a legkeskenyebb olyan sávot a síkon, mely lefed a pontok közül legalább $n/2+1$ darabot, ennek a középvonala lesz a robusztus regressziós egyenes. (A „legkeskenyebb" itt azt jelenti, hogy *y* irányú szélessége a legkisebb.)

*5. ábra. Regressziós egyenes robusztus illesztése*




Megjegyzendő, hogy a robusztus módszer nemcsak egy, hanem akár (*n*/2)–1 darab pont személyes viselkedésére is érzéketlen lehet.

A hagyományos módon illesztett regressziós egyenes paraméterei egyszerűen számíthatók még akkor is, ha nem áll rendelkezésünkre számítógép. Ez azért van, mert az eltérés-négyzetösszeg minimalizálása egy könnyen kezelhető kétváltozós függvény minimumkeresésének a problémájára vezet. A keresett minimumhely pedig tömör, zárt alakban megadható. Az alternatívaként felkínált robusztus eljárásra sokkal nehezebb egzakt formulát találni. Ez azért van, mert a *medián* függvény matematikailag nehezen kezelhető. A kívánt robusztus regressziós egyenest jobb híján csak különféle optimumkeresési eljárások segítségével, iteratív módon találhatjuk meg, esetleg akkor is csak bizonyos hibával. Ez számítógép használata nélkül rendkívül bonyolult és hosszadalmas procedúra. Lényegében ez a fő oka annak, hogy a gyakorlati munkában legtöbbször a hagyományos módon számolunk. A mai számítógépekkel azonban már szinte egyformán gyorsan megoldható mindkét fajta egyenesillesztés. Ezért figyelembe véve a robusztussággal járó nyilvánvaló előnyöket – érdemes a második módszert használni. Miután meghatároztuk a

robusztus regressziós egyenest, nézzük meg az egyes pontok reziduumait. Tekintsük azokat a pontokat, amelyek reziduumai jelentősen eltérnek a többi pontra jellemző értékektől (azaz a reziduumok halmazában outliernek minősíthetők valamilyen eljárás alapján). Ezek a pontok vagy hibás mérésből származnak, vagy esetükben más típusú kapcsolat van a vizsgált két változó között, mint a pontok zöménél, esetleg érdemes lehet rájuk egy újabb robusztus regressziós eljárást végrehajtani.

## KÖVETKEZTETÉSEK

A korábban leírt hatásosabb módszerek jól használhatók arra, hogy csökkentsük egy-egy rétegben a túlbecslés mértékét, amennyiben azt valóban egy erősen kiugró érték mintába kerülése okozza. Néhány esetben azonban óvatosnak kell lennünk. Ha magában a kérdéses rétegben nincs jelen kiugró érték, akkor is előfordulhat, hogy a minta legnagyobb eleme outliernek tűnik. Ez lehet a helyzet, ha például egyötödös kiválasztási arány mellett a mintába kerül a második legnagyobb rétegbeli elem, de a második legnagyobb mintaelem a teljes réteg 12. eleme. Ilyenkor a mintában kiugróan nagynak tűnik a legnagyobb elem, azt outliernek minősíti az általunk használt teszt. A vázolt esetben a réteg legnagyobb elemei alulreprezentáltak lehetnek, míg az eljárás az outlierként azonosított elem felszorzási súlyának mérséklésével csökkenti a rétegbeli becslést. Ezáltal előfordulhat, hogy egy amúgy is alulbecsült réteg még inkább alulbecsültté válik. Ráadásul minél sarkítottabban jelentkezik az a probléma, azaz minél inkább alulreprezentált a réteg felső része, annál inkább outliernek tűnik a legnagyobb mintaelem, annál erősebben csökkentjük felszorzási súlyát, ezért annál inkább alulbecsült lesz a réteg. Így ilyenkor még nagyobb hibát okoz a becslés további drasztikus csökkentése.

Előfordulhat olyan eset is, hogy egy olyan rétegben, amely nem tartalmaz kiugró értéket, a minta eloszlása olyan, hogy a nagyobb rétegelemek túlreprezentáltak, ennek következtében pedig a teljes réteg is túlbecsült. Ilyenkor kívánatos lenne csökkenteni a túlbecslést, azonban az outlier-teszt nem azonosít kiugró értéket, hiszen aránylag sok hasonló nagyságrendű elem van jelen a mintában.

Figyelemre méltó, hogy nem csak akkor jelentkezhetnek a fenti problémák, ha a véletlen mintavétel kritériumai sérülnek. Ha nem is túl gyakran, de az esetek mintegy 10 százalékában pusztán a véletlen szeszélyei létrehoznak olyan mintát, melynél az outlierteszt a fenti okok miatt megbukik. Tekintettel arra, hogy sok mintaréteg van, akár tucatnyi rétegben is jelentkezhet ez a probléma. Ha bizonyos rétegeket összevontan kezelünk, akkor csökkenthetjük ezeknek a kellemetlen jelenségeknek az előfordulási valószínűségét, egyúttal azonban előfordulhat, hogy az összevonás következtében az egyedi rétegek problémáit elfedjük.

Az outlierek kezelése során felmerülő problémák előrevetítik, hogy hosszabb távon érdemes lehet bizonyos szervezeteket eleve kiemelten kezelni a reprezentatív megfigyelés rendszerén belül. (Ez a KSH adatgyűjtéseinek jó részénél már gyakorlat.) Ha előre kiválasztjuk és az adatgyűjtésbe bevonjuk azokat a szervezeteket, melyek nagyságuknál fogva potenciális outlierek lehetnek, akkor ezeknek az adatait teljeskörűen számíthatjuk be a becslésbe, ezzel megelőzve a felmerülő problémákat.

Amennyiben a jövőben a teljeskörűsítéshez használt becslési módszer megváltozik, indokolt lehet az outlier-kezelő eljárás felülvizsgálata. Egyes becslési módszerek – pél-

dául a hányadosbecslés – felhasználnak korábbi időszakokra vonatkozó többletinformációt is. Ezt érdemes lehet az outlierek azonosításakor is figyelembe venni.

További nehézség, hogy az outlier-kezelő módszerek csak a túlbecsléseket hivatottak kezelni, az alulbecsléseken nem tudunk javítani velük. Így, ha statisztikánk eleve alulbecsült, akkor még ha a fent vázolt problémás rétegek nem is fordulnak elő, és csak olyan rétegekben korrigáljuk a becslést, ahol valóban túlbecsült volt a kérdéses mutató, akkor is rontunk a helyzeten, hiszen csak növelni tudjuk az alulbecslés mértékét.

Tegyük fel, hogy becslésünk relatív hibája 1–2 százalékos. Némely réteg alulbecsült, mások felülbecsültek. Egy-egy rétegben a becslés hibája jóval jelentősebb lehet, mint a teljes sokaság esetében. Ezek a hibák azonban a különböző rétegek átlagában nagyjából kiegyenlítik egymást. Egy outlier-teszttel, még ha csökkentjük is a túlbecslések hibáját, a teljes sokaság becslését ronthatjuk, mégpedig előre nem látható mértékben, hiszen egy-egy réteg becslésének a hibája jelentősen ingadozhat. Gondot jelenthet az is, hogy egy enyhe mértékű tendenciózus túlbecslést csökkenthetünk ugyan, de ezáltal az idősorban egy törés következik be, melyet a módszertani váltás okoz. Ezért indokolt lehet kisebb lépésekben, évről évre finomítani az outlier-kezelési technikát, valamint ez alatt az átmeneti periódus alatt párhuzamosan az eredeti módszerrel is elkészíteni a becslést.

Láttuk, hogy vannak olyan eljárások, amelyek valamilyen eloszlási modell alapján dolgoznak, és vannak olyanok, amelyek modell-függetlenek. Ha túl keveset tudunk a sokaságról ahhoz, hogy valamilyen előfeltevéssel élhetnénk az eloszlási modell tekintetében, akkor nehéz objektív outlier-szűrő módszert találni. Ilyenkor mindig nagy szerepet kap a tapasztalat, illetve az elérhető segédinformációk szakértői értékelése abban, hogy milyen tesztet használjunk és annak eredményeit milyen szigorúsággal értékeljük.

## IRODALOM


BARNETT, V. – LEWIS, T. [1984]: *Outliers in statistical data*, 2nd ed. Wiley. John Wiley and Sons Ltd. New York.

GRUBBS, F. E. [1969]: Procedures for detecting outlying observations in samples. *Technometrics.* 11. évf. 1. sz. 1–21. old.

http://www.graphpad.com/calculators/GrubbsHowTo.cfm

http://www.itl.nist.gov/div898/handbook/eda/section3/eda35h.htm

HULLIGER, B. – BEGUIN, C.: *Detection of multivariate outliers by a simulated epidemic*. http://webfarm.jrc.cec.eu.int/ETK-NTTS/Papers/final_papers/68.pdf

MUNOZ-GARCIA, J. – MORENO-REBOLLO, J. L. – PASCUAL-ACOSTA, A. [1990]: Outliers: A formal approach. *International Statistical Review.* 58. évf. 3. sz. 215–226. old.

ROUSSEEUW, P. J. – ZOMEREN, B. C. [1990]: Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association.* 85. évf. 411. sz.

VERMA, S. P. [1997]: Sixteen statistical tests for outlier detection and rejection in evaluation of international geochemical reference materials: Example of Microgabbro PM-S. *Geostandards Newsletter.* 21. évf. 59–75. old.


## SUMMARY


The distributions in business statistics are typically very skew. That is why the detection and treatment of outliers is a very important task. In a stratified sampling scheme we are interested in both a good population estimate and in relatively good estimates for single strata. This poses the need of a simultaneous outlier detection algorithm. This can be done by a modified Grubbs-type method. However we must not accept the result of any outlier-test automatically without any critic. There are several reasons to say that. It seems to be that the opinion of an expert is sometimes as important as a good detection-algorithm.

# A NEMVÁLASZOLÁS ELEMZÉSE A MUNKAERŐ-FELVÉTELBEN

GYÖRGY ERIKA

A tanulmány a nemválaszolás jellemzőit vizsgálja meg a Központi Statisztikai Hivatal 2003 első negyedéves munkaerő-felvételén, felhasználva a népszámlálás és a korábbi munkaerő-felvételek adatait. A nemválaszolással kapcsolatos elméleti háttér és definíciók rövid ismertetése után foglalkozik a munkaerő-felvételben a nemválaszolás három típusának: a megtagadás, a nincs kapcsolat és az egyéb nemválaszolás jelenségének alakulásával. Logisztikus regressziós modellekkel elemzi az egyes modellbe bevont magyarázó változóknak a megtagadásra és a nincs kapcsolat jelenségére való hatását.

TÁRGYSZÓ: Nemválaszolás. Munkaerő-felvétel. Logisztikus regresszió.

A Központi Statisztikai Hivatal (KSH) munkaerő-felmérése 2003 januárjától új mintával dolgozik, amelyről már több tanulmány is megjelent (*Mihályffy* [2000], *Lakatos–Mihályffy* [2003]), melyek részletesen foglalkoznak az új minta mintavételi tervével, súlyozási eljárásával. A 2003. első negyedéves munkaerő-felvétel speciális abban a tekintetben, hogy ekkor egy új mintavételi terv alapján mintacserére, új összeírók és részben megváltozott kérdőív alkalmazására került sor. Ebben az időszakban már rendelkezésre álltak a 2001. évi feldolgozott népszámlálási adatok is, kézenfekvőnek mutatkozott tehát az igény, hogy az új munkaerő-felvétel mintáján tapasztalt meghiúsulások vizsgálata a népszámlálási adatok alapján történjen. A munkaerő-felvétel nemválaszolási jellegzetességeiről már korábban is készültek kutatások, elemzések (*Marton* [1995], *Marton–Varga* [2000]), ezek azonban a jelen elemzéssel szemben más módszereken, illetve nem a népszámlálás és a munkaerő-felvétel összekapcsolásán alapultak.

## A MUNKAERŐ-FELVÉTEL MINTÁJÁRÓL RÖVIDEN – A NEMVÁLASZOLÁS TÜKRÉBEN

A 2003 januárjától működő új munkaerő-felvétel mintájának egyik fontos jellemzője, hogy 2004 második negyedévéig két részből állt. A minta egyik része, az ún. régi mintarész még az 1990-es népszámlálási mintakeretet használta, a másik, ún. új mintarész azonban teljesen az új, 2001-es népszámláláson alapult. (*Mihályffy* [2004]). A régi mintáról az új mintára való átmenet ily módon fokozatosan valósult meg.

A régi és az új mintában közös 204 település mellett a régi mintából kilépett 571 település, és a helyükre belépett 480 település. A településcsere folyamatosan ment végbe a mintában. Míg januárban a közös részen levő településeken a régi mintarész ötszöröse volt az új mintarésznek, ez negyedévente egyhatoddal növekedett az új mintarész javára, így 2004. második negyedévétől a minta kizárólag új részből áll.

Mindezek az alábbiak szerint lehetnek összefüggésben a nemválaszolással:

– Az új mintába frissen bekerült településeken új összeírókat alkalmazott a KSH, ami hatással lehet a nemválaszolásra.

– A munkaerő-felvételben minden háztartás maximum hat egymást követő negyedévig vesz részt a felvételben, majd azt követően kirotálódik abból. Ezt a hat negyedévet a következőkben hullámoknak, az aktuális negyedévet a kérdezés sorszámának nevezzük. A minta kialakításakor a kiválasztott címek megoszlása egyenletes a hullámok szerint. Ez az új mintára is teljesült technikailag, ám valójában az új mintába került háztartások tulajdonképpen első alkalommal kerülnek kapcsolatba a felvétellel. Így az új mintába kiválasztott címek első hullámosnak tekinthetők abban az értelemben, hogy ezeknek a háztartásoknak a viselkedése hasonló az első hullámba került háztartásokéhoz. Ez a megkülönböztetés azért fontos, mert a nemválaszolás általában az első hullámban a legnagyobb, a negyedévek elteltével folyamatosan csökken (*Marton* [1995], *Marton–Varga* [2000]). Aki tehát egy korábbi hullámban már válaszolt az összeíró kérdéseire, az a későbbi hullámokban kisebb eséllyel lesz nemválaszoló. A nemválaszolás elemzésekor a tényleges kérdezési sorszámokat érdemes vizsgálni a technikai sorszámok helyett. A 2003. év első negyedéves mintájának ezen sorszámok alapján való megoszlását láthatjuk az 1. táblában.

1. tábla

*A 2003. I. negyedéves munkaerő-felvételbe kiválasztásra került címek megoszlása*

| Időszak | Kérdezés sorszáma | Kiválasztott címek száma (darab) | | | | |
|---|---|---|---|---|---|---|
| | | Régi mintarész | Új mintarész | | Összesen | |
| | | | Technikai kérdezési sorszám szerint | Tényleges kérdezési sorszám szerint | Technikai kérdezési sorszám szerint | Tényleges kérdezési sorszám szerint |
| 2003. január | 1 | | 2 292 | 8 204 | 2 292 | 8 204 |
| | 2 | 1 235 | 1 191 | | 2 426 | 1 235 |
| | 3 | 1 169 | 1 178 | | 2 347 | 1 169 |
| | 4 | 1 121 | 1 182 | | 2 303 | 1 121 |
| | 5 | 1 122 | 1 176 | | 2 298 | 1 122 |
| | 6 | 1 117 | 1 185 | | 2 302 | 1 117 |
| *2003. január összesen* | | *5 764* | *8 204* | *8 204* | *13 968* | *13 968* |
| 2003. február | 1 | | 2 292 | 8 206 | 2 292 | 8 206 |
| | 2 | 1 243 | 1 182 | | 2 425 | 1 243 |
| | 3 | 1 166 | 1 176 | | 2 342 | 1 166 |
| | 4 | 1 129 | 1 185 | | 2 314 | 1 129 |
| | 5 | 1 120 | 1 180 | | 2 300 | 1 120 |
| | 6 | 1 110 | 1 191 | | 2 301 | 1 110 |
| *2003. február összesen* | | *5 768* | *8 206* | *8 206* | *13 974* | *13 974* |
| 2003. március | 1 | | 2 261 | 8 104 | 2 261 | 8 104 |
| | 2 | 1 239 | 1 171 | | 2 410 | 1 239 |
| | 3 | 1 168 | 1 164 | | 2 332 | 1 168 |
| | 4 | 1 135 | 1 176 | | 2 311 | 1 135 |
| | 5 | 1 104 | 1 164 | | 2 268 | 1 104 |
| | 6 | 1 103 | 1 168 | | 2 271 | 1 103 |
| *2003. március összesen* | | *5 749* | *8 104* | *8 104* | *13 853* | *13 853* |
| *2003. I. negyedév* | | *17 281* | *24 514* | *24 514* | *41 795* | *41 795* |

## A NEMVÁLASZOLÁSRÓL

Nemválaszolásról (nonresponse) akkor beszélünk, ha egy mintában, cenzusban, egyes változók vonatkozásában, vagy netán semmilyen vonatkozásban nem sikerül információt nyerni az adatszolgáltatótól. A nemválaszolásnak tehát két fő típusa van, beszélhetünk egyrészt tétel szintű (item nonresponse), másrészt egység szintű nemválaszolásról (unit nonresponse). (*Biemer–Lyberg* [2003])

A tétel szintű nemválaszolás esetében az adatszolgáltatótól ugyan használható információt nyerünk egyes változókra nézve, de néhány változó esetében hiányos adatokat kapunk. Egység szintű nemválaszolás pedig akkor következik be, ha a megkérdezett egyáltalán nem szolgáltat információt, azaz a mintavételi egység szintjén jelentkezik a nemválaszolás. A továbbiakban ez utóbbit elemezzük.

Ismeretes a nemválaszolás következő csoportosítása (*Covar–Rancourt* [2003]):

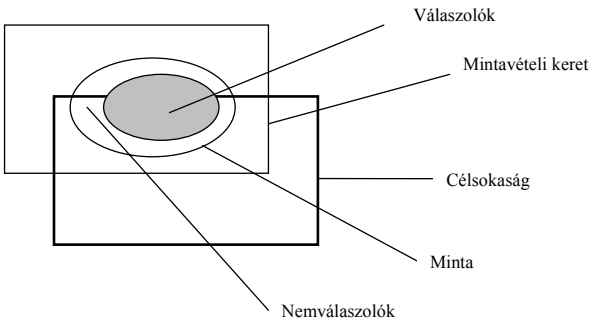
– Teljesen véletlenszerű adathiány. Ebben az esetben nincs különbség a válaszolók és a nemválaszolók között, így a nemválaszolás nem visz szisztematikus torzítást a becslésekbe. A nemválaszolás nem a megkérdezett változó miatt következik be. Ha egy felvételben megkérdezik például a háztartás létszámát, akkor a nemválaszolás valószínűleg nem emiatt a változó miatt történik.

– Véletlenszerű adathiány. Ennél a típusnál az adathiány nem a megkérdezett változóval van összefüggésben. Ha egy felvételben megkérdezik a nőket a terhességeik, majd a gyermekeik számáról, akkor a nemválaszolás valószínűleg nem önmagában a megkérdezett változók miatt következik be, hanem például azért, mert ezekből következtetni lehet egy újabb változóra (például az elvégzett abortuszok számára), amely végül is a nemválaszolást eredményezi.

– Nem véletlenszerű adathiány. Ekkor az adathiány kizárólag a megfigyelt változónak köszönhető. Ez utóbbi eset a legveszélyesebb, ilyenkor a nemválaszolás szisztematikusan torzíthatja a becsléseket. Tipikus példa erre a jövedelemmel kapcsolatos felvétel, ahol egyértelműen a megkérdezett változó miatt következhet be nemválaszolás.

Az 1. ábra a nemválaszolókat egy általános mintavételi elméleti rendszerben helyezi el *(Lundström–Sarndal* [2002]).

*1. ábra. A célsokaság, a mintavételi keret, a minta és a nemválaszolás kapcsolata*




A nemválaszolás azért rejt veszélyeket magában, mert a mintavételi hiba növekedéséhez és torzított becslésekhez vezethet. Az előbbi az elemszám csökkenése miatt következhet be, az pedig, hogy a becslés végül a nemválaszolás következtében torzított lesz-e vagy sem, függ egyfelől a nemválaszolás mértékétől, valamint a válaszadók és

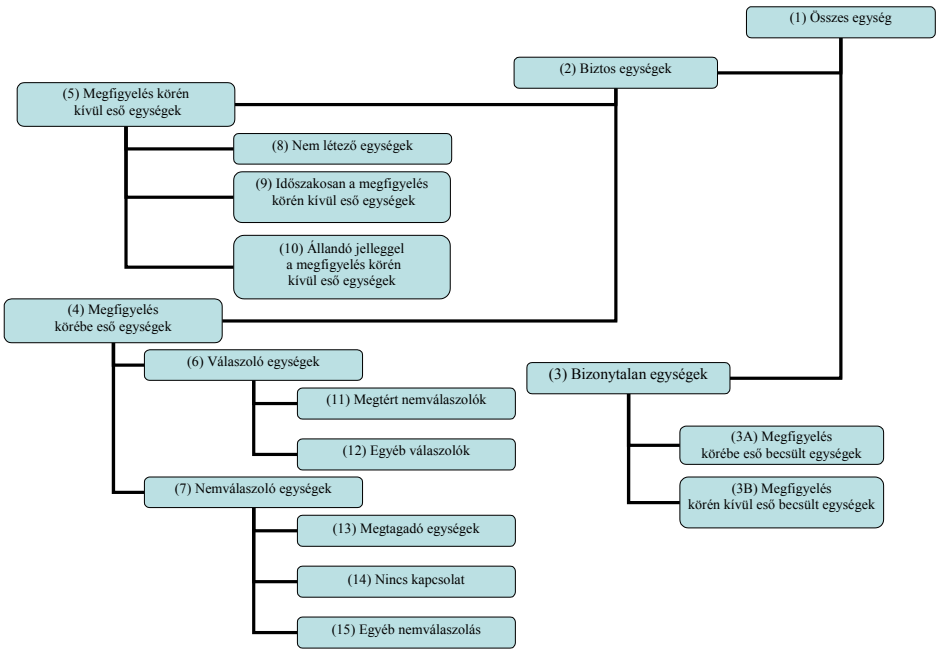
nemválaszolók jellemzőinek különbözőségétől (*Platek* [1986]). Ha rendelkezésünkre állnak a nemválaszolókra vonatkozó ismeretek, akkor ez segíthet a nemválaszolás okainak feltárásában, és hozzásegíthet olyan pótlólagos információkhoz, amelyek segítségével csökkenthető a nemválaszolás okozta torzítás – vagyis javulhat az adatfelvétel és a végeredményként kapott statisztikák minősége, ami napjainkban egyre alapvetőbb követelmény a statisztikai hivatalokkal szemben (*Laaksonen* [1995]). A munkaerő-felvételről az Eurostatnak kötelezően küldendő minőségi jelentések (quality reports) külön fejezetet szentelnek a nem mintavételi hibának, azon belül is a nemválaszolás mértékének és kezelési módjának (*Eurostat* [2001]).

A következőkben ismertetjük a nemválaszolással kapcsolatos, a Kanadai Statisztikai Hivatalban alkalmazott definíciókat, kidolgozásuk *Hidiroglou, Drew* és *Gray* [1993] nevéhez fűződik.

A nemválaszolási és válaszolási arányokat a szerzők olyan változók hányadosaként definiálják, amelyek a válaszolás vagy nemválaszolás megadott kategóriáit reprezentálják egy vizsgált területen. Ezek a változók lehetnek súlyozatlanok vagy súlyozottak. Az általuk kidolgozott rendszer mind a lakossági, mind a gazdasági felvételekre alkalmas, és a nemválaszolást elsősorban a mintavételi egység szintjén definiálja (bár a tétel szintű nemválaszolásra is adaptálható).

A 2. ábra bemutatja a szerzők által képviselt elméleti rendszert, a mintavételi egységek adatgyűjtési folyamat során való csoportosítását.

*2. ábra. A válaszolók és nemválaszolók összetétele az adatgyűjtési folyamat során (Hidiroglo–Drew–Gray [1993])*



(1) Az *összes egységet* azok az egységek alkotják, amelyekről az adatfelvétel folyamatának kezdete előtt rendelkezésre álló információk alapján feltehető, hogy azok a célsokasághoz tartoznak. Értéke lehet súlyozott és súlyozatlan is.

(2) A *biztos egységek* azok, amelyeknek célsokasághoz való tartozása vagy nem tartozása az adatfelvétel folyamatának befejeződéséig ismert.

(3) A *bizonytalan egységek* státusza az adatfelvétel folyamatának befejeződéséig nem határozható meg. Innen:

$$\text{Biztos egységek aránya} = \frac{\text{Biztos egységek száma}}{\text{Összes egység száma}}$$

(4) A *megfigyelés körébe eső egységeket* figyelik meg az adatgyűjtés során.

$$\text{Megfigyelés körébe eső egységek aránya} = \frac{\text{Megfigyelés körébe eső egységek száma}}{\text{Biztos egységek száma}}$$

(5) Hasonlóan, a *megfigyelés körén kívül eső egységeket* nem figyelik meg az adatgyűjtés folyamán.

$$\text{Megfigyelés körén kívül eső egységek aránya} = \frac{\text{Megfigyelés körén kívül eső egységek száma}}{\text{Biztos egységek száma}}$$

(6) A *válaszoló egységek* azok a megfigyelés körébe eső egységek, amelyek az adatgyűjtés befejeződésének időpontjáig válaszoltak, és ennek során „használható" információt nyújtottak. A használható információ fogalmát a szerzők azokra a válaszolókra is érvényesnek tekintik, akik/amelyek csak részleges információt nyújtottak (item nonresponse). Azt javasolják azonban, hogy a kérdőív kitöltöttségének vonatkozásában legyen kijelölve egy olyan küszöb, amely alatt az egységek nemválaszolóknak tekinthetők.

A *válaszadási arány* az elemzés céljától függően többféleképpen is definiálható. A szerzők elsősorban a következőt részesítik előnyben:

$$\text{Válaszadási arány} = \frac{\text{Válaszadó egységek száma}}{\text{Megfigyelés körébe eső egységek száma} + \text{Bizonytalan egységek száma}}$$

Mivel a bizonytalan egységek egy része a megfigyelés körén kívül eső egységek közé tartozik – amelyektől az adatgyűjtés során értelemszerűen nem szerzünk információt –, ezért a fenti képlet alulbecsüli a valós válaszolási arányt, konzervatív becslést adva ezáltal a mintavételi keret és az adatgyűjtés folyamatának minőségére.

Egy alternatív definíciót is felajánlanak, itt a nevezőben csak a megfigyelés körébe eső elemek száma található. Ez az arány kimondottan az adatgyűjtési folyamat hatékonyságát méri.

(7) A *nemválaszoló egységek* a megfigyelés körébe eső egységeken belül azok, amelyek eddig egyik kategóriába sem kerültek. A nemválaszolási arány a válaszolási arány komplementere.

$$\text{Nemválaszolási arány} = \frac{\text{Nemválaszoló egységek száma} + \text{Bizonytalan egységek száma}}{\text{Megfigyelés körébe eső egységek száma} + \text{Bizonytalan egységek száma}}$$

A bizonytalan egységek számlálóból és nevezőből való elhagyásával egy további definíció is felírható a nemválaszolási arányra. Hasonló módon lehetséges egy definíció a bizonytalan egységek tovább bontásával is, a megfigyelés körébe eső becsült és a megfigyelés körén kívül eső becsült egységekre is.

(11) *Megtért megtagadók* azok az egységek, akik bár megtagadták a válaszolást az adott vagy egy korábbi periódusban, az utókövetés során mégis sikerült őket válaszolásra bírni. A *megtért megtagadási arány* annak az erőfeszítésnek a sikerét méri, amelyet arra fordítanak, hogy az adatgyűjtés időszakában a megtagadókat válaszolásra bírják.

$$\text{Megtért megtagadási arány} = \frac{\text{Megtért megtagadó egységek száma}}{\text{Megtért megtagadó egységek száma} + \text{Megtagadó egységek száma}}$$

(12) Azokat a válaszolókat, akik nem megtérített megtagadók, egyszerűen *egyéb válaszolóknak* nevezik.

(13) A *megtagadók* azok a nemválaszoló egységek, amelyekkel sikerült ugyan kapcsolatba lépni, de visszautasították a kérdőívben való közreműködést.

$$\text{Megtagadási arány} = \frac{\text{Megtagadók száma}}{\text{Megfigyelés körébe eső egységek száma}}$$

(14) A *nincs kapcsolat* esetei közé azok a megfigyelés körébe eső egységek tartoznak, amelyekkel nem sikerült kapcsolatba lépni. Lakossági felvételeknél ilyenek például azok a lakások, amelyek lakói időszakosan távol vannak, valamint azok a háztartások, amelyekből senki sem volt otthon az összeíró ottjártakor.

$$\text{Nincs kapcsolat aránya} = \frac{\text{Nincs kapcsolat eseteinek száma} + \text{Bizonytalan egységek száma}}{\text{Megfigyelés körébe eső egységek száma} + \text{Bizonytalan egységek száma}}$$

(15) Az *egyéb nemválaszolók* közé azok az egységek tartoznak, akik speciális körülmények miatt (például nyelvi problémák) nem válaszoltak vagy nem szolgáltattak használható információt.

$$\text{Egyéb nemválaszolási arány} = \frac{\text{Egyéb nemválaszolók száma}}{\text{Megfigyelés körébe eső egységek száma}}$$

## NEMVÁLASZOLÁSI JELLEMZŐK A 2002–2003. ÉVI MUNKAERŐ-FELMÉRÉSEKBEN

A következő néhány tábla és ábra a *Hidirouglou* és szerzőtársai által kidolgozott fogalmak alapján mutatja be a munkaerő-felvétel negyedéves mintáin 2002-től 2003 év végéig a nemválaszolási jellemzők alakulását, súlyozatlan arányokat használva.

A munkaerő-felvételben háztartásokra és egyénekre is rögzítenek nemválaszolási kódokat. Mivel azonban ez utóbbi kevésbé jellemző, valamint a háztartás egy tagjának a nemválaszolása a munkaerő-felvételben maga után vonja a teljes háztartás nemválaszolását, elemzésünkben csak a háztartás szintjén történő nemválaszolással fog-

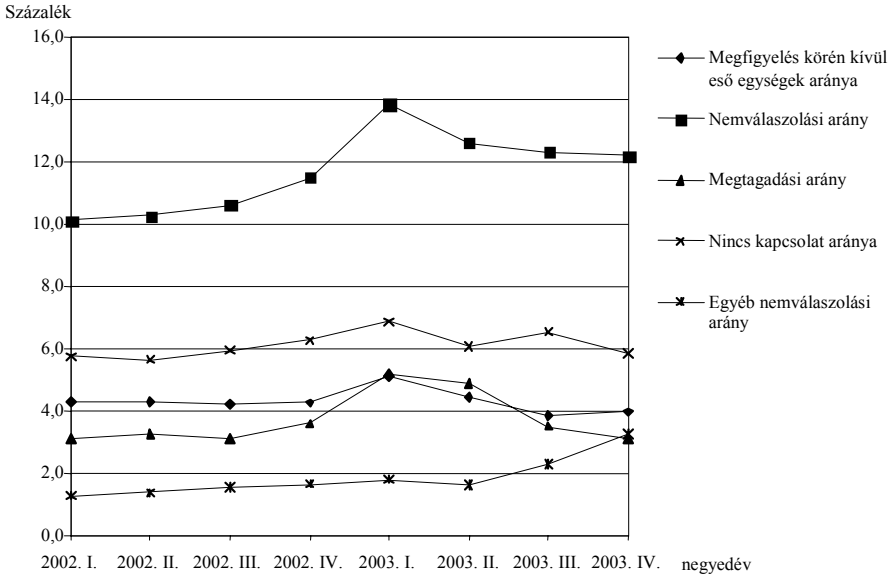
lalkozunk. A továbbiakban ezért megtagadók és nemválaszolók alatt mindig háztartásokat értünk, és a nincs kapcsolat jelenségét is a háztartásra értelmezzük.

2. tábla

*Súlyozatlan nemválaszolási arányok a 2002–2003-as munkaerő-felvételekben*
(százalék)

| Időszak (negyedév) | Megfigyelés körébe eső egységek aránya | Megfigyelés körén kívül eső egységek aránya | Válaszadási arány | Nemválaszolási arány | Megtagadási arány | Nincs kapcsolat aránya | Egyéb nemválaszolási arány |
|---|---|---|---|---|---|---|---|
| 2002. I. | 95,7 | 4,3 | 89,9 | 10,1 | 3,1 | 5,8 | 1,2 |
| 2002. II. | 95,7 | 4,3 | 89,7 | 10,3 | 3,2 | 5,7 | 1,4 |
| 2002. III. | 95,8 | 4,2 | 89,4 | 10,6 | 3,1 | 5,9 | 1,6 |
| 2002. IV. | 95,7 | 4,3 | 88,5 | 11,5 | 3,6 | 6,3 | 1,6 |
| 2003. I. | 94,9 | 5,1 | 86,1 | 13,9 | 5,2 | 6,9 | 1,8 |
| 2003. II. | 95,6 | 4,4 | 87,4 | 12,6 | 4,9 | 6,1 | 1,6 |
| 2003. III. | 96,1 | 3,9 | 87,7 | 12,3 | 3,5 | 6,5 | 2,3 |
| 2003. IV. | 96,0 | 4,0 | 87,8 | 12,2 | 3,1 | 5,8 | 3,3 |

*3. ábra. Súlyozatlan nemválaszolási arányok a 2002–2003-as munkaerő-felvételekben*




A nemválaszolási arány összetevőit – a megtagadási arányt, a nincs kapcsolat arányát és az egyéb nemválaszolási arányt – vizsgálva megállapítható, hogy 2003 első negyedévében mindhárom arány megnőtt, majd ezt követően a megtagadási arány és a nincs kapcsolat aránya csökkenésnek indult. Ennek eredményeként a nemválaszolási arány 2003 első negyedévtől csökkenést mutat, de 2003 év végére még így is magasabb szintre áll be, mint a 2002-es év végi érték.

A 3. tábla a nemválaszolási arányokat a hullámok függvényében mutatja be.

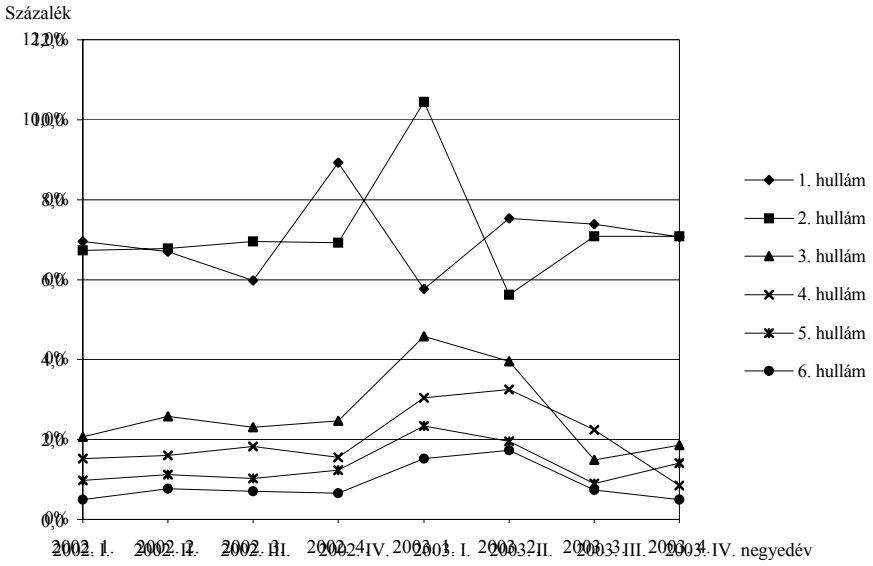
3. tábla

*A 2003. első negyedéves munkaerő-felvétel mintájának összetétele és a nemválaszolási arányok*
(százalék)

| Minta-rész | Kérdezés gyakorlati sorszáma | Megtagadási arány | Nincs kapcsolat aránya | Egyéb nemválaszolási arány | *Nemválaszolási arány összesen* |
|---|---|---|---|---|---|
| Új | 1. hullám | 5,8 | 7,6 | 1,1 | *14,5* |
| | 2. hullám | 10,4 | 9,1 | 2,5 | *22,0* |
| | 3. hullám | 4,6 | 6,4 | 2,1 | *13,1* |
| Régi | 4. hullám | 3,0 | 5,0 | 3,1 | *11,1* |
| | 5. hullám | 2,3 | 4,3 | 3,0 | *9,6* |
| | 6. hullám | 1,5 | 4,2 | 3,4 | *9,1* |

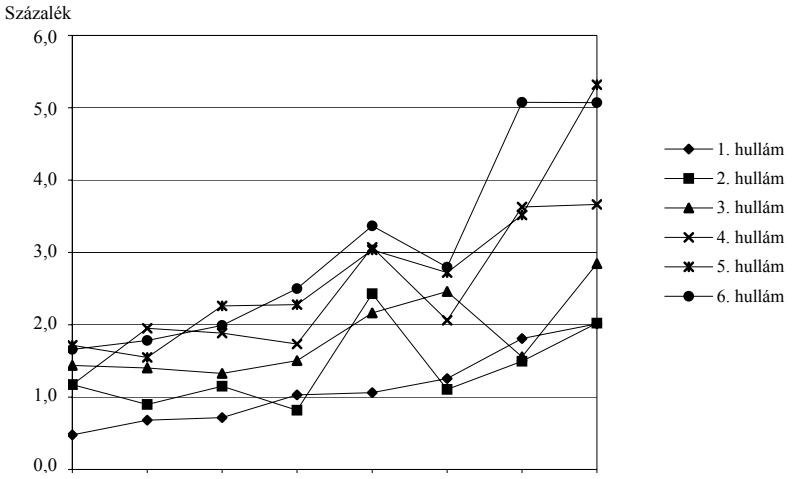
Bár a megfigyelt időszakban az új mintarészbe tartozó valamennyi háztarás először került kapcsolatba a felvétellel – gyakorlatilag első hullámos háztartásnak tekinthető –, mégis itt a nemválaszolási arány mindegyik típusa kisebb volt, mint a régi mintarészben az első illetve második hullámban tapasztalt értékek. Az idő függvényében a megtagadási arány és a nincs kapcsolat aránya a 4-5. ábrában bemutatottaknak megfelelően alakult.

*4. ábra. Megtagadási arány a különböző hullámokban a 2002–2003 évi munkaerő-felvételekben*



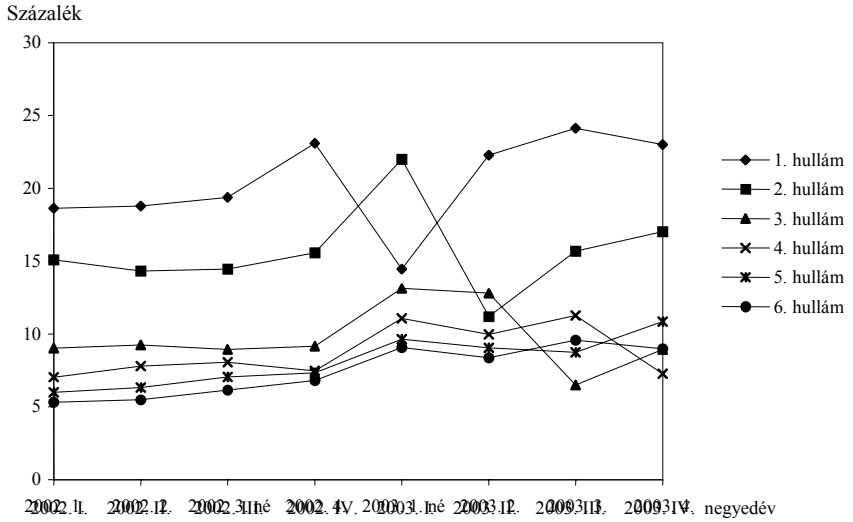

A 2002. év végéről 2003. év elejére növekedés tapasztalható a megtagadási arányokban a 2-6. hullámig (tehát a régi mintarészben). A második negyedévtől néhány hullám esetében megkezdődik a megtagadási arányok csökkenése, majd ezek 2003 év végére megközelítik a 2002-es év eleji szintet. Az első hullámban (azaz az új mintarészen) feltűnően kedvezően alakult a megtagadási arány, értéke a 2002. negyedik negyedéves 8,9 százalékról 5,8 százalékra csökkent. A következő negyedévben aztán növekedésnek indul (7,5%), majd ezt követően enyhén csökkenve eléri a 2002-es első negyedévi szintet (7,1%).

*5. ábra. Az egyéb nemválaszolás aránya a különböző hullámokban a 2002–2003. évi munkaerő-felvételekben*




Amíg a megtagadási arányok alakulásához teljesen hasonló tendencia tapasztalható a nincs kapcsolat aránya esetében, addig az egyéb nemválaszolás alakulása eltér ezektől. Az egyéb nemválaszolás (5. ábra) minden hullámban növekedést mutat, míg azonban az első hullámban szereplő háztartások esetében csak szerény növekedést, addig a többi hullámban felkeresett háztartás esetében nagyfokú, hirtelen megugrást tapasztalunk. A második negyedévre a legtöbb esetben visszaesés állapítható meg, de az év végére az egyéb nemválaszolási arány jócskán meghaladja mindegyik hullámban az előző évi értéket.

*6. ábra. A nemválaszolás aránya a különböző hullámokban a 2002–2003. évi munkaerő-felvételekben*



A 6. ábra a teljes nemválaszolási arányt mutatja be, tekintet nélkül arra, hogy az melyik kategóriába (megtagadás, nincs kapcsolat, egyéb nemválaszolás) esik.

A teljes nemválaszolásról elmondható, hogy az – 2003. első negyedévét kivéve – az első hullámban a legnagyobb. Minél több hullámban szerepel egy háztartás a felvételben, általában annál jobban csökken a nemválaszolási arány. Az első hullámban tapasztalható legnagyobb nemválaszolási arány azért érthető, mert ilyenkor a legnehezebb a háztartásokat meggyőzni arról, hogy hat egymást követő negyedévben válaszoljanak a felvétel kérdéseire. A legkisebb arány pedig azért várható a 6. hullámban, mert ha egy háztartás már egymást követő 5 negyedévben válaszolt a felvételre, akkor már nagy valószínűséggel az utolsó hullámban is válaszolni fog. A 2002–2003. év során tapasztaltak tehát nagyjából meg is feleltek ezeknek a várakozásoknak, azonban 2003 első negyedéve ebből a szempontból kivételesnek mondható. Az ehhez tartozó első hullámot (vagyis az új mintarészt) a szokásosnál jobb nemválaszolási arányok jellemzik. Ezek a háztartások még a második alkalommal felvételben szereplő háztartásokhoz képest is kisebb nemválaszolási aránnyal rendelkeznek.

## A NEMVÁLASZOLÁS KEZELÉSE

A nemválaszolást számos módon kezelhetjük. Ha a nemválaszolók összetétele, jellemzői hasonlítanak a válaszolókéhoz, akkor a nemválaszolás hatása kiküszöbölhető – a nemválaszolási rátát figyelembe véve – egy nagyobb minta kiválasztásával. Ha azonban a válaszolók alapjában véve különböznek a nemválaszolókról, akkor a nemválaszolás jelentős torzítást vihet a becslésekbe. Ilyenkor nem elegendő pusztán a minta elemszámának emelése, a probléma differenciáltabb megközelítésmódot igényel. Megfelelő technikákat felhasználva, ha ismételten kapcsolatba lépünk a nemválaszolókkal, elérhetjük, hogy közvetlen tőlük szerezzük be a szükséges információkat. Ha ez nem lehetséges, akkor külső információk is felhasználhatók, például adminisztratív forrásokból. A felsoroltakon kívül a hiányzó adatok pótlása (vagy imputálása), és az átsúlyozás szolgálhatnak még eszközül a nemválaszolás kezelésére.

A nemválaszolókról azonban nem könnyű információkat gyűjteni. A gyűjtés történhet a nemválaszolók ismételt megkérdezésével, vagy más információforrások (egyéb felvételek, adminisztratív adatok) felhasználásával. Ez utóbbi források a munkaerő-felvétel nemválaszolásának elemzésekor a 2001. évi népszámlálás, illetve a korábbi munkaerő-felvételek adatai voltak, amelyek összekapcsolásától remélhető volt, hogy minél több ismeretet szerezhetünk a nemválaszolókról.

Abban az esetben, ha a nemválaszoló háztartást sikerült már válaszadásra bírni valamelyik előző hullámban, akkor ezek a korábbi munkaerő-felvételben szereplő adatok alapul szolgáltak a nemválaszoló háztartások adatainak imputálásához. Ha korábbi munkaerő-felvételek adatait nem lehetett felhasználni, mert a háztartás egyik megelőző hullámban sem válaszolt, akkor a népszámlálás adatait használtuk.

Az elemzés elvégzéséhez a népszámlálás következő változóit imputáltuk:

– Demográfiai tényezők: háztartásban élők kora, neme, családi állapota, kiskorúak száma;
– Társadalmi jellemzők: háztartásban élők iskolai végzettsége, gazdasági aktivitása, foglalkozási viszonya;
– Lakás adatai: a lakás nagysága, komfortossága, a lakóövezet jellege;
– Település-fejlettségi mutatók.

Az összefüggések feltárásához az összeírók jellemzőit is felhasználtuk: az összeírók nemét, korát, végzettségét és tapasztalatát tartalmazó változókat.

Mivel a nemválaszolást mint háztartás szintjén tapasztalt jelenséget szerettük volna vizsgálni, ezért gyakran háztartás szintű változókat hoztunk létre, hogy beazonosítsuk a nemválaszoló háztartások jellemzőit, összetételét. Ilyen volt például a háztartásban élők együttes iskolai végzettsége, amikor is klaszteranalízis segítségével osztottuk be a háztartásokat az alapfokú, középfokú és felsőfokú kategóriákba, létrehozva ezzel egy, a háztartást jellemző komplex iskolai végzettség változót.

Ugyancsak komplex mutatóval ragadtuk meg a háztartásban élők aktivitását. Ebben az esetben azt vizsgáltuk, hogy az aktivitás vagy az inaktivitás aránya a nagyobb a háztartásban, és ennek megfelelően tekintettük a háztartást komplexen aktívnak vagy inaktívnak.

A háztartásban élők családi állapotát is komplex mutatóval fejeztük ki. A háztartás komplex családi állapotát házasnak tekintettük, ha abban volt házas. Nőtlennek/hajadonnak, ha nem volt a háztartásban házas, de volt legalább egy nőtlen-hajadon személy. Özvegy komplex családi állapotot akkor feltételeztünk, ha a háztartásban nem volt házas és nőtlen-hajadon személy, tehát csak özvegyekből és/vagy elváltakból állt a háztartás. Végül elváltnak a kizárólag elvált személy(ek)ből álló háztartást tekintettük.

Feltettük, hogy a háztartások elemzésben felhasznált főbb jellemzői (nem, kor, iskolai végzettség, kiskorúak száma) a két év során nem változtak. (Ennek az elfogadására sajnos az adott körülmények miatt szükség volt, ez azonban egy olyan feltétel, amelyen változtatva a későbbiekben fejleszteni lehetne az elemzésünkön.)

Az összetételbeli torzulásokat a mintavételi tervnek megfelelő súlyozással és kalibrálással igyekeztük kiküszöbölni. A nemválaszolás elemzése során ezért kalibrált nemválaszolási arányokkal fogunk dolgozni. Mivel a súlyozás háztartási szinten történik, és az elemzéshez a háztartások jellemzőit használtuk fel, ezért megállapításainkat a háztartásokra tettük.

## A NEMVÁLASZOLÁS ELEMZÉSÉRE ALKALMAZOTT LOGISZTIKUS REGRESSZIÓS MODELLEK FELÉPÍTÉSÉNEK SZEMPONTJAI

Elemzéseinket egy többváltozós elemzési módszer, a dichotom logisztikus regresszió alkalmazásával végeztük. (A logisztikus regressziós modellről lásd bővebben *Hajdu* [2003], *Agresti* [2002], *Székelyi–Barna* [2003].)

A továbbiakban a nemválaszolás két típusára, a megtagadásra, illetve a nincs kapcsolatra épített modelleket mutatjuk be.[1]

Tekintsük elsőnek a megtagadás esetét. Ekkor jelölje az $Y$ eredményváltozó 0 értéke a válaszolást, 1 a megtagadást. Az $Y$ tehát diszkrét, kategória kimenetű, dichotom eredményváltozó. A megtagadás valószínűségét jelölje $P(Y=1)$, a válaszolás valószínűségét pedig $P(Y=0)=1-P(Y=1)$.

A logisztikus regresszió azonban nem közvetlenül a két, egymást kölcsönösen kizáró kategória valószínűségeivel számol, hanem a valószínűségek hányadosával, az ún. odds-

[1] Az egyéb nemválaszolásra épített modell illeszkedése – valószínűleg a befolyásoló tényezők heterogenitása miatt – nem bizonyult kielégítőnek, így ennek ismertetésétől eltekintünk.

aránnyal más néven esélyhányadossal. Esetünkben az odds-arány a megtagadás és a válaszolás valószínűségének aránya:

$$\text{Odds} = \frac{P(Y=1)}{1-P(Y=1)}.$$

A logisztikus regressziós modellben az odds logaritmusa az $x_1$, $x_2$, …, $x_k$ magyarázó változók lineáris kombinációja:

$$\log \frac{P(Y=1)}{1\text{-}P(Y=1)} = b_0+b_1x_1+\ldots+b_kx_k.$$

$$\text{Ebből következően } \frac{P(Y=1)}{1\text{-}P(Y=1)} = \text{Exp}(b_0 + b_1x_1 + \ldots + b_kx_k).$$

A logisztikus regressziós modell eredménye az egyes események bekövetkezési valószínűsége, amelynek kalkulációs alapja tehát az odds-arány. A regressziós paraméterek értelmezésére az Exp($b_i$) (a továbbiakban odds-arány mutató vagy multiplikátor) szolgál, amely az $x_i$ magyarázó változó egységnyi abszolút növekedésének az odds-arányra gyakorolt parciális multiplikatív hatását mutatja a többi változó rögzített értéke mellett. Az Exp($b_i$) multiplikátor értéke függ a viszonyítás alapjától. Ez utóbbi a referenciakategória, amelyet minden $x_1$, $x_2$, ..., $x_k$ magyarázó változó esetében meg kell adnunk.

A Függelék 1. táblája alapján egy budapesti III. kerületi háztartáshoz tartozó odds-arány mutató értéke 17,114 a megtagadás 1. modellje alapján, ahol az egyéb településen (nem Budapesten és nem megyei jogú városban) élő háztartásokat jelöltük ki referenciakategóriának. Ez azt jelenti, hogy egy III. kerületi háztartásnál a megtagadásnak a válaszoláshoz viszonyított esélye – ceteris paribus – várhatóan több mint 17-szer akkora, mint egy egyéb településen élő háztartásé.

A továbbiakban bemutatjuk a nemválaszolás különböző típusaira alkotott három-három modellt, amelyek hierarchikusan egymásra épülnek. Erre azért volt szükség, mert a logisztikus regressziós modelleknél egy változóra jellemző multiplikátor számítása úgy történik, hogy azt a többi változó hatásától – azok modellbe való bevonásával – megtisztítjuk. Az egymással ok-okozati kapcsolatban lévő változóknak a modellben történő együttes szerepeltetése eredményezheti azt, hogy az okozó változó hatását alulbecsüljük, mivel az ahhoz kapcsolódó egyes hatások átvándorolhatnak a többi, újonnan bevont változó multiplikátoraiba. Éppen ezért, az okozó változó hatását egy korábbi, az okozati változót nem tartalmazó modellben tudjuk jól mérni.

Az első modell mindegyik esetben a következő változókat tartalmazza (a változók a népszámlálást követő két év elteltével viszonylag stabilnak tekinthetők):

– háztartásban élők demográfiai adatai: nem szerinti megoszlás, korcsoport szerinti megoszlás, iskolai végzettség szerinti összetétel;
– összeíró neme, kora, iskolai végzettsége, tapasztalata;
– település (Budapest kerületei, megyei jogú városok és egyéb települések);
– kérdezés sorszáma.

A második modellt már kiegészítettük olyan változókkal, amelyek az első modellbe bevont változókkal ok-okozati viszonyban állnak, és ugyancsak állhatnak kapcsolatban a nemválaszolással. Ezek a következők:

– háztartásban élők családi állapot szerinti összetétele;
– háztartásban nevelt kiskorúak száma;
– háztartásban élők aktivitása;
– vállalkozók száma a háztartásban.

Ezek a változók már kevésbé mondhatók stabilnak. A háztartás családi állapot szerinti összetétele és a gyerekszám kapcsolatban lehet a háztartás kor szerinti megoszlásával, hiszen a kor nagyban meghatározhatja a családi állapotot, és a gyerekvállalás is korhoz köthető az egyén életében. A háztartásban élők aktivitása és a vállalkozók száma a háztartásban pedig függhet az iskolai végzettségtől. Az 1. modellbe bevont változók odds-arány mutatói módosulni fognak a későbbi modellekben, az új változók modellbe való bevonásának következtében.

A harmadik modellbe bevont változók a lakással és a lakókörnyékkel állnak kapcsolatban. Ezek:

– lakás nagysága;
– lakás komfortossága;
– lakóövezet jellege.

Ezek azért csak a legutolsó modellbe kerültek – bár hatásuk feltehetőleg jelentős –, mert szerintünk ok-okozati kapcsolatban állnak a háztartás aktivitásával, esetleg a vállalkozói léttel, mindazzal, ami még a jövedelemre enged következtetni. A magasabb jövedelműek valószínűleg nagyobb eséllyel élnek jobb lakóövezetben és nagyobb lakásban.

A felsorolt modellek mindegyikében jelezzük egy, két vagy három csillaggal az egyes ismérvekhez és azok változataihoz tartozó szignifikanciaszinteket az alábbiak szerint

$$\text{Szignifikancia} = \begin{cases} *, \text{ ha } p \text{ érték} < 0{,}01 \\ **, \text{ ha } 0{,}01 <= p \text{ érték} < 0{,}05 \\ ***, \text{ ha } 0{,}05 <= p \text{ érték} < 0{,}1 \end{cases}$$

Szintén közöljük az egyes változókhoz tartozóan a modellbe bevonás sorrendjét a stepwise algoritmus alapján, amely megmutatja, hogy a változók milyen sorrendben gyakorolnak szignifikáns hatást a modell illeszkedésének javulására.[2]

Bár a legjobb illeszkedést mindegyik esetben a legtöbb változót tartalmazó, harmadik modellben sikerült elérni, azonban éppen a változók közti ok-okozati kapcsolatok miatt az első két modell magyarázata is fontosnak ígérkezett, így azokat is bemutatjuk a végeredménynek tekinthető harmadik modell mellett.

Természetesen még számos más háztartási változót is kialakíthattunk volna, igyekeztünk azonban jól interpretálható változókat választani. A modellekben csak főhatásokat vizsgáltunk, a kereszthatásokat nem vontuk be a modellekbe, elsősorban értelmezési nehézségek miatt.

[2] Számításainkat az SPSS 11.5 programcsomaggal végeztük.

A modellekben a kiválasztott referenciakategóriákat az egyes modelleket bemutató táblázatok után tüntettük fel.

A modellek illeszkedésének a jóságát *Hajdu* [2003] alapján az ún. pszeudó $R^2$ mutatóval mértük.[3] Ennek értéke azt mutatja meg, hogy a vizsgált modell a – magyarázó változó nélküli, csak tengelymetszettel paraméterezett – null modell illeszkedését hány százalékkal javítja a maradék nélkül magyarázó tökéletes vagy szaturált modellhez képest (a pszeudó $R^2$ értéke szélső esetekben 0 vagy 1 lehet).

## A MEGTAGADÁST MEGHATÁROZÓ TÉNYEZŐK A LOGISZTIKUS REGRESSZIÓS MODELLEK ALAPJÁN

A Függelék 1. táblájában szereplő három modellben a háztartásban élők *nem* szerinti megoszlása szignifikánsan járult hozzá a modell javításához, a megtagadási magatartásra gyakorolt hatás azonban csekély mértékű.

Az 1. modell alapján a háztartásban élők *korcsoport* szerinti megoszlást figyelembe véve a középkorúakat és időseket tömörítő háztartások megtagadási kockázata kisebb, mint az olyan háztartásoké, amelyekben csak idősek vannak. A többi korosztályt tartalmazó háztartások odds-arány multiplikátora nagyobb: főleg a középkorú tagokból álló háztartásoknak igazán nagy az esélye a megtagadásra, ők több mint másfélszer nagyobb eséllyel lesznek megtagadók, mint az idős háztartások. A fiatalabbakat tömörítő háztartások szintén rosszabbul válaszolnak, mint az idősek.

A 2. modell alapján az idősekre jellemző kedvezőbb, alacsonyabb megtagadási multiplikátor azért nő meg, mert az már az aktivitásnál és a családi állapotnál megmutatkozik, a változóknak a modellbe való bevonásával az időskorúakra jellemző kedvező megtagadási tulajdonságok egy része átvándorol ezen új változók multiplikátoraiba.

A 3. modell szerint – a lakás és lakókörnyék változók bevonása után – a legmagasabb megtagadási multiplikátorral rendelkező kategória a középkorú háztartásoké. Ezek odds-arány mutatója még mindig több mint másfélszer nagyobb a referenciakategóriánál, ráadásul úgy, hogy az összes többi modellbe bevont változó hatása itt már nem jelentkezik. A legutolsó modell szerint tehát ők azok, akik a kor szerinti megoszlást tekintve a legkevésbé nyitottak a válaszolásra.

Mindhárom modell szerint a háztartásban élők *iskolai végzettség* szerinti megoszlása alapján a megtagadási kockázat az alapfokú végzettségűeket tömörítő háztartások esetében a legmagasabb. A felsőfokú végzettségű egyéneket tartalmazó háztartások pedig „jobb" válaszolóknak bizonyulnak mind a középfokú, mind az alapfokú végzettségű háztartásoknál. Ez egybecseng a gyakorlati tapasztalatokkal, amelyek azt mutatják, hogy a felsőfokúak kevésbé bizalmatlanok az adatgyűjtéssel szemben, és gyakrabban válaszolnak, mint az alacsonyabb iskolai végzettséggel rendelkezők.

A *település* kiemelkedő fontosságú tényező a megtagadás szempontjából. Ez a változó döntően befolyásolja a különböző modellek illeszkedését. A megkérdezett háztartások

[3] Pszeudó $R^2 = 1 - \frac{GF\_CHI^2_{tárgy}}{GF\_CHI^2_{null}} = 1 - \frac{\ln L_{tárgy} - \ln L_{szaturált}}{\ln L_{null} - \ln L_{szaturált}}$, ahol a $GF\_CHI^2_{tárgy}$ az aktuális tárgymodell tökéletes modellhez való közelségét kifejező, a $GF\_CHI^2_{null}$ pedig a null modell és szaturált modell távolságát kifejező Goodness-of fit chi-négyzet statisztika értéke.

Budapesten a XXII. kerületet kivéve minden kerületben, és a megyei jogú városokban is nagyobb eséllyel válnak megtagadókká, mint a referencia kategóriának választott egyéb településen élő háztartások. Az első modell szerint a leginkább megtagadó kerületek a III., XIV., IV. és a VII. kerület, amelyek mindegyikében több mint tízszeres a háztartások megtagadási kockázata az egyéb településen élő háztartásokénál.

A 2. modellben, az aktívak és a vállalkozók hatásának modellbe való bevonása után, a településekhez tartozó megtagadási multiplikátorok csökkennek, de a sorrend a kerületek között e tekintetben kevéssé változik. A harmadik modellben, amelyben a lakásjellemzők hatásától tisztítottuk meg az odds-arány mutatókat, néhány kerületnek az esélye már nagyobb a válaszolásra, mint az egyéb településeken élő háztartásoknak. Ezeknél a kerületeknél kifejezetten a jobb lakáskörülményeknél fordul elő magasabb megtagadási kockázat, ennek egyik látható jele, hogy például a harmadik kerületben a megtagadás lakásjellemzőktől megtisztított multiplikátora jelentősen csökkent. Ez azt jelenti, hogy ebben a kerületben jellemzően nagyobb, komfortosabb lakások és jobb lakóövezetek fordulnak elő, amelyek növelik a megtagadás odds-arány szorzóját. A 3. modellben egyedül a XXIII. kerületben nőtt a megtagadás multiplikátora. Ez azt jelenti, hogy ebben a kerületben jellemzően fordultak elő olyan nagyságú és komfortú lakások, lakóövezetek, amelyek alacsonyabb megtagadási szorzót jelentenek, és amelyek kedvező hatásaitól ebben az esetben a XXIII. kerület megtagadási multiplikátorát megtisztítottuk.

Főleg a III., XIV., XV., VII., IV. és XIII. kerületekre jellemző, hogy döntően a lakásjellemzők (amelyek egyértelműen a jövedelmi helyzetre vezethetők vissza) hatottak a megtagadásra, azaz itt valószínűleg a magas jövedelműeket sikerült kiválasztani a mintába. A gazdagabb budai kerületeknél is természetesen megmutatkozik a lakásjellemzők hatása a megtagadásra, de végül az odds-arány mutató nem lesz nagy, mivel már az 1. modell szerint sem itt volt a megtagadás kockázata a legmagasabb. Ennek egyik lehetséges oka, hogy a megtagadáshoz személyes kontaktusra van szükség, ezekben a kerületekben viszont lehetséges, hogy az összeíró még csak be sem tudott jutni a lakásba, kapcsolatba sem tudott lépni az adott háztartással. Itt, ahogyan azt majd a későbbiekben kiderül, a nincs kapcsolat multiplikátora lesz kiemelkedően magas.

Az, hogy a háztartás hányadik *hullámban* szerepelt a mintában (vagyis, hogy mi a *kérdezés sorszáma*), nagyon erősen hat a megtagadásra. Legnagyobb megtagadási kockázata az 1. hullámban szereplő háztartásoknak van, ők mintegy 13-szor nagyobb eséllyel tagadják meg a választ, mint a 6. hullámban szereplő háztartások. A megtagadási kockázat csökken, ha egy háztartás több hullámban szerepelt a felvételben. A lakóövezet és lakásnagyság hatásának modellbe való bevonása után észrevehető, hogy az 1. hullámban szereplő háztartások megtagadási kockázata jelentősen lecsökken a többi hullámhoz képest. Ennek oka, hogy a lakóövezet jelleg és a lakásnagyság eltérő megoszlást mutat ebben a hullámban (tehát az új mintában) a többi hullámhoz (régi mintához) képest. A kereszttáblás megoszlások szerint az első és a többi hullám között jelentős eltérés, hogy a régi mintában családi házas lakóövezetek, az új mintában falusias övezetek fordulnak elő nagyobb arányban. A falusias övezethez képest a családi házas övezeteknek kisebb a megtagadási kockázata. Így az 1. hullámban szereplő háztartások alacsonyabb multiplikátora arra vezethető vissza, hogy az új mintában a lakóövezetek megoszlása terén eltolódás figyelhető meg a családi házas övezettől a falusias lakóövezet felé. Még a lakóövezetek modellbe való bevonása után is magasak a megtagadási odds-arány muta-

tók, főleg az új mintarészen, ami az összeírók szerepének fontosságára hívja fel a figyelmet: az ő munkájuk fontos szerepet tölt be a megtagadások alakulásában.

A lakásnagyság tekintetében is jelentős a különbség az első és a többi hullám között. Amíg az első hullámban a többi hullámhoz képest több a 2 szobás lakás konyhával, addig a 2-6. hullámban a 3 szobás lakások a gyakoribbak. A nagyobb lakásokban lakó háztartások pedig – ahogyan az később látható lesz – nagyobb odds-arány multiplikátorral rendelkeznek.

Kevésbé befolyásolja a megtagadást az *összeíró neme*, bár férfi összeíró esetén a megtagadás kockázata nagyobb. A *fiatal összeíróknál* a 2. modell szerint a legkisebb a megtagadási multiplikátor, az idős összeíróknál pedig itt a legnagyobb. Ha viszont a modellbe a lakókörnyék és lakással kapcsolatos változókat is bevonjuk, akkor látszik a multiplikátorok változásából, hogy az idősebb összeírók írnak össze a legnehezebb, magas megtagadási kockázatú lakóövezetekben és lakásokban, a fiatal összeírók alacsonyabb multiplikátora a kisebb megtagadási kockázatot jelentő lakóövezeteknek, lakásoknak köszönhetők. A modell szerint minél magasabb az *összeíró végzettsége*, annál kisebb az általa összeírt háztartás megtagadásának az odds-arány mutatója. Azok a háztartások, amelyeket tapasztalatlan, vagy ellenkezőleg, legalább két év gyakorlattal rendelkező összeíró keresett fel, rendelkeznek a legkisebb megtagadási eséllyel.

A háztartásban élők *családi állapot* szerinti összetételénél a referenciakategória a házasokat tartalmazó háztartás volt, ehhez képest azoknak a háztartásoknak, amelyekben van nőtlen vagy hajadon, de nincs házas, magasabb a megtagadási kockázatuk. Az özvegyeket tartalmazó háztartásoknak (amelyek nem tartalmaznak már sem házasokat, sem nőtleneket/hajadonokat) a referenciakategóriánál kisebb a megtagadási kockázatuk, és ugyancsak kevésbé megtagadók a csak elváltakat tartalmazó háztartások. A 3. modell azt mutatja, hogy a házasokon kívül minden más típusú háztartás odds-arány mutatója alacsonyabb.

A felállított modellek szerint, a háztartásban nevelt kiskorúak száma alapján látható, hogy a *három vagy több kiskorút* nevelő háztartások rendelkeznek a legkisebb megtagadási kockázatattal. A kiskorút nevelő háztartások megtagadási esélye pedig lényegesen kisebb, mint a kiskorút nem nevelő háztartásoké. A 3. modellben az odds-arány mutatók lakásjellemzőktől való megtisztítása során az egy kiskorút nevelő háztartások multiplikátora csökkent a kiskorút nem nevelő háztartásokhoz képest, a két kiskorút nevelő háztartások szorzója nem nagyon változott, míg a három vagy több kiskorút nevelők odds-arány mutatója nőtt. Ez azt jelenti, hogy a kevesebb kiskorút nevelő háztartások nagyobb része él olyan nagyságú és komfortos lakásban és lakóövezetben, amelyben nagyobb a megtagadás esélye, míg a több gyermeket nevelők inkább rosszabb körülmények között élnek, amelyhez kisebb megtagadási odds-arány mutató járul.

A 2. modell alapján az *aktív* háztartásoknak az inaktívaknál nagyobb a megtagadási kockázata. A *vállalkozót* tartalmazó háztartások megtagadási kockázata pedig nagyobb azokhoz a háztartásokhoz képest, amelyekben nincs vállalkozó, és a vállalkozók számának növekedésével pedig tovább nő a multiplikátorok értéke, vagyis a megtagadás kockázata.

A *lakás nagyságát* a harmadik modellben vontuk be. Ez az ismérv a modellben szereplő tizenhat változó közül második legerősebbként volt befolyással a modell illeszkedésére. A lakásnagyság és a megtagadási kockázat közti kapcsolat szembetűnő. Az 1 vagy 2 szobás konyha nélküli lakásokban élő háztartások megtagadási kockázata kisebb, mint a 2 szoba konyhás lakásokban élőké. A nagyobb lakások nagyobb megtagadási koc-

kázattal rendelkeznek, ami a jövedelem és a megtagadás kapcsolatát mutatja a referencia kategóriához képest. Ugyancsak nagyobb a megtagadási multiplikátora az 1 szoba konyhás lakásoknak, ami azért lehetséges, mert sok aktív, fiatal házaspár vagy egyedül élő nőtlen/hajadon ilyen garzonlakásban él.

A *lakás komfortossága* szintén hat a megtagadási kockázatra. Az összkomfortos lakásokhoz képest a komfortosakban élő háztartások odds-arány mutatója kedvezőbb, de a romló körülmények, a félkomfortos, komfort nélküli, valamit szükség és egyéb lakásokban élő háztartások nagyobb eséllyel megtagadáshoz vezetnek.

A bevont magyarázó változók közül a legjobban a *lakóövezet jellege* határozza meg a modell illeszkedését, ez a változó hat legerősebben a megtagadásra. A referenciakategóriának választott falusias lakóövezetnél a családi házas, valamint a csoportos beépítésű külterületi övezet, és az üdülő-, ipari-, üzemi területek, valamint a szociális szempontból nem megfelelő, egyéb övezetek multiplikátora kisebb. A magányos beépítésű külterületi övezetek megtagadási kockázata viszont több mint háromszorosa a falusias övezetnek. A városias övezetek ennél is jobban, a falusias övezetben élőkhöz képest majdnem hatszor akkora eséllyel tagadják meg a választ. A villanegyed vagy villanegyed jellegű társasházi lakónegyedek megtagadási kockázata igen magas, multiplikátoruk értéke közel 7. Meglepő, hogy a leginkább megtagadó háztartások a lakótelepiek, ahol több mint nyolcszoros az odds arány mutató értéke egy falusias övezethez képest.

Az odds-arány mutatók értéke következhet abból is, hogy a megtagadáshoz személyes kapcsolat létesítésére van szükség, ezért abban az esetben, ha az összeírónak senki nem nyit ajtót, pedig vannak otthon, az a nemválaszolásnak a nincs kapcsolat kategóriáját jelenti. Érdemes ezért a megtagadásra kapott modell eredményeit összevetni a nincs kapcsolat modelljeivel is.

Az első modell illeszkedése meglehetősen gyenge, a pszeudó $R^2$ mutató 0,128, míg a második modellé 0,132. Az első modell illeszkedése nem mondható jónak, és az aktivitás, vállalkozói foglalkozás, valamint a háztartásban élők családi állapota és a háztartásban nevelt kiskorúak számának változói sem javítják a modellt jelentős mértékben. A harmadik, lakásjellemzőkkel bővített, végső modell pszeudó $R^2$ értéke azonban már 0,278-es, jó illeszkedést jelez. Azt mondhatjuk, hogy az így kialakított 3. modell adekvát, megfelelően eltávolodik a null modelltől. Látható, hogy ez csakis a legutoljára bevont változóknak köszönhető, amelyek egyértelműen a háztartás jövedelmi helyzetére utalnak.

Összefoglalva a fentieket, felépített modellek alapján a megtagadást leginkább meghatározó tényezők a lakóövezet jellege, a lakás nagysága, a kérdezés sorszáma, a település jellege, valamint ötödikként az összeíró tapasztalata, és csak ezután következnek a lakás komfortossága, majd a háztartás demográfiai és egyéb társadalmi jellemzői.

## A NINCS KAPCSOLATOT MEGHATÁROZÓ TÉNYEZŐK A LOGISZTIKUS REGRESSZIÓS MODELLEK ALAPJÁN

A nincs kapcsolat logisztikus modelljei (lásd Függelék 2. tábla) hasonlítanak a megtagadás modelljeihez. A három modell illeszkedése ebben az esetben kicsit rosszabb, a végső modell pszeudó $R^2$ mutató értéke 0,242. Ezeknél a modelleknél egyes esetekben kategóriaösszevonásokra volt szükség, hogy elkerüljük a túl alacsony cellánkénti elemszámokat.

A modellek alapján a háztartásban élők *nem* szerinti megoszlása alapján elmondható, hogy a nincs kapcsolat odds-arány mutatója a csak férfiakból álló háztartások esetében a legmagasabb.

A háztartásban élők *kor* szerinti megoszlásában a legkisebb a fiatal háztartások kockázata az idős háztartásokhoz viszonyítva, legnagyobb a középkorúaké, hasonlóan a megtagadás modelljeinél tapasztaltakhoz. A lakás- és lakóövezet változók bevonása után a harmadik modell szerint a fiatalok, valamint a legidősebbek a legjobban elérhetők. A 3. modellben a kor szerinti megoszlást tekintve a többi háztartástípus a multiplikátort tekintve már nem különbözik egymástól.

A háztartás *iskolai végzettség* szerinti megoszlásában nem látható nagy különbség a nincs kapcsolat tekintetében. Legnagyobb eséllyel a középfokú, legkisebb eséllyel a felsőfokú végzettségűekkel létesíthető kapcsolat.

Az, hogy a megkérdezett háztartás milyen *településen* él, erős kapcsolatban áll a nincs kapcsolat jelenségével. Ebből a szempontból leginkább a budapesti VIII., III., XII., V., XIV., XV., IV., X., VII., I. vagy XIII. kerületben élő háztartások a legkockázatosabbak, - a felsorolt sorrendben-, ezeknél nagyobb eséllyel járt kudarccal a kapcsolat létesítésére tett kísérlet, mint az egyéb településeken élő háztartásoknál. A modell szerint a VIII. kerületben az egyéb településeken élőkhöz képest több mint hétszeres annak az esélye, hogy nem lehet egy háztartással kapcsolatot létesíteni. A többi budapesti kerületekben élő háztartás az egyéb településeken élőknél kisebb kockázattal esik a nincs kapcsolat kategóriába.

A megyei jogú városokban élő háztartások esetében is kisebb a nincs kapcsolat multiplikátora az egyéb településen élő háztartásokénál. A 2. modellben nem változnak jelentősen az odds-arány mutatók, a budapesti kerületek nagy részében kissé csökkennek. A harmadik modellben a VIII., III., XII., XV., XIV., XVII., XXIII., IV. kerület az, ahol még a lakás és lakóövezetek változóinak modellbe való bevonása után is nagyobb a kapcsolat létesítés kudarcának az esélye, mint az egyéb településeken. A VIII. kerületben, például, leszámítva azt a hatást is, hogy itt több a rosszabb minőségű lakóövezetben élő inaktív, kevesebb a vállalkozót tartalmazó, sok kiskorút nevelő háztartás, négyszeres a nincs kapcsolat odds-arány mutatója az egyéb településeken élőkhöz képest.

A *kérdezés sorszámának* vizsgálata esetében a megtagadáshoz hasonló mintázathoz jutunk. A kérdezés sorszáma még a legbővebb harmadik modellben is harmadik legerősebbként gyakorol hatást az eredményváltozóra. Minél több hullámban szerepel egy háztartás a felvételben, annál nagyobb a vele való kapcsolatbalépés sikerének esélye. Nyilvánvalóan az a körülmény jut itt érvényre, hogy aki egyszer már részt vett a felvételben, az a következő alkalommal nagyobb eséllyel fog abban részt venni. A második modellben szinte alig változnak az odds-arány mutatók. A harmadik modell estében viszont ugyanazt tapasztaljuk, mint a megtagadásnál: a régi és az új minta eltérő jellegzetességeket mutat, az új mintában (az első hullámban) szereplő háztartások összességében jobban elérhetők bármelyik későbbi hullámban részt vevő háztartásnál, ha a lakás és lakókörnyék változók hatását kontroll alatt tartjuk. Mint már említettük ennek hátterében az áll, hogy a két mintarész más megoszlást mutat lakóövezetek szerint: a régi mintában (a 2–6. hullámban) nagyobb arányban fordul elő a családi házas lakóövezet, az új mintában pedig a falusias övezet. A falusias övezetben lakó háztartásokkal a családi házas övezetekben élőkkel összevetve nagyobb eséllyel létesíthető kapcsolat.

A lakásnagyság tekintetében is jelentős a különbség az első és a többi hullám között: míg az első hullámban inkább a 2 szoba konyhás lakás a gyakoribb, addig a 2–6. hullámban a 3 szobás lakásokból van több. A lakásnagyság növekedésével pedig a nincs kapcsolat kockázata is nő.

Az *összeírók jellemzői* szintén befolyásolják a háztartásokkal való sikeres kapcsolatbalépés esélyét. A férfi összeírók által megkérdezett háztartások esetében magasabb a nincs kapcsolat multiplikátora, még a lakásjellemzők bevonása után is (tehát leszámítva azt, hogy ezek az összeírók főleg olyan lakóövezetben, lakásokban írnak össze, ahol amúgy is kisebb esélye van a háztartásokkal való kontaktusnak). Idős összeírók esetében a nincs kapcsolat odds-arány mutatója kisebb, a középkorú összeírók esetében pedig nagyobb, mint a referencia-kategóriának választott fiatal összeírók esetében. Az összeírók végzettségének növekedésével a nincs kapcsolat multiplikátora megnő. A megtagadásnál megfigyeltekkel éppen ellentétesen, az egy éves tapasztalattal rendelkező összeíróknál a legalacsonyabb a nincs kapcsolat odds-arány mutatója, míg a tapasztalatlan összeírók, valamint a két vagy több év tapasztalattal rendelkező összeírók nagyobb eséllyel kódolnak nincs-kapcsolatot. Ez lehet azért is, mert a legtapasztaltabbak és a kezdők nem olyan rámenősek, nem próbálnak meg olyan intenzíven személyes kapcsolatot építeni, mint a legfeljebb egy éves tapasztalattal rendelkező összeírók.

A *családi állapot* szerinti összetétel esetében az özvegyeket tartalmazó háztartások esetében a legkönnyebb kapcsolatot kialakítani, másfelől legnehezebb a nőtleneket/hajadonokat tartalmazó háztartásokkal. Ha nevelnek a háztartásban *kiskorút*, akkor szintén kisebb a nincs kapcsolat multiplikátora, mint a kiskorút nem nevelő háztartásoknál.

*Az aktív* háztartások növelik a nincs kapcsolat odds-arány mutatójának értékét, és a *vállalkozók számának* növekedésével is nő a multiplikátor értéke, vagyis a kapcsolat létesítés kudarcának esélye.

A modell javításához a *lakás nagysága* szignifikánsan járul hozzá, a legbővebb modellben ez a sorban a második olyan változó, amely a legnagyobb mértékben javítja a modell illeszkedését.

A nagyobb (4, 5, 6 vagy több szobás) lakásokban lakó háztartásokkal való kapcsolatlétesítés sikertelenségének esélye nagyobb, mint a 2 szoba konyhás lakásokban élőké. A 3 szoba konyhás lakásokban élőkkel viszont könnyebbnek bizonyult a kapcsolat létesítése, a kisebb lakások multiplikátora pedig nagyjából azonos értéket vesz fel.

A *lakás komfortosságának* vizsgálatánál látszik, hogy a komfortos, félkomfortos és összkomfortos lakásokban lakó háztartások odds-arány mutatója egy körül ingadozik, míg a komfort nélküli lakásokban élő háztartásoknál a nincs kapcsolat kockázata kisebb. A szükség és egyéb lakásokban élő háztartásokban fordul elő a legnagyobb eséllyel a kapcsolatlétesítés kudarca.

A *lakóövezet jellege* az egyik legfontosabb, a modell illeszkedését a leginkább befolyásoló tényező. A családi házas övezetet kivéve minden más lakóövezetben egynél magasabb a nincs kapcsolat multiplikátora. Kimagasló, a falusias jellegű övezetben lakó háztartásokhoz képest tizenegyszeres a kapcsolat hiányának odds-arány szorzója a villanegyed és villanegyed jellegű társasházi lakónegyedekben lakó háztartásokban. Látható, hogy a megtagadás helyett a nincs kapcsolat a jellemzőbb meghiúsulási forma ebben a lakóövezetben, itt a tipikus tehát az, hogy az összeírók még csak kapcsolatba sem tudnak lépni a háztartásokkal.

Kiemelkedően magas még mindig a lakótelepek, valamint a városias lakóövezetek meghiúsulási kockázata a falusias jellegű lakóövetekhez képest (7,6 és 6,7-szeres multiplikátorok). Az üdülőterületek, ipari, üzemi területek, szociális szempontból nem megfelelő és egyéb területeken megkérdezett háztartások odds-arány mutatója szintén nagyon magas, ezek esetében négyszer nagyobb a kapcsolat hiányának multiplikátora. A magányos ill. a csoportos beépítésű külterületi övezetekben lakó háztartásoknál is magasabb a nincs kapcsolat kockázata, mint a falusias lakóövezetben élők esetében.

A modellbe bevont változók közül a következő öt járult hozzá döntően a modell illeszkedésének javításához: lakóövezet jellege, lakás nagysága, összeíró tapasztalata, kérdezés sorszáma, lakás komfortossága, település.

*

A nemválaszolás két típusára felállított modellek hozzásegítettek a munkaerőfelvételben tapasztalt nemválaszolás jelenségének megismeréséhez, megértéséhez. Rámutattak arra, mely változók vannak nagyobb hatással a megtagadásra és a nincskapcsolat jelenségére. A lakóövezet és lakásjellemzőkön túl ide tartoztak a kérdezés sorszáma, valamint az összeírók tulajdonságai, s csak ezután következtek a háztartások egyéb jellemzői. Bár a nemválaszolás hatása nem küszöbölhető ki teljesen, egyes módszerekkel csökkenthető, korrigálható. A legszerencsésebb azonban az, ha ilyen korrekciókra a lehető legkisebb mértékben kerül sor, és az adatokat közvetlenül az adatszolgáltatóktól sikerül beszerezni. Az elállított adatok minőségét nagyban meghatározza az adatgyűjtési folyamat minősége, ami rámutat az összeírók munkájának óriási szerepére: már az adatgyűjtés fázisa döntő hatással lehet a KSH által előállított statisztikák minőségére.

Természetesen a nemválaszolással kapcsolatos elemzések tovább fejleszthetők, akár egyéb jellemzők vizsgálatával, akár más módszerek alkalmazásával. Az ilyen típusú elemzések hozzásegíthetnek a nemválaszolási jelenségek előrejelzéséhez, amelyre így tudatosabban fel lehetne készülni. Szintén alapul szolgálhatnának a nemválaszolás okozta torzítás becsléséhez, végeredményben pedig a kapott adatok minőségének javulásához.

## FÜGGELÉK

1. tábla

*A megtagadás logisztikus modelljei*

| Változó megnevezése | 1. modell | | | 2. modell | | | 3. modell | | |
|---|---|---|---|---|---|---|---|---|---|
| | Exp($b_i$) | Szignifikancia | Stepwise algoritmus estén a változó bevonásának sorrendje | Exp($b_i$) | Szignifikancia | Stepwise algoritmus estén a változó bevonásának sorrendje | Exp($b_i$) | Szignifikancia | Stepwise algoritmus estén a változó bevonásának sorrendje |
| Háztartásban élők nem szerinti megoszlása | | * | 9 | | * | 11 | | * | 13 |
| Csak férfiakból álló háztartás | 0,987 | *** | | 1,162 | * | | 1,068 | * | |
| Férfi többségű háztartás | 0,954 | * | | 0,988 | | | 0,989 | | |
| Női többségű háztartás | 1,002 | | | 1,083 | * | | 1,108 | * | |
| Csak nőkből álló háztartás | 1,104 | * | | 1,163 | * | | 1,188 | * | |

*(A tábla folytatása a következő oldalon.)*

*(Folytatás.)*

| Változó megnevezése | 1. modell | | | 2. modell | | | 3. modell | | |
|---|---|---|---|---|---|---|---|---|---|
| | Exp($b_i$) | Szignifikancia | Stepwise algoritmus estén a változó bevonásának sorrendje | Exp($b_i$) | Szignifikancia | Stepwise algoritmus estén a változó bevonásának sorrendje | Exp($b_i$) | Szignifikancia | Stepwise algoritmus estén a változó bevonásának sorrendje |
| Háztartásban élők korcsoportja | | * | 4 | | * | 4 | | * | 7 |
| Fiatalok | 1,363 | * | | 1,200 | * | | 1,290 | * | |
| Fiatalok és középkorúak | 1,638 | * | | 1,271 | * | | 1,276 | * | |
| Középkorúak | 1,782 | * | | 1,453 | * | | 1,629 | * | |
| Középkorúak és idősek | 0,694 | * | | 0,519 | * | | 0,605 | * | |
| Háztartásban élők iskolai végzettség szerinti megoszlása | | * | 5 | | * | 6 | | * | 12 |
| Középfokú végzettségű háztartás | 0,953 | * | | 0,881 | * | | 0,890 | * | |
| Felsőfokú végzettségű háztartás | 0,811 | * | | 0,788 | * | | 0,830 | * | |
| Település | | * | 1 | | * | 2 | | * | 4 |
| Budapest 1. kerület | 2,696 | * | | 2,539 | * | | 0,710 | * | |
| Budapest 2. kerület | 4,146 | * | | 3,868 | * | | 1,081 | * | |
| Budapest 3. kerület | 17,114 | * | | 15,938 | * | | 4,892 | * | |
| Budapest 4. kerület | 12,838 | * | | 12,523 | * | | 3,186 | * | |
| Budapest 5. kerület | 3,092 | * | | 2,871 | * | | 0,655 | * | |
| Budapest 6. kerület | 2,735 | * | | 2,523 | * | | 0,811 | * | |
| Budapest 7. kerület | 12,585 | * | | 11,999 | * | | 3,311 | * | |
| Budapest 8. kerület | 4,787 | * | | 4,728 | * | | 2,169 | * | |
| Budapest 9. kerület | 2,110 | * | | 2,013 | * | | 0,461 | * | |
| Budapest 10. kerület | 6,424 | * | | 6,282 | * | | 1,983 | * | |
| Budapest 11. kerület | 4,662 | * | | 4,583 | * | | 1,811 | * | |
| Budapest 12. kerület | 3,834 | * | | 3,733 | * | | 1,560 | * | |
| Budapest 13. kerület | 9,313 | * | | 9,061 | * | | 3,052 | * | |
| Budapest 14. kerület | 14,211 | * | | 13,912 | * | | 4,242 | * | |
| Budapest 15. kerület | 8,050 | * | | 7,754 | * | | 3,577 | * | |
| Budapest 16. kerület | 2,680 | * | | 2,619 | * | | 1,943 | * | |
| Budapest 17. kerület | 4,094 | * | | 3,873 | * | | 2,405 | * | |
| Budapest 18. kerület | 1,226 | * | | 1,234 | * | | 0,642 | * | |
| Budapest 19. kerület | 3,214 | * | | 3,053 | * | | 1,912 | * | |
| Budapest 20. kerület | 1,429 | * | | 1,353 | * | | 1,106 | * | |
| Budapest 21. kerület | 2,212 | * | | 2,096 | * | | 1,096 | ** | |
| Budapest 22. kerület | 0,421 | * | | 0,412 | * | | 0,262 | * | |
| Budapest 23. kerület | 1,180 | ** | | 1,145 | *** | | 1,716 | * | |
| Megyei jogú város | 2,099 | * | | 2,040 | * | | 0,894 | * | |
| Kérdezés sorszáma | | * | 2 | | * | 1 | | * | 3 |
| 1. | 13,251 | * | | 13,358 | * | | 1,115 | * | |
| 2. | 10,842 | * | | 10,837 | * | | 6,794 | * | |
| 3. | 4,130 | * | | 4,123 | * | | 3,359 | * | |
| 4. | 1,975 | * | | 1,974 | * | | 1,711 | * | |
| 5. | 1,468 | * | | 1,461 | * | | 1,413 | * | |
| Összeíró neme | | * | 6 | | * | 9 | | * | 15 |
| Férfi összeíró | 1,168 | * | | 1,182 | * | | 1,116 | * | |
| Összeíró korcsoportja | | * | 8 | | * | 13 | | * | 16 |
| Középkorú összeíró | 0,996 | | | 1,001 | | | 0,876 | * | |
| Idős összeíró | 1,178 | * | | 1,175 | * | | 0,813 | * | |
| Összeíró végzettsége | | * | 7 | | * | 12 | | * | 11 |
| Érettségivel rendelkező összeíró | 0,866 | * | | 0,859 | * | | 0,735 | * | |
| Felsőfokú végzettségű összeíró | 0,978 | | | 0,965 | ** | | 0,892 | * | |

*(A tábla folytatása a következő oldalon.)*

*(Folytatás.)*

| Változó megnevezése | 1. modell | | | 2. modell | | | 3. modell | | |
|---|---|---|---|---|---|---|---|---|---|
| | Exp($b_i$) | Szignifikancia | Stepwise algoritmus estén a változó bevonásának sorrendje | Exp($b_i$) | Szignifikancia | Stepwise algoritmus estén a változó bevonásának sorrendje | Exp($b_i$) | Szignifikancia | Stepwise algoritmus estén a változó bevonásának sorrendje |
| Összeíró tapasztalata | | * | 3 | | * | 3 | | * | 5 |
| Tapasztalatlan összeíró | 0,439 | * | | 0,436 | * | | 0,427 | * | |
| 2 vagy több éves tapasztalattal rendelkező összeíró | 0,724 | * | | 0,724 | * | | 0,804 | * | |
| Családi állapot szerinti összetétel | | | | | * | 5 | | * | 8 |
| Vannak nőtlenek, hajadonok a háztartásban, de nincsenek házasok | | | | 1,101 | * | | 0,93 | * | |
| Vannak özvegyek a háztartásban, de nincsenek nőtlenek-hajadonok és házasok | | | | 0,834 | * | | 0,672 | * | |
| Csak elváltak vannak a háztartásban | | | | 0,656 | * | | 0,598 | * | |
| Háztartásban nevelt kiskorúak száma | | | | | * | 7 | | * | 9 |
| 1 | | | | 0,743 | * | | 0,724 | * | |
| 2 | | | | 0,749 | * | | 0,752 | * | |
| 3 és több | | | | 0,702 | * | | 0,741 | * | |
| Háztartásban élők aktivitása | | | | | * | 8 | | * | 14 |
| Aktív háztartás | | | | 1,211 | | | 1,101 | | |
| Vállalkozók száma a háztartásban | | | | | * | 10 | | * | 10 |
| 1 | | | | 1,150 | * | | 1,252 | * | |
| 2 és több | | | | 1,292 | * | | 1,393 | * | |
| Lakás nagysága | | | | | | | | * | 2 |
| 1 szobás konyha nélkül | | | | | | | 0,260 | * | |
| 1 szobás konyhával | | | | | | | 1,610 | * | |
| 2 szobás konyha nélkül | | | | | | | 0,853 | * | |
| 3 szobás konyhával | | | | | | | 1,179 | * | |
| 4 szobás konyhával | | | | | | | 1,891 | * | |
| 5 szobás konyhával | | | | | | | 1,910 | * | |
| 6 és több szobás konyhával | | | | | | | 1,071 | * | |
| Lakás komfortossága | | | | | | | | * | 6 |
| Komfortos | | | | | | | 0,636 | * | |
| Félkomfortos | | | | | | | 1,024 | *** | |
| Komfort nélküli | | | | | | | 1,045 | ** | |
| Szükség és egyéb lakás / üdülőhelyiség | | | | | | | 1,400 | * | |
| Lakóövezet jellege | | | | | | | | * | 1 |
| Családi házas | | | | | | | 0,613 | * | |
| Lakótelep | | | | | | | 8,361 | * | |
| Városias | | | | | | | 5,696 | * | |
| Villanegyed vagy villanegyed jellegű társasházi lakónegyed | | | | | | | 6,794 | * | |
| Magányos beépítésű külterületi övezet | | | | | | | 3,797 | * | |

*(A tábla folytatása a következő oldalon.)*

*(Folytatás.)*

| Változó megnevezése | 1. modell | | | 2. modell | | | 3. modell | | |
|---|---|---|---|---|---|---|---|---|---|
| | Exp($b_i$) | Szignifikancia | Stepwise algoritmus estén a változó bevonásának sorrendje | Exp($b_i$) | Szignifikancia | Stepwise algoritmus estén a változó bevonásának sorrendje | Exp($b_i$) | Szignifikancia | Stepwise algoritmus estén a változó bevonásának sorrendje |
| Csoportos beépítésű külterületi övezet | | | | | | | 0,752 | * | |
| Üdülőterület, ipari, üzemi terület, szociális szempontból nem megfelelő, egyéb övezet | | | | | | | 0,478 | * | |
| Konstans | 0,006 | * | | 0,007 | * | | 0,138 | * | |
| Pszeudó $R^2$ | 0,128 | | | 0,132 | | | 0,278 | | |

*Megjegyzés*. A referenciakategóriák rendre a következők: azonos számú férfit és nőt tartalmazó háztartás; időseket tartalmazó háztartás; alapfokú végzettségűeket tartalmazó háztartás; egyéb településen (nem megyei jogú városban és Budapesten) élő háztartás; 6. kérdezési hullám; női összeíró; fiatal összeíró; érettségivel nem rendelkező összeíró; 1 éves tapasztalattal rendelkező összeíró; házasokat tartalmazó háztartás; kiskorút nem nevelő háztartás; inaktív tagokat tartalmazó háztartás; vállalkozó személyt nem tartalmazó háztartás; 2 szoba, konyhás lakás; összkomfortos lakás; falusias jellegű lakóövezet.

2. tábla

*A nincs kapcsolat logisztikus modelljei*

| Változó megnevezése | 1. modell | | | 2. modell | | | 3. modell | | |
|---|---|---|---|---|---|---|---|---|---|
| | Exp($b_i$) | Szignifikancia | Stepwise algoritmus estén a változó bevonásának sorrendje | Exp($b_i$) | Szignifikancia | Stepwise algoritmus estén a változó bevonásának sorrendje | Exp($b_i$) | Szignifikancia | Stepwise algoritmus estén a változó bevonásának sorrendje |
| Háztartásban élők nem szerinti megoszlása | | * | 4 | | * | 3 | | * | 7 |
| Csak férfiakból álló háztartás | 1,740 | * | | 1,616 | * | | 1,596 | * | |
| Férfi többségű háztartás | 0,971 | * | | 1,024 | * | | 1,105 | * | |
| Női többségű háztartás | 1,187 | * | | 1,264 | * | | 1,331 | * | |
| Csak nőkből álló háztartás | 1,375 | * | | 1,423 | * | | 1,302 | * | |
| Háztartásban élők korcsoportja | | * | 3 | | * | 7 | | * | 10 |
| Fiatalok | 0,952 | * | | 0,704 | * | | 0,795 | * | |
| Fiatalok és középkorúak | 1,593 | * | | 1,240 | * | | 1,310 | * | |
| Középkorúak | 1,824 | * | | 1,347 | * | | 1,381 | * | |
| Középkorúak és idősek | 1,404 | * | | 1,128 | * | | 1,255 | * | |
| Háztartásban élők iskolai végzettség szerinti megoszlása | | * | 8 | | * | 11 | | * | 16 |
| Középfokú végzettségű háztartás | 0,935 | * | | 0,901 | * | | 0,951 | * | |
| Felsőfokú végzettségű háztartás | 1,163 | * | | 1,067 | * | | 1,052 | * | |
| Település | | * | 2 | | * | 2 | | * | 5 |
| Budapest 1. kerület | 1,101 | * | | 1,021 | | | 0,568 | * | |
| Budapest 2. kerület | 0,603 | * | | 0,577 | * | | 0,385 | * | |
| Budapest 3. kerület | 3,764 | * | | 3,567 | * | | 3,275 | * | |

*(A tábla folytatása a következő oldalon.)*

*(Folytatás.)*

| Változó megnevezése | 1. modell | | | 2. modell | | | 3. modell | | |
|---|---|---|---|---|---|---|---|---|---|
| | Exp($b_i$) | Szignifikancia | Stepwise algoritmus estén a változó bevonásának sorrendje | Exp($b_i$) | Szignifikancia | Stepwise algoritmus estén a változó bevonásának sorrendje | Exp($b_i$) | Szignifikancia | Stepwise algoritmus estén a változó bevonásának sorrendje |
| Budapest 4. kerület | 1,220 | * | | 1,267 | * | | 1,080 | * | |
| Budapest 5. kerület | 1,714 | * | | 1,600 | * | | 0,956 | | |
| Budapest 6. kerület | 0,635 | * | | 0,604 | * | | 0,330 | * | |
| Budapest 7. kerület | 1,154 | * | | 1,116 | * | | 0,831 | * | |
| Budapest 8. kerület | 7,464 | * | | 7,494 | * | | 4,072 | * | |
| Budapest 9. kerület | 0,274 | * | | 0,267 | * | | 0,169 | * | |
| Budapest 10. kerület | 1,189 | * | | 1,199 | * | | 0,964 | | |
| Budapest 11. kerület | 0,688 | * | | 0,688 | * | | 0,693 | * | |
| Budapest 12. kerület | 3,608 | * | | 3,579 | * | | 3,110 | * | |
| Budapest 13. kerület | 1,101 | * | | 1,101 | * | | 0,927 | * | |
| Budapest 14. kerület | 1,609 | * | | 1,611 | * | | 1,370 | * | |
| Budapest 15. kerület | 1,591 | * | | 1,539 | * | | 1,649 | * | |
| Budapest 16. kerület | 0,745 | * | | 0,771 | * | | 0,851 | * | |
| Budapest 17. és 23. kerület | 0,769 | * | | 0,761 | * | | 1,228 | * | |
| Budapest 18. kerület | 0,103 | * | | 0,108 | * | | 0,137 | * | |
| Budapest 19. kerület | 0,438 | * | | 0,423 | * | | 0,740 | * | |
| Budapest 20. kerület | 0,729 | * | | 0,717 | * | | 0,918 | *** | |
| Budapest 21. kerület | 0,519 | * | | 0,511 | * | | 0,630 | * | |
| Budapest 22. kerület | 0,701 | * | | 0,710 | * | | 0,844 | * | |
| Megyei jogú város | 0,403 | * | | 0,418 | * | | 0,874 | * | |
| Kérdezés sorszáma | | * | 1 | | * | 1 | | * | 3 |
| 1. | 9,272 | * | | 9,385 | * | | 0,879 | * | |
| 2. | 4,210 | * | | 4,208 | * | | 2,973 | * | |
| 3. | 2,219 | * | | 2,222 | * | | 1,856 | * | |
| 4. | 1,501 | * | | 1,502 | * | | 1,232 | * | |
| 5. | 1,186 | * | | 1,194 | * | | 1,095 | * | |
| Összeíró neme | | * | 6 | | * | 6 | | * | 12 |
| Férfi összeíró | 1,320 | * | | 1,338 | * | | 1,262 | * | |
| Összeíró korcsoportja | | * | 7 | | * | 9 | | * | 8 |
| Középkorú összeíró | 1,108 | * | | 1,109 | * | | 0,928 | * | |
| Idős összeíró | 0,692 | * | | 0,684 | * | | 0,486 | * | |
| Összeíró végzettsége | | * | 5 | | * | 5 | | * | 6 |
| Érettségivel rendelkező összeíró | 1,062 | * | | 1,059 | * | | 0,860 | * | |
| Felsőfokú végzettségű összeíró | 1,100 | * | | 1,092 | * | | 0,969 | ** | |
| Összeíró tapasztalata | | * | 9 | | * | 12 | | * | 13 |
| Tapasztalatlan összeíró | 1,162 | * | | 1,160 | * | | 1,184 | * | |
| 2 vagy több éves tapasztalattal rendelkező összeíró | 1,278 | * | | 1,277 | * | | 1,412 | * | |
| Családi állapot szerinti összetétel | | | | | * | 8 | | * | 14 |
| Vannak nőtlenek, hajadonok a háztartásban, de nincsenek házasok | | | | 1,195 | * | | 1,036 | * | |
| Vannak özvegyek a háztartásban, de nincsenek házasok és nőtlen-hajadonok | | | | 0,793 | * | | 0,770 | * | |
| Csak elváltak vannak a háztartásban | | | | 1,130 | * | | 0,920 | * | |

*(A tábla folytatása a következő oldalon.)*

*(Folytatás.)*

| Változó megnevezése | 1. modell | | | 2. modell | | | 3. modell | | |
|---|---|---|---|---|---|---|---|---|---|
| | Exp($b_i$) | Szignifikancia | Stepwise algoritmus estén a változó bevonásának sorrendje | Exp($b_i$) | Szignifikancia | Stepwise algoritmus estén a változó bevonásának sorrendje | Exp($b_i$) | Szignifikancia | Stepwise algoritmus estén a változó bevonásának sorrendje |
| Háztartásban nevelt kiskorúak száma | | | | | * | 13 | | * | 15 |
| 1 | | | | 0,902 | * | | 0,870 | * | |
| 2 | | | | 0,904 | * | | 0,925 | * | |
| 3 és több | | | | 0,714 | * | | 0,779 | * | |
| Háztartásban élők aktivitása | | | | | * | 4 | | * | 11 |
| Aktív háztartás | | | | 1,271 | * | | 1,229 | * | |
| Vállalkozók száma a háztartásban | | | | | * | 10 | | * | 9 |
| 1 | | | | 1,209 | * | | 1,260 | * | |
| 2 és több | | | | 1,578 | * | | 1,882 | * | |
| Lakás nagysága | | | | | | | | * | 2 |
| 1 szobás konyha nélkül | | | | | | | 0,899 | * | |
| 1 szobás konyhával | | | | | | | 1,033 | * | |
| 2 szobás konyha nélkül | | | | | | | 1,172 | * | |
| 3 szobás konyhával | | | | | | | 0,254 | * | |
| 4 szobás konyhával | | | | | | | 1,483 | * | |
| 5 szobás konyhával | | | | | | | 1,867 | * | |
| 6 és több szobás konyhával | | | | | | | 2,090 | * | |
| Lakás komfortossága | | | | | | | | * | 4 |
| Komfortos | | | | | | | 0,929 | * | |
| Félkomfortos | | | | | | | 1,086 | * | |
| Komfort nélküli | | | | | | | 0,502 | * | |
| Szükség és egyéb lakás / üdülőhelyiség | | | | | | | 1,423 | * | |
| Lakóövezet jellege | | | | | | | | * | 1 |
| Családi házas | | | | | | | 0,820 | * | |
| Lakótelep | | | | | | | 7,679 | * | |
| Városias | | | | | | | 6,709 | * | |
| Villanegyed vagy villanegyed jellegű társasházi lakónegyed | | | | | | | 11,242 | * | |
| Magányos beépítésű külterületi övezet | | | | | | | 2,698 | * | |
| Csoportos beépítésű külterületi övezet | | | | | | | 2,202 | * | |
| Üdülőterület, ipari, üzemi terület, szociális szempontból nem megfelelő, egyéb övezet | | | | | | | 4,009 | * | |
| Konstans | 0,009 | * | | 0,010 | * | | 0,085 | * | |
| Pszeudó $R^2$ | 0,098 | | | 0,102 | | | 0,242 | | |

*Megjegyzés*. A referenciakategóriák rendre a következők: azonos számú férfit és nőt tartalmazó háztartás; időseket tartalmazó háztartás; alapfokú végzettségűeket tartalmazó háztartás; egyéb településen (nem megyei jogú városban és Budapesten) élő háztartás; 6. kérdezési hullám; női összeíró; fiatal összeíró; érettségivel nem rendelkező összeíró; 1 éves tapasztalattal rendelkező összeíró; házasokat tartalmazó háztartás; kiskorút nem nevelő háztartás; inaktív tagokat tartalmazó háztartás; vállalkozó személyt nem tartalmazó háztartás; 2 szoba, konyhás lakás; összkomfortos lakás; falusias jellegű lakóövezet.

## IRODALOM

AGRESTI, A. [2002]: *Categorical data analysis.* John Willey & Sons. New Jersey.
BIEMER, P. P. – LYBERG, L. E. [2003]: *Introduction to survey quality.* John Willey & Sons. New Jersey.

COVAR, J. – RANCOURT, E. [2003]: Workshop on editing and imputation of survey data. Berlin.
EUROSTAT [2004]: *How to make a quality report?* Assesment of quality in statistics. Version 2. Luxembourg.
HAJDU O. [2003]: *Többváltozós statisztikai számítások.* Központi Statisztikai Hivatal. Budapest.
HAVASI É. [1997]: Válaszmegtagadó háztartások. *Statisztikai Szemle*. 75. évf. 10. sz., 831–843. old.
HAVASI É. – MARTON Á. [1998]: Nonresponse in the 1996 income survey (Supplement to the microcensus). In: *Nonresponse in survey research. Proceedings of the eighth international workshop on Household Survey Nonresponse 24–26 September 1997.* Mannheim.
HIDIROGLOU, M. A – DREW, J. D. – GRAY, G. B. [1993]: A Framework for measuring and reducing nonresponse in surveys. *Survey Methodology*. 19. évf. 1. sz. 81–94. old.
KAPITÁNY B. – SPÉDER ZS. [2004]: Szegénység és depriváció. Társadalomszerkezeti összefüggések nyomában. *Életünk forduló-pontjai. Műhelytanulmányok 4.* Budapest.
LAKATOS J. – MIHÁLYFFY L. [2003]: Az új népszámlálási módszerek hatása a munkaerő-felmérésre. *Statisztikai Szemle.* 81. évf. 12. sz. 1045–1053. old.
LAAKSONEN, S. [1992]: Handling household survey non-response data. Statistical Research Reports 13. The Finnish Statistical Society. Helsinki.
LAKSONEN, S. [1995]: Nonresponse – an essential indicatot for survey quality, Seppo Laaksonen. In: *International perspectives on nonresponse. Proceedings of the sixth international workshop of hosehold survey nonresponse 25-27 October 1995.* Statistics Finland. Helsinki.
LUNDSTRÖM, S. – SARNDAL, C. E. [2002]: *Estimation in the presence of nonresponse and frame imperfections*. Statistics Sweden. Stockholm.
MARTON Á. [1995]: Nonresponse in the Hungarian Household Surveys. In: *International perspectives on nonresponse. Proceedings of the Sixth International Workshop of Household Survey Nonresponse 25–27 October 1995.* Statistics Finland. Helsinki. 148–153. old.
MARTON Á. – VARGA A. [2000]: A KSH munkaerő-felmérésének néhány kérdése. KSH munkaanyag.
MIHÁLYFFY L. [2000]: Címregiszteren alapuló lakossági minták terve. *Statisztikai Szemle.* 78. évf. 10–11. sz. 873–892. old.
MIHÁLYFFY L. [2004]: A munkaerő-felmérés mintavételi terve. Központi Statisztikai Hivatal. Munkaanyag.
PLATEK, R. [1986]: *A survey practitioner's notion of nonresponse by, 1986-10-2, Promemorior Fran U/Stm, NR 26.* Statistiska Centralbyran. Stockholm.
SZÉKELYI M. – BARNA I. [2003]: *Túlélőkészlet az SPSS-hez. Többváltozós elemzési technikákról társadalomkutatók számára.* Typotex Kiadó. Budapest.

## SUMMARY

This study analyses the nonresponse features in the Labour Force Survey conducted by the Hungarian Central Statistical Office in the first quarter of 2003 using the data of Census and former LFS. After a short introduction into the theoretical framework of nonresponse, the paper deals with the three main types of the nonresponse: refusal, no-contacts and other nonresponse, and uses logistic regression to analyse how the different explanatory variables effect the refusal and no-contact mechanisms.

# A MINŐSÉG A HIVATALOS STATISZTIKÁBAN*

SZÉP KATALIN – VIGH JUDIT

A statisztika minőségével foglalkozó átfogó fejlesztőmunka már a nyolcvanas években megkezdődött és egyre több ország statisztikai hivatalában elterjedt. Időközben azonban a minőség fogalma változott. A nagy nemzetközi szervezetek (OECD, IMF) kialakították saját minőségi rendszerüket. Az Eurostat 2003-ra véglegesítette a minőséggel kapcsolatos alapdokumentumokat, megfogalmazta ajánlásait, és a minőségi követelmények néhány felvételnél már jogszabályi szinten is megjelentek. Jelen tanulmányban az Eurostat kutatásaira és ajánlásaira támaszkodva mutatjuk be a minőségi keretrendszer elemeit: a termékminőség dimenzióinak mérési lehetőségeit, a statisztikai munkafolyamatok minőségi megközelítését, majd a minőség-ellenőrzés, minőségirányítás módszereit. A Központi Statisztikai Hivatal statisztikai munkájában hagyományosan szem előtt tartja számos minőségi szempont érvényesítését. Stratégiai célkitűzései közé emelte a minőség fejlesztését. Az Eurostat számára rendszeresen készítünk minőségjelentéseket, de még előttünk áll a statisztikai minőségfejlesztés átfogó programjának kidolgozása.

TÁRGYSZÓ: A statisztika minősége. Termékminőség. Minőségirányítás.

A gazdaságilag fejlett és élen járó statisztikai rendszerrel rendelkező országokban a statisztikai hivatalok már jó húsz évvel ezelőtt közzétették a statisztikai termékeik elvárható minőségi jellemzőit, s azóta több ország is felsorakozott melléjük. A hivatalos statisztika minőségi kérdései rendszeresen szerepelnek a nemzetközi statisztikai konferenciákon. A statisztika minőségével foglalkozó legutóbbi konferencia 2004 májusában, Mainzban volt.

A Központi Statisztikai Hivatal munkájában szem előtt tartja számos minőségi szempont érvényesítését, a minőség fejlesztését stratégiai célkitűzései közé emelte, a Magyar Statisztikai Társaság konferencián tárgyalta a minőség kérdését, az érintett főosztályok pedig rendszeresen küldik az Eurostat számára egyes felvételek minőségjelentéseit, de mind ez ideig nem született még meg egy átfogó program. Mára megérett a helyzet a statisztika minőségével kapcsolatos eddigi munkák, az elért eredmények áttekintésére. A következőkben erre teszünk kísérletet.

A történeti áttekintés első részében (1-2. fejezet) a minőség fogalmának alakulását tárgyaljuk, majd az Európai Unió statisztikai hivatala, az Eurostat ehhez kapcsolódó tevékenységét ismertetjük. Írásunk 3. fejezetében – az Eurostat kutatásaira és ajánlásaira

* Köszönetet mondunk *Marton Ádámnak* a cikk megírásához nyújtott sokoldalú támogatásáért, értékes tanácsaiért.

támaszkodva – bemutatjuk a minőségi keretrendszer elemeit: a termékminőség dimenzióinak mérési lehetőségeit, a statisztikai munkafolyamatok minőségi megközelítését, majd a minőség-ellenőrzés, minőségirányítás módszereit. Végül a 4. fejezetben, a magyar hivatalos statisztika minőségi kérdéseiről szólunk.

## 1. A MINŐSÉG FOGALMÁNAK ALAKULÁSA A HIVATALOS STATISZTIKÁBAN

A minőség fogalma a statisztikában nem könnyen definiálható, mert időben változó, függ a felhasználó igényeitől, de a statisztikai termék jellegétől is. Kezdetben a statisztika minőségét a statisztikai adat pontosságával azonosították. Később ez a fogalom fokozatosan bővült újabb jellemzőkkel, mint tartalom, relevancia, időszerűség stb. Ugyanakkor egyes statisztikai szervezetek szakemberei a minőség értelmezését az egyes termékekről fokozatosan kiterjesztették a statisztikai folyamatokra, sőt egyes országokban a statisztikai hivatalok szervezetének irányítására is.

Előljáróban idézzük a hivatalos statisztika minőségének tudományos definícióját a nemzetközi szakirodalomban jól ismert enciklopédiából. *E. Elvers* és *B. Rosén* svéd statisztikusok írták az *Encyclopedia of Statistical Sciences* (1999) harmadik kötetében a hivatalos statisztika minőségének fogalmáról szóló részt. Megállapításuk szerint a minőségre leggyakrabban alkalmazott megközelítés a teljes körű minőség fogalmán alapul, melynek fő összetevői a következők.

– Egy termék minőségét elsősorban az határozza meg, hogy mennyiben elégíti ki a termék felhasználójának jelenlegi és várható igényeit.

– A minőségfogalomnak tükröznie kell a termék minden olyan tulajdonságát, mely alapján a felhasználó dönthet a termék minőségéről.

Az idézett *Encyclopedia* szerint a hivatalos statisztika minőségének fő elemei: a tartalom, a pontosság, az időszerűség, a koherencia és összehasonlíthatóság, valamint a hozzáférhetőség és átláthatóság. A szerzők a fejezet végén egy rövid történeti áttekintést is adnak az egyes statisztikai hivataloknak és nemzetközi szervezeteknek szerepéről a hivatalos statisztika minőségének fejlesztésében.

A kezdeteket az amerikai statisztikusok[1] felvételekkel kapcsolatos módszertani munkái jelölték ki. A Kanadai Statisztikai Hivatal 1978-ban, a Svéd Statisztikai Hivatal 1979-ben adta közre a statisztikai termékek minőségének bemutatásáról szóló útmutatóját. Az amerikai, a kanadai és a svéd statisztikai hivatalok a statisztikai adataik minőséggel kapcsolatban elért eredményei alapján az ENSZ Statisztikai Hivatala és az Európai Gazdasági Bizottság 1983-ban kiadott egy módszertani útmutatót (*Guidelines for Quality…* [1983]). Az útmutató tervezetének a KSH Nemzetközi kapcsolatok osztályán készült fordításából idézzük: „…a hivatalok a közzétett statisztikai adatok mellett jelentessék meg az adatforrásokat, a főbb fogalmakat, a vizsgálandó alapsokaság leírását, a felvétel tervét és módszertanát, a becslési módszereket. Ezen felül publikálják az általános hibabecslést és a főbb hibaforrásokat, a nemválaszolási arányokat és hibákat, a mintavételi hibákat, a

[1] Az Egyesült Államok szövetségi kormánya statisztikai intézményeiben, főként a Népszámlálási Hivatalban (Bureau of Census) folytak jelentős fejlesztő munkák.

mérési, feldolgozási és egyéb típusú hibákat, az összehasonlíthatóság problémáit, valamint a minőségre vonatkozó egyéb információk elérhetőségét".

A Kanadai Statisztikai Hivatal a minőségi mutatókról szóló első kiadványát és a minőségértékelést részletező útmutatóját 1985-ben adta ki. Az időközben felhalmozott tapasztalatoknak, a technológiai és irányítási területen alkalmazott eredményeknek köszönhetően azóta a negyedik, átdolgozott kiadást jelentette meg (*Statistics Canada Quality Guidelines* [2003]). Az említett kiadványokban a statisztikai információ minőségét általánosan úgy definiálják mint a felhasználásra való alkalmasságot (fitness for use), melynek hat alkotóeleme vagy más néven dimenziója van: relevancia, pontosság, időszerűség, hozzáférhetőség, értelmezhetőség és koherencia.

A *Journal of Official Statistics* című svéd folyóirat 2001. évi 1. száma döntően a statisztikai adatok minőségével kapcsolatos vitát tartalmazta (*Havasi–Marton* [2002]). A vitaindítóban *R. Platek* és *C.-E. Särndal*, a Kanadai Statisztikai Hivatal tanácsadói sorra vették a minőség különféle jellemzőit, de nem találtak egységes, általánosan elfogadott definíciót (*Platek–Särndal*, [2001a]).

A beérkezett hozzászólások közül csak egy-két véleményt idézünk. A vitában *Elvers* és *Nordberg* svéd statisztikusok azt hangsúlyozták, hogy a minőség megítélése a felhasználótól függ, de a pontosság megítélése a statisztikus nélkül lehetetlen (*Elvers–Nordberg* [2001]). A legjobb minőség elérése a munka tervezésével kezdődik, és a végrehajtással folytatódik. Ez utóbbihoz kapcsolódik a teljes körű minőségirányítás (Total Quality Management – TQM) vagy a folyamatos minőségfejlesztés (Continuous Quality Improvement – CQI) koncepciója, amely a folyamat legjobb minőségű végrehajtására törekszik. A Kanadai Statisztikai Hivatal főstatisztikusa, *Fellegi Iván*, a nemzetközileg elismert szakember azt a már több fórumon kifejtett véleményét hangsúlyozta, miszerint a statisztikai hivatalok életben maradása egyrészt az adatszolgáltatókkal szemben tanúsított magatartástól (főként az egyedi adatok bizalmas kezelése, a megfelelő adatvédelem és az adatszolgáltatói terhek optimális kezelése), másrészt, az információk hitelességétől függ, melynek a pontosság fontos, de csak egyik része. A hitelesség a statisztikai hivatalok számára élet-halál kérdés, mert az adat belső értéke és felhasználhatósága a statisztikai rendszer hitelességétől függ (*Fellegi* [1996]). A hitelességnek legalább négy jellemzője van: a pontosság, átláthatóság, politikamentes objektivitás és az adatok fontossága (relevancia). Fellegi hozzászólásának terjedelmes része a „statisztikai szolgálat minősége kontra egyedi statisztikák minősége" témáról szólt (*Fellegi* 2001). *P. Nanopoulos*, aki 2001-ben az Eurostat statisztikai információkért felelős vezetője volt, kétségbe vonta az ISO-ra alapozott megközelítés alkalmasságát a hivatalos statisztika minőségének kezelésére. Véleménye szerint (*Nanopoulos* [2001]) csak elméletileg igaz, hogy a felhasználó definiálja a statisztikus számára a minőségi szabványt. Sajnos, a felhasználók többsége nem hajlandó az általa keresett adat minőségére vonatkozó explicit szabványok megadására. Így a statisztikusoknak egyedül kell dönteniük, anélkül, hogy ismernék a paramétereket és az adatokban meglevő hibák következményeit. Ezzel kapcsolatban felmerül a kérdés, hogy egy statisztikai hivatalban mennyire követhető a felhasználói igények kielégítésének elve úgy, hogy ugyanakkor független, tudományos szempontból pedig objektív intézmény maradjon. A vitaindító cikk szerzői viszontválaszukban (*Platek–Särndal*, [2001b]) azt emelték ki, hogy a statisztikusokat ezúton szerették volna felkészíteni a jelen új kihívásaira (informatikai társadalom, globalizáció, új gazdaság). A nemzeti statisztikai

hivataloknak biztosítaniuk kell az adatszolgáltatók és a felhasználók számára a bizalmat, a tekintélyt és a hitelességet. A statisztikák minőségének többféle értelmezése és mérése létezik, a TQM alkalmazása fokozatosan felváltja a minőség szűkebb fogalmát: a pontosságot.

Az elmúlt évtizedekben a minőség fogalmának meghatározására különböző, de a fő tényezőket tekintve hasonló megoldások születtek. A Svéd Statisztikai Hivatal minőségdefiníciója megegyezik az *Encyclopedia* már idézett meghatározásával. A Kanadai Statisztikai Hivatal hat dimenzióval jellemzi a statisztikai minőséget: relevancia, pontosság, időszerűség, hozzáférhetőség, értelmezhetőség és koherencia (*Statistics Canada Quality Guidelines* [2003]).

Az Európai Statisztikai Rendszer (ESR) a statisztika minőségének meghatározásakor az ISO szabványban (*ISO* [1996]) definiált minőség fogalmából indult ki, azaz „…a minőség egy szolgáltatás, illetve termék azon tulajdonsága, illetve jellegzetessége, hogy milyen mértékben felel meg a deklarált vagy feltételezett elvárásoknak". Az Eurostat legutóbbi módszertani anyagában a statisztikai minőség legújabb fogalmának hat kritériuma van: relevancia, pontosság, időszerűség és (időbeli) pontosság, hozzáférhetőség és átláthatóság, összehasonlíthatóság, valamint koherencia (*Eurostat* [2003b]). 2002-ig volt egy hetedik kritérium is: a statisztikai adatok teljessége (completeness), amely arra a mértékre utal, hogy a publikált számok minden olyan információt tartalmaznak-e, amelyet a felhasználók igényelnek. Ez a hetedik kritérium 2003-ra kikerült az Eurostat módszertanából (*Eurostat* [2003f]).

Az IMF minőségdefiníciója öt ismérvet tartalmaz, de mindezt bevezeti egy 0-diknak nevezett „a minőség előfeltételeit" (szakmaiság, etikai normák, átláthatóság) tartalmazó tétel. Az öt ismérv: objektivitás, módszertani megalapozottság, pontosság és megbízhatóság, szolgálatkészség, hozzáférhetőség (*Laliberté et al*. [2003]).

Az OECD definíciója tartalmazza a legtöbb (nyolc) ismérvet a statisztikai minőség jellemzésére, ezek: relevancia, pontosság, időszerűség, pontosság (időbeli), hozzáférhetőség, értelmezhetőség, koherencia, hitelesség (*Giovannini* [2003]).

## 2. AZ EUROSTAT TEVÉKENYSÉGE A STATISZTIKAI MINŐSÉG TERÉN

A világban felhalmozott tapasztalatok alapján, és az Európai Unió egyes tagállamai statisztikai hivatalainak támogatásával az Eurostat 1994-ben indított kutatásokat a minőség mérésére (*Linden–Sonnberg* [2002]). Az Eurostaton belüli minőségpolitika első dokumentumai 1996-ban készültek el (*Eurostat* [1996a], [1996b]), az első tanfolyamra 1998-ban került sor (*Franchet* [1998]), a felhasználók statisztikával kapcsolatos véleményének felmérését 1999-ben kezdték el. Ezt követően a minőség mérésével és jelentésével kapcsolatos módszertani munkák megvitatása következett. Egy-egy statisztikai felmérés minőségi kérdéseinek vizsgálata az adott témára és meghatározott időre kialakított „Task Force"-ok keretében történt (mint például a munkaerő-statisztikai felvétel, az éves gazdaságszerkezeti felvétel, nemzeti számlák, fizetési mérleg stb.), melyek eredményeként az adott terület statisztikai felvételével kapcsolatban rögzítették a minőség jelentéstartalmát és a minőségértékelés kritériumait. 1999 és 2003 között öt idevágó jogszabály született. Az Eurostatban 2002-re már a statisztika 30 részterületére elkészültek a szabvány minőségjelentések.

Az adatminőség általánosabb módszertani kérdéseinek tisztázására Minőségértékelési munkacsoport (Working Group on the Assessment of Quality in Statistics) alakult 1998-ban. Feladatkörébe tartozott a minőség definícióinak harmonizálása, a minőségjelentések standardizálása, a módszertani kérdésekkel foglalkozó tevékenységek koordinálása és a minőségjelentések alkalmazása az egyes területeken. Többéves tevékenysége során a munkacsoport kidolgozta és 2003-ban véglegesítette a minőségre vonatkozó alapdokumentumokat, melyek a következők.

– A minőség definíciója (*Definition of Quality Statistics*, Eurostat [2003b]).
– A minőség statisztikai kérdéseinek fogalomtára (*Glossary „Quality in Statistics"*, *Eurostat* [2003c]).
– Standard/szabványosított minőségjelentés (*Standard Quality Report*, *Eurostat* [2003a]).
– Kézikönyv a minőségjelentés készítéséről (*Handbook „How to make a Quality Report"*, *Eurostat* [2003f]).

Ezek a munkák teremtették meg a fogalmi és koncepcionális harmonizáció alapját. Ugyanis az egyes szakterületek a minőségi munkacsoporttal párhuzamosan fejlesztették ki saját a minőségkoncepciójukat és mérési módszereiket, és ehhez kezdetben még nem volt meg a közös, referenciaként szolgáló alapdokumentum. Az érintett területek képviselőinek részvételével a munkacsoport rendszeresen áttekintette az egyes szakterületeken a minőségjelentésekkel kapcsolatosan végzett munkát. 2003-ban például a külkereskedelmi statisztika, a pénzügyi mérlegek, az éves gazdaságszerkezeti- (Structural Business Survey – SBS), a munkaerő- (Labour Force Survey – LFS), a közösségi innováció- és munkaerőköltség-statisztika területén. A munkacsoport foglalkozott más nemzetközi szervezetek minőségkoncepciójával, azok összehasonlításával, a munka összehangolásával, így például 2002-ben az OECD, 2003-ban az IMF minőségértékelési gyakorlatát vetették össze a terület szakértői.

Az Eurostatban a minőséggel kapcsolatos fejlesztő munka azonban nem zárult le ekkor, további kutatások és gyakorlati munkák folytak 2003-ban is. Ebben a tárgykörben készültek el például a következő tanulmányok: A statisztikai célú adminisztratív adatok minőségértékelése (*Eurostat* [2003g]); a Gyorsbecslések minőségértékelése (*Eurostat* [2003m]); a Minőség és a metaadatok (*Eurostat* [2003h]); a Standard/szabványosított minőségmutatók (*Eurostat* [2003e]).

### *A minőség szakértői csoportjának (LEG[2] on Quality) tevékenysége*

Az Eurostat keretében folyó módszertani munka új szakasza kezdődött a minőség szakértői csoportjának megalakulásával. A LEG-et a Svéd Statisztikai Hivatal javaslatára a Statisztikai Programbizottság (SPC[3]) 1999 márciusában hívta életre azzal a feladattal, hogy az ESR számára minőségfejlesztési ajánlásokat fogalmazzon meg. A csoport vezetője *Lars Lyberg*, a Svéd Statisztikai Hivatal nemzetközi hírű munkatársa lett, mellette a téma 16 ismert szakértője: három svéd, két holland, két Eurostatban dolgozó, két angol, két német, két portugál és egy-egy francia, olasz, spanyol és görög szakember dolgozott együtt. Kilenc

[2] Leadership Expert Group.
[3] A Statisztikai Programbizottság tagjai a tagországok statisztikai hivatalainak vezetői, elnöke az Eurostat főigazgatója. Feladata az Európai Bizottság támogatása a középtávú statisztikai programok általános koordinálásában; javaslatokat tesz a középtávú statisztikai programok kidolgozására (célok kijelölése, a szükséges intézkedések, módszertani kérdések megtárgyalására); részt vesz a jogszabályok megalkotásában; évente jelentést készít a EU-ban folyó statisztikai munkáról.

ülés után jelentésüket 2001 márciusában, egy Stockholmban rendezett nemzetközi konferencián bocsátották vitára, majd azt az SPC 2001 szeptemberében elfogadta.

Végleges jelentésében (*Eurostat* [2001]) a LEG 150 oldalon tárgyalja a minőség helyzetét az európai statisztikai rendszerben. Az anyag tartalmazza a nemzeti statisztikai hivatalok gyakorlatát, az adatminőség jellemzőit, a minőség és a felhasználók kapcsolatát, az értékelés eszközeit, a dokumentálást és egy 22 pontból álló ajánlást,[4] valamint az ESR minőségügyi keretrendszerét, a statisztikai termelő folyamatok minőségének standardizálását és fejlesztését, végül a teljes körű minőségirányítás alkalmazását az egyes nemzeti statisztikai hivatalokban és az Eurostatban.

A LEG minőségre vonatkozó ajánlásainak 22 pontos jegyzéke (lásd a Mellékletet) lényegében tömör összefoglalása a jelentés első fejezetében foglaltaknak, melyben a minőséggel kapcsolatos teendők felsorolása található a nemzeti statisztikai hivatalok és az európai statisztikai rendszer valamennyi tagja számára. A Szakértői csoport országonkénti jelentést kér a statisztikai termékek és folyamatok minőségi jellemzőiről, az adatszolgáltatókkal és az adatfelhasználókkal való kapcsolatról, a hivatalokban jelenleg alkalmazott legjobb módszer (Current Best Method – CBM) gyűjteményéről, a dokumentálás és információkezelés, terjesztés helyzetéről. A továbbiakban javasolja, hogy a hivatalok erősségeik és gyenge pontjaik felmérését követően hozzák nyilvánosságra feladatmeghatározásukat (mission statement), tájékoztatási és minőségügyi irányelveiket (quality declaration), készítsenek cselekvési programokat. Az ESR tagjai minőségfejlesztésre és teljesítményértékelésre használják az Európai Minőségfejlesztési Alapítvány (üzleti) kiválóság modelljét, (EFQM). A jegyzék a szervezetek minőségirányítási rendszerének megvalósításához feltételként írja elő a személyi állomány minden tagjának oktatását és továbbképzését. Az információterjesztés érdekében szorgalmazza, hogy kétévente szervezzenek konferenciát[5] a módszertannal és a minőséggel kapcsolatos témákban, valamint alapítsanak egy kétévente adományozható minőségdíjat a teljesítmények elismerésére. Végül a 22. pont arról szól, hogy az ajánlások végrehajtására alakuljon egy „végrehajtási szakértői csoport" (LEG on Implementation), mely úgy tekinthető mint az SPC számára felállított minőségügyi tanácsadó testület. E csoport fő feladata kezdetben az információgyűjtés, később az ajánlott tevékenységek végrehajtásának koordinálása. Együtt kell működnie a Minőségértékelési munkacsoporttal, hiszen a végrehajtás sikere az Európai Statisztikai Rendszer tagjainak aktív részvételétől függ. A végrehajtási csoportnak 2004-ben kell a zárójelentését az SPC elé terjeszteni.

### *Az Európai Unió jogi szabályozása a statisztika minőségéről*

Az EU-ban jelenleg még nincs átfogó, az ESR egészére érvényes jogszabály a statisztikai termékek minőségéről. A közösségi statisztika megteremtésének feltételeit, az ESR működési alapelveit az 1997-ben aláírt Amszterdami Szerződés 285. cikkelye tartalmazza. Az 1997-ben elfogadott 322/97. számú Tanácsi rendeletet[5] az Unió „statisztikai törvényének" is szokták nevezni, melynek az volt a célja, hogy jogi keretet biztosítson a kö-

[4] List of LEG on Quality Recommendations, Final Report of the Leadership Expert Group (LEG) on quality, Annex 3. (*Eurostat* [2001]).

[5] Az első ilyen konferenciát a Német Statisztikai Hivatal szervezte Mainzban, 2004. május 24. és 26. között.

[5] Council Regulation (EC) No 322/97 of 17 February 1997 on Community Statistics. Official Journal L 052, 22/02/1997.

zösségi statisztikák rendszeres előállítása számára. A rendelet indoklásában egyrészt az Amszterdami Szerződésre, másrészt az ENSZ Statisztikai Bizottságának 1994-ben elfogadott „A hivatalos statisztika alapelvei" c. dokumentumára (*Szilágyi* [2004]) hivatkozott. E rendelet 10. cikkelye megállapítja: „Annak érdekében, hogy a közösségi statisztikák etikai és szakmai szempontból egyaránt a lehető legjobb minőségűek legyenek, a következő elvek kell, hogy érvényesüljenek: pártatlanság, megbízhatóság, relevancia, költséghatékonyság, adatvédelem és átláthatóság".

1997 óta több jogszabály is érintette a közösségi statisztikával kapcsolatos minőségi kérdéseket, melyek úgy is tekinthetők, hogy részben megteremtették a minőségértékelés alapjait, mivel egyrészt előírták az összehasonlítás és a pontosság javítását biztosító módszertani standardok alkalmazását, másrészt az Európai Központi Bank adatigényeinek részletesebb, pontosabb megfogalmazásával és a rövidebb határidők igényével felgyorsították a jogszabályok fejlesztési munkáit. Mindez komoly kihívást jelent az évközi statisztikai felmérések felelős szakemberei számára mind az Eurostat, mind a tagországok szintjén. Végül, 1998 és 2003 között öt olyan rendelet született, mely érintette egy-egy terület adatgyűjtésének minőségértékelését is. Ezek közül a munkaerő reprezentatív felvételének EU-n belüli megszervezéséről szóló 577/98. számú Tanácsi rendelet[6] 3. cikkelye a minta reprezentativitására vonatkozóan fogalmazott meg előírásokat. Ugyancsak 1998-ban lépett életbe az évközi gazdaságszerkezeti statisztikára vonatkozó 1165/98. számú Tanácsi rendelet[7], melynek 10. cikkelye az adatgyűjtés minőségértékelését írta elő a tagországok számára. A 1618/1999. számú Bizottsági rendelet[8] teljes egészében az éves gazdaságszerkezeti statisztikai felmérés minőségértékelési kritériumairól szólt. A 2000-ben meghozott 452/2000. számú Bizottsági rendelet[9] az 530/1999. számú Tanácsi rendelet végrehajtásáról, közelebbről a munkaerőköltség éves gazdaságszerkezeti felmérésének minőségértékeléséről szólt. A 1216/2003. számú Bizottsági rendelet[10] a munkaerő költség-index teljes minőségértékelését részletezte, az összes minőségkritérium szempontjából. A fentieken kívül is vannak olyan jogszabályok, melyek explicit módon tartalmaznak minőségre vonatkozó információs igényeket, mint például a hulladék-statisztikára, a vasúti közlekedésre, a jövedelem- és életszínvonalra, a piaci áron számított bruttó nemzeti jövedelem harmonizálására vonatkozó 2002-ben és 2003-ban meghozott Európai Parlamenti és Bizottsági rendeletek. Ezen jogszabályok és a különböző munkacsoportokban 1998 óta folyó minőségértékelési módszertani fejlesztések között kölcsönhatás volt, melynek egyik eredményeként a szakértők 2003 végére javaslatot tettek az Eurostat minőségjelentéseinek véglegesítésére.

## 3. A MINŐSÉGI KERETRENDSZER ELEMEI
### (AZ EUROSTAT AJÁNLÁSAI ÉS AZ IRÁNYÍTÁSÁVAL ÉS AJÁNLÁSAI ALAPJÁN VÉGZETT KUTATÁSOK)

A statisztikai hivatalok „terméke", a felhasználók számára átadott, publikált statisztikai adat. Értelemszerűen a minőség legkidolgozottabb fogalma – az üzleti élet analógiájá-

[6] *Official Journal of the European Communities* L 77/3 14.3.1998.
[7] *Official Journal of the European Communities* L 162/1 5.6.1998
[8] *Official Journal of the European Communities* L 162/1 5.6.1998
[9] *Official Journal of the European Communities* L 55/53 29.2.2000
[10] *Official Journal of the European Communities* L 169/37 8.7.2003

ra – a *statisztikai termékekre* vonatkozik. (Lásd „A termékminőség dimenziói" című alfejezetet.)

A termékminőség a végtermék minősége, ami a termelési (előállítási) folyamat során alakul ki. Amennyiben csak a termékminőséget mérjük, akkor szükség esetén az esetleges korrekcióra csak újrafeldolgozással, utólagos javítással kerülhet sor, ami költséges és időigényes. Magától értetődő a termelési folyamat és a termékminőség összefüggése, amiből következik, hogy a *folyamatok minőségével* is érdemes foglalkozni. (Lásd „A statisztikai folyamatok minőségéről" című alfejezetet.)

A minőséget erősen befolyásolja, miként épül be a minőségi szemlélet a szervezet irányításába. A minőség javításának előfeltétele a minőség mérése, majd a következő a minőség értékelése, továbbá az értékelés eredményeire alapozott fejlesztési döntések. „Az értékelési eszközök" című alfejezetben bemutatjuk, hogy miként vonatkoztathatjuk ezeket az egyes termékekre, egyes folyamatokra, de a rendszer működése akkor lesz teljes, ha az egész szervezetre kiterjesztjük a folyamatos jobbítás elvét. Ezzel elérkeztünk a TQM-hez, a *minőségirányítási* rendszerekhez, melyek célja képessé tenni a szervezetet arra, hogy folyamatosan javuló minőségű termékeket állítson elő. A minőségirányítási rendszerekről az ezt követő alfejezetben foglalkozunk.

*A termékminőség dimenziói*

A termékminőség összetevői alkotják a minőségvektor dimenzióit. A dimenziók jellemzésére nem mindig találunk mennyiségi ismérvet. Gyakran előfordul, hogy csak a mérni kívánt dimenzióval kapcsolatban álló, mérhető változót (proxit) használhatunk, más esetben csak minőségi értékítéletre hagyatkozhatunk. Például a mintavételi hiba mennyiségi ismérv, a nem mintavételi hiba egyes elemeiről pedig csak minőségi értékítéletünk lehet.

A statisztikák minőségének jellemzésére a rendszeresen összeállított minőségjelentés szolgál. A minőségről szóló jelentések felhasználói:

– a statisztika előállításáért felelős intézmény vezetői,
– az intézményen belüli felhasználók,
– az intézményen kívüli felhasználók.

Természetesen az egyes felhasználói csoportok felhasználási célja és statisztikai ismerete különbözik, következésképp a minőségjelentéssel szemben is eltérők az igényeik. A legrészletesebb, legszakszerűbb minőségjelentésre az első csoportnak van szüksége, hiszen erre alapozva hozhatnak döntéseket. Az intézményen belüli felhasználóknak is elsődlegesen azt kell megítélniük, hogy a termék alkalmas-e, felhasználható-e adott szakmai célra. A külső felhasználók nem alkotnak homogén csoportot, megtalálható köztük a statisztikai ismeretekkel nem rendelkező, a felhasznált adatok minőségére csak pár pillanatot szánó felhasználó éppúgy, mint az elmélyült ismeretekkel rendelkező kutató.

Az Eurostat külön „Task Force"-ot (munkacsoportot) hozott létre az egyes dimenziókat jellemző, mérhető mutatók kialakítására. 2003-ban elkészült a termelésorientált mutatókra vonatkozó javaslat, a következő feladat a felhasználóorientált mutatók kidolgozásának elindítása. A minőséggel foglalkozó munkacsoportban és az egyes szakstatisztikai területeken

is folyamatban van az indikátorok fejlesztése. Azokban a statisztikai hivatalokban, ahol minőségirányítási rendszert hoztak létre, kidolgozták a saját belső minőségjelentésüket, ami lehetővé tette a nemzeti szempontok, helyi sajátosságok érvényesítését.

A következőkben felvázoljuk az Eurostat általános sémája szerinti minőségdimenziók jellemzőit és mérési lehetőségeiket.

Az első (és általában mindenféle felsorolásban első) dimenzió a *relevancia* (relevance), amely azt jelenti, hogy tartalmát tekintve a statisztikai adat kielégíti-e a felhasználói igényeket. Ennek megállapításához ismerni kell a tényleges és potenciális felhasználókat, igényeiket és a felhasználók visszajelzését a szóban forgó adatról. *T. Dalenius* [1985] a releváns hivatalos statisztikáról írott cikkében az alábbi meghatározást adta a relevanciáról: „A releváns statisztika nem más, mint egy valódi probléma statisztikai problémává való átalakítása, melynek megoldása hozzájárul a valós probléma megoldásához". A relevancia jellemzésére le kell írjuk, hogy a felhasználók legszélesebb körének mire és mennyire hasznos a statisztika. Először a felhasználókról kell információt gyűjteni (kik azok, hányan vannak, milyen fontosak), másodszor az igényeikről (kinyilvánított, ki nem nyilvánított, közeljövőben várhatóan felmerülő), végül értékelni kell, hogy milyen mértékben sikerült az igényeiknek megfelelni. A felhasználók csoportosítása és jellemzése különböző lehet az egyes termékeknél, bizonyos statisztikák eltérő fontosságúak a felhasználók különböző csoportjai számára. A felhasználók között lehetnek nemzetközi szervezetek, az EU intézményei, országos és regionális intézmények (például minisztériumok, parlament, bíróságok, a Magyar Nemzeti Bank), vállalatok, társadalmi szervezetek, a média képviselői, kutatók, diákok, érdeklődő egyének/állampolgárok, továbbá szervezeten belüli felhasználók.

A statisztikusok számára a felhasználói igények azonosítása, leírása összetett feladat. Egyrészt a felhasználók változnak, váratlan, rövid távú események erősen befolyásolják őket. Következésképp a felhasználói igények változékonyak, nehezen kiszámíthatók. Továbbá a felhasználók számára a legtöbb esetben a statisztika önmagában nem jelent felhasználásra kész információt, hanem csak alapanyag a társadalmi-gazdasági elemzéshez. A felhasználók jellemzően többnyire saját szaknyelvükön fejezik ki kívánságaikat, azt, hogy milyen kérdésekre várnak választ, milyen célból van szükségük az információra. Ezeket az igényeket a statisztikusoknak le kell fordítani statisztikai terminológiára: meg kell fogalmazni a statisztikai koncepciót, a változókat, a megfelelő kérdést a kérdőíven, vagyis meg kell határozni a célnak megfelelő statisztikai terméket. A közgazdasági, társadalmi fogalmak, jelenségek lefordítása mérhető statisztikai mutatókra nem mindig egyszerű, gondoljunk csak a versenyképesség, szegénység vagy a társadalmi kirekesztettség kérdéseire.

A második kritérium a *pontosság* (accuracy), ami nem egyéb, mint a statisztikai munkafolyamatban a becsült érték és a valódi, de ismeretlen sokasági érték közötti eltérés/hiba. A statisztikai hiba magában foglalja az egész statisztikai munkafolyamat alatt keletkező hibákat a koncepció kialakításától az eredmények közléséig. Ha ismernénk a teljes hibát és annak összetevőit, be lehetne azonosítani a jelentős hibaforrásokat és tenni lehetne csökkentésükért. Azonban a teljes hiba modellje még elméletileg sem áll rendelkezésre, az egyes hibafajták nehezen különíthetők el és még nehezebben mérhetők. Megkülönböztetünk mintavételi hibát és nem mintavételi hibát.

*A mintavételi hiba* számításának kiterjedt irodalma van, és ezeket a számításokat számos közforgalomban levő szoftver segíti (például: SUDAAN, WESVAR, CLAN,

STATA, GES, GSSE). Az Eurostat áttekintő kiadványt készített az ajánlott hibaszámítási módszerekről (*Eurostat* [2002]).[11]

Például a munkaerőfelvétel esetében kiadványainkban a mintavételi hibát (sampling error) közöljük, ami az adott mutató mértékegységében kifejezve tájékoztatja a felhasználót a hiba mértékéről. Az Eurostat számára a minőségjelentésben a relatív szórást (coefficient of variation) adjuk meg a fontosabb mutatókra, mivel ez lehetővé teszi a mutató abszolút nagyságától függetlenül a pontosság nemzetközi összehasonlítását százalékos formában.

A gyakorlatban általában túlhangsúlyozzák a mintavételi hiba jelentőségét és kevés figyelmet szentelnek a nem mintavételi hibára. Ennek elsődleges oka a nem mintavételi hibák azonosítási és mérési nehézsége. Sok esetben csak kvalitatív értékítéletig juthatunk el.

*A nem mintavételi hiba* mind a mintán alapuló, mind a teljes körű felvételeket érinti. Ilyenek a:

– lefedettségi hibák,
– mérési hibák,
– feldolgozási hibák,
– nemválaszolásból eredő hibák,
– modell-feltételezések hibái.

A lefedettségi hiba a célsokaság és a keretsokaság eltérése. A lefedettségi hibák nem mindig derülnek ki. A felvételben minden megfigyelt egység esetében ellenőrizni kell, hogy a keretben a rá vonatkozó információ helyes-e, valóban a célsokasághoz tartozik-e. Ebből következtethetünk az egész keret ilyen jellemzőire. Ez a módszer azonban csak a lefedettségi többlet feltárására alkalmas. A lefedettségi hiány feltárása nehezen általánosítható, specializált módszereket igényel. Például egy gazdaságstatisztikai felvételnél a mintavételi keret a gazdasági szervezetek regisztere, ebben is az 5–49 fős szervezetek. A $t$+1 évben végrehajtott felvétel mintáját $t$ évben választjuk ki, amikor az sok tekintetben még a vállalatok $t$–1 évi állapotát tükrözi. Így a felvételkor lefedettségi többletet jelent az időközben 5 fő alatti létszámúvá vált vállalat. Ez a többlet a felvételkor azonosítható és kezelhető a megfigyelt körben. Ezzel szemben az 5 fő fölé növekedett vállalatokat nem tudjuk figyelembe venni, így ez lefedettségi hiányt okoz.

A mérési hiba az adatgyűjtés során elkövetett hiba, aminek következménye, hogy a rögzített adat eltér a tényleges értéktől. A felmérés eszköze (kérdőív, telefonos interjú), a megkérdezett vagy a kérdező egyaránt lehet a hiba okozója. A mérési eszköz és a kérdezőbiztos hatása különböző feltételek melletti ismételt kérdezéssel, véletlen kísérletekkel mérhető (például kérdőívváltozatok lekérdezése, különböző kérdezőbiztosokkal történő kérdezés). A válaszadói hatás nehezebben mérhető, itt is alkalmazható ismételt lekérdezés vagy független adatforrások felhasználásával végzett ellenőrzés. Hazai tapasztalatok is alátámasztják, hogy a kérdőíven egy kérdés átfogalmazása, érthetőbbé tétele alapvetően megváltoztathatja a válaszolási arányt és a beérkező válaszokat.

Az editálás azonosítja az adatok inkonzisztenciáját, ami jellemzően hibát jelent. A hiba nem szükségszerűen mérési hiba, lehet feldolgozási hiba is (kódolási vagy adatbeviteli

[11] A különböző hibaszámítási módszerek és eljárások elméleti, és valós adatbázisokon végzett gyakorlati összehasonlító vizsgálatát célozta a komplex felvételek varianciabecslési módszereivel foglalkozó kutatás (DACSEIS) (www.dacseis.de).

hiba). Az editálás során hibásnak talált rekordok aránya az eredeti adatok és nem feltétlen a végleges adatok minőségét jellemzi.

A feldolgozási hiba az adatgyűjtést követő feldolgozási munka (a szöveges válaszok kódolása, adatok számítógépbe való bevitele, az adatok ellenőrzése az inkonzisztencia és outlierek kiszűrése céljából (editálás), a hiányzó vagy hibásnak minősített adatok pótlása (imputálás)) során elkövetett hiba. A feldolgozási hibákat kísérletek segítségével mérhetjük, például a kérdőívek újrakódolásával vagy más munkafolyamatok ismételt elvégzésével. Az így feltárt hibák javíthatók, így ezeket a hibaszámításnál figyelmen kívül hagyhatjuk, de ezek segítségével becsülhetjük a hibaarányokat a vizsgálaton kívül maradt adatokon. A kérdőívek újrakódolása költséges, így az ilyen vizsgálatok ritkák. Az újrakódolással végzett foglalkozásikód-ellenőrzések eredményei az mutatták, hogy a legrészletesebb kategóriák esetében az azonos besorolás aránya 50–70 százalék között van, míg a magasabb aggregátum szintjén ennél sokkal kedvezőbb a kép.

Számítógéppel támogatott adatfelvételnél a feldolgozási munkaszakaszok egy része is megvalósul, mialatt az adatszolgáltató megadja a válaszokat, így mind a mérési, mind a feldolgozási hiba esélye csökken.

A nemválaszolás[12] vagy meghiúsulás jellemzésére a válaszadási arány két típusa számítható: az egységszintű válaszolási arány, ami a legalább részben adatot szolgáltató egységek aránya az adatszolgáltatásra kijelölteken belül, valamint a tételválaszolási arány, ami a felvételben szereplő minden változóhoz számítható, és a tételre adott válaszok arányát mutatja az összes felvételre kijelölt adatszolgáltatókon, vagy a legalább néhány tételre válaszoló adatszolgáltatókon belül. A válaszadási arányt 1-ből kivonva megkapjuk a nemválaszolási arányt. A súlyozott válaszadási arány a mintaelemek súlyát használja fel, az értéksúlyozott válaszadási arány pedig kiegészítő információkkal súlyoz. Abban az esetben, ha a kiegészítő információ korrelál a felmérni kívánt változóval, úgy az utóbbi mutató jobban mutatja a nemválaszolás hatását az eredményekre.

A nemválaszolás tipikusan olyan minőségi aldimenzió, ami részben kvalitatív információkkal (nemválaszolás okai, a nemválaszolás csökkentésére tett erőfeszítések, az imputáció módszerei, ezek figyelembevételének módja a becslésnél, hibaszámításnál), részben kvantitatív információkkal (nemválaszolási arány, válaszolók és nemválaszolók összehasonlító jellemzése, a nemválaszolás hatása a becslésre) jellemezhető.

A modell-feltételezések hibáiról akkor beszélünk, ha az alkalmazott modellek[13] feltételei nem teljesülnek. Ha a feltételezéseket kellően ellenőrizzük, ez a hiba elkerülhető. Azt, hogy a kapott statisztika mennyire robusztus a feltételek teljesülésére, érzékenységvizsgálatok segítségével vizsgálhatjuk.

A harmadik kritérium vagy dimenzió az *időszerűség* (timeliness), és az időbeli *pontosság* (punctuality). Az időszerűség, (máshol gyorsaságnak fordítják) azt az időtartamot jelöli, amely a jelenség felmerülése (referenciadátuma) és az erről készített felmérés befejezése/adatátadása között van. (Például a negyedéves GDP-adat a tárgynegyedévet követően hány nap múlva jelenik meg.) Az időbeli pontosság pedig egy előre bejelentett közlési időpont (a tájékoztatási naptárban publikált határidő) és a tényleges közlés időpontja közötti eltérés.

[12] A nemválaszolási hiba típusairól bővebben ír *György Erika* e számban megjelenő cikkében 747–772. old.

[13] Ilyen eljárások a kalibrálás, a regressziós becslés, a szezonális kiigazítás, az igazodási pontokhoz való igazítás (benchmarking).

Mindkét mutató könnyen mérhető, de értékelésük nem ennyire egyszerű. Az időszerűség a statisztika relevanciájával is összefüggésben van, hiszen a jelenségtől időben távolodva annak jelentősége csökken, egy éves adat közlése évekkel később, vagy egy havi adat közlése több hónappal később már nem számít releváns információnak. Az Eurostat egyre szigorítja az időszerűségre vonatkozó követelményeit, melynek teljesítése az egész adatgyűjtési, feldolgozási folyamat újratervezését teszi szükségessé a nemzeti statisztikai hivatalokban.

A gyors adatszolgáltatás mérsékli a pontosság követelményének betarthatóságát, például a késve érkező adatok nem vehetők figyelembe, kevesebb információ áll rendelkezésre, nincs idő elvégezni bizonyos ellenőrzéseket. Azoknál a statisztikáknál, ahol a gyorsaság a pontosság rovására is biztosítandó a felhasználók mindkét igényét szem előtt tartva, egy adott statisztikára, időbeli sorrendben, ún. gyorsbecslés, előzetes adat és végleges adat is publikálásra kerül, egyre növekvő pontossággal (például a hazai GDP-számítások ilyen rendszerben készülnek).

A negyedik minőségi dimenzió a *hozzáférhetőség* és *érthetőség* (accessibility, clarity). A felhasználó számára fontos szempont, hogy tudjon az információ meglétéről, és ismerje a hozzáférés módját. Ezt nagyban meghatározza egy hivatal tájékoztatási szolgáltatásának színvonala, de függ az alkalmazott technológiától is. Az érthetőség az adatokhoz kapcsolódó információkra vonatkozik (dokumentáció, módszertani magyarázatok, metaadatok, ábrák stb.).

A hozzáférhetőséget a tájékoztatási csatornák sokfélesége (papíralapú, CD, internet), az adatszolgáltatás formájának változatossága (mikro-, makroadat), a felhasználók terhei, azaz a költségek és az az idő jellemzi, amíg a szükséges statisztikához hozzájutnak. A legjobb elérhetőséget azonban az adatvédemi szempontok egyidejű érvényesítésével kell biztosítani.

A felhasználók rendelkezésére bocsátott dokumentáció teszi lehetővé, hogy a felhasználók megfelelően értelmezzék és használják az adatokat. Az érthetőség úgy is értelmezhető, mint a metaadatok relevanciája. A metaadatok teljessége a tartalomra vonatkozó standard címszavak összeállításával teljesíthető. Az érthetőség közvetlenül a felhasználók megkérdezésével jellemezhető.

Az ötödik dimenzió az *összehasonlíthatóság* (comparability). A statisztikai adatok időbeli, területi, régiók, országok közötti összehasonlításának biztosítása az Európai Unióban is kiemelt minőségi szempont. Az összehasonlíthatóság relatív kategória, mindig a vizsgálat céljától függ. Kiemelt jelentősége van az EU és olyan nemzetközi szervezetek minőségi megközelítésében, melyek maguk nem végeznek felméréseket, hanem a tagországok statisztikai információit hasznosítják.

A földrajzi összehasonlíthatóság mérése két különböző megközelítésen alapulhat: egyrészt ha a számot egy standardhoz hasonlítjuk, ami lehet egy európai norma vagy egy ország modellként szolgáló felvétele. A másik megközelítés, ha több országból van adatunk, akkor az összes lehetséges páros összehasonlítást elvégezzük, és végül összesítjük az eredményeket. Az összehasonlításnál a metaadatok összehasonlításából kell kiindulni, majd a feltárt eltérések mértékét kell megbecsülni a rendelkezésre álló adatok segítségével. Viszonylag széleskörűen alkalmazott eljárás a tükörstatisztika, ami szokásos gyakorlat a külkereskedelem, és a turizmus terén vagy szállítási statisztikában. A különböző országokban az adott jelenségre kapott statisztikákat állítják szembe egymással, például két

ország egymás közötti külkereskedelmére vonatkozó adatokat. Bár jellemzően a koherencia vizsgálatára használják, a földrajzi összehasonlításban is hasznos.

Az időbeni összehasonlíthatóság gyakran csorbul; vegyük csak a leggyakoribb okokat: a megfigyelési egység, referenciaidőszak (üzleti év/naptári év), számviteli fogalmak változása, a jogi szabályozás (például adószabályok) változása, osztályozási rendszerek, földrajzi besorolások változása, és akkor még nem is beszéltünk a mérési, becslési módszerek változásáról. Finnország például 1983-ban a munkaerő-felvételnél áttért a postai úton történő megkérdezésről a személyes megkérdezésre, és ez a foglalkoztatottak számára százezerrel nagyobb értékbecslést eredményezett. Néhány esetben ezek az eltérések valamilyen algoritmus alapján végzett transzformációval korrigálhatók. Ennek azonban az az előfeltétele, hogy a változások hatása elkülöníthetően mérhető legyen, hogy ezt a folyamatot a változások bevezetése előtt megtervezzék.

Az EU új tagországai és a csatlakozásra váró országok esetében a jelenleg alkalmazott módszerek harmonizációja volt a központi feladat, ezért az időbeli összehasonlíthatóság, azaz a saját múlttal való összevethetőség nemegyszer áldozatul esett a földrajzi összehasonlíthatóság követelményének (az EU harmonizáció biztosítása miatt kevesebb erőforrás jutott a korábbi idősorok átdolgozására).

A hatodik dimenzió a statisztika *koherenciája* (coherence), mely azt biztosítja, hogy a különböző felhasználási célokból vagy különböző forrásokból előállított adatok kombinálhatók. Például a munkatermelékenység számlálójának és nevezőjének ágazati lefedettsége azonos, mindkét adat ugyanarra az aggregációs szintre vonatkozik. A koherencia javítható a közös osztályozási rendszerek, mintakeretek, átfogó fogalmi rendszerek/keretek (pl. a nemzeti számlákra az ESA) használatával.

A koherencia és az összehasonlíthatóság is két adatsor egymáshoz való viszonyát jellemzi, nem véletlen, hogy egyes országok a két dimenziót együtt kezelik. Az összehasonlíthatóság azonban jellemzően két független sokaságra vonatkozó statisztika összehasonlítására, míg a koherencia ugyanarra, vagy egy nagyrészt hasonló sokaságra vonatkozik. Előbbi a metaadatok alapján, míg az utóbbi, a koherencia az adatokban feltárt inkonzisztenciák alapján ítélhető meg. Konzisztenciavizsgálatokat leggyakrabban az előzetes és a végleges adatok, az éves és az évközi (havi, negyedéves) adatok között, valamint az azonos társadalmi-gazdasági területre vonatkozó különböző statisztikák, továbbá bizonyos szakstatisztikák és a nemzeti számlák között szoktak végezni. A koherencia vizsgálatára a következő gyakorlat terjedt el: egy alkalmas mutatóval jellemezzük a konzisztenciát, illetve annak hiányát (abszolút százalékos hiba vagy torzítatlan százalékos hiba), ha ennek mértéke egy előre megadott mértéket meghalad, célszerű a konzisztencia átfogó vizsgálata.

Több kísérlet is történt arra, hogy a minőségdimenziók jellemzőit *egyetlen minőségindexbe* tömörítse. Eddig ezek a kísérletek nem jártak eredménnyel, legutóbb, 2003 őszén az Eurostat-munkacsoport is elutasított egy ilyen javaslatot. Az elutasítás oka nemcsak a különböző mutatók aggregálásának technikai nehézségeiben rejlik, hanem elsődlegesen abban az elvi problémában, hogy a különböző termékek esetében az egyes minőségi dimenziók különböző jelentőségűek. A rövid távú konjunktúramutatók esetében az időszerűségnek nagyobb a súlya, mint a pontosságnak, míg más mutatóknál (például a végleges nemzeti számla-adatoknál) nem ez a helyzet. Nem lehet tehát egységes súlyrendszert kialakítani. További probléma, hogy a gyakorlatban a minőségi dimenziók közül nem min-

dig a legfontosabbakat, hanem a könnyen mérhetőket mérik. Így ezek torzított képet is adhatnak a termék minőségéről.

A minőségjelentés összeállításában az adott felvétel felelőseinek, minőségi és módszertani szakértőknek és informatikusoknak kell részt venniük. A minőségjelentés időszerűsége (minél előbb rendelkezésre álljon), továbbá az erőforrásokkal való takarékosság érdekében a jelentést célszerű nem utólag, hanem lehetőség szerint, az adatelőállítás folyamatával párhuzamosan elvégezni. Az adatfeldolgozást gondos tervezési munkának kell megelőznie: elő kell állítani és tárolni kell a minőségmutatóhoz szükséges információkat. Ilyenek például: a nyers adat, a javított adat, az elvégzett editálások és ezek alapja, a válaszadás dátuma, melyekre azért van szükség hogy a nemválaszolás és a korrekciós módszerek, a felvételezési eljárásra vonatkozó információk stb. elemezhetők legyenek. Így biztosítható, hogy a statisztikai munka során a minőséget jellemző mutatók értéke automatikusan, a folyamat szerves részeként kerüljön becslésre, kiszámításra. Esetenként a minőségi mutatók értéke lényeges hibát jelezhet, amit azonnal korrigálni kell.

*A statisztikai folyamatok minőségéről*

Az előző fejezetekben a statisztikai termékek minőségéről és mérési problémáiról esett szó. A statisztikákat számos, kölcsönösen összefüggő munkafolyamat eredményeként állítják elő, mint például a felvétel tervezése, az adatgyűjtés, a kódolás, az adatbevitel, az editálás, az imputálás, a teljeskörűsítés stb. A munkafolyamatok nagymértékben meghatározzák a végső kibocsátásra, közlésre kerülő statisztika minőségét. Ha a statisztika minőségének növelése a cél, oda kell figyelnünk a statisztikák előállítási folyamataira is. A statisztikák minőségvizsgálati eredményei és az előállítás során lezajlott folyamatok közötti kapcsolat feltárása alapján be kell azonosítani a termékminőség szempontjából legfontosabb folyamatokat. A statisztikák minőségének lényeges elemeit meghatározó munkafolyamatokat értékteremtő folyamatoknak nevezik.

Egyes statisztikai hivatalok összeállították saját listájukat, melyek átfogják a statisztikai munka szakaszait a felvétel tervezésétől kezdve az eredmények archiválásáig. Közelebbről ezek a következők: az adatgyűjtés elindításáról szóló döntés, az adatgyűjtéstervezés, az adminisztratív adatokhoz való hozzáférés, a mintavételi terv kialakítása, végrehajtása, az adatgyűjtés végrehajtása, editálás, validálás, a változók képzése és kódolás, súlyozás és becslés, elsődleges eredmények elemzése (beleértve a minőségjelentést), indexszámítás, idősorelemzés, felfedés elleni védelem, adatok és metaadatok közzététele, archiválás. Ez a részletezés azért lényeges, mert ezeknek a folyamatoknak a minőségi követelményeire figyelve tehetjük a legtöbbet a termékminőség érdekében. A hatékonyság, a termelékenység, a robusztusság, a rugalmasság és az átláthatóság növelése a cél, amikor a folyamat minőségét javítjuk.

A statisztikai munkaszakaszok, folyamatok standardizálása kézenfekvő. A standardizálás azt jelenti, hogy egy-egy munkafolyamat végrehajtása azonos séma szerint, azonos módon történik az egész szervezeten, vagy akár az egész hivatalos statisztikai szolgálaton belül.

A standardizálás pozitív hatásának tekinthető, hogy megkönnyíti a folyamatok dokumentálását, az új dolgozók betanítását és a dolgozók belső rotációját, az új fejlesztések kidolgozását és bevezetését, továbbá az információtechnológiai eszközök hatékony al-

kalmazását (például közös szoftver alkalmazása). Ugyanakkor, az adott speciális terület sajátosságaira is tekintettel kell lenni, mert ami az esetek nagy részében elfogadható, néhány esetben nem. Például a szezonális kiigazításnál általános gyakorlat, hogy a paraméterrögzítésre évente egyszer, az év utolsó adatának beérkeztekor kerül sor. Ugyanakkor, ha az alapadatban rendszeresen év közben történik revízió, akkor indokolt, hogy a paraméterrögzítés időpontját a revízióhoz igazítsuk.[14]

A leggyakrabban használt standardizálási eszközök: a jelenlegi legjobb módszerek, ajánlott módszerek, minimális standardok, minőségi útmutatók, ellenőrzőlisták.

A *jelenlegi legjobb módszer* (Current Best Methods – CBM) az egyes folyamatok jelenleg legjobbnak tudott módszereinek leírása. Az olasz statisztikai hivatal, az ISTAT ilyen módszertani kézikönyvet dolgozott ki a felvételtervezés, a kérdőívtervezés, a kérdezési technikák, a mintavételi tervek, a varianciabecslés, az adatminőség ellenőrzése és az ábrázolás folyamataira. A svéd statisztikai hivatalban az editálás, az ábrázolás, a nemválaszolás csökkentése, a projektmunka, a felfedhetőség ellenőrzése, a nemválaszolás korrekciója és a kognitív tesztelés folyamataira dolgoztak ki módszertani ajánlásokat. A CBM garanciát nyújt a folyamat minőségére, csökkenti a megvalósítási különbségeket, egyúttal a gyakorlat dokumentációjaként is szolgál.

Az Eurostat megbízásából létrejött minőséggel foglalkozó LEG ajánlásai alapján a CBM kidolgozásának kézenfekvő szintje a nemzeti szint és nem valószínű, hogy európai szinten ilyenek kidolgozhatók és elfogadhatók. Az egyes országokban eltérők a kulturális hagyományok, a törvények és a szabályok, a rendelkezésre álló regiszterek és a külső adatforrások. Ezek a különbségek ezért más és más országonkénti megközelítést indokolnak.

Az *ajánlható módszerek* (Recommended Practices – RP) a bizonyítottan eredményesen alkalmazható módszerek gyűjteménye, mely ismertetésük mellett fontosabb jellemzőiket is tartalmazza. Az összeállítás célja, hogy az adott feladatra alkalmazható lehetséges módszerek áttekintése alapján segítse a statisztikusokat, azt a legjobb eljárást kiválasztani, amely az adott körülményekhez a legjobban illeszkedik.

A *minimális standardok* (Minimum Standards – MS) azokat a feltétlenül teljesítendő feltételeket rögzítik, amelyeket az egyes statisztikai folyamatoknak ki kell elégíteniük. Ezek jellemzően arra vonatkoznak, hogy a statisztikai munkának tudományos alapelvekre kell épülnie, és az ezek közötti választás az eredményt is befolyásolja. Az Európai Unió szintjén a statisztikai közösségi vívmányok (acquis communautaire = regulations and gentlements agreemens) tekinthetők minimális standardnak.

A *minőségi útmutató* (Quality Guidelines – QG) tekinthető a legelterjedtebben alkalmazott standardizálást célzó módszernek. A minőségi útmutató megfogalmazza a legfontosabb statisztikai munkaszakaszokra azokat az alapelveket, melyek mentén minőségi statisztika állítható elő. Mint ilyen, jó kiindulási alap is lehet a további munkálatokhoz, például az ajánlható módszerek összeállításához.

Az *ellenőrzőlisták* (Checklist) az egyes munkaszakaszok elvégzendő részfolyamataira, vagy/és számszerű jellemzőire, kritériumokra vonatkoznak.

Az EU 5-ös kutatási keretprogramjának finanszírozásában, az Eurostat szervezésében számos statisztikai módszertani kutatási program folyik, melyek eredményei közvetlenül

[14] Lásd *Bauer Péter* és *Földesi Erika* szezonális kiigazításról írt tanulmányát ebben a számban (691–704. old.).

vagy közvetetten, de hasznosíthatók a standard eszközök kidolgozásában (*Szép–Trajtler* [2004]).

*A minőségértékelés eszközei*

A statisztikai termékek minőségdimenzióinak és a munkafolyamatok minőségének mérése nem öncélú tevékenység. A cél az, hogy a kapott eredmények értékelése a további feladatok tervezésének és megvalósításának alapjául szolgáljon. Az értékelést rendszeres időközönként célszerű elvégezni, előre tervezett és ismert módszerekkel és szempontok szerint.

Az *értékelés* során a statisztikai munkafolyamatot, a statisztika jellemzőit vetjük össze bizonyos elvárásokkal, és vizsgáljuk az eltérések indokoltságát. A legegyszerűbb, leggyorsabb, ha egy ellenőrzőlista kérdésein megyünk végig, és a teljesítést vagy nem teljesítést jelöljük – ilyet alkalmaznak az Egyesült Királyság Statisztikai Hivatalában, a Holland Statisztikai Hivatalban és Németországban. Az elvégzett munkafolyamat lépéseit összevethetjük az érvényben levő szabályzatok, előírások követelményeivel vagy az előre elkészített folyamatleírással, dokumentációval. A minőségi dimenziók mutatóinak értékét összehasonlíthatjuk a viszonyítási alapként választott korábbi, vagy más országokban mért, vagy normaként előírt minőségjelentés értékeivel. Az értékelés szakértelem- és időigényes feladat. Ezért leggyakrabban az *önértékelés* módszerét használják, amikor is az értékelést az adott termékért, folyamatért felelős vezető, szakértő (azaz az érintett), előzetesen kidolgozott, elfogadott szempontok szerint végzi el.

Az *auditálást* egy erre a célra létrehozott csoport végzi. Az értékelésben érintett hivatali egység felelős vezetője csak külső támogatóként kapcsolódik a csoporthoz, mivel a cél a független és objektív értékelés végrehajtása. Gyakorlati okokból és a folyamatosság érdekében célszerű, ha egy hivatali szervezeten belül számos belső auditor működik. Az auditorokat többnyire a hivatal munkatársai közül választják ki, külső cég oktatja és készíti fel őket az auditálás feladatára. A belső auditorok alkalmazásának előnye a nagyobb helyismeret és a korábbi tapasztalatok hasznosítása, a szervezeten belüli összefüggések és kapcsolatok ismerete. Az auditálási csoport munkájában külső szakértők, a tudomány, az adatfelhasználók és az adatszolgáltatók képviselői is részt vehetnek, ami növeli az auditálás eredményének szervezeten kívüli elismertségét.

Akár önértékelésről, akár auditálásról legyen szó, az értékelés célja a vezetés tájékoztatása a jelenlegi helyzetről, ami az ezekre alapuló további fejlesztések, munkatervek leírásával kiegészítve informálja a vezetést. Az értékelések célja nem a számonkérési célú hibakeresés, hanem valós helyzetértékelés a jövőbeni munkaterv megalapozására, a folyamatos fejlesztés, a jobbítás megvalósítása. Amiatt is fontos a negatív tapasztalatok, kudarcok rögzítése – amivel szemben minden résztvevőnek természetes ellenérzése van –, mert ez segítség mások számára, más területeken is a jövőben hasonló kudarcok elkerülésére.

Fontos előfeltétel, hogy az érintettek jó előre ismerjék az értékelés szempontjait, célját és az értékelési rendszert, szabályokat. A kanadai statisztikai hivatalban, és más országokban – köztük néhány új tagországban – is az a szokásos eljárás, hogy a fontosabb statisztikákra éves rendszerességgel készül önellenőrzés, és nagyobb időközönként (például ötévente) külső szakértők bevonásával átfogó auditálás.

*A minőségirányítási rendszerek*

A termékek, illetve folyamatok minőségellenőrzése hasznos, mert a végtermék minőségére koncentrál, a termékek elvárt jellemzői egyértelműen előre rögzítettek, így világosak a követelmények a termelők számára és a felhasználók számára egyaránt. A rendszer azonban nem eléggé dinamikus, nincs beépített kényszer a folyamatos fejlesztésre, és nincs közvetlen hatása a termelőszervezet felépítésére, működésére.

Azok a (köztük statisztikai) szervezetek, melyek teljesítményük javítására törekszenek, gyakran alkalmaznak különböző minőségirányítási megközelítéseket. Ilyenek például a minőség-ellenőrzés (Quality Control), a teljes körű minőségirányítás, az EFQM-modell, az ISO-szabványcsalád.

A már korábban is említett *teljes körű minőségirányítás*[15] egy olyan irányítási filozófia, amely a minőséggel kapcsolatos alapelvek megvalósítását célozza. A hangsúly a folyamatok állandó felügyeletén és javításán van annak érdekében, hogy a felhasználóknak értéket szolgáltassunk, és ebbe a szervezet minden tagját bevonjuk. Az elgondolás azonban nem ad gyakorlati útmutatást a követendő eljárásra, hanem a különböző, a következőkben példaként bemutatott minőségirányítási modellek biztosítják azt.

Az Európai Minőségbiztosítási Alapítvány (European Foundation for Quality Management) nevű magán nonprofit szervezet kifejlesztett egy TQM-alapú modellt, az *EFQM kiválósági modellt*. Az EFQM kiválósági modell a TQM szempontjai lefedésére kilenc főkritériumot és harminckét alkritériumot azonosít. A kritériumok közül öt vonatkozik a szervezetre magára, ezért ezeket „lehetőségteremtőnek" (enablernek) nevezzük. Ezek lefedik a vezetői példamutatást, a politikát és stratégiát, az embereket, a partneri kapcsolatokat és forrásokat, valamint a folyamatokat. A másik négy kritérium arra vonatkozik, amit a szervezet elér, ezek az „eredmények" címszó alá kerülnek. A felhasználók eredményei, az emberek eredményei, a társadalom eredményei és a szervezet kulcsteljesítményeinek eredményei tartoznak ide. Ez rugalmas rendszer, ami nem igényel olyan részletes dokumentációt, mint az ISO-megközelítés.

A kiegyensúlyozott stratégiai mutatószámrendszer *Balanced Score Card* (BSC), egy olyan eszköz, amelynek segítségével egy szervezet teljesítményét kiegyensúlyozott módon mérhetjük. A szervezet jövőképéből kiindulva arra keresi a választ, hogy a stratégiai célokat hogyan lehet lebontani a szervezeti egységek szintjére. A küldetés és a stratégia a teljesítményindikátorok átfogó rendszerére transzformálódik. Mindez együttesen alkotja a mérési-irányítási rendszer keretét.

Az *International Organisation for Standardisation* (ISO) ellenőriz és igazol, de nem tekinthető egy ellenőrzésen alapuló minőségi rendszernek. Az ISO minőségi filozófiájának kiindulópontja a következő: annak megfelelően cselekszik, amit írásban lefektetett, azt írta le, amit csinál. A folyamatnak átláthatónak kell lennie. Rendszeresen ellenőrzik, hogy a gyakorlat megfelel-e ezeknek az elveknek. Az ISO-modell olyan központi értékeket képvisel, mint a felhasználó-orientáltság, vezetői példamutatás és a teljes körű részvétel, folyamatirányultság és folyamatos jobbítás. Ebben az értelemben nincs nagy különb-

[15] A teljes körű minőségirányítás (TQM) fogalma a BS 4778 szabvány 2. rész 5.4 fejezete szerint: „Az összes olyan tevékenységet felölelő vezetési filozófia, amelyek segítségével az ügyfelek, a közösség igényeit és elvárásait, valamint a szervezet célkitűzéseit a lehető leghatékonyabb és költségkímélő módon lehet kielégíteni úgy, hogy végsőkig kihasználjuk azt a lehetőséget, amely az összes alkalmazottnak a folyamatos jobbításra való törekvésében rejlik".

ség az ISO és az EFQM között. A megvalósításban van a különbség. Az ISO-modellben a szervezetnek mindent hivatalosan írásos formában dokumentálni kell. Ezek teljesítését az ISO-standardoknak megfelelően rendszeresen ellenőrizni kell. Az ISO 9000 kitűzi a célt a szervezet, a termék vagy szolgáltatás, a speciális folyamatok számára. Definiálja a minőségkoncepciót és a megvalósítás irányelveit. Az ISO 9000:2000 az ISO 9000 felülvizsgált változata. Az új változat több figyelmet szentel a fejlesztésnek, javításnak, így közelebb van az EFQM modelljéhez.

*Üzleti folyamatok újratervezése* (Business Process Redesign – BPR) rendszer célja – más teljes minőségirányítási rendszerekhez hasonlóan – a jobbítás. A BPR újrakezdést jelent. A BPR nem olyan jobbítást céloz, amely az alapstruktúrát változatlanul hagyja és csak fejlesztési változtatásokat végez. Alapfilozófiája, hogy arra a kérdésre kell válaszolni miszerint: „Ha ma kezdeném újraindítani ezt a vállalkozást, jelenlegi ismereteim alapján, az adott technológia mellett, akkor milyen lenne?" Egy céget újratervezni azt jelenti, hogy félrelökjük a régi rendszereket, és mindent elölről kezdünk. A BPR és a minőség-program számos közös témával foglalkoznak. Mindkettő elismeri a folyamatok jelentőségét, mindkettő a felhasználói igényekből indul ki, és abból fejti vissza a tennivalókat. A kettő mégis alapjaiban különbözik. A minőségi program a szervezet jelenlegi keretei között dolgozik, és abban gondolkodik, hogy a jövőben is azt kell csinálni mint eddig, csak jobban. Az újratervezés áttörést jelent, nem a jelenlegi folyamatok jobbítását, hanem azok helyettesítését teljesen új folyamatokkal. A valóságban azonban nincs mindig, vagy legalábbis nem túl gyakran van szükség ilyen radikális változtatásokra, továbbá ez az alapvető újratervezés erőforrásigényes folyamat. A gyakorlatban a BPR összekapcsolása a kisebb jobbító lépésekkel jól működő kombinációt eredményezhet.

Ha nem is minden szervezet fogadja el a TQM-megközelítést, de a jó minőségű termék-előállítás, a folyamatos jobbítás, a mérés, a kísérletezés és a felhasználók bevonása általánosan elfogadott alapelvek. Minden statisztikai hivatalnak jó minőségű, megbízhatóan használható terméket kell előállítania alacsony költséggel. Ehhez javasolja a minőséggel foglalkozó LEG-csoport az EFQM kiválósági modelljét azoknak a hivataloknak, akik eddig nem alkalmaztak mást. Ezt a modellt sok statisztikai hivatalban és számos kormányzati szervnél sikeresen alkalmazzák. A modell alapkoncepciója nagyon hasonló más kiválósági modellekéhez, úgymint eredményirányultság, felhasználók a figyelem középpontjában, vezetői példamutatás, folyamatokon és tényeken alapuló irányítás, emberek bevonása és képzése, folyamatos tanulás, innováció és jobbítás, kapcsolatépítés, közös felelősség.

## 5. A STATISZTIKA MINŐSÉGE A KÖZPONTI STATISZTIKAI HIVATALBAN

Mint az előző fejezetekben láttuk a minőség sokféle szempontból vizsgálható. A magyar Központi Statisztikai Hivatal adatai és szervezetének teljesítménye általában jó minősítést kapott az elmúlt évtizedekben. Elegendő utalni az utóbbi évek szakértői értékelő anyagaira, az Eurostat szakértőinek állásfoglalásaira (*Fellegi–Ryten* [2001]. A történeti áttekintésben leírtakhoz hasonlóan a magyar statisztikai hivatal munkájában is a pontosság mérése, javítása tekinthet vissza a legrégibb múltra. A társadalomstatisztikában a rep-

rezentatív lakossági felvételek körében, a KSH munkatársai nagy tapasztalatokkal rendelkeztek. A rendszeres reprezentatív felvételeknél mind a lakossági, mind a gazdasági felvételeknél végeztek mintavételi hibaszámítást, a Háztartási Költségvetési Felvételeknél (HKF) a hetvenes évek elején ezek a kiadványok részét képezték. A nem mintavételi hibák számbavétele az utóellenőrzések keretében történt például a mezőgazdasági összeírásokat és a népszámlálásokat követően. A reprezentatív felvételek módszertani kérdései közül a megbízhatóság témakörét tekintette át *Marton Ádám* a hazai és nemzetközi publikációk alapján (*Marton* [1991]). Tanulmányában felhívta a figyelmet a gazdaságstatisztika reprezentatív felvételeivel kapcsolatos nehézségekre, valamint részletesen bemutatta a mintavételi és nem mintavételi hiba becslésének újabb módszereit. A reprezentatív felmérésekre alapozott árstatisztikák mintavételi hibájának becslésére a hatvanas évek óta publikált cikkeket, többek között *Éltető Ödön* [1959], *Marton Ádám* ([1960], [1971]), *Szilágyi György* [1988] és *Telegdi László* [1990]. Az MTA Statisztikai Bizottsága 2003-ban tárgyalta az árindexek számítási gyakorlatát, ahol Marton Ádám előadásában kitért az árindexek minőségének mérési problémáira is. Arra hívta fel a figyelmet, hogy a mintavételi hiba becslése a jó kezdetek után nem vált általános gyakorlattá a hivatalban, pedig szükség lenne a mintavételi és nem mintavételi hibák számítására, mert ezek adhatnának jó alapot a módszertan további fejlesztéséhez.

A Magyar Statisztikai Társaság 2000-ben konferenciát szervezett a „Minőség a statisztikában" címmel (*Kovács S-né* [2001]), ahol a hazai előadók a gazdaság- és társadalomstatisztika területén tárgyalták a minőség aktuális helyzetét Magyarországon, az Eurostat szakértői pedig az elméleti kérdések ismertetésén kívül, beszámoltak az Eurostatban és a tagországokban folyó minőségfejlesztési tevékenységek eredményeiről.

A rendszerváltozás következményeként a gazdaságstatisztikában pár év alatt robbanásszerűen megnőtt az adatszolgáltatók száma, a korábban gyakori teljes körű felmérések nagy részét lehetetlen volt végrehajtani. Átalakult az adatgyűjtés rendszere, megnőtt a mintavételes megfigyelések aránya. Mindez új feladatok elé állította a statisztikusokat. Erre az időszakra esett az Európai Unió statisztikai módszertanának megismerése és a hazai módszertanok harmonizálása az Eurostat előírásai szerint. *Waffenschmidt Jánosné* [2001] a területi statisztika példáján mutatta be az 1990 előtti és utáni évek statisztikai rendszerének jellemzőit és egy területi statisztikus tapasztalatai alapján hasonlította össze a statisztika minőségi kritériumait a vizsgált időszakokban. A minőség javítását szolgáló módszerek között javasolta az egész szervezet javítására irányuló szemlélet érvényesítését.

A KSH 2000-ben közétett középtávú fejlesztési stratégiájának első fejezetében a hivatal általános céljai közé sorolta, hogy az általa nyújtott statisztikai információk:

– „legyenek hitelesek, szakszerűek, és objektívek;
– legyenek tudományosan és módszertanilag megalapozottak;
– tegyék lehetővé az időbeli és térbeli összehasonlítást;
– nyújtsanak sokoldalú képet a társadalom, a gazdaság és a környezet állapotáról, a bekövetkezett változásokról…" (*A Központi* … [2001]).

A stratégia harmadik fejezetében vázolt jövőkép azt feltételezte, hogy az évtized közepére már a statisztikai információk megfelelnek a pontosság és megbízhatóság követelményeinek. „Rendszeressé, folyamatossá válik az adatok minőségének belső kontrollja", de a minőség komplex fejlesztés koncepciója még nem készült el a KSH-ban.

A továbbiakban rövid áttekintést adunk arról, hogy mi történt a KSH-ban az utóbbi években a statisztikai termékek és folyamatok minőségének fejlesztése területén.

*A statisztikai termékek minőségi kritériumai*

A statisztikai termékek első helyen említett minőségi kritériuma a *relevancia*, vagyis a felhasználók igényeinek megfelelő tartalmú statisztika előállítása. Nem véletlen, hogy a KSH vezetésének kérésre végzett hivatali átvilágítás (*Fellegi–Ryten* [2001]) eredményeképpen készült szakértői jelentésben megfogalmazott javaslatok között első helyen szerepelt a hivatal és a felhasználók közötti kapcsolat fejlesztése. Ennek érdekében szakmai tanácsadó testületek felállítását javasolták, jobb, közérthetőbb kommunikációt, a hivatal elemző tevékenységének növelését és a KSH-nak nagyobb jogkört a minisztériumi statisztikai programok véleményezésében. A vizsgálat eredményeként 2002-ben munkacsoportok alakultak az egyes területek részletesebb vizsgálatára, mint például: a KSH szerepének erősítése az államigazgatás rendszerében; a KSH központi szervezetének hatékonysága; a hivatal külvilággal való kapcsolata; a megyei igazgatóságok helyzete; a humánpolitika és a képzés a KSH-ban. A munkacsoportok bizonyos eredményeinek hasznosítása 2002-ben kezdődött el. Ennek eredményeként a KSH az utóbbi években nagyobb figyelmet fordított a statisztikát felhasználókra: elégedettségüket, adatigényeiket kérdőívvel mérték fel, főbb felhasználócsoportjaiknak és a média szakembereinek tájékoztatót tartottak a statisztika egy-egy területének módszertanáról, új fórumok alakultak a felhasználókkal való együttműködés erősítésére.

A KSH-ban egyre több területen válik gyakorlattá a reprezentatív felmérések *pontosságára*, a mintavételi és nem mintavételi hibára vonatkozó adatok publikálása, a kiadványok módszertani fejezetei kibővülnek a felvételek pontosságáról informáló tényekkel. A közölt adatok pontossági szűrése – legalábbis a mintavételi hiba alapján – azt eredményezte, hogy egyes kiadványokban a táblák részletezettsége csökkent, például a háztartás-statisztikában a hivatal a jövedelemdecilisenkénti adatközlésről áttért a jövedelemkvintilis szerinti bontásra. A többek között a nem mintavételi hibák csökkentése érdekében 1997-ben létrejött az Összeírás-kommunikációs és -képzési osztály (ÖSZKO) célja az interjú típusú adatfelvételek komplex minőségjavítása, a kérdőívek szakmai-módszertani szerkesztése, felhasználóbaráttá tétele, az egységes összeíró-kiválasztási és -képzési rendszer kidolgozása, megszervezése, illetve oktatási tankönyveinek megírása az adatfelvételek lakossági kommunikációs stratégiáinak kidolgozása.

Az utóbbi években a statisztika majd minden területén csökkent az *adatgyűjtési és feldolgozási idő*, ami elsősorban az infrastruktúra teljes megújulásának volt köszönhető. A KSH az utóbbi tíz évben előre közreadta tájékoztatási naptárát, melyhez viszonyítva elenyésző esetben volt késedelmes adatközlés. 1996 óta a KSH honlapján megjelennek az erre vonatkozó információk. A törekvés a határidők rövidítésére természetes felhasználói elvárás, de az Európai Unióban kifejezett igény, mert az Unió intézményei (különösen az Európai Központi Bank) gazdaság- és társadalompolitikai intézkedéseikhez egyre több, jó minőségű évközi és éves információt igényelnek egyre korábbi időpontra (ezzel kapcsolatosan megnőtt a különböző területek jelzőszámai iránti igény is).

A statisztikai adatokhoz és elemző anyagokhoz való *hozzáférés* területén erőteljes fejlődésnek vagyunk tanúi az utóbbi években. A hivatal szokásos papíralapú publikációi

mellett jelentősen fejlesztették az Internet által nyújtott tájékoztatási gyakorlatot. 2004 júniusától a KSH honlapján elérhető elektronikus tájékoztatási rendszerből már ingyenesen letölthetők a STADAT-rendszer valamennyi blokkja, a tájékoztatási adatbázis, az osztályozások, a nómenklatúrák, a statisztikai kérdőívek és gyorstájékoztatók. A nemzetközi gyakorlatnak megfelelően ugyancsak hozzáférhető lesz a jövőben ezen a csatornán egyre több kiadvány és adatbázis. Ugyanakkor a hivatal felelőssége megnő ezen információk értelmezése, elemzési célú felhasználása területén. (Itt kell megjegyezni, hogy sok tennivaló van még hazánkban a statisztikai műveltség fejlesztése terén.) Nagyon fontos például, hogy az elemzésekhez tartozó módszertani magyarázatok sehol se hiányozzanak, és lehetőség legyen segítség nyújtására, ha értelmezhetőségi kérdéseket vetnek fel a felhasználók.

Az Európai Unióhoz való csatlakozási folyamat egyik állandó feladata volt a statisztikánk harmonizálása a tagországok által kialakított módszertan alapján. E területen a KSH teljesítette teendőit, de az utóbbi évekre vonatkozó, térben összehasonlítható tartalmú adatok kidolgozása háttérbe szorította a hosszabb időszakra vonatkozó *összehasonlítható adatok* előállítását. Ezért bizonyos összehasonlítható tartalmú múltbeli idősorok előállítása még sok munkát igényel a tapasztalt szakstatisztikusoktól.

A statisztikák *koherenciájának* biztosítása az egyik legnehezebb statisztikusi tevékenység, mert például a különböző területeken kialakult hagyományos módszertanok összeegyeztetése országonként is nehéz feladat.

*A KSH-ban folyó statisztikai munka folyamatairól*

A Központi Statisztikai Hivatalban 1995-től több területen elkezdődött a statisztikai munka egyes eljárásainak standardizálása, a korábbi „íratlan" szabályokat felváltotta a módszertani előírások szabályzatokba foglalása. Ezek közül az egyik a módszertani füzetek megírásának szempontjaival,[17] a másik pedig a tájékoztatáshoz kapcsolódó módszertani leírásokkal[18] tartalmával kapcsolatban tartalmazott előírásokat. Mindkét anyagban követelményként fogalmazódott meg a reprezentatív felvételek minőségének értékelésére használható mutatók számítása és publikálása. A tájékoztatás módszertani előírásai között szerepel többek között: a reprezentatív felvételen alapuló adatok esetén a kiválasztási módszerek rövid leírása; az alkalmazott matematikai statisztikai módszer fő jellemzői; az adatok minősége, illetve hibája, a válaszadási arányok. Más kérdés, hogy utóbb a szabályzat gyakorlati megvalósítása elhúzódó folyamatnak bizonyult. A statisztikai munka standardizálása 2004-ben folytatódott, mely keretében elkészült a KSH egységes szezonális kiigazítási gyakorlatáról szóló szabályzat,[19] mely a folyamatos karbantartás mellett a minőség-ellenőrzés jövőbeli megvalósítását is megalapozza.

*

A KSH 1988 óta részt vesz az Eurostat különböző szervezeteiben és projektjeiben folyó minőségre vonatkozó módszertani fejlesztési munkákban. Ezek eredményeire ala-

[17] V/1995. (SK 1996. 1.) KSH szabályzat a módszertani füzetek megírásának szempontjairól. http://kshintra/intra/gyorsele/jogtar/kshkiajo/anyagok/1995_ev/V_1995/V-95-modszfuz.doc

[18] VI/1995. (SK 1996. 1.) KSH szabályzat a tájékozódáshoz kapcsolódó módszertani leírások tartalmáról http://kshintra/intra/gyorsele/jogtar/kshkiajo/anyagok/1995_ev/VI_1995/VI-95-tajek-modsz.doc

[19] I/2004. (SK 2.) KSH szabályzat a KSH egységes szezonális kiigazítási gyakorlatáról.

pozva a KSH illetékes részlege rendszeresen elkészíti az előírt standard minőségjelentéseket, amihez a Statisztikai mintavételi és módszertani osztály szakemberei nyújtanak segítséget. Így például az Iparstatisztikai főosztályon elkészültek az Éves gazdaságszerkezeti felmérés (SBS) 2000. és 2001. évi Minőségjelentései, az Életszínvonal és emberi erőforrás főosztályon pedig a Munkaerő-felmérés 2001. és 2002. évi Minőségjelentései, valamint a Háztartási Költségvetési Felvétel Minőségjelentése.

A KSH-ban, az utóbbi évtizedben megvalósított informatikai fejlesztések új technikai lehetőségeket teremtettek a statisztikai termékek minőségfejlesztésének más típusú megközelítésére. Elegendő az adattárházra, a metaadatbázisokra, a gazdasági szervezetek adatgyűjtés-szervezési rendszerére utalni. A közeljövőben mindezek jó alapot szolgáltathatnak a *statisztikai folyamatok* minőségfejlesztésére is.

Az eddigi áttekintésből látható, hogy a KSH-ban sok területen történtek lépések a statisztika minőségének fejlesztése érdekében, de ezek egymástól elszigetelt tevékenységek voltak saját kezdeményezésből, vagy az Eurostatból érkező előírás teljesítésére. Az egyes szervezeti egységek jelenleg a náluk folyó munkaszakaszok (adatgyűjtés, informatikai feldolgozás, szakfőosztályok tevékenysége) minőségének fejlesztésén dolgoznak, de a helyi optimumok nem szükségképpen eredményezik az előállított statisztikai termék minőségének optimumát, ha nincs egy egységes, összehasonlíthatóságot biztosító mérési, minőségértékelési rendszer. A közeljövő feladata a hivatalban a statisztikai minőség rendszerszemléletű megközelítése.

Több fejlett statisztikai rendszerrel rendelkező ország statisztikai hivatalának minőségfejlesztési projektjein dolgozó szakemberek álláspontja (*Marker–Morganstein* [2004]) szerint egy statisztikai hivatalban a folyamatos minőségfejlesztés sikeres megvalósításához a következő hat tényező szükséges: a felső vezetés elkötelezettsége, megfelelő szervezeti struktúra, önfenntartó működés, kiváló kommunikáció, team-munka és folyamatos fejlesztés.

A legfrissebb információk (*Eurostat* [2004b]) szerint az új tagországok közül Csehország, Lengyelország, és Litvánia statisztikai hivatalai intenzív kétoldalú kapcsolatokat létesítettek az EU korábbi tagországainak hivatalaival (Finnország, Svédország, Németország) a minőségfejlesztés kérdései komplex kezelésének kidolgozása céljából. A Litván Statisztikai hivatal 2002–2004. évi (*Strategy of Statistics*… [2004]) és az előkészületben levő 2005–2007. évi stratégiájában is külön fejezet tartalmazza a minőségirányítási feladatokat.

A Központi Statisztikai Hivatal jelenleg formálódó stratégiájában a prioritások közé került a minőségértékelés és minőségbiztosítás is. Egyrészt be kell építeni az egyes fejlesztési projektek programjába a minőségbiztosítási követelményeket, másrészt fel kell építeni a minőségbiztosítási rendszert. Az eredményekről a későbbiekben beszámolunk.

## MELLÉKLET

### A SZAKÉRTŐI CSOPORT (LEG) MINŐSÉGGEL KAPCSOLATOS AJÁNLÁSAI[20]

*1*. Minden Nemzeti Statisztikai Intézetnek/hivatalnak (NSI) minőségjelentést kell készítenie (statisztikai) termékeiről az Európai Statisztikai Rendszer (ESR) minőségdimenziói és ezek részletes bontásai szerint.

*2*. Az ESR minőségi dimenzióinak és ezek részletesebb csoportjainak mérhetőségét tovább kell fejleszteni.

[20] List of LEG on Quality Recommendations, Quality in the European Statistical System – the way forward, 1 Final report of the Leadership Expert Group (LEG) on quality, Annex 3, 24-25.

*3*. A folyamatmérések létfontosságúak a fejlesztési munka számára. Kézikönyvet kell készíteni a kulcsfontosságú folyamatváltozók azonosításáról, azok méréséről és a mérés elemzéséről.

*4*. Az ESR valamennyi szervezetének a minőség javítására rendszerszemléletű megközelítést kell alkalmaznia. Az ESR tagjainak fejlesztési munkájuk alapjaként az Európai Minőségfejlesztési Alapítvány (üzleti) kiválóság modelljét (European Foundation for Quality Management EFQM excellence model) kell használniuk, kivéve, ha már más, hasonló modellt alkalmaznak.

*5*. A NSI-knek törekedniük kell az adatszolgáltatókkal való kapcsolataik javítására, és fel kell mérni, hogy az adatszolgáltatók tisztában vannak-e a feladataikkal. Különös figyelmet kell fordítani az adatszolgáltatói terhek csökkentésére és az adatszolgáltatók körében a statisztikának a társadalomban játszott szerepével kapcsolatos ismeretek növelésére.

*6*. Az ESR tagjainak szolgáltatásiszint-szerződéseket[21] kell kidolgozniuk a fő programjaikhoz kapcsolódóan.

*7*. A felhasználói elégedettség felmérésének tervezésére, végrehajtására és elemzésére vonatkozóan projekt kidolgozását kell kezdeményezni.

*8*. Minden ESR-tagnak jelentést kell benyújtani a felhasználók-adatelőállítók párbeszédének jelenlegi helyzetéről, beleértve annak leírását, hogy a felhasználókat bevonták-e a tervezési folyamatba. Össze kell gyűjteni és az ESR-tagok rendelkezésére kell bocsátani azokat a jó példákat, amelyek a felhasználók ismereteit fejlesztik a minőség problémákról.

*9*. Az ESR legfontosabb erősségeiről és gyenge pontjairól mélyreható vizsgálatot kell végezni. E vizsgálat megállapításai alapján cselekvési programot kell készíteni.

*10*. A NSI-knek a leggyakoribb eljárásaikra ki kell fejleszteniük a jelenlegi legjobb módszereket (CBM). Készüljön egy kézikönyv a legjobb módszerek kidolgozásáról, beleértve azok megalkotását, terjesztését, végrehajtását és felülvizsgálatát. A meglevő és lényeges legjobb módszereket össze kell gyűjteni és terjeszteni kell az ESR-ben.

*11*. A statisztikák előállítására javasolt gyakorlati alkalmazásokból egy jegyzéket kell összeállítani. A munkát néhány terület legjobb gyakorlati alkalmazásainak kidolgozásával kell kezdeni, ezt követné az ESR keretében történő megvalósíthatóság tesztelése.

*12*. Az ESR-tagoknak használniuk kell a létező jó információkezelési és terjesztési gyakorlati alkalmazások jegyzékét, melyet a szakértői csoport (LEG) készít és fontolják meg a belső használatú bevezetésüket.

*13*. A jelenlegi ESR információs rendszerrel kapcsolatos felhasználói igényeket felül kell vizsgálni és az Eurostat mostani adatbázisát ennek megfelelően ki kell bővíteni. Az információs rendszer irányítási elveit ki kell dolgozni.

*14*. Az ESR módszertani és minőségi témáiról kétévente konferenciát kell szervezni.

*15*. Az adatfelvételek vezetői számára az ESR-ben ki kell dolgozni egy egyszerű önértékelési program általános feladatjegyzékét.

*16*. Át kell tekinteni a különböző szintű és célú értékelési/vizsgálati módszereket (belső, külső, egy időpontra vonatkozó, folyamatos vagy gördülő, gyors és alaposabb módszerek, mint az EFQM-értékelés) és ajánlásokat kell készíteni az ESR részére.

*17*. Az ESR tagjainak tanulmányozniuk kell munkatársaik helyzetértékelését. Ennek egyik módja, hogy felméréseket végeznek a személyi állomány véleményéről.

*18*. Az ESR tagjainak jelentésben kell elemezniük saját dokumentálásaik helyzetét. Az erről készült jelentésnek tartalmaznia kell egy akciótervet, a fejlesztések és ezen belül a prioritások ütemtervét.

*19*. Minden ESR-tagnak nyilvánosan hozzáférhető dokumentumokban kell leírnia feladatmeghatározását (mission statement), közzétételi/publikációs/tájékoztatási és a minőségpolitikáját.

*20*. A személyi állomány minden tagját képezni kell a minőségi munkára, különböző oktató programokat kell használni a különböző állománycsoportok számára. Minden ESR-tagnak ki kell dolgoznia ilyen oktató programot. Erősíteni kell az európai szintű képzést.

*21*. A hivatalos statisztika területén egy kétévente adományozható ún. minőségdíjat kell alapítani. A díj adható egy fejlesztésiprojekt-csapatnak, egy jó innovációs ötletért, egy jól működő ESR-szervezetnek, vagy egy statisztikaiprogram-csapatnak.

*22*. Szükség van egy Végrehajtási Szakértői Csoportra (LEG Implementation Group), amely összehangolja a Statisztikai Programbizottság (SPC) ajánlásai nyomán folyó tevékenységeket.

[21] Service Level Agreement (SLA): a szolgáltatás igénybevevőjével kötött olyan szerződés, amely az elvárt szolgáltatási szintet, minőséget és egyéb feltételeket, jogi kitételeket tartalmazó formában rögzíti.

FORRÁS- ÉS IRODALOMJEGYZÉK

*A Központi Statisztikai Hivatal középtávú fejlesztési stratégiája* [2001]. *Statisztikai Szemle*. 79. évf. 1. sz. 84–90. old.
*Az EFQM letölthető dokumentumai* http://www.efqm.org
BAGÓ E. [2000]: *Minőség a gazdaságstatisztikában*. A Magyar Statisztikai Társaság a „Statisztikai minősége" című konferenciájára készített előadás. Kézirat.
BECKER, R. – KOVACS, K. – DE VRIES, W. [2003]: *Official statistics, capacity and quality*. The 54th Session of ISI, August Berlin. https://www.isi-2003.de
BRACKSTONE, G. [1999]: Managing data quality in a statistical agency, Statistics Canada. *Survey Methodology*. 25. évf. 2. sz. 139–149. old.
BROECKE, M. [2000]: *Quality in statistics*. A Magyar Statisztikai Társaság a „Statisztikai minősége" című konferenciájára készített előadás. Kézirat.
DALENIUS, T. [1985]: Relevant official statistics. Some reflections on conceptual and operational issues. *Journal of Official Statistics*. 1. évf. 1. sz. http://www.jos.nu/Articles/article.asp
DOBBS, J., – GIBBINS, C. ET AL. [1998]: *Reporting on data quality and process quality*. http://www.amstat.org
ELVERS, E. [2004]: *Survey quality*. Definitions, assessments and measurements. Short Paper. http://www.std.lt
ELVERS, E. – NORDBERG, L. [2001]: Comment. *Journal of Official Statistics*. 17. évf. 1. sz. 1–20. old. http://www.jos.nu
ELVERS, E. – ROSÉN, B. [1999]: *Encyclopedia of Statistical Sciences*. Quality concept for official statistics. Update Volume 3, A Wiley-Interscience Publication, John Wiles & Sons, Inc. New York – Chiclster – Weinheim – Brisbane – Singapore – Toronto (621–629. old.) http://dsbb.imf.org
ÉLTETŐ Ö. [1959]: A reprezentatív módszerrel nyert árindex hibájának számítása. *Statisztikai Szemle*. XXXVII. évf. 2. sz. 147–163. old.
*Eurostat* [2001]. Quality in the European Statistical System – the way forward. European Commission, Eurostat, Luxembourg.
*Eurostat* [1996a]. Quality in business statistics. Eurostat/D3/Quality/96/02-final for structural business statistics.
*Eurostat* [1996b]. How to measure quality of statistics based on administrative data or estimations. Eurostat/D3/Quality/96/10rev.1.
*Eurostat* [1998]. Definition of quality in statistics. Eurostat/A4/Quality/98/General Definition.
*Eurostat* [2002]: Monographs of official statistics. Variance estimation methods in the European Union. Collection: Research in official statistics. Theme 1 General statistics. 70 old. Eurostat 2002 edition/KS-CR-02-001-EN-C.
*Eurostat* [2004a]: Regulations concerning fifth domain on statistics: Labour Force Survey, Short-term Business Statistics, Structural Business Statistics, Labour cost statistics, structural statistics on earnings. http://forum.europa.eu.int
*Eurostat* [2004b]: Minutes of last Phare SMGSC meeting November 2003.Sixth Meeting of the MGSC 18-19 March 2004 .
*Eurostat* [2003a] Standard quality report. Working Group „Assessment of quality in statistics". Luxembourg.
*Eurostat* [2003b]: Definition of quality statistics. Methodological documents, Working Group „Assessment of quality in statistics". Luxembourg.
*Eurostat* [2003c]: Glossary „Quality in statistics". Methodological documents, Working Group „Assessment of quality in statistics". Luxembourg.
*Eurostat* [2003d]: Quality report form. Working Group „Assessment of quality in statistics". Luxembourg.
*Eurostat* [2003e]: Standard quality indicators, producer-oriented. Working Group „Assessment of quality in statistics". Luxembourg.
*Eurostat* [2003f]: Handbook „How to make a quality report". Methodological Documents, Working Group „Assessment of quality in statistics". Luxembourg.
*Eurostat* [2003g]: Quality assessment of administrative data for statistical purposes. Working Group „Assessment of quality in statistics". Luxembourg.
*Eurostat* [2003h]: Quality and metadata. Working Group „Assessment of quality in statistics". Luxembourg.
*Eurostat* [2003i]: Structural business statistics – quality action plan. Working Group „Assessment of quality in statistics". Luxembourg.
*Eurostat* [2003j]: Quality report on Labour Force Statistics. Working Group „Assessment of quality in statistics". Luxembourg.
*Eurostat* [2003k]: Structural Indicators – quality profile and long-term assessment. Working Group „Assessment of quality in statistics". Luxembourg.
*Eurostat* [2003l]: Summary quality report for the Labour Cost Survey. Working Group „Assessment of quality in statistics". Luxembourg.
*Eurostat* [2003m]: Quality reporting of flash estimates. Working Group „Assessment of quality in statistics".
*Eurostat* [2003n] Foreign trade statistics – quality report. Working Group „Assessment of quality in statistics". Luxembourg.
FELLEGI, I. P. [1996]: A hatékony statisztikai rendszer jellemzői. *Statisztikai Szemle*. 74. évf. 10. sz. 789–804. old. (Rövidített fordítás a Characteristics of an effective statistical system (1996) *International Statistical Review* 2. számában megjelent cikkből.)
FELLEGI, I. P. [2001]: Comment. *Journal of Official Statistics*. 17. évf. 1. sz. 151–155. old. http://www.jos.nu
FELLEGI, I. P. – RYTEN, J. [2001]: A magyar statisztikai rendszer szakértői vizsgálata. *Statisztikai Szemle.* 80. évf. 2. sz. 107–185. old. http://www.ksh.hu
FRANCHET, Y. – GRÜNEWALD, W. [2002]: Eurostat's approach to quality. *The Statistics Newsletter*. 8. sz.
FULL, S. – JONES, N. [2002]: Towards a framework for quality measurement and reporting within the Office for National Statistics (ONS). GSS Methodology Conference. http://www.statistics.gov.uk
*Fundamental principles of official statistics* http://unstats.un.org
GIOVANNINI, E. [2003]: The OECD quality framework. *The Statistics Newsletter*. 14. sz.

*Guidelines for quality presentations*. Prepared for users of statistics. United Nations. Statistical Commission and Economic Commission for Europe, Conference of European Statisticians. Meeting on Statistical Methodology, 21-24 November 1983. (Útmutató-tervezet a statisztikai adatok köréről és minőségéről a felhasználók részére adandó tájékoztatás elkészítéséhez [1983]. Kézirat.)

HAVASI É. [2001]: A minőség fő kérdései a háztartási költségvetési felvétel megújításának tapasztalatai alapján – munka- és vitaanyag. Kézirat.

HAVASI É. – MARTON Á. [2002]: Vita a statisztika minőségéről. (A *Journal of Official Statistics*, 17. évf. 1. számában e témában megjelent cikkek ismertetése) *Statisztikai Szemle*. 80. évf. 1. sz. 67–74. old.

HOLT, T. [2000]: Measuring and managing quality: processes and issues. Paper for IMF conference on quality of statistics, Seoul, December. http://www.nso.go.kr

HOLT, T. – JONES, T. [1998]: Quality work and conflicting quality objectives. Paper for the 84th DGINS conference in Stockholm 28-29 May 1998. http://www.statistics.gov.uk

*ISO* 8402-1986, 3.5 A minőségirányítási rendszer bevezetése. *Minőségirányítás 1997* [1996]. Az Informatikai Tárcaközi Bizottság (ITB) ajánlásai. Miniszterelnöki Hivatal. Budapest. http://www.itb.hu

DR. KOVÁCS S-NÉ [2001]: Konferencia a statisztika minőségéről. *Statisztikai Szemle*. 79. évf. 1. sz. 90–95. old.

LALIBERTÉ, L. (IMF) – GRÜNEWALD, W. – PROBST, L. (Eurostat) [2003]: Data quality: a comparison of IMF's Data Quality Assessment Framework (DQAF) and Eurostat's quality definition. (Draft version) The OECD/IMF Workshop Assessing and Improving Statistical Quality of November 5-7. http://www.oecd.org

LINDEN, H. [2000]: Quality assessment of statistics in Europe. A Magyar Statisztikai Társaság a „Statisztikai minősége" című konferenciájára készített előadás. Kézirat.

LINDEN, H. – SONNEBERG, H. [2002]: Assessment of data quality for comparisons across countries: Eurostat's experiences and the Leadership Group on Quality (LEG) Recommendations. Joint UNECE/Eurostat Work Session on Statistical Metadata. 6-8 March 2002, Luxembourg.

LYBERG, L. [2003]: Definitions and measurements of survey quality. http://www.gallup-europe.be

LYBERG, L. et. al. (szerk.) [1997]: *Survey measurement and process quality*. Wiley Series in Probability and Statistics, John Wiley & Sons, Inc.

LYBERG, L. [2000]: Recent advances in the management of quality. Paper for IMF conference on quality of statistics, Seoul, December 2000 http://www.nso.go.kr

LYBERG, L. – JAPEC, L. – BIEMER, P. [1998]: Quality improvement in survey – a process perspective. http://www.amstat.org

Marker, D. A. – Morganstein, D. R. [2004]: Keys to successful implementation of continuous quality improvement in a Statistical Agency. *Journal of Official Statistics*. 20. évf. 1. sz. 125–136. old.

MARTON Á. [1961]: A reprezentatív módszer alkalmazásának néhány kérdése a külkereskedelmi árindexszámításban. *Statisztikai Szemle*. XXXIX. évf. 2. sz. 147–159. old.

MARTON Á. [1971]: A reprezentatív módszer alkalmazása a kiskereskedelmi árindexek kiszámításánál. *Statisztikai Szemle*. 49. évf. 2. sz. 167–184. old.

MARTON Á. [1991]: *A reprezentatív felvételek megbízhatósága*. KSH Könyvtár és Dokumentációs Szolgálat. Budapest.

*Minőségirányítás* [1997]. Az Informatikai Tárcaközi Bizottság (ITB) ajánlásai. Miniszterelnöki Hivatal, Informatikai Koordinációs Iroda. Budapest. http://www.itb.hu

*Monographs of official statistics* [2002]. Variance estimation methods in the European Union. Collection: Research in official statistics. Theme 1 General statistics. 70 old. Eurostat 2002 edition / KS-CR-02-001-EN-C

NANOPOULOS, P. [2001]: Comment. *Journal of Official Statistics*. 17. évf. 1. sz. 77–86. old. http://www.jos.nu

*OECD* [2004]: Main results of 2003-2004 quality reviews. http://www.oecd.org

PLATEK, R. – SÄRNDAL, C.-E. [2001a]: Can a statistician deliver? *Journal of Official Statistics*. 17. sz. 1. sz. 1–20. old. http://www.jos.nu

PLATEK, R. – SÄRNDAL, C.-E. [2001b]: Rejoinder. *Journal of Official Statistics*. 17. évf. 1. sz. 113–127. old. http://www.jos.nu

*Quality Glossary* [2004]. The American Society for Quality. http://www.asq.org

SAEBOE, H. V. – BYFUGLIEN, J. – JOHANNESSEN, R. [2003]: Quality issues at statistics Norway. *Journal of Official Statistics*. 19. évf. 3. sz.287–303. old.

*Statistics Canada Quality Guidelines* [2003]. Fourth edition, Statistics Canada. http://www.statcan.ca

*Statistics Canada's Quality Assurance Framework* [2002]. Statistics Canada. http://www.statcan.ca

*Strategy of Statistics Lithuania* 2002–2004. Statistics Lituania. Statisticos Departamentas. http://www.paris21

SZÉP K. – TRAJTLER G. [2004]: Tájékoztató az EPROS munkaértekezletéről. *Gazdaság és Statisztika*. 16. (55.) évf. 3. sz. 66–72. old.

SZILÁGYI GY. [2004]: A hivatalos statisztika alapelveinek érvényesülése és etikája. *Statisztikai Szemle*. 82. évf. 5. sz. 453–461. old.

SZILÁGYI GY. [2000]: *A statisztika minőségének néhány elméleti kérdése*. A Magyar Statisztikai Társaság a „Statisztikai minősége" című konferenciájára készített előadás. Kézirat.

TELEGDI L. ET AL. [1990]: *Az árindexek mintavételi hibájának számítása; alkalmazás a kiskereskedelmi árindexre*. Statisztikai Módszertani Füzetek, 32. Budapest.

VIRÁGH E. [2001]: Az adatfelvételek „minőségbiztosításának" feladatai: próba, szervezés, ellenőrzés. Kézirat. KSH, Budapest.

DE VRIES, W. [2000]: Solid structures. The quality of official statistics; Institutional factors, Paper for IMF Conference on quality of statistics. Seoul. December 2000 http://www.nso.go.kr

WAFFENSCHMIDT J.-NÉ [2002]: Minőség a területi statisztikában. Az MST Területi Statisztikai Szakosztályának „A statisztika fél évszázada" című konferenciáján (Keszthely, 2002. június 13-14.) elhangzott előadás szerkesztett változata a Területi Statisztika című folyóirat 2002. évi júliusi, szeptemberi és novemberi számában jelentek meg http://www.mstnet.hu

WAFFENSCHMIDT J.-NÉ [2001]: *Adatgyűjtés és az adatok minősége*. *Statisztikai Szemle*. 79. évf. 9. sz. 741–751. old.

SUMMARY

The extensive activity for the development of quality in statistics started already in the eighties and is spreading more and more in the statistical offices of different countries. In the meantime, the concept of the quality has also broadened. The great international organizations – the OECD and the IMF – have developed their own systems of quality. The Eurostat developed, by 2003, the basic documents relative to quality, along with its recommendations, and the quality requirements appeared in the case of certain surveys even in some legal acts.

We present the main elements of the quality framework on the basis of the research and recommendations of the Eurostat: the possibility of measuring the dimensions of product quality, the qualitative approach to the statistical working processes, further the methods of quality control and quality management.

The HCSO has traditionally kept in view in its statistical activity the enforcement of a number of quality aspects and the improvement of the quality was included to the strategic aims. The HCSO systematically sends to the Eurostat the quality reports but the set up of a comprehensive program for the development of quality in statistics has not yet been launched.

## SZEMÉLYI HÍREK

**Kitüntetés.** A Központi Statisztikai Hivatal elnöke *Bablina Erzsébetnek*, a Pénzügy-statisztikai főosztály osztályvezetőjének, aki közel két évtizedes statisztikai módszertani munkájával jelentősen hozzájárult a hazai kormányzati szektorra vonatkozó statisztika Eurostat által is elismert fejlesztéséhez; *Czibulka Zoltánnak*, a Népszámlálási főosztály vezetőjének, aki közel négy évtizedes tevékenysége alatt valamennyi népszámlálásban részt vett, a gyakorlati irányítási feladatok mellett a nemzetiségi statisztika országosan ismert kutatójává vált, és kiemelkedő munkát végzett az 1990. és a 2001. évi népszámlálás munkafázisaiban; *Fazekasné Kovács Katalinnak*, a Fogyasztás- és felhalmozás-statisztikai főosztály főosztályvezető-helyettesének, statisztikai főtanácsadónak, a közel három évtizedes hazai makrostatisztika fejlesztése területén végzett eredményes és magas színvonalú munkájáért; *dr. Futó Ivánnak*, az Adó- és Pénzügyi Ellenőrzési Hivatal informatikai elnökhelyettesének, a hazai számítástechnika kiemelkedő személyiségének, aki jelentős eredményeket ért el az adóbevallással kapcsolatos számítógépes, illetve internetes támogatásának kidolgozásában és bevezetésében; *Irtzl Károlynénak*, az Informatikai főosztály főosztályvezető-helyettesének, aki több mint három évtizede elkötelezett híve a statisztikai adatok hasznosításának, a külső és belső adatigények mind teljeskörűbb kielégítésének. Kezdeményezésére kezdték el a főosztályon a piaci igényeket egyre jobban kiszolgáló CD-termékek kifejlesztését és önálló terjesztését; *Kerekes Ottónénak*, a Tájékoztatási főosztály ny. osztályvezetőjének, a statisztikai elemzések, publikációk készítésében, a szakmai tájékoztatási tevékenységben kifejtett igényes, alapos és kötelességtudó munkájának elismeréseként; *dr. Lindnerné dr. Eperjesi Erzsébetnek*, az Életszínvonal- és emberierőforrás-statisztikai főosztály osztályvezetőjének, statisztikai főtanácsadónak, a munkaügyi statisztika területén tudományos szakértelemmel végzett hazai és nemzetközi tevékenysége elismeréseként; *dr. Paksi Andrásnak*, a Semmelweis Egyetem Általános Orvostudományi Kara tudományos tanácsadójának, több évtizedes tudományos és oktatói tevékenysége, valamint – a Hivatallal együttműködve – a magyar egészségügyi ellátórendszer statisztikai adatainak értékelése és elemzése területén végzett kiemelkedő tevékenysége elismeréseként; *Sándor Istvánnak*, a KSH Jász-Nagykun-Szolnok Megyei Igazgatósága igazgatójának, két évtizeden át végzett területi statisztikai tájékoztatás fejlesztése, valamint a területi gazdasági fejlődési folyamatok elemzése terén kifejtett munkája elismeréseként

*Fényes Elek Emlékérmet*

adományozott.

## SZERVEZETI HÍREK – KÖZLEMÉNYEK

A **Magyar Statisztikai Társaság** Statisztikatörténeti Szakosztálya 2004. május 17-én tartotta szakmai üléssel egybekötött tisztújító taggyűlését a KSH Nagytanácstermében. Az ülést *dr. Faragó Tamás*, a Szakosztály elnöke vezette, és a következő előadások hangzottak el: *dr. Heinz Ervin–dr. Lakatos Miklós*: A Központi Statisztikai Hivatal szerepe a német lakosság kitelepítésében; *Czibulka Zoltán*: A névjegyzékek adatai. Az előadásokat *dr. Tóth Ágnes* és *Dr. Klinger András* hozzászólása követte. A szakmai program után megtartották a Szakosztály tisztújító gyűlését. A résztvevők ismételten dr. Faragó Tamást választották meg a Szakosztály elnökévé, dr. Lakatos Miklóst pedig a Szakosztály titkárává. A vezetőségi tagok létszáma szavazategyenlőség miatt egy fővel emelkedett, így a vezetőségi tagok névsora a

következő: *Joubert Kálmán*, *Kalmár Ella*, *Kapros Tiborné*, *Marton Ádám*, *Vukovich Gabriella*.

**Látogatás az Osztrák Statisztikai Hivatalban.** *Dr. Bálint Csabáné*, a KSH EU Integrációs és nemzetközi kapcsolatok főosztály vezetője, valamint *Vándorné Gálos Katalin* 2004. május 12-én látogatást tettek Bécsben, az Osztrák Statisztikai Hivatalban. A találkozást a magyar hivatal munkatársai kérték, hogy konzultálhassanak az osztrák hivatal tapasztalatairól a csatlakozást követő EU-koordi-náció megszervezésében, különös tekintettel a tagállami működéssel kapcsolatos kezdeti nehézségekre. Az osztrák hivatal részéről – *Kutzenberger* elnök úron kívül – *Werner Holzer* főigazgató assszisztens, a Minőségbiztosítási osztály vezetője, *Reinhold Schwarzl* főigazgató-helyettes, a Makrogazdasági részleg vezetője, *Martin Bauer*, a Társadalom- és oktatásstatisztikai egység helyettes vezetője, *Brigitte Grandizs*, a Nemzetközi kapcsolatok részleg vezetője, valamint *Elisabeth Sachs*, a Nemzetközi kapcsolatok részleg munkatársa vettek részt. A magyar küldöttség megismerhette a hivatal szervezeti struktúráját és a munkák megosztását, különös tekintettel a Nemzetközi kapcsolatok részlegére. Az osztrák hivatal elnöke felajánlotta, hogy amennyiben a magyar KSH igényt tart rá, szívesen megosztják a projektmunkáról és a költségelszámolásról szerzett tapasztalataikat.

**ASA-ülés Luxembourgban.** A Mezőgazdasági Szektorelemzésekkel Foglalkozó Bizottság (Comitee for Agricultural Sector Analysis – ASA) 2003. május 6–7. között tartotta ülését Luxembourgban, ahol a KSH részéről *Laczka Éva* vett részt. Az ülésen megtárgyalták a CAP-reform (Common Agricultural Policy – Általános Mezőgazdsági Politika) és az európai bővítés hatását az agrárstatisztikára. Ezt követően a Munkacsoportok értékelésére került sor, azok vezetőinek beszámolója alapján. Számos kérdést vizsgáltak meg a különböző csoportok, és a bizottság áttekintést kapott a folyamatban lévő TAPAS-akciókról (Technical Action Plan for Improving Agricultural Statistcs – Mezőgazdasági Statisztikák Gyakorlati Cselekvési Terve), 2004 első fázisáig bezárólag.

**A Népszámlálás 2001** kiadványsorozat 17. kötete tábláinak első csoportja a háztartások lakáskörülményeinek változását több évtizedes, 1960-ig visszanyúló, számos ismérvet feldolgozó adatsorral és -kombinációkkal követi. A táblák másik csoportja a 2001. évi népszámlálás eredményeinek felhasználásával mutatja be a háztartások lakáskörülményeit, részletes, esetenként többdimenziós feldolgozásban. A szöveges értékelés összefoglalja a háztartások lakáskörülményeinek legfőbb jellemzőit, felhívja a figyelmet az időbeli változásokra, a fontosabb összefüggésekre, és részletesen kifejtett támpontot ad az adatok további értékeléséhez, elemzéséhez. A kötetet a használt fogalmak magyarázatát tartalmazó fejezet zárja.

(Népszámlálás 2001. 17. A háztartások lakáskörülményei. Központi Statisztikai Hivatal. Budapest. 2004. 226 old.)

Megjelent a *Gazdaság és Statisztika* című folyóirat 2004. évi júniusi száma.

# STATISZTIKAI IRODALMI FIGYELŐ

## KÜLFÖLDI STATISZTIKAI IRODALOM

### A STATISZTIKA ÁLTALÁNOS ELMÉLETE ÉS MÓDSZERTANA

DASHEN, M. – FRICKER, S.:

NYITOTT KATEGORIÁLIS KÉRDÉSEK HATÁSA AZ ADATOK MINŐSÉGÉRE

(Understanding the cognitive processes of open-ended categorical questions and their effects on data quality.) – *Journal of Official Statistics*, 2001. 4.sz. 457–477. old.

Általánosan elterjedt az a nézet, amely szerint az, ahogyan az emberek értelmezik a kérdéseket, hatással van arra, hogy milyen módon jutnak el a válaszhoz.

Az adatfelvételek során gyakran alkalmaznak kategoriális kérdéseket, mivel így mérsékelhetők a kérdezett terhei, és időt lehet megtakarítani. Ha túl sok eldöntendő kérdést teszünk fel, a kérdezett hajlamos gyakran nemmel válaszolni. Ha kategoriális kérdéseket teszünk fel, ezzel csökkenthetjük a „nem" válaszok arányát.

Az elemeket gyakran a felhasználó, és nem a kérdezett szempontjai szerint csoportosítják, ami félreértésekhez vezethet. A bizonytalanság csökkentése és az adatok minőségének javítása érdekében gyakran listát is mellékelnek a kategoriális kérdések mellé. Ebben az esetben a kérdezettek általában azt feltételezik, hogy ami nem szerepel a listán, az nem is tartozik az adott kategóriába. Ez a módszer tehát csak akkor működik jól, ha a lista tartalmazza az összes lehetőséget.

Bár egy ilyen lista kétségkívül hasznos, nem minden esetben alkalmazható (például telefonos kérdezésnél). Érdekes módon, a szemtől szembe kérdezés során vagy önkitöltős kérdőíveknél is gyakran találkozunk azzal a szemlélettel, hogy nem mellékelnek listát az ilyen típusú kérdések mellé, mondván, hogy az megnöveli a válaszadás időtartamát.

A lista hiánya alkalmassá teszi a kategoriális kérdéseket arra, hogy feltárjuk, hogyan hat a kérdezett értelmezése az adatok minőségére. Lista nélkül a kérdezettek saját megítélésükre és tapasztalataikra hagyatkoznak a kategoriális kérdések értelmezésében. A válaszadók gyakran bizonytalanok abban, hogy az a kritérium, amit ők a beletartozás kritériumaként meghatároztak, vajon helyes-e vagy sem. Ha a kritérium nem helyes, akkor nem megfelelő elemeket is beleérthetnek a kategóriába (hamis pozitív válaszok), és kizárhatnak odatartozó elemeket (hiányzó válaszok), amelyek egyaránt hatással lehetnek az adatok minőségére.

A nyitott kategoriális kérdéseket gyakran használják a különböző adatfelvételek során, és sok kutatás vonatkozik arra, hogyan használják az emberek a listát, hogy világosabbak legyenek számukra a kategóriák. Alig van arra vonatkozó kutatás, miként értelmezik az emberek az ilyen típusú kérdéseket lista nélkül. Az ismertetett tanulmány ezt a hiányt igyekszik pótolni.

A cikk a ruházattal, élelmiszerrel és számítógépekkel és tartozékaikkal kapcsolatos vásárlási szokásokra vonatkozó telefonos interjúk során általánosan alkalmazott kategóriákkal foglalkozik. A besorolási kritériumok megállapításánál a szerzők három elméletre koncentráltak: a fizikai hasonlóság, valamilyen lényeges tulajdonság (essence), illetve valamilyen cél (goal) alapján való besorolásra. (Például, ha a „női ruhák" kategóriát nézzük, akkor a külső hasonlóság alapján ide sorolható minden olyan ruha, amely egy felső részből és egy hozzá tartozó alsó

*Megjegyzés.* A *Statisztikai Irodalmi Figyelő* rovatot a Központi Statisztikai Hivatal Könyvtár és Dokumentációs Szolgálat állítja össze. A rovat minden hónapban *Külföldi Statisztikai Irodalom* fejezetet (külföldi statisztikai és demográfiai könyvek és cikkek ismertetését *Rettich Béla* szerkesztésében), páratlan hónapban általában *Bibliográfiát* (a könyveket az MSZ 3423/2–84, az időszaki kiadványokat az MSZ 3424/2–82 szabvány szerinti feldolgozásban), páros hónapokban *Külföldi folyóiratszemlét* tartalmaz.

szoknya részből áll, de önmagában egy alj nem tartozik bele. Ha a besorolás alapja egy belső tulajdonság megléte, a női ruhák esetében ez a tulajdonság lehet az, hogy valamilyen formális alkalomra való ruha, így például ide tartozik a kosztüm mint munkahelyi viselet, a koktélruha, a színházi ruha stb., de nem tartozik ide a pólóruha.) A célorientált gondolkodás két típusát különböztették meg. A „valamivel együtt jár" (to accompany) típusú gondolkodás esetén a „kávé" kategóriába sorolhatják a cukrot, a tejet, a kiskanalat stb., mert ezek kapcsolatban vannak a kávéval. A „készítéshez szükséges" (to make) típusú gondolkodás esetén ugyanennél a kategóriánál megjelenhet a filter, a víz, a babkávé stb., mert ezek a kávé elkészítéséhez kellenek. Ezek az elméletek megfelelően dokumentáltak, és jól alkalmazhatók ebben az esetben.

A tanulmány célja kettős: egyrészt arra keres választ, hogy a kérdezettek szisztematikusan választják-e ki az adott kategóriába tartozó elemeket, másrészt annak lehetőségeit keresi, hogyan lehet csökkenteni a hibák előfordulásának valószínűségét.

Két vizsgálat eredményeit mutatják be a szerzők. Mindkét vizsgálat során önként jelentkezőkkel dolgoztak, akiknek a részvételért fejenként 25 dollárt fizettek. A résztvevők átlagéletkora 49 év volt az első vizsgálatban, 45 év a másodikban, és mindkét vizsgálat résztvevői átlagosan 16 osztályt végeztek.

Az első vizsgálat során arra kérték a résztvevőket, hogy egy adott kategóriához (például kávé) tartozó minden elemet írjanak le, ami eszükbe jut, és indokolják meg, hogy miért gondolják odatartozónak. Például, egy kérdezett mondhatja azt, hogy a „tejszín" a „kávé" kategóriába tartozik, mert ő mindig tesz tejszínt a kávéjába (ez a „valamivel együtt járó" típusú válaszok közé tartozik, mivel a kérdezett azért sorolta ide, mert a kávéjával együtt mindig fogyaszt tejszínt is.

Önkitöltős módszerrel dolgoztak, és nem korlátozták a kérdezettek rendelkezésére álló időt. A kapott eredményeket megvizsgálták abból a szempontból, hogy mennyire egyezett a kérdezés tervezőinek az elképzelésével, illetve megnézték, hogy a válaszadók milyen indoklás alapján sorolták be az elemeket egy-egy kategóriába.

Az eredmények azt mutatták, hogy a kapott válaszok nagymértékben eltértek a várttól, tehát a kérdezettek nem a kérdezők, illetve a kérdőív tervezőinek szándékai szerint értelmezték a kérdéseket, ezek az eltérések azonban nem voltak véletlenszerűek.

A válaszok indoklása szerint kialakított négy kategória némileg eltért az előzetes elképzelésektől: a kérdezettek a besorolásnál egyáltalán nem alkalmazták a fizikai hasonlóság elvét, leggyakrabban pedig a szűken vett, szó szerinti (literal) értelmezés alapján való besorolás fordult elő (ez utóbbi esetben például a „kávé" kategóriába olyan elemek tartoznak, mint koffeinmentes kávé, az instant kávé, a presszókávé stb.). A szó szerinti értelmezés vezetett leginkább az eredeti elképzeléseknek megfelelő eredményekhez. A kérdezettek 64 százaléka, vagyis közel kétharmada egyetlen módszert használt, amikor a kategóriákhoz tartozó elemeket felsorolta.

A második vizsgálat során – az első vizsgálat eredményeire támaszkodva – négy csoportot alakítottak ki az előző válaszadási típusok szerint, és a résztvevők feladata az volt, hogy előre megadott szempont szerint sorolják fel a kategóriához tartozó elemeket. Összességében nagyon hasonló eredményeket kaptak, mint az első vizsgálatban, ez pedig azt jelenti, hogy az előző vizsgálatban kapott eredmények megbízhatók és valóban bepillantást engednek abba, hogy az emberek hogyan értelmezik a kategóriák elnevezését.

Ennek a két vizsgálatnak nagyon fontos gyakorlati következményei vannak: világossá vált, hogy ugyanazon kérdések eltérő értelmezése valóban mérési hibát okozhat. Tehát, ha a kategóriák megnevezését (titles) nem kíséri egy rövid instrukció arról, hogy hogyan kell értelmezni, akkor valószínűleg nem lesz konzisztens az eredmény. Ez a munka segítséget nyújthat az instrukciók megfogalmazásában, azzal, hogy egyszerre vizsgálja a válaszokat és indoklásukat, így kiválasztható és megfogalmazható az a módszer, amely a leginkább kívánatos eredményhez vezet.

A jövőben azt is meg kell vizsgálni, hogy három fontos tényező – a felmérés célja, a válaszadó szakértelme és a felvétel formája – hogyan befolyásolja a kérdezett válaszadását. Egy adott kategória értelmezése függ a felmérés észlelt céljától (például, ha a „kávé" egy droghasználatra vonatkozó kérdőívben szerepel, aligha a tejszínt fogják társítani hozzá). A válaszadó szakértelme nyilván befolyásolja a felsorolt elemek számát és típusát (például a „fotográfiai felszerelés" kategóriába egy profi fotós valószínűleg több különböző típusú kamerát, lencsét és filmet sorol fel, mint egy amatőr fotós). Az adatfelvétel körülményei – például, ha siet a kérdezett, ami a telefonos interjúknál elég gyakran előfordul – szintén befolyásolják a válaszadást. E három tényező együttes figyelembevétele hozzásegíthet annak még árnyaltabb megértéséhez, hogy az emberek hogyan értelmezik a nyitott kérdéseket, és ezáltal hozzájárulhat az adatminőség javítására szolgáló hatékony módszerek kidolgozásához is.

(Ism.: *Földházi Erzsébet*)

ZÜHLKE. S. – ZWICK, M., – SCHARNHORST, S.:

## KUTATÁSI CÉLÚ MIKROSTATISZTIKAI ADATOT SZOLGÁLTATÓ NÉMETORSZÁGI HÁLÓZAT

(Die Forschungsdatenzentren der Statistischen Ämter des Bundes and der Länder.) – *Wirtschaft und Statistik*, 2003. 10. sz. 906–911. old.)

A tudományos kutatás sokrétű statisztikai adatokat igényel a vizsgált komplex gazdasági, társadalmi folyamatokról, és ehhez az informatika megfelelő feltételeket nyújt. Korunkban a gazdasági és társadalmi elemzés sokféle csoportra vonatkozóan használja fel a lehető legrészletesebb információkat, hosszabb idősorok formájában is. A hagyományos, főleg nyomtatott táblákra épített statisztikai adatközlések ilyen célokra nem elegendők. A tudományos, valamint szakmai elemzéseket végzők sajátos módszertani, tartalmi előírásaiknak megfelelő statisztikai adatokat igényelnek.

A német statisztikai szolgálat olyan anonim jellegű mikroadatok felhasználását is lehetővé teszi, amelyekből nem lehet visszakövetkeztetni az adatszolgáltatókra. Az oktatás és a kutatás német szövetségi minisztériuma által felkért szakmai testület 2001-ben ajánlást állított össze (Wege zu einer besseren informationellen Infrastruktur. Baden-Baden, 2001), és a javaslatok között szerepelt a statisztikai adatokat szolgáltató kutatóközpont létesítése (Forschungsdatenzentren). A hivatalos statisztikai szolgálat 2001-ben létesített ilyen központi kutatóhelyet, majd a helyi kutatóhelyek a német szövetségi tartományi statisztikai hivatalokban is megkezdték működésüket, 2002 márciusától. Ilyen szervezeti egység az ország minden tartományában, országosan tizenhat helyen működik.

A cikk történelmi visszatekintést ad arról, ahogy a kutatók korábban hivatalos adatokhoz juthattak a statisztikai törvény előírása szerint. A mikroadatok tudományos célú kiadásának szabályai nem szerepeltek kiemelten a statisztikai törvény első, 1953-as szövegében, az érdekeltek aligha vetetettek fel akkor ilyen lehetőséget. Viszonylag ritkán mutatkozott ilyen adatigény, egészen a hetvenes évek közepéig. A kutatók az 50-es, a 60-as, valamint 70-es években úgy érhették el a formai módon anonimmá tett statisztikai adatokat, hogy azokon nem módosítottak, azonban leválasztották az adatok egyedi azonosítóit (az átadás például név, cím, személyazonosító szám nélkül történhetett). Ilyen adatháttérrel a hetvenes években kiterjedt elemzések készültek, ezek alapozták meg a szociálpolitikai döntések átfogó mutatóit is (Sozialpolitisches Entscheidungs- und Indikatorsystem – SPES).

A szerzők utalnak arra, hogy a hivatalos statisztika a mikrocenzus, valamint a mintavételes háztartási jövedelemfogyasztási felmérés alapján szolgáltatott anonim mikroadatokat az ilyen kutatásokhoz. A népszámlálási törvény előkészítéséhez a kutatóknak átadták az 1970-es németországi népszámlálás hasonló módon anonimmá tett mikroadatait.

A német szövetségi adatvédelmi törvényt 1977-ben fogadták el, figyelembe véve az informatika gyors fejlődéséből eredő kockázatokat. A szövetségi statisztikai törvény következő, 1980-ban elfogadott változata is pontosabb előírásokat tartalmaz, az adatvédelmi szabályokkal összhangban. Büntetőjogi következményekkel fenyegeti azt, aki megsérti a személyes adatok védelmének szabályait. Az 1980-as német statisztikai törvény (11. § (5) bekezdés) az adatkezelés szabályait is kifejti, ennek alapján a hivatalos mikroadatok elérhetősége is tisztázott.

A szerzők kiemelik az abszolút anonim adatok kedvezőtlen (az információvesztésből eredő) hatásai kapcsán, hogy itt két alkotmányos jog érvényesülését kell vizsgálni. Az egyik az információs önrendelkezés, a másik a tudomány szabadságának joga. Az anonim adatok hatékonyabb elérése csak a statisztikai törvény 1987-ben elfogadott szabályai szerint teremt összhangot e két jog között.

A németországi népszámlálással kapcsolatos 1983-as alkotmánybírósági döntést is figyelembe vette az a törvényi előírás, amely biztonságosan nem zárja ki ugyan az érintett azonosításának visszaállítását, de a mikroadat kiadását engedi, ha az ilyen célú adatok összekapcsolása csak aránytalanul nagy ráfordításokkal lenne megvalósítható (ez a „ténylegesen anonim adat” fogalma). A kilencvenes évek közepén alakultak ki a ténylegesen anonim mikroadatok átfogó előírásai, szabványai, és azóta a szigorú feltételeket teljesítő kutatók átvehetik a törvény előírásainak megfelelően kezelt egyedi információkat is, például a háztartási felvételekből.

A cikk megállapítja, hogy a gazdaságstatisztika terén sokkal kevésbé megoldott a mikroadatok kiadása, és kiemeli a sajátos akadályokat. Több előterjesztés vizsgálta a kilencvenes években a megoldás lehetőségeit, és ezek között szerepelt a tudományos kutatáshoz szükséges statisztikai adatok átadására jogosított kutatóhelyek kialakítása. A mikroadatok csaknem 90 százalékával a szövetségi tartományok statisztikai hivatalai rendelkeznek, amelyek az adatok feldolgozását, tárolását végzik. Tekintettel az egynél több tartományt (például Németország egészét) átfogó elemzésekre, bármely helyi adatállomáson elérhetővé tették a teljes országos adatállományt.

A kutatók áttekinthetik a vizsgált adatkör minőségi jellemzőit, az adatok gyűjtésének, feldolgozásá-

nak módszertanát. Ehhez az internetre alapozott metaadatbázis-rendszer áll rendelkezésükre a kutatóhelyen, ahol szigorú jogosultsági szabályok szerint érhetők el a mikroadatok, és ezzel az anonim jelleg (túl az adatazonosítók leválasztásán) garantálható.

A szerzők, az „abszolút" és a „ténylegesen" anonim mikroadatok szabályai mellett kifejtik a konkrét kutatási célhoz kötött helyi lekérdezésekre, valamint az ellenőrzött távolsági adatfeldolgozásokra vonatkozó adatvédelmi előírásokat.

Az aggregált adat abszolút anonim jellegű. Akkor is teljesül ez a feltétel, ha az egyedi adat minden azonosítóját eltávolítják. Az ilyen közhasználatú adathoz (public use file – PUF) bárki hozzáférhet. Ilyen formában érhetők el például a németországi időmérleg-felvétel, valamint a szociális segélyezés statisztikai adatállományai. További adatállományokat is elérhetnek majd a kutatók közhasználatú lekérdezéssel, különös tekintettel a felsőoktatás adatigényeire. A szerzők utalnak az ilyen (ún. „campus-file") fejlesztésekre, azzal a kifejezett céllal, hogy hallgatók sokféle statisztikai adatra alapozhassák elemzéseiket, például az 1998-as mikrocenzus alapján.

A kutatók az előbbinél kisebb információveszteséggel használhatják fel a ténylegesen anonim mikroadatokat, amennyiben az ismérvekhez tartozók megállapítása aránytalanul nagy idő-, munkaerő- és költségfelhasználást igényel. Ilyen ténylegesen anonim adat (szakkifejezéssel a scientific use file – SUF) csak tudományos kutatási célokra szolgáltatható a hivatalos statisztikákból. A cikk utal azokra a statisztikai eljárásokra, amelyekkel elfogadható mértékű információveszteséggel elvégezhető a ténylegesen anonim adatok előkészítése.

Megfelelő védőintézkedéseket is tesznek, hogy meggátolják az azonosítások helyreállítását, például más adatforrásokból nyerhető kiegészítő információkkal. A kutatók leggyakrabban a mikrocenzus, a mintavételes jövedelem- és fogyasztás-felvételek, valamint az időmérleg-felvétel mikroadatait igénylik. Ezt az adatkínálatot – kutatóhelyek fejlesztései révén – fokozatosan bővítik, például meghatározott munkaügyi (bér- és kereset-) felvételekkel, egészségügyi statisztikákkal. Olyan bér- és keresetfelvételt is elérhetővé tesznek ilyen ténylegesen anonim adatokkal, amelyben a munkaadók és a munkavállalók összekapcsolhatók (employer-employee files).

Az előbbi ténylegesen anonim adatállományok viszonylag nagy ráfordításokkal alakíthatók ki, ezért a kevésbé igényelt, vagy csak körülményesen átalakítható adatkörökre egy harmadik megoldást alkalmaznak. Az anonim jellegű feldolgozás itt célhoz kötött, meghatározott konkrét kutatási területre vonatkozik, így csak bizonyos csoportosító ismérvek szerint nem azonosíthatók a felmérés eredményei. Ez is ténylegesen anonim adatokat eredményez, de nem a szabvány minden előírása szerint. A jogosult vendégkutató csakis a hivatalban érheti el a kért mikroadatokat.

Ez a helyhez kötöttség gyakorlatilag kizárja, hogy az azonosításhoz kiegészítő információkat is felhasználhassanak. Az ilyen helyben elérhető mikroadatokban több eredeti információ megőrizhető, mint az előbbi két adatalakítással. A vendégkutatók jelenleg Berlinben, Bonnban, valamint a Hivatal székhelyén Wiesbadenben vehetnek igénybe ilyen helyi szolgáltatást, azonban más tartományi hivatalokban is terveznek ilyen kutatóhelyeket. Közismert számítógépes alkalmazások (SAS, SPSS) vehetők igénybe a helyben végzett elemzésekhez.

A cikk negyedik lekérdezési lehetőségként ismerteti a távoli kutatóhelyeknek átadott mikroadatokat. A hálózaton átvehető, formai módon anonim jellegűvé tett adatokat a felmért egyedi adatoktól elválasztva tárolják. Az adatot igénylők olyan alkalmazásokat dolgoznak ki (syntax-script), amelyek megfelelnek a vizsgálati eljárásuknak. A felkért hivatalos statisztikai szervezet ezeket felhasználva állítja össze az igény szerinti adatállományokat. Itt is a közismert (SAS, SPSS) programokkal történhet a hozzáférés a statisztikai adatbázisokhoz, de a kutató közvetlenül nem tekinthet meg egyedi adatot.

A „scientific use files" lényegében csak meghatározott kutatói kör számára elérhető, viszont az ilyen távoli lekérdezésekre bármelyik kutatóhely igényt tarthat, akár országhatáron túlról is. Nem tudományos célokra is igényelhetők a mikroadatok. A lekérdezés előkészítéseként a Hivatal nyilvánosságra hozza adatbázisainak szerkezetét, és a kutatók erre alapozva dolgozzák ki lekérdező programjaikat. A metainformáció (az adattartalom nélkül) tartalmazza az adathelyek leírását, az adatkörök struktúráját és kódjeleit, az adatok hosszát és jellegét stb.

A távoli kutatóhely jelenleg olyan átalakított adatokhoz juthat, amelyek a mikrocenzus, valamint a bér- és keresetstatisztika felméréseire alapozottak. Az ilyen, hálózaton igényelt mikroadatok feldolgozása és ellenőrzése viszonylag nagy munkaigénnyel jár, hogy az adatvédelmi előírásokat betarthassák. A szerzők utalnak a jelenleg kézzel végzett műveletek automatizálását célzó fejlesztésekre és követendő célként említik a dániai on-line adatfeldolgozások szolgáltatásait, valamint a Luxemburgból származó mikroadatokkal végzett (Luxembourg Income Study, Luxembourg Employment Study – LIS/LES) munkaügyi elemzések internetes adathátterét. A fejlesztés célja, hogy a jelszóval rendelkező távoli kutatók automatikus adatértékeléseket indíthassanak. A fel-

dolgozás eredményeit (mint a dán statisztikai szolgálat kínálatában) a szerver e-mail útján, automatikusan továbbítja a távoli kutatóhelyre.

A kutatók, az előbbi adatlekérdezés megoldásával megbízhatják a statisztikai szolgálatot, külön térítés ellenében. Az igényelt mikroadatok körében a felek állapodnak meg, és az ehhez szükséges feldolgozások programjait a statisztikai szolgálat készíti el. A konkrét előírásokat követő feldolgozások eredményeit a titoktartási szabályok betartásával adják ki, anélkül, hogy a felhasználó bármilyen kapcsolatba kerülne a mikroadatokkal.

A cikk hivatkozik a 2002 nyarán végzett felmérésre, amely keretében a kutatók véleményét kérték az igényelt mikroadatok fontosabb felhasználási területeiről. Összesen 700 választ értékeltek, ebből mintegy 600 potenciális felhasználó szerint szükségesek a hivatalos statisztika mikroadatai a kutatási tevékenységeikhez. Viszonylag széles adatkört vázoltak fel a kutatóktól beérkezett igények alapján, és ezek szerint jelölhetők meg a fejlesztések középtávú programjának prioritásai. Alig mutatkozott érdeklődés a vendégkutatók helyi munkavégzése iránt, a többség a saját munkahelyén kívánja elérni a mikroadatokat, vagy abszolút, vagy ténylegesen anonim jelleggel. A mikroadatok távoli kutatóhelyekről elérhető on-line szolgáltatásaira is viszonylag kevesen mutattattak igényt.

(Ism.: *Nádudvari Zoltán*)

## TÁRSADALOMSTATISZTIKA – DEMOGRÁFIA

CLARKE, J.– SALT, J.:

### A MUNKAENGEDÉLYEK ÉS A KÜLFÖLDI MUNKA AZ EGYESÜLT KIRÁLYSÁGBAN

(Work permits and foreign labour in the UK: a statistical review.) – *Labour market trends*, 2003. 11. sz. 563–574. old.

A munkaengedélyek rendszere a munka céljából történő bevándorlás szabályozásának fő mechanizmusa az Egyesült Királyságban az elmúlt években több új sémával egészült ki. Nagy-Britanniában először az első világháború idején korlátozták a külföldiek munkavállalását. 1919-20-ban szabályozták a Brit Nemzetközösséghez nem tartozó országokból érkezettek foglalkoztatási feltételeit. Ezután, egészen az 1962-es Nemzetközösségi Bevándorlási Törvényig semmilyen formában nem ellenőrizték a Nemzetközösségből érkező külföldiek munka céljából történő bevándorlását.

1945 után a háború sújtotta európai gazdaság újjáépítéséhez Nagy-Britanniában is nagy szükség volt munkaerőre. 1945 és 1950 között mintegy 170 ezer Kelet-Európából érkezett menekültet foglalkoztattak, rajtuk kívül 136 ezer nyugat-európai érkezett ebben az időben munkaengedéllyel. A munkaengedélyek kibocsátása a második világháború után ingadozó képet mutatott. A hatvanas évek végéig növekvő trend érvényesült. Az ötvenes években a legtöbb engedélyt képzetlen, vagy kevéssé képzett munkások kapták, akiket a hazai szolgáltató ágazatában foglalkoztattak. 1950 és 1955 között a nevelőnőknek kiadott engedélyek száma megduplázódott, 2400-ra növekedett.

1972. január 1-jétől az Európai Gazdasági Közösségen kívülről érkező képzetlen és kevéssé képzett munkaerő nem kapott munkaengedélyt. Az 1971-es Bevándorlási Törvény tovább szigorította az ellenőrzést azzal, hogy a Nemzetközösség polgárai számára ugyanolyan feltételek esetén biztosított munkaengedélyt, mint az Európai Gazdasági Közösségen kívülről érkezők számára. Ezután csak speciális foglalkozásúak és képzettségűek juthattak engedélyhez.

Az ezt követő időszakban jelentősen csökkent a munkaengedélyek kibocsátása és a nyolcvanas évek elején 15 ezret tett ki. 1982-ben a hosszú távra szóló engedélyek száma csak 5700 volt, majd a nyolcvanas évek közepétől emelkedett a kibocsátás, mely 1990-ben 30 ezer engedélyt adtak ki. 1994 után hirtelen emelkedés következett be, 2002-ben a második világháború után a legtöbb, 129 041 kérvényt hagytak jóvá.

A munkaengedélyek kibocsátásának mai rendszere szélesebb körű, mint bármikor. A 2001-es választások után az ügyintézés az Oktatási és Szakképzési Minisztériumtól átkerült a Belügyminisztériumhoz, ami azt jelentette, hogy a többi migrációval kapcsolatos kérdéssel együtt kezelik. Az állami politika célja, hogy a gazdasági migrációnak azt a részét támogassa, amely a hazai munkaerő-kínálatból hiányzik. Ezt szolgálja a munkaengedélyek rendszere, mely a két fokozatú fő séma kiegészítéseként új programokat vezetett be a szakmák spektrumának mindkét végén. A rendszernek négy fő eleme van:

– a rendszer fő része (mely magába foglalja a munkaengedélyeket, az első engedélyeket és az Oktatás és Munkatapasztalatok sémáját);
– a szezonális mezőgazdasági dolgozók rendszerét;
– a szektorok szerinti rendszert;
– a magasan képzett be- és kivándorlók programját.

A rendszer fő része alapján az 1971-es Bevándorlási Törvény szerint adott munkáltatónak, adott foglalkoztatott részére és adott munkakörre biztosították az engedélyt. Az 1981-es felülvizsgálat csak kisebb módosításokat eredményezett, melyek a gazdasági recessziót, a magas munkanélküliséget igyekeztek csökkenteni. Az 1980-as évek javuló gazdaságában a politika egyre inkább a vállalatokat kezdte támogatni az általános foglalkoztatás helyett. A továbbképzésbe fektetett összegek csökkenése erősítette a versengést a munkáltatók részéről bizonyos szakmák iránt, és egyúttal a munkaerő mobilitását is növelte, hazai és nemzetközi szinten egyaránt. Ezzel egyidejűleg egyre nagyobb jelentősége lett a szektorok között a szolgáltatásoknak, melyek többsége globálisan működik. Ezt a folyamatot gyorsította a pénzügyi dereguláció, ami nagy keresletet támasztott a pénzügyi munkahelyek betöltéséhez szükséges munkaengedélyek iránt. A bizonyos szakmákért folytatott nemzetközi verseny élesedésével nagyobb rugalmasságra és mobilitásra lett szükség a munkaerőpiacon. Úgy tűnt, hogy a munkaengedélyek rendszere akadályokat gördített a külföldi munkaerő alkalmazása elé és költségessé tette azt. A rendszer 1989-ben megkezdett újabb felülvizsgálata 1991 októberétől már jelentős változásokat hozott. Bevezették a kérvények feldolgozásának kétfokozatú rendszerét. Az első fokozatban egy egyszerűsített eljárás keretében hagyták jóvá azokat a kérvényeket, amelyek egyértelműek voltak és kielégítették a szükséges feltételeket. Ilyenek voltak például a felsővezetői posztok, vagy azok, amelyekben munkaerőhiány mutatkozott. A többi kérvény jogosságát a második fokozatban kellett ismételten igazolni a külföldi jelöltek alkalmazásához. Bevezették a „kiemelten fontos alkalmazott" kategóriát, ami a magas szintű, nyelvekkel és kultúrával kapcsolatos foglalkozásokat jelentette. A változások fő célja az volt, hogy könnyebb legyen a magasan képzett külföldi vállalati szakemberek átvétele.

A növekvő nemzetközi verseny az Egyesült Királyság gazdaságát is versenyképességének fokozására késztette. A bevándorlási törvény 2000-es felülvizsgálata egyszerűbb, átláthatóbb, költségtakarékosabb rendszert eredményezett, mely rugalmasabb és jobban megfelel a munkáltatók kívánalmainak. A kiemelten fontos kategória bekerült a rendszer fő részébe. A változások legfontosabb következménye az volt, hogy a kérvények elintézése hetek helyett csak napokat vett igénybe, ami jelentős előnyt jelentett Nagy-Britannia számára. Egy másik jelentős változás, az ún. szektorpanelek létrejötte, melyet a Munkaengedélyek Hivatala valamint a munkáltatók és a szakszervezetek képviselői működtetnek, és amelyek feladata a szakmahiányok természetének és mértékének figyelése és megfelelő javaslatok kidolgozása. Mivel a munkaengedélyekről szóló statisztikai adatok gyűjtésének és feldolgozásának módszere 1995-től jelentősen megváltozott, a statisztikai elemzés időszakának kezdete is ez az év lehet. Az összes kérvény magába foglalja a munkaengedélyek (a munkaadóknak adott engedély, adott külföldi személyek foglalkoztatására), a meghosszabbítások, az első engedélyek (a már Nagy-Britanniában élő, de munkaengedéllyel nem rendelkező külföldi dolgozóknak adott engedély), a foglalkozásváltozások és a mellékfoglalkozások iránti beadványokat. Számuk 1995-től 2002-ig minden évben jelentősen nőtt, és a hét év alatt 38 617-ről 155 216-ra emelkedett. A legnagyobb növekedés 42 százalékos volt és 1999-ről 2000-re következett be az információs és kommunikációs technológia boomjának köszönhetően, illetve a betöltetlen orvosi állások miatt. A jóváhagyások trendje az adott időszak alatt nagyjából megfelel a kérvények változásainak. A jóváhagyási ráta 83 (2002) és 92 százalék (2000) között mozgott. A jóváhagyott beadványok száma évről évre növekedett, az 1995-ös 32704-ről a 2002-es 129 041-re, a legnagyobb arányban, 47 százalékkal 2000-ben.

Emelkedett a visszautasítások száma is: 4811-ről 13 773-ra a hét év során. A 185 százalékos emelkedés jóval alacsonyabb, mint ami az összes kérvény és a jóváhagyott kérvények számában következett be. Ebből következik, hogy a visszautasítások arányában csökkenő trend mutatkozott (1995-ben 13, 2001-ben 6 százalék). 2002-ben viszont az összes kérvény 9 százalékát utasították vissza. A visszautasítások okai között a legnagyobb csoportot az egészségügyi okok alkotják. Az összes jóváhagyás 71 százaléka hosszú távú volt, több mint egy évre szólt, 29 százaléka pedig rövidebb időszakra.

A rendelkezésekre álló legfrissebb adatok szerint 2003 január és augusztus között 112 462 kérvényt adtak be, melynek 82 százalékát hagyták jóvá. Ha ezeket az adatokat átszámítjuk egy évre, a kérvények és jóváhagyások számában 9, illetve 7 százalékos növekedést látunk, és a visszautasítások számában is erős emelkedés tapasztalható. A benyújtott kérvények számának havonkénti vizsgálata azt mutatja, hogy a 2003. márciusi jelentős növekedést áprilisban hirtelen csökkenés követte. Ennek az volt az oka, hogy a munkáltatónak áprilistól fizetnie kell a beadott engedélyek után. Ez a hatás azonban csak egy hónapig fejtette ki hatását, hiszen júniusban és júliusban már magasabb is volt a benyújtott engedélyek száma, mint márciusban.

Ha gazdasági ágazatok szerint vizsgáljuk a munkaengedélyek és az első engedélyek együttes

számát, akkor 2002-ben vezető helyen az egészségügy és orvosi szolgáltatások állnak 24 százalékkal, ezt követik a számítógépes szolgáltatások 17 százalékkal, az adminisztrációs, üzleti és menedzseri szolgáltatások (13%), az oktatási és kulturális tevékenységek (8 %), a pénzügyi szolgáltatások (8%). Ez meglehetősen különbözik az 1995-ös képtől, amikor az adminisztrációs és üzleti szolgáltatások (17%) álltak az első helyen, majd a pénzügyi szolgáltatások (13%), a szórakozási és szabadidős szolgáltatások (12%), a kiskereskedelmi szolgáltatások (12%) és a különböző iparágak volt a sorrend. A változások tükrözik a gazdasági ágazatok jelentőségének változásait. Az információs és kommunikációs technológiai ágazat 2002 szeptemberében került le a betöltetlen állásokkal rendelkező ágazatok listájáról.

A munkaengedélyek és az első engedélyek foglalkozások szerinti vizsgálata alapján 2000 és 2002 között az összes jóváhagyott engedélynek legalább a felét a főiskolai diplomát igénylő és műszaki foglalkozások tették ki. Arányuk a 2000-es 58-ról 2002-re 50 százalékra csökkent, bár az abszolút számokat tekintve 37 193-ról 44 319 emelkedett a kiadott engedélyek száma. A kategórián belül három csoport emelkedik ki: az egészségügyi foglalkozások (24%), melyek az ápolónőket és az olyan orvosi foglalkozásokat foglalják magukba, mint a röntgenológus, vagy a fizikoterápiás szakemberek; a számítógépes elemzők/programozók és más információtechnológiához kapcsolódó foglalkozások (11%) és a művészettel, szórakoztatással és sporttal kapcsolatos foglalkozások (6%).

Az egyetemi diplomát igénylő foglalkozások jelentik a második legnagyobb kategóriát, mintegy 24 százalékkal 2002-ben. Az ebbe a csoportba tartozó legfontosabb foglalkozások: a mérnökök és más műszaki szakemberek (2002-ben 11 százalék), például szoftvermérnökök, számítógépes és elektronikai mérnökök, illetve az egyetemi diplomás tanárok és az orvosok. A kiadott engedélyek száma mindhárom alcsoportban emelkedett 2000 és 2002 között, különösen a tanárok esetében (1464-ről 5814-re). Az egyetemi diplomások csoportja a kiadott engedélyek közel negyedét tették ki 2002-ben, míg a menedzserek és adminisztrátorok csoportjában a 2000-es 21-ről 2002-re 8 százalékos csökkenés következett be. Bár a menedzserek és adminisztrátorok csoportjában csökkent a kiadott engedélyek száma, a speciális menedzserek esetében a 2000 és 2002 között 1038-ról 1362-re emelkedett a külföldi foglalkoztatottak száma, de a közbeeső évben elérte az 1850 főt is.

A kibocsátott engedélyek nagy többsége viszonylag kisszámú foglalkozásra szólt. Csak nyolc olyan foglalkozás van, amire 5000-nél több engedélyt adtak ki. Ez azt mutatja, hogy a munkaengedély-rendszer speciális foglalkozástípusokra koncentrált.

A jóváhagyott munkaengedélyek és első engedélyek nemzetiség szerinti elemzése azt mutatja, hogy 1995 és 2002 között jelentős változások mentek végbe ebből a szempontból. Az időszak kezdetén az Egyesült Államokból érkezett dolgozók kaptak a legnagyobb arányban munkaengedélyt, melyek száma az összes engedély egyharmadát tette ki. Ez az arány ugyan 2002-re 11 százalékra mérséklődött, az abszolút számok viszont növekedést mutatnak. A japán bevándorlók aránya szintén csökkent, az 1995-ös második helyről a hetedik helyre estek vissza az időszak végére, amikor az összes jóváhagyott munkaengedélynek csak 3 százalékát kapták. A 2002-es adatok szerint a legtöbb jóváhagyott munkaengedélyt indiaiak kapták, részesedésük 21 százalék volt, ami jelentős növekedés az 1995-ös 8 százalékhoz képest. Számottevően növekedett a Fülöp-szigetekről, Dél-Afrikából és Malajziából érkezettek aránya is. A statisztikai adatok elemzése azt is mutatja, hogy az évenkénti változások időnként ellentmondtak az általános trendnek.

Az aggregált adatok nem mutatnak összefüggést az országok és a foglalkozások között, a részletesebb elemzés azonban ezekre is fényt derít. Megállapítható, hogy a Fülöp-szigetekről, Zimbabweból és Nigériából érkezettek főként a főiskolai diplomát igénylő egészségügyi szakmákban helyezkedtek el, a mérnökök és műszaki alkalmazottak, valamint a számítógépes szakemberek főleg indiaiak voltak, menedzserként és adminisztrátorként pedig főleg az Egyesült Államokból és Japánból bevándorlók kívántak az Egyesült Királyságban dolgozni. A szakmák szemszögéből nézve a főiskolai szintű végzettséggel rendelkező egészségügyiek mintegy egyharmada érkezett a Fülöp-szigetekről, a mérnökök és műszaki alkalmazottak 70 százaléka, a számítógépes szakembereknek pedig 80 százaléka Indiából. A vizsgált időszak alatt végbement lényeges változás például, hogy míg 2000-ben a Kínából érkezők 14 százaléka helyezkedhetett el a vendéglátóiparban munkaengedéllyel, addig 2002-ben ez a szám már 21 százalék volt. Az ápolónők között a Fülöp-szigetekről érkezettek aránya 50-ről 33 százalékra zsugorodott.

A nemzetközi vállalatok gyakran irányítják alkalmazottaikat a cég más országokban lévő munkahelyeire. A vállalaton belüli átvételek céljára a nyolcvanas években az összes beadvány mintegy felét hagyták jóvá, arányuk 2002-ben csak 30 százalék volt.

2002 januárjában indították a Magasan képzett bevándorlók programját (Highly Skilled Migrant Programme – HSMP), melynek célja, hogy Nagy-Britanniába csalogassa a jó képességekkel és nagy tapasztalatokkal rendelkezőket, hogy ott akár vállalatoknál, akár önállóan dolgozzanak. Az ily módon érkezettek először egyéves tartózkodási engedélyt, kapnak, amit további három évre meghosszabbítanak, és ezután megkapják a letelepedési engedélyt. A munkaengedély rendszer fő részétől eltérően ennél a programnál nincs szükség arra, hogy egy munkáltató konkrét munkakört jelöljön meg, az engedélyt a külföldi magánszemélynek adják saját kérésére. Ausztráliához és Kanadához hasonlóan Nagy-Britanniában is pontrendszert vezettek be. A sikeres jelentkezéshez a jelentkezőnek először be kell mutatni, hogy képes arra, hogy szakmai karrierjét az Egyesült Királyságban folytassa, majd bizonyítania kell, hogy öt területen legalább 75 pontot ért el: iskolai végzettség, munkatapasztalat, kereset, a munkájában eddig elért eredmények és a HSMP-program prioritása a jelentkezés során. 2002. február 1. és 2003. július 31. között 4861 fő kívánt e program keretében Nagy-Britanniában dolgozni, és 2978 kérelmet fogadtak el (61%). Az elfogadott engedélyek között a pénzügyi foglalkozások (könyvelők, bankügyekkel és befektetéssel foglalkozók), a vállalati menedzserek, az információtechnológiai szakemberek és az egészségügyi foglalkozások dominálnak. Lényeges terület még ebből a szempontból a tudományos kutatás és a marketing. Kis számuk miatt az e-program keretében Nagy-Britanniába érkezettek nem gyakorolnak komoly hatást a brit munkaerőpiacra. A rendszer mégis fontos eleme a brit foglalkoztatási politikának, hiszen hangsúlyozza, hogy az Egyesült Királyság kifejezetten támogatja, hogy magasan kvalifikált szakemberek munkahelyként és lakóhelyként Nagy-Britanniát válasszák.

Az időszakos mezőgazdasági dolgozók rendszerét közvetlenül a második világháború után vezették be, pótlólagos munkaerőforrást teremtve az időszakos mezőgazdasági munkákhoz. Bár a programban résztvevők száma emelkedett az évek során, elvei és jellegzetességei nagyjából ugyanazok maradtak. A résztvevők főleg 18-25 éves egyetemi hallgatók, akiket ún. „operátorok" verbuválnak és helyeznek el a farmokra, biztosítják őket, hogy megfelelő munkabért és szállást kapnak. A résztvevők számának szabályozására kvótát állítottak fel, ami a kilencvenes években 10 000-ről 15 200-ra emelkedett, 2003-ra pedig 25 000 fő a jóváhagyott létszám. A programot korábban a Belügyminisztérium irányította, nemrég azonban a Munkaengedélyek Hivatalához került át. A 2002-es felülvizsgálat után több javaslatot fogalmaztak meg a program hatékonyabbá tételére, melyek 2004 januárjában fognak életbe lépni. A legfontosabb változások az operátorok szerepét érintik. A kvótákat az operátorok ajánlatai, a farmerek adott évre szóló becslései, illetve az előző években ténylegesen foglalkoztatottak száma alapján fogják meghatározni. Az operátoroknak három évre előre meg kell határozni a szezonális mezőgazdasági munkások iránti szükségletet, hogy mód legyen hosszabb távú tervezésre. Az operátorok nem helyezhetik a résztvevőket más közvetítők felügyelete alá. 2002-ben 19 372 fő érkezett a program keretében, a legtöbben (25%) Lengyelországból, Ukrajnából (20%) és a balti államokból (18%). Összességében mintegy 10 ezren érkeztek az EU-hoz 2004-ben csatlakozó országokból, ahonnan 2004 májusától szabad munkaerő áramlás lesz Nagy-Britanniába. Jelenleg még nem lehet világosan látni, hogy mi lesz a következménye a csatlakozásnak.

A szektor alapú rendszer legújabb változatát 2003 májusában vezették be az alacsonyabb képzettséget igénylő foglalkozásokban mutatkozó munkaerőhiány pótlására, kezdetben a gazdaság két szektorában (az élelmiszer feldolgozásban és a vendéglátásban, azaz a szállodaiparban és a vendéglátóiparban). A kvóta 2004-ig mindegyik ágazat esetében 10 ezer fő, az engedélyeket a munkáltatóknak kell kérni az először foglalkoztatottak számára. A kérvényeket akkor hagyják jóvá, ha a munkakör hazai munkaerővel nem tölthető be. A 18–30 éves külföldi foglalkoztatott legfeljebb egy évig dolgozhat. Az engedéllyel rendelkezők nem hozhatják magukkal házastársukat, és az engedély lejárta után el kell hagyniuk az országot. A rendszer továbbfejlesztésének két útja van. Az egyik a rendszer kiterjesztése más, rövid távú munkaszerződéseket alkalmazó ágazatokra, a másik pedig, hogy az egész rendszert beolvasszák a munkaengedélyek rendszerének fő részébe. Ez utóbbi esetben az arra alkalmas foglalkozások felkerülnének a fő részben szereplő hiánylistákra. Jelenleg kevés statisztikai adat áll rendelkezésre a szektoralapú rendszerről. 2003 májusától augusztus 6-ig 2559 kérvényt nyújtottak be, melyből 82 százalék a vendéglátóiparban történő foglalkoztatásra irányult. Ezen belül a legkeresettebb iparág a húsipar volt, a jelentkezők körülbelül kétharmada itt akart elhelyezkedni. Nemzetiségük szerint vizsgálva az élen az ukránok állnak 24 százalékkal, majd a lengyelek (18%), a szlovákok (13%) és a csehek (11%) következnek. A csatlakozó EU-országokból érkezett a jelentkezések 55 százaléka, míg a többi kelet-közép-európai ország 37 százalékkal részesedett. Ebben az esetben sem lehet előre látni, hogy mi

lesz a csatlakozás következménye, de a kezdeti adatok szerint az adott országokban készség mutatkozik az alacsonyabb képzettségi szintet igénylő munkák vállalására. Egy másik jellegzetesség, hogy a munkáltatók kevéssé használják ezt a lehetőséget a munkaerőhiány pótlására, legalább is a bevezetés utáni első két hónapban. Ennek oka lehet, hogy nincs szükség külföldi munkaerőre, bár ennek csekély a valószínűsége. Inkább arról lehet szó, hogy a munkáltatók kevéssé ismerik ezt a lehetőséget, illetve illegálisan foglalkoztatják a külföldi munkaerőt, kevesebb bérért és szegényesebb ellátás mellett.

Összefoglalva elmondható, a munkaengedélyek rendszere fontos eszköz a Nagy-Britanniába irányuló bevándorlás szabályozására, melyen keresztül egyre több ember érkezik az országba a világ egyre több országából. Legfontosabb célja a hiányzó munkaerő pótlása a megfelelő szakmákban. A Munkaengedélyek Hivatala a munkáltatókkal együtt értékeli a munkaerőpiac helyzetét. Az utóbbi tíz évben a legkeresettebb szakmák közé kerültek az információtechnológiával kapcsolatos foglalkozások, a legtöbb munkaerőt adó országok pedig egyre inkább az olyan államokból kerülnek ki, mint India, Kína, Malajzia, vagy a Fülöp-szigetek. A munkaengedélyek rendszerének komoly kiterjesztését jelenti 2000 óta a szezonális mezőgazdasági munkások alkalmazása és az alacsonyabb képzettséget igénylő szektor alapú rendszer, amelynek célja az illegális foglalkoztatás visszaszorítása volt.

Nehéz előre jelezni a munkaengedélyek segítségével történő nemzetközi munkaerőmozgást, magát az engedélyezés rendszerét azonban a változásoknak megfelelően folyamatosan tökéletesíteni kell.

(Ism.: *Dévai Péter*)

## GAZDASÁGSTATISZTIKA

JEPIHINA, A. V.:

### ÖSSZOROSZORSZÁGI MEZŐGAZDASÁGI ÖSSZEÍRÁS

(O problemah provedenija vszerosszijszkoj szel'szkohozjajsztvennoj perepiszi.) – *Voproszi sztatisztiki*. 2003. 3. sz. 12–14. old.

Az Orosz Föderáció mezőgazdaságában lényeges változások mentek végbe a XX. század 90-es éveiben. Megtörtént a mezőgazdasági termelés új közgazdasági feltételeinek kialakítása; tökéletesedtek a földviszonyok; aktivizálódott az egyéni szektor fejlődése. A végrehajtott reformok nyomán megteremtődtek a sokszektorú gazdaság alapjai.

Oroszországban a 90-es évekig a hatalmas mezőgazdasági árutermelő üzemek voltak az alapvető gazdasági egységek, amelyek a termelés háromnegyedét adták. A 90-es években a földek újraelosztásával megváltozott a mezőgazdaság szerkezete, bővült a mezőgazdaság egyéni szektora.

Az Orosz Föderáció mezőgazdaságában a termelők három alapvető csoportja alakult ki:

– mezőgazdasági szervezetek (termelőszövetkezetek, részvénytársaságok, különféle társulások stb.), amelyek körében a nagy és a közepes méretű vállalatok száma 24 800. Egy vállalatra átlagosan 6,4 ezer hektár mezőgazdasági terület, 2,6 ezer hektár vetésterület, 600 darab szarvasmarha, 311 sertés, 167 juh és kecske jut. E vállalatokban a dolgozók átlagos létszáma 171 fő;

– paraszti (farmer-) gazdaságok, az árutermelők e kategóriája a 90-es években jelent meg az Orosz Föderációban. E gazdaságok száma 4,4 ezerről 265,5 ezerre növekedett 1990-től 2001-ig. Az ilyen gazdaságok használatában lévő mezőgazdasági terület 0,1 millióról 15,6 millió hektárra nőtt az említett időszakban. A farmergazdaság földterületének átlagos mérete 62 hektár;

– a lakosság háztáji területe átlagosan 0,4 hektár (ebben a csoportban körülbelül 16 millió család található), valamint a 19 millió kollektív, egyéni kertészeti-konyhakertészeti parcella átlagosan 0,09 hektár.

2001-ben a mezőgazdasági termékek előállításában a mezőgazdasági szervezetek súlya 44 százalék volt. A lakosság gazdaságainak aránya 52 százalékot, a paraszti (farmer-) gazdaságoké 4 százalékot tett ki.

Az új közgazdasági viszonyok fejlődésével a központi tervgazdálkodásról a piacgazdaságra való átmenet során a statisztika által tradicionálisan alkalmazott teljes megfigyelés módszere egy sor objektív oknál fogva, közöttük költségessége miatt, nem volt hatékony. A mezőgazdaság sokszektorúságának, az egyéni szektor szerepe növekedésének, a mezőgazdasági szervezetek felbontásának viszonyai között a statisztikai megfigyelés tökéletesítésének egyik alapvető iránya a reprezentatív módszer gyakorlati alkalmazásának bővítése volt.

Jelenleg az orosz mezőgazdasági statisztikában a reprezentatív megfigyelések rendszere már kialakult különféle kiválasztási típusok alkalmazásával, tekintettel a mezőgazdasági termelés sajátosságaira és a megfigyelés tárgyára. Így a paraszti gazdaságok mezőgazdasági tevékenységének statisztikai megfigyelése az egycélú rétegzett kiválasztás módszerének alkalmazásával történik az alapvető növényi

kultúrák vetésterülete vagy a fajonkénti állatállomány esetében.

A lakosság gazdaságaiban folytatott mezőgazdasági tevékenység statisztikai megfigyelése a magánszektorban levő mezőgazdasági kultúrák vetésterületének, a gyümölcsös és bogyós ültetvényeknek, továbbá az állatállománynak az időszakos összeírásán alapul, valamint azon, hogy reprezentatív módon megfigyelik a háztartásokat, támaszkodva a falusi közigazgatás földügyi és földbirtok-rendezési bizottságainak közléseire. A lakossági gazdaságok reprezentatív kiválasztásánál a négylépcsős valószínűségi kiválasztási módszert alkalmazzák.

A mezőgazdasági szervezetek és a paraszti gazdaságok esetében a reprezentatív nagyságok kialakításának alapjául szolgál a „mezőgazdasági árutermelők" elnevezésű alregiszter, amely a „mezőgazdasági vállalatok", a „paraszti (farmer-) gazdaságok" és a „mezőgazdasági kisvállalatok" regisztereit egyesíti. A lakosság gazdaságai esetében a földadót fizetőket tartalmazó listák szolgálnak alapul.

A lakosság gazdaságainak reprezentatív megfigyelése az egyik alapvető forrás – többek között – a mezőgazdaságról szóló makroökonómiai mutatók számításához, az élelmiszermérlegek kialakításához, a mezőgazdaság termelési volumenének meghatározásához, az alapvető élelmiszerek lakossági fogyasztásának megállapításához.

A piaci viszonyok fejlődésével, az agrárszféra sokszektorúságának kialakulásával és a reprezentatív statisztikai megfigyelés alkalmazásának bővülésével mind időszerűbbé válik az összeírás, összhangban a FAO világcenzusának programjával és figyelembe véve az orosz mezőgazdaság nemzeti sajátosságait.

Az Orosz Föderáció középtávú (2002-2004. évekre szóló) társadalmi-gazdasági fejlesztési programja előirányozza az összeírás lebonyolításához szükséges szervezési-módszertani intézkedések megvalósítását.

Ilyen méretű mezőgazdasági összeírás gyakorlatilag első ízben fog megtörténni hosszú évek után Oroszországban. Hasonló munkát legutóbb 1920-ban szerveztek. A későbbiekben specifikus mezőgazdasági összeírások voltak. Közöttük megemlítendő a mezőgazdasági növények vetésterületének összeírása (1964-ben, 1976-ban és 1985-ben), az összes gazdaságkategóriára vonatkozóan a gyümölcs- és bogyósültetvény-, valamint szőlőskert-összeírás 1970-ben és 1984-ben, valamint a lakosság gazdaságaiban tartott állatállomány felmérése (utoljára 1996. január 1-jén). Az összes legutóbbi összeírás gyakorlatilag a mezőgazdaság reformja előtt történt.

Az új összeírás megszervezése lehetővé teszi, hogy olyan adatokat kapjanak a mezőgazdasági termelés alapvető jellemzőiről az új közgazdasági viszonyok között, amelyek bázisként használhatók a megfelelő mutatók összeírások közötti alakulásának értékeléséhez, amelyekkel elvégezhető a reprezentatív felmérések esetében az általános nagyságrendek aktualizálása, és végül amelyekkel a mezőgazdaság állapotára vonatkozó statisztikai megfigyelés mutatóinak rendszere bővíthető.

A mezőgazdasági összeírás eredményei nagy jelentőségűek lesznek a hatékony agráripari politika kidolgozásához.

A világcenzus keretében lebonyolított összeírás növeli a nemzeti statisztika jelentőségét, lehetővé teszi az orosz mezőgazdaság fejlődéséről szóló komplex információ integrálását a globális információs rendszerbe, ezzel elősegíti a világcenzus mint legfontosabb nemzetközi szakmai esemény céljainak elérését.

*Az összeírás előkészítése és lebonyolítása*. A Goszkomsztat, Oroszország Állami Statisztikai Bizottsága hozzákezdett az összeírás előkészítéséhez. A nemzetközi együttműködés keretében 2002-ben tanulmányozták az Egyesült Államok, Lengyelország és Észtország tapasztalatait a mezőgazdasági összeírások lebonyolítása terén. „A mezőgazdasági összeírás szervezése és lebonyolítása a FAO ajánlásai alapján" c. témában a Goszkomsztat és a területi statisztikai szervek szakemberei tanulmányokat folytattak az Egyesült Államok Mezőgazdasági Minisztériumának nemzeti agrárstatisztikai szolgálatánál a STASYS-program keretében.

Jelenleg már rendelkezésre áll a kidolgozott összeírási technikai feladat, és pályáztatás folyik a konzultációs szolgáltatások nyújtására az említett témában a STASYS-programnak megfelelően. 2003 végére tervezték az összeírás normatív-jogi dokumentumainak elkészítését, az összeírás programjának összeállítását, az összeírási listák és a módszertani útmutatások, továbbá a kérdezőbiztosok képzési programja és módszere kidolgozását, valamint próbamegfigyelések elvégzését, végül a program összeállítását az összeírás eredményeinek összegzésére és közzétételére.

Az orosz mezőgazdasági statisztika 2001-2005. évekre szóló fejlesztési programja keretében kidolgozták az összoroszországi mezőgazdasági összeírás előkészítésének és lebonyolításának szervezési mintatervét.

*Az összeírás lebonyolításával kapcsolatos módszertani és szervezési problémák*. A következő módszertani jellegű kérdéscsoportok különíthetők el.

– Az összeírás lebonyolításának jogi bázisa. Mivel az Orosz Föderációban hiányzik a statisztikai

tevékenységet szabályozó törvényhozás, célszerű a mezőgazdasági összeírások lebonyolításáról szóló olyan speciális szövetségi törvény kidolgozása, amely meghatározza a szövetségi végrehajtó hatalmi szervek és a válaszadók kötelezettségeit, az összeírás programját, a költségvetési források kiutalásának rendjét, az összeírás során gyűjtött adatok bizalmas jellegét.

– A mezőgazdasági összeírás statisztikai egységének meghatározása. A statisztikai megfigyelés egységének a különböző gazdálkodási formákhoz tartozó mezőgazdasági termék előállítókat célszerű tekinteni, azaz a mezőgazdasági szervezeteket, a paraszti (farmer-) gazdaságokat, a nem jogi személyiségű egyéni vállalkozókat és a lakosság gazdaságait. A mezőgazdasági összeírás statisztikai egységének meghatározása a nemzetközi gyakorlatban magában foglalja számos ismérv szerint a gazdaság nagyságának összeírási küszöbét. A mezőgazdasági termelők kiválasztásánál az adott küszöb megállapítása rendszerint – a megfigyelt egységek optimális száma esetén – a mezőgazdasági összeírással átfogja a termelés 95-99 százalékát. Minthogy Oroszországban az átlagosan 0,09 hektáros kis gazdaságokra, kerti és konyhakerti parcellákra jut a burgonya és a zöldség termelésének 20-25 százaléka, a gyümölcsök és bogyósok termésének körülbelül fele, a lehető legkisebb küszöböt szükséges megállapítani a földparcella és az állatállomány vonatkozásában.

– Az összeírandó mezőgazdasági termelőket tartalmazó listák kialakításának szintjei. A lehető legteljesebb körű számbavétel céljából a listák összeállításánál a következő információforrások használhatók: a Goszkomsztat rendszerében használatos alregiszterek („mezőgazdasági vállalatok", „paraszti (farmer-) gazdaságok", „mezőgazdasági kisvállalkozások"); a Roszzemkadasztr (orosz állami földkataszter) területi szerveinek listái a földadó-fizetőkről; azoknak a falusi és városi (települési) igazgatási szerveknek a gazdasági nyilvántartásai, amelyeknek a területén mezőgazdasági települések vannak.

– A mutatók rendszerének meghatározása az összeírás programja számára. A FAO vonatkozó dokumentumainak megfelelően a következő mutatókat célszerű az összeírási programba foglalni: *a*) információk a gazdaságról (azonosítók, osztályozási ismérvek, tájékoztatási célú információk); *b*) a tulajdonos és a háztartás jellemző adatai (nem, kor, családi állapot, szakképzettség, a háztartásban élők száma, rokonsági viszonyok); *c*) foglalkoztatás (a gazdaságilag aktív tagok, beleértve a beszámolási időszakban foglalkoztatottakat és a nem foglalkoztatottakat, a szakmai és szakképzettség szerinti összetétel, az állandóan és az időszakosan foglalkoztatott mezőgazdasági munkások); *d*) földerőforrások (összes földterület, a földbirtoklás formája, a földek összetétele, a mezőgazdasági földbirtokok területe és szerkezete, a védett talajterület, az öntözött és a lecsapolt terület); *e*) mezőgazdasági növényi kultúrák (a növényi kultúrák területe, az évelő ültetvények területe, felhasznált műtrágya és szervestrágya, peszticidek); *f*) állatok (az élőállatok és baromfi állománya, az alkalmazott állattenyésztési rendszer); *g*) gépek és felszerelések (gépek és felszerelések állománya, birtoklásuk formája); *h*) épületek és más építmények (állományuk és a birtoklás formája); *i*) egyéb tevékenységi formák.

Megfelelő pénzügyi források esetén az összeírás programja kiegészíthető egy sor egyéb mutatóval (például az árutermelés mértéke, a költségszerkezet stb.).

Az Orosz Föderációban a mezőgazdaság sokszektorúságával összefüggésben fennáll a speciális összeírási listák különféle gazdaságkategóriák szerinti kidolgozásának problémája.

*Az összeírás módszerei és szakaszai*. Az összeírás szervezésének legfontosabb problémája a mezőgazdasági termelők tevékenységéről való információszerzés módszereinek meghatározása. Egy sor országban csak a teljes körű összeírást alkalmazzák. A nemzetközi gyakorlatban alkalmazzák a teljes körű és a reprezentatív megfigyelés módszereinek kombinációját is. A reprezentatív regisztrációnál csak a sokaságok elemeinek egy részét figyelik meg, ami lehetővé teszi a megfigyelési program kiszélesítését.

A FAO ajánlásainak megfelelően az összeírás beszámolási időszaka lehet a naptári év vagy a mezőgazdasági év, és az összeírás több szakaszban mehet végbe.

(Ism.: *Balogh András*)

# KÜLFÖLDI FOLYÓIRATSZEMLE

DEMOGRAFIE
revue pro výzkum populačního vývoje

A CSEH STATISZTIKAI HIVATAL
FOLYÓIRATA

2004. ÉVI. 1. SZÁM

*McDonald, P. F.*: Önfenntartó termékenység az állami politikán keresztül: a választási lehetőségek.
*Kocourková, J.*: Gyermekek támogatása Csehországban: preferenciák és a valóság.
*Sotkovsky, I. – Tvrdy, L.*: A természetes és migrációs növekedés a morva-sziléziai településeken, 1992–2001.
*Ales, M.*: Termékenység és házasságkötés.

Journal of the Royal Statistical Society

AZ ANGOL KIRÁLYI STATISZTIKAI
TÁRSASÁG FOLYÓIRATA
(A SOROZAT)

2004. ÉVI 2. SZÁM

*Charlton, J.*: Automatikus adateditálási és imputálási módszerek és az EUREDIT projekt.
*Redfern, P.*: A 2001-es cenzus egy alternatív nézete és a jövőbeni népszámlálás.
*Petrakos, G. és mások:* Az adatellenőrzés specifikálásának új módjai.
*Béguin, C. – Hulliger, B.:* Többváltozós outlierek kijelzése nem teljes felvételi adatokban: egy epidemikus algoritmus és transzformált rangkorrelációk.
*Manzari, A.:* Adateditálási és imputálási módszerek kombinálása: egy kísérleti alkalmazás népszámlálási adatokra.
*DiZio, M. és mások:* Bayesi hálózatok imputáláshoz.
*Chambers, R.- Hentges, A.- Zhao, X.:* Robusztus automatikus módszerek outlierek és hibák kijelzésére.
*Longford, N.T.:* Hiányzó adatok és kisterületi becslések az angol munkaerő-felvételben.

Economie Statistique 31

A FRANCIA GAZDASÁGI
ÉS PÉNZÜGYMINISZTÉRIUM ÉS A STATISZTIKAI
ÉS GAZDASÁGKUTATÓ INTÉZET FOLYÓIRATA

2003. ÉVI 367. SZÁM

*Beffy, P. O. – Bonnet, X. – Monfort, B. – Darracq-Parriés, M.*: MZE, egy makroökonómiai modell az eurózónára.
*Desplatz, R. – Jamet, S. – Passeron, V. – Romans, F.*: Keresetvisszatartás az 1980-as évek óta Franciaországban.
*Crépon, B. – Duhautois, R.*: Lassulás a termelékenységben és állások újbóli kihelyezése: két növekvő trend.
*Kalugina, E. – Najman, B.*: Munka és szegénység Oroszországban: objektív értékelések és szubjektív érzékelések.

2003. ÉVI 368. SZÁM

*Prouteau, L. – Wolff, F. C.*: Informális szolgáltatások háztartások között: az önkéntes munka egy kevésbé ismert vonatkozása.
*Minni, C. – Topiol, A.*: Hogyan foglalkoznak a vállalatok az öregedő személyzettel?
*Aubert, P.*: Az ötvenévesek a magán kereső foglalkoztatásban.
*Aubert, P. – Crépon, B.*: Az idősebb munkások termelékenysége: előzetes becslés.

STATISTICAL SCIENCE

AZ EGYESÜLT ÁLLAMOK
MATEMATIKAI STATISZTIKAI INTÉZETÉNEK
FOLYÓIRATA

2003. ÉVI 3. SZÁM

Kegyeletadás John. W. Tukey-nak.
Emlékezés John. W. Tukey-ra.
*Fernholz, L. T. – Morgenthaler, S.:* Beszélgetés John. W. Tukey-val.
*Fokianos, K. – Kedem, B.:* Regressziós elmélet kategoriás idősorokra.
*Sutradhar, B. C.*: Diszkrét longitudinális válaszok regressziós modelljeinek áttekintése.
*Lindsay, B. G. – Qu, A.*: Következtetési függvények és kvadratikus score-tesztek.

statistika
EKONOMICKO - STATISTICKÝ ČASOPIS

A CSEH STATISZTIKAI HIVATAL
FOLYÓIRATA

2004. ÉVI 2. SZÁM

*Fischer, J. – Slégrová, H.*: A cseh állami statisztikai szolgálat az EU küszöbén - kötelezettségek, kihívások és remények.
*Vojtisek, P.*: A Központi Bank statisztikái – harmonizálás az EU-szabványokkal.
*Zahradnik, P. – Jedlicka, J.*: A reál és nominál konvergencia néhány kérdéséről az EU-val kapcsolatban.
*Jilek, J.*: Az EU strukturális mutatói mint kihívás.
*Mandel, M. – Tomsik, V.*: Az átalakuló országok konvergenciája az EU irányában, a belső és külső egyensúly szerint vizsgálva.

*Marcek, D.*: A háztartások időfelhasználásának előrejelzése ökonometriai és állapot-tér modellekre alapozva.

*Hanzlová, D.*: A külföldi leányvállalatok statisztikája és a francia gyakorlat.

## Statistische Nachrichten

AZ OSZTRÁK KÖZPONTI STATISZTIKAI HIVATAL FOLYÓIRATA

2004. ÉVI 4. SZÁM

2001-es épület és lakás összeírás: fő eredmények Karintiára.

2001-es épület és lakás összeírás: fő eredmények Alsó-Ausztriára.

2002-es anyagráfordítás felvétel a bányászatban és feldolgozóiparban.

Rövid távú statisztikák – kereskedelem és szolgáltatások – új fogalmak.

A termelékenységi indexek új számítása (2000 = 100).

Szabadidős és üzleti utak 2003. második negyedévében.

Kereseti adó statisztikák, 2002.

Kormányzati kiadások funkciók (COFOG) szerint az EU tagországokkal összehasonlítva.

2004. ÉVI 5. SZÁM

2001-es népszámlálás: a foglalkoztatottak gazdasági tevékenységeinek jellemzői.

Foglalkoztatottság és munkanélküliség körzetek szerint 2004. január végén.

2002-es kereseti adó statisztikák: társadalomstatisztikai perspektíva.

Az osztrák Euro vásárlóereje külföldön.

2001-es épület és lakás összeírás: Felső-Ausztriára vonatkozó fő eredmények.

Szőlő szüret, bor készlet és bortárolási kapacitás 2003-ban.

A vállalatok beszámolási kötelezettségei a Statistik Austria által végrehajtott statisztikai felvételek esetén 2003-ban.

Szabadidős és üzleti utak 2003. harmadik negyedévében.

Gépjármű állomány 2003-ban.

Polgári repülés 2003-ban.

Regionális elszámolások NUTS 3. régiós szinten 1995-től 2001-ig.

Külkereskedelem 2003. januártól decemberig: előzetes eredmények.

## WIADOMOŚCI STATYSTYCZNE

A LENGYEL STATISZTIKAI FŐHIVATAL FOLYÓIRATA

2004. ÉVI 2. SZÁM

*Milo*, *W. et al.*: Gondolatok a természeti vagyonról.

*Michalak*, *J.:* Deming filozófiájának statisztikai szempontjai a minőség kezelésében.

*Jerczynska*, *M.* . Üzleti ciklus a kiskereskedelemben.

*Gieraltowska*, *U. – Patek*, *E.:* A beruházási portfóliókockázat elemzése.

*Borowski*, *P.:* Munkanélküliségi felvétel a struktúra csoportosítás módszerével például a Lublini vajdaságban.

*Wieczorek*, *P.:* Aránytalanság az EU társadalmi és gazdasági fejlettségében a bővítés előtt és után.

*Bak*, *I.:* Nemzetközi idegenforgalom a világban.

## Wirtschaft und Statistik

A NÉMET SZÖVETSÉGI STATISZTIKAI HIVATAL FOLYÓIRATA

2003. ÉVI 11. SZÁM

*Klumpen, D. – Köhler, S.*: Napjaink igényei a hivatalos statisztika iránt.

*Meyer, I. – Timm, U.*: Közösségi statisztikák a jövedelemről és életkörülményekről (EU-SILC).

*Wernicke, I. H.*: Partnerség mezőgazdasági statisztikában Litvánia és Németország között.

*Linkert, K.*: A karácsonyi vásárlás fontossága a kiskereskedelemben.

*Pfaff, H.*: Személyzet a nővéri gondozásban.

*Schoer, K. – Becker, B.*: A környezetgazdasági elszámolás és a környezetstatisztika válogatott eredményei.

*Lüttinger, P. – Lechert, Y. – Breiholz, H.*: A Mikrocenzus Tudományos Felhasználású Adatállományai felhasználóival folytatott második interjú eredményei.

2003. ÉVI 12. SZÁM

*Brachinger, H. W.*: A Szövetségi Statisztikai Hivatal 2003. évi Gerhard Fürst-díja.

*Jörger, N.*: Strukturális mutatók – a haladás mérése a lisszaboni stratégia hatáskörén belül.

*Klatt, G.*: Törvények megvalósításának elfogadására vonatkozó eljárások (comitológia).

*Sommer, B. – Voit, H.*: Népesség alakulás, 2002.

*Emmerling, D.*: Válások, 2002.

*Höh, H.*: Az építőipar Németországban.

*Vorndran, I.*: A lakáshelyzet Németországban, 2002 április.

*Haug, H. F. – Revermann, C.*: A kutatásra és kísérleti fejlesztésre vonatkozó statisztikák összehasonlítása.

*Frank-Bosch, B.*: Az épületárak indexének új számítása 2000-es bázison.

*Vorholt, H.*: Kereseti struktúrák Németországban.

2004. ÉVI 1. SZÁM

*Buchwald, W.*: Az előző fogyasztóiár-indextől az újig.

*Hartmann, N.*: GDP, 2003.

*Ehling, M. – Linz, S. – Minkel, H.*: A statisztika nemzetközi harmonizálása – alapok és példák a háztartás-statisztikai szektorból.

*Ebigt, S. – Sturm, R. – Volkmann, S.*: „A hivatalos statisztika által a gazdaságra nehezedő teher jelentőségéről" kezdeményezett tanulmány.

*Fischer, I.*: Online felvételek a belkereskedelemben, valamint a szálloda- és vendéglátóiparban.

*Blang, D.*: Az Intrastat automatizálásában elért haladás Európában.

*Reim, U.*: Kombinált szállítás, 2002.

*Kaufen, S.*: Közszolgálat nyugdíjasok 2003. január 1-jén.

*Hass, H. J. – Gross, S.*: Tíz javaslat a statisztikai információs infrastruktúra további fejlesztéséről a német ipar szempontjából.

2004. ÉVI 2. SZÁM

*Schulze, W.*: „Twinning Project" a Lengyel Statisztikai Hivatallal: példa a kétoldalú adminisztratív partnerségre.

*Czajka, S.*: Önértékelések a minőségkezelésben.

*Fritsch, S. – Lüken, S.*: Foglalkoztatottság Németországban.

*Riede, T. – Sacher, M.*: Munkaerő-piac Németországban – az új ILO statisztika első komponense.

*Schmidt, P. – Waldmüller, B.*: A „speciális célú felvétel" koncepciója és eredményei a vállalat statisztikai fogalmával kapcsolatban.

*Petrauschke, B. – Pesch, K. H.*: A strukturális felvétel eredményei a szolgáltatási szektorban, 2001.

*Blumöhr, T. – Walsemann, U.*: A mezőgazdaság Németországban, 2003.

*Eberth, F.*: Külkereskedelem a FÁK országaival.

*Haustein, T. – Krieger, S.*: A menedékjogot kérők köztámogatására és segélyezésére vonatkozó statisztikáinak eredményei, 2002.

*Deckl, S. – Krebs, T.*: A háztartások felszereltsége tartós fogyasztási cikkekkel és lakáskörülményei.

*Gold, M.*: Igazodási költségek: a túlórák egyik oka.

2004. ÉVI 3. SZÁM

*Bierau, D. – Reim, U.*: A szállítási statisztikai törvény módosítása.

*Opfermann, R.*: Termékkísérő szolgáltatások és statisztikai lefedettségük.

*Haug, S.*: Migráció Németország és az EU-hoz társuló közép- és kelet-európai országok között.

*Mehlin, I.*: Szőlőművelés és borstatisztika Németországban.

*Grillmaier, G.*: Trendek a nagykereskedelemben, 2003.

*Decker, J.*: Trendek a kiskereskedelemben, 2003.

*Fischer, I.:* Trendek a szálloda- és vendéglátóiparban.

*Gehle, S.:* Külkereskedelem az EU-val.

*Dietz, O.:* A központi, regionális és helyi hatóságok személyzeti kiadásai.

*Heinz, W.:* A kriminológia társadalmi és kulturális alapjai.

2004. ÉVI 4. SZÁM

*Greulich, M.*: A gazdasági osztályozások felülvizsgálata 2007-ig – közbenső beszámoló.

*Namislo, D.*: Európai Parlamenti választások, 2004.

*Fritz, J.*: Az európai strukturális és foglalkoztatási mutatók becslése.

*Angele, J.*: Csődök Németországban.

*Spörel, U.*: Belső idegenforgalom, 2003: stabilizáció a kedvezőtlen keretfeltételek ellenére.

*Loschky, A.*: Külkereskedelem országok szerint, 2003.

*Fischer, R.*: Kereskedelmi légi szállítás, 2003.

*Weinmann, J. – Zifonun, N.*: Egészségügyi kiadások és személyzet, 2002.

*Pfaff, H.*: Fogyatékosság és egészség.

*Kolvenbach, F. J.*: Állami ifjúsági jóléti támogatás fiatal nagykorúaknak.

*Rehm, H.*: Közpénzek, 2003.

*Niese, M.*: Gyenge vállalatok – erős piac.

## AZ OROSZ ÁLLAMI STATISZTIKAI BIZOTTSÁG FOLYÓIRATA

2004. ÉVI 2. SZÁM

*Ulianov, I. S.:* Jövedelmezőség és állóeszköz-beruházások.

*Khanin, G. I. – Fomin, D. A.:* A mezőgazdaság jövedelmezőségének alternatív becslése 2001-ben.

*Gulidov, A. D. – Golovanov, Yu. K.:* Az orosz Goskomstat információs és feldolgozó rendszerének fejlesztése 2003-ban.

*Gulidov, A. D. -Golovanov, Yu. K. – Proshletsov, S. V.:* Az orosz Goskomstat információs és feldolgozó rendszerét támogató alrendszerek tervezése és fejlesztése.

*Sychev, E. B.:* Az összorosz népszámlálás adatfeldolgozási tapasztalatai és ajánlások statisztikai feladatokra való használatukhoz.

*Kozlov, M. P.:* A mezőgazdasági termelők gazdasági helyzete makroökonómiai növekedés mellett.

*Kolesnikova, A. A.:* Az orosz lakosság befizetései lakásépítési és kommunális komplex szolgáltatásokra, valamint szociális védelemre a régiókban.

*Krotov, Yu. E.:* A gazdasági biztonság fő mutatói a nyizsnij–novgorodi régióban.

*Aichepsheva, R. P.:* Az egyéni vállalkozókra vonatkozó összorosz gazdasági cenzus előkészítő munkája.

*Ageenko, A. A. – Novikova, E. L.:* A háztartási elszámolás mint információforrás a falusi lakosságra és gazdaságokra vonatkozóan.

*Kornileva, S. S.:* A vállalatok külkereskedelmi tevékenységének statisztikái regionális szinten.

*Timofeeva, R. A.*: Közigazgatási reform a területi statisztikai hivataloknál.

2004. ÉVI 3. SZÁM

*Kremlev, N. D. – Rozenberg, D. K.*: A regionális elszámolási rendszer mint a gazdasági fejlődés tükre középszinten.

*Zamaraev, B. A.*: A „kereskedelmi profit (veszteség)" kutatása Oroszország gazdaságára 1996 és 2003 között.

*Semenova, D. S.*: A valutaalapok tükröződése Oroszország fizetési mérlegében.

*Gorbacheva, T. L.*: A 17. ICLS a munkaerő statisztika fejlődésének aktuális problémáiról.

A munkaerő statisztika koncepciója a gazdasági reform feltételei mellett.

*Miticheva, O. N. – Khlestunova, T. N.*: A szervezetek kiadásai a munkaerő fenntartására a Vologda régióban 2002-ben.

*Popov, A. D.*: Az üres állás, mint a modern munkaerő-státus jellemzője.

Az Orosz Föderáció fő társadalmi és gazdasági mutatói 1998 és 2003 között.

*Ovsjannikova, I. I. – Zabelina, A. A.*: Észak angarai terület: a fejlődés problémái, trendjei és kilátásai.

*Gorjacheva, V. G. és mások*: Nyizsnij-novgorodi régió: a család és a népesség reprodukció problémái.

2004. ÉVI 4. SZÁM

*Rajjskaja, N. N. – Sergienko, Ja. V. – Frenkel', A. A.*: A fizetőképes kereslet fő dinamikai tényezőinek elemzése az ipari ágazatokban.

*Kabanova, T. A. – Koncevich, O. V. – Shvandar, K. V.*: A világ kedvező árupiaci konjunktúrájának hatása az orosz gazdaság külgazdasági szektorára.

*Eliseeva, I. I. – Shhirina, A. N. – Kapralova, E. B.*: Az árnyék tevékenység volumenének definiálása makrogazdasági mutatók alapján.

*Fedorova, E. A.*: Az feketegazdaság elszámolása az SNA-ban: „pró" és „kontra".

*Frolova, E. B. – Mukhanova, O. A.*: Országos felvétel a háztartások jólétéről és társadalmi programokban való részvételéről.

*Kashina, O. N.*: Koncepcionális alapelvek és módszerek a társadalmi kockázatok mérésére és prognózisára, valamint a társadalmi biztonság integrált információs rendszerének létrehozására.

*Glushhenko, G. I.*: A nem hivatalos pénzátutalások rendszerei: történet, fejlődés és kilátások.

*Shadrova, N. V. – Semchenko, N. I.*: Átmenet a gazdasági tevékenységek összoroszországi osztályozása felé.

*Pashinceva, N. I.*: A településstatisztika szervezési szempontjai az Orosz Föderációban.

*Emel'janov, V. V. – Samojjlova, M. A.*: A településstatisztika problémái a helyi önelszámolási reform feltételei mellett.

*Simchera, V. M.*: Oroszország gazdasága 2003-ban: a növekvő reményű átalakulások eredményei.

*Kuznecova, E. V.*: Oroszország és a kelet-európai országok gazdasági fejődése a XX. század végén és a XXI. század elején.

A FÁK országainak gazdasága 2003-ban.