Statisztikai Szemle

A KÖZPONTI STATISZTIKAI HIVATAL
TUDOMÁNYOS FOLYÓIRATA

88. ÉVFOLYAM 7–8. SZÁM 2010. JÚLIUS–AUGUSZTUS

ISSN 0039 0690

Megjelenik havonta egyszer
Főszerkesztő: dr. Lakatos Miklós
Osztályvezető: Dobokayné Szabó Orsolya
Kiadja: a Központi Statisztikai Hivatal
A kiadásért felel: dr. Vukovich Gabriella
2010.132 – Xerox Magyarország Kft.

Szakreferensek: Farkas János (társadalomstatisztika),
dr. Hajdu Ottó (módszertan), Laczka Sándorné dr. (gazdaságstatisztika)
Szerkesztők: Bartha Éva, dr. Kondora Cosette, Visi Lakatos Mária
Tördelőszerkesztők: Bartha Éva, Simonné Káli Ágnes
Internet szerkesztése: Bada Ilona Csilla

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.hu
Kiadó: 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 (1008 Budapest, Orczy tér 1).
Előfizethető közvetlen a postai kézbesítőknél, az ország bármely postáján,
valamint e-mailen (hirlapelofizetes@posta.hu) és faxon (303-3440).
További információ: 06-80-444-444
Előfizetési díj: fél évre 6000 Ft, egy évre 10 800 Ft
Beszerezhető a KSH Könyvesboltban. Budapest II., Fényes Elek u. 14–18. Telefon: 345-6789

Tartalom

Tanulmányok

Műhely

Fórum

Szakirodalom

Bevezető*

A *Statisztikai Szemle* évek óta jelentet meg tematikus számokat a gazdaság- és társadalomstatisztika különböző területeiről. A mostani arra vállalkozik, hogy a statisztika e két fő ágazatában alkalmazott új módszertani megoldások közül mutasson be néhányat. A statisztikai módszerek fejlődését világszerte a tudományos élet megkülönböztetett érdeklődése övezi. Korunk egyre bonyolultabb társadalmi és gazdasági folyamatait csak ezek folyamatos fejlesztésével, gyakorlatba történő gyors bevezetésével lehet nyomon követni.

Az első tanulmány néhány példán keresztül áttekinti a statisztikai termelési folyamatokra és az európai statisztikai hivatalokra kidolgozott minőségügyi keretrendszer gyakorlati megvalósítását, részletesen ismerteti a Központi Statisztikai Hivatal e rendszerrel kapcsolatos fejlesztési eredményeit, jövőbeni feladatait.

A lakossági felvételeknél egyre gyakoribbá válnak a vegyes (hibrid) módszerrel végrehajtott adatfelvételek. A második tanulmány az online és a személyes adatfelvételek módszertani megoldásait ismerteti, megvizsgálja, hogy mire alkalmas a hibrid kutatás és választ próbál adni arra a kérdésre, hogy mi az oka e kutatási forma felértékelődésének.

A harmadik tanulmány, áttekinti azokat a legfontosabb statisztikai mintavételi eljárásokat, melyek olyan esetekben alkalmazhatók, amikor nem áll rendelkezésre mintavételi keret. Az eredeti szakirodalmi példákon túl foglalkozik egy, az erdélyi populációra készült reprezentatív kutatás gyakorlati megvalósításának tapasztalataival is.

Közvetlen gyakorlati jelentősége van a negyedik tanulmánynak, mely a kiskereskedelmi forgalom havi megfigyelésének reprezentatív módszertanát ismerteti a 2000-es években, a megfigyelés jellemzőivel, a rétegezéssel, a rétegenkénti mintanagyság meghatározásával és a minta kiválasztásával foglalkozik, továbbá a hiányzó adatok pótlását, a felhasznált alternatív becslési módszereket, a becslések helyességének vizsgálatát tárgyalja.

A következő tanulmány a matematikai statisztika területén a statisztikai kapcsolatok mérési skála által meghatározott típusai – variancia, korreláció, asszociáció, látencia – mérésének sokváltozós mérőszámait tekinti át, az elemezendő mátrixok sajátértékeinek tükrében.

A hatodik tanulmány szintén fontos területet érint, a mikroszimulációs modellek néhány kérdését mutatja be, felhasználva a KSH által készített Háztartások Költségvetési Felvételének (HKF) adatait, javítva a konzisztenciát az egyéb adatforrásokból származó makrogazdasági adatokkal.

A hetedik tanulmány azokat az eloszlásokat, illetve eloszláscsaládokat tekinti át, amelyek alkalmasak lehetnek változatos alakú – lehetőség szerint a ferdeség/csúcsosság síkot minél jobban lefedő – eloszlásból származó minták generálásához.

Szintén a gyakorlati munkát segíti a nyolcadik tanulmány, mely egy Excel-környezetű parancsfájl részletes használati-értelmezési útmutatója a lineáris regressziós analízis metódusához.

Napjaink izgalmas módszertani kérdése, hogy a szezonális kiigazítás miképpen működik a gazdasági válság közepette. A tematikus szám két rövidebb írást közöl, melyek adat-előállítói, illetve felhasználói szemmel tárgyalják ezt a témát.

Reméljük, hogy a *Statisztikai Szemle* ezen száma hozzájárul napjaink legfontosabb statisztikai módszertani kérdéseinek jobb megértéséhez, és segít értelmezni egy-egy jól körüljárható módszertani problémát.

Statisztikai Szemle Szerkesztősége

* A Szerkesztőség ezúton mond köszönetet *dr. Szép Katalinnak*, a KSH főosztályvezetőjének és *dr. Hajdu Ottónak*, a Corvinus Egyetem tanszékvezető egyetemi docensének, folyóiratunk szakreferensének a tematikus szám összeállításához nyújtott értékes tanácsaiért, észrevételeiért.

Minőségügyi keretrendszerek a statisztikai hivatalokban*

Földesi Erika,
a KSH főosztályvezető-helyettese
E-mail: Erika.Foldesi@ksh.hu

Kajdi László,
a KSH tanácsosa
E-mail: Laszlo.Kajdi@ksh.hu

Mag Kornélia,
a KSH főosztályvezető-helyettese
E-mail: Kornelia.Mag@ksh.hu

Szép Katalin,
a KSH főosztályvezetője
E-mail: Katalin.Szep@ksh.hu

Vigh Judit,
a KSH szakmai tanácsadója
E-mail: Judit.Vigh@ksh.hu

A hivatalos statisztika minőségfogalmáról 2004-ben megjelent tanulmányt követően (*Szép–Vigh* [2004]) időszerűvé vált, hogy néhány európai statisztikai hivatal példáján áttekintsük a statisztikai termelési folyamatokra és egy konkrét statisztikai hivatalra kidolgozott minőségügyi keretrendszerek gyakorlati megvalósítását. Jelen írás második része a magyar Központi Statisztikai Hivatalban (KSH) 2005 óta folyó minőségügyi keretrendszer kiépítésének fejlesztési eredményeiről, valamint a jövőbeni feladatokról szól.

TÁRGYSZÓ:
Minőségbiztosítás.
Statisztikai módszertan.
Statisztikai intézmény.

* A szerzők ezúton mondanak köszönetet módszertanos kollégáiknak, különösen *Katona Szilviának* és *Kővári Zsoltnak*, akik munkájukkal elősegítették a cikk második részében bemutatott fejlesztések megvalósulását, továbbá azon szakstatisztikus kollégáknak, akik – főként a tesztelések során – észrevételeikkel, kritikáikkal, hasznos tanácsaikkal támogatták a fejlesztéseket.

Az üzleti világhoz hasonlóan, a hivatalos statisztika minőségügyi fejlődése esetében is természetes folyamat volt kezdetben a statisztikai adat (végtermék) minőségének definiálása, mérése, értékelése. Ezt követte az adott végterméket előállító „termelési folyamat" minőségének meghatározása, mérése és értékelése a hibák megelőzése érdekében, majd az ebből következő fejlesztési tervek végrehajtása (minőségbiztosítás). A statisztikai hivatalok szívós erőfeszítéseket tettek alaptevékenységük (a jó minőségű adatok előállítása a felhasználók már kifejezett, valamint még nem ismert igényei szerint) minőségbiztosításának megalapozására. Az élen járó statisztikai hivatalok az adatminőségre és az adat előállítási folyamatának szakaszaira a minőséget garantáló keretrendszereket dolgoztak ki. Az utóbbi két évtizedben számos statisztikai hivatal adaptálta a Kanadai Statisztikai Hivatal 1985 óta működtetett minőségirányítási irányelveit és az 1997-ben bevezetett minőségbiztosítási keretrendszerét (*Statistics Canada* [2002], [2009]). Ugyanebben az időszakban, a hivatalos statisztika szervezetei és a nemzetközi szervezetek (ENSZ, IMF, EU) egy sor intézményi szintű minőségügyi keretrendszert fejlesztettek ki, melyek hatással voltak a statisztikai hivatalok minőségügyi fejlődésére. A nemzetközi dokumentumokban a statisztikai hivatalok egészére vonatkozó minimumkövetelményeket határoztak meg. Ilyen általános minőségügyi keretrendszernek tekinthetők:

– a hivatalos statisztika alapelvei, ENSZ Statisztikai Bizottsága, 1994 (*KSH* [2004]);

– az Európai statisztikai rendszer minőségügyi deklarációja, 2001 (*KSH* [2005b]);

– a Szakértői Csoport (Leadership Expert Group on Quality – LEG) minőséggel kapcsolatos ajánlásai, 2001 (*KSH* [2005a]);

– adatminőség-értékelési keretrendszer (Data Quality Assessment Framework – DQAF) (*Carson* [2001], *IMF* [2003]);

– az Európai statisztikai rendszer gyakorlati kódexe (*KSH* [2005c]).

A felsorolt dokumentumok közül az elsőről tudni kell, hogy megalkotását az ENSZ Európai Gazdasági Bizottsága kezdeményezte a kilencvenes évek elején, amikor a kelet-európai országokban bekövetkezett rendszerváltozás miatt szükségessé vált a statisztikai hivatalok szerepének újradefiniálása. Ez a tartalmában és terjedelmében is általános elvi nyilatkozat arról szól, hogy milyen feladataik és jogosultságaik vannak a hivatalos statisztika szervezeteinek, azaz milyen szakmai és etikai követelményeket kell teljesíteniük a megbízható, jó minőségű információk biztosítása érdekében.

A második és harmadik dokumentumot az EU vezető minőségügyi szakértőinek munkacsoportja (Leadership Expert Group on Quality – LEG) fogalmazta meg azzal a céllal, hogy jövőképet vázoljon fel, meggyorsítsa a statisztikai hivatalok számára a minőségfejlesztést, hosszú távra kijelölje a megvalósítandó feladatokat (*EC* [2002]).

Az Európai Statisztikai Rendszer (ESR) minőségügyi deklarációja jövőképet ad, és a megvalósításához szükséges alapelveket a minőségirányítás alapelvei szerint rögzíti.

A LEG-csoport minőségre vonatkozó ajánlásai 22 pontban részletezik az EU statisztikai hivatalai előtt álló minőségfejlesztési feladatokat, az élen járó statisztikai hivatalok tapasztalatai alapján.

Az IMF hét statisztikai területre alkalmazott adatminőség-értékelési keretrendszere az ENSZ alapelveire támaszkodik oly módon, hogy a lehető legrészletesebb információkat kéri az egyes országoktól a begyűjtött statisztikai adatok minőségének sokoldalú értékelésére.

Végül, az Európai Unió tagállamaira érvényes Európai statisztikai rendszer gyakorlati kódexe (Gyakorlati kódex)[1] az eddigi legátfogóbb és legrészletesebb minőségügyi követelményrendszer, mely a hozzá csatolt önértékelő kérdőívvel és az önértékelés ellenőrzését biztosító konzultatív szakértői vizsgálattal (peer review) kötelezővé teszi a tagállamok statisztikai hivatalainak átfogó értékelését. A Gyakorlati kódex követelményeinek jövőbeni teljesítését erősíti az a tény, hogy a 2009-ben elfogadott közösségi statisztikáról szóló EU-rendelet hivatkozásként tartalmazza a Gyakorlati kódexről szóló 2005. évi ajánlást (*Európai Unió Hivatalos Lapja* [2009]).

Tanulmányunk első részében áttekintést adunk arról, hogy az EU-országok statisztikai hivatalai milyen minőségügyi keretrendszereket működtetnek, a második részben pedig bemutatjuk, hogy a magyar KSH hol tart saját rendszerének kiépítésében.

1. A statisztikai hivatalok minőségügyi keretrendszerei

Az Európai Unió statisztikai hivatalainak minőségfejlesztési tevékenysége három szakaszra osztható. Az első időszakban (1993 és 1998 között) a skandináv országok statisztikai hivatalai a kanadai és az észak-amerikai statisztikai hivatalok eredményei alapján, egy amerikai tanácsadó cég[2] segítségével megismerkedtek a teljes körű mi-

[1] A Gyakorlati kódexnek a fent felsorolt keretrendszerekkel való megfelelését tárgyaló anyagok elérhetők: http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/code_of_practice/related_quality_initiatives (Elérés dátuma: 2010. június. 23.)

[2] A Westat Inc. az Egyesült Államok egyik munkatársi tulajdonban levő kutatási cége, hosszú évek óta sok ügyféllel dolgozott világszerte a teljes körű minőségirányítás (total quality management – TQM) bevezetésén, így nagy tudásra tettek szert a módszerek és eszközök területén. Ráadásul a Westat ugyanakkor egy nagy statisztikai cég, jól ismeri a statisztikai hivatalok feladatait és problémáit.

nőségirányítás (total quality management – TQM)[3] alapelveinek hivatali alkalmazásával.

A második szakaszban (1998 és 2005 között) a legtöbb minőségügyi tapasztalattal rendelkező Svéd Statisztikai Hivatal kezdeményezésére létrejött az EU vezető minőségügyi szakértőinek munkacsoportja. A LEG-csoport 2001-ben elfogadott minőségfejlesztési ajánlásainak teljesítéséről országonkénti felmérések készültek, a legutóbbi 2009-ben.

Az EU keretében céltudatos minőségfejlesztési tevékenység folyt az utóbbi két évtizedben. Először meghatározták a statisztikai termék és folyamat minőségének jellemzésére és mérésére alkalmas összetevőket és mérőszámokat. Ezután jogszabályokban rögzítették az egyes adatfelvételekhez előírt minőségi kritériumokat, nemzetközi összefogással standard dokumentumokat, értelmező szótárat és kézikönyveket dolgoztak ki és jelentettek meg a minőségügyi kultúra elterjesztése érdekében.[4] Végül, az ezredfordulóra eljött az idő a minőségnek a statisztikai termelési folyamaton túlmutató, átfogó jellegű megközelítésére, azaz a hivatali szintű teljes körű minőség definiálására, az üzleti világban jól ismert TQM elveinek adaptálására.

A minőségfejlesztés harmadik szakaszában, 2005-ben elkészült és ajánlás formájában megjelent a Gyakorlati kódex.

1.1. A kezdetek

Az Amerikai Statisztikai Társaság által szervezett vitaülésen *Colledge* és *March* [1991] a Kanadai Statisztikai Hivatal munkatársai felvázolták a statisztikai hivatalok számára alkalmas minőségirányítási keretrendszer elemeit. Felidézték, hogy a minőség iránti elkötelezettség nagyban hozzájárult Japán üzleti sikereihez, csakúgy, mint az észak-amerikai vállalatok utóbbi húsz évben tapasztalt versenyképességének javításához. A minőségirányítás sokféle néven jelenhet meg (teljes körű minőségirányítás, integrált minőségirányítás, folyamatos minőségfejlesztés), de lényegében mindegyik ugyanarra az üzleti gondolkodásra vezethető vissza, ami *Deming*, *Juran* és *Crosby* munkáiban megtalálható. Az ő alapelveik között nincs ellentmondás, „jóllehet, számos próféta van, de csak egy minőségirányítási vallás létezik" (*Colledge–March* [1991] 724. old.).

[3] „A teljes körű minőségirányítás olyan vezetési filozófia és vállalati gyakorlat, amely a szervezet céljainak érdekében a leghatékonyabb módon használja fel a szervezet rendelkezésére álló emberi és anyagi erőforrásokat. Az Amerikai Egyesült Államokban, az 1980-as évek közepén fogalmazták meg az alapelveit, sok ponton a Japánban kialakult minőségmenedzsment módszerekre és szemléletre alapozva. A TQM felülről, vezetői szintről építkezik. Átfogja az egész szervezet működését; nemcsak a folyamatokra terjed ki, hanem az irányításra és az erőforrásokra is. A hangsúlyt a vevői elégedettségre és a szervezeti működés folyamatos fejlesztésére helyezi." (*Kövesi–Topár* [2009] 16. old.)

[4] A minőség standard dokumentumai elérhetők: http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/quality_reporting (Elérés dátuma: 2010. június. 23.)

A Kanadai Statisztikai Hivatal stratégiai tervezési és szakmai tapasztalatai alapján a szerzők sorra vették azokat a minőségirányítási alapelveket, melyek egy statisztikai hivatal számára is fontosak lehetnek. Röviden felsorolva az alábbiakat tárgyalták:

– a belső és külső felhasználók egyaránt fontosak;
– a beszállítókat/adatszolgáltatókat a folyamat részeként kell kezelni;
– eleve jó minőséget kell előállítani;
– az összes folyamatot folyamatosan fejleszteni kell;
– ki kell alakítani a minőségmérés eszközeit és standardjait;
– első a felső vezetés elkötelezettsége, a minőségpolitika kialakítása;
– ki kell alakítani az integrált minőségirányítási struktúrát (minőségügyi tanács, minőségügyi szervezeti egység, minőségfejlesztő csoportok);
– a minőségfejlesztéshez elengedhetetlen a munkatársak teljes körű bevonása, képzése és elismerése;
– mindehhez fejlett belső és külső kommunikációra van szükség.

Európában a *Svéd Statisztikai Hivatal* úttörő szerepet játszott a minőségfejlesztésben. Jan Carling, a Svéd Statisztikai Hivatal akkori vezetője 1993-ban döntött a TQM-re alapozott minőségfejlesztés bevezetéséről (*Carling* [2002]). Ebben a munkában igénybe vették a Westat Inc. amerikai tanácsadó cég segítségét is, amely a folyamatos minőségfejlesztés területén rendelkezett amerikai és ausztrál tapasztalatokkal, és ezeket az ismereteket később a finn, a norvég, a holland és a dán statisztikai hivatalokkal is megosztotta (*Marker–Morganstein* [2004]). A tanácsadó cég szerint többnyire akkor kezdtek a minőségre figyelni a statisztikai hivatalok, amikor hirtelen és alaposan megváltoztak működési feltételeik, vagy a költségvetési megszorítások vagy az adatgyűjtési piac liberalizálása következtében.

A Svéd Statisztikai Hivatal a minőségügyi munka áttekintése és rendszerezése érdekében öt szervezeti célt fogalmazott meg: a felhasználói igények kielégítése, a termékek és a folyamatok minőségének fejlesztése, a hivatalos statisztikai rendszer vezetése és fejlesztése, a jó kapcsolatok fejlesztése az adatszolgáltatókkal, végül a munkatársak szakképzettségéről és elkötelezettségükről való gondoskodás. A legtöbb kezdeti TQM-tevékenység fejlesztési projekt volt. A folyamatok és a felhasználói kapcsolatok fejlesztésére, az erősségek és a gyenge pontok meghatározására, valamint a célok és stratégiák megfogalmazására összpontosítottak. 1994 és 1999 között hozzávetőleg 150 fejlesztési projektet alakítottak ki, az egész hivatalt átfogó teamek formájában (*Lisai* [2007]).

A *Finn Statisztikai Hivatalt* 1996-ban kezdték átalakítani a termékorientáltról folyamat-orientált szervezetté, ebben segítségükre volt az említett amerikai tanácsadó cég. Két év alatt 120 főnek oktattak alapfokú TQM ismereteket, ezzel egyidejűleg 10 projektet indítottak, valamint 1996-ban és 1998-ban elvégezték a hivatal önértékelését a Finn Minőségdíj kritériumai alapján (*Statistika Centralbyran* [2001] 33. old.).

A *Holland Statisztikai Hivatal* 1996-ban átfogó, egy évtizedre szóló minőségügyi programot kezdett. Amikor a 2000-ig tartó üzleti tervet készítették, fő célkitűzésük volt a minőségügyi rendszerek bevezetése az összes szervezeti egységben. A keretrendszer előzetes irányelveit 1997-ben dolgozták ki, a standardizált modellt 1998-ra véglegesítették. Ezzel egyidejűleg vezették be a statisztikai auditálás rendszerét, melynek célja annak feltárása, hogyan működik a minőségirányítás a statisztikai szervezeti egységekben, valamint hogyan lehet fejleszteni a statisztikai termékek és folyamatok minőségét. Az auditálási gyakorlatukat később több hivatal is alkalmazta. A hivatalban több minőségirányítási rendszert is megvizsgáltak. Úgy találták, hogy az ISO-szabványok[5] bevezetése elég költséges és bürokratikus lenne, továbbá nem kívántak befektetni egy olyan rendszerbe, amely az akkor drámai átalakulási szakaszban levő folyamatokat állandósítaná, ezért egy egyszerűbb, gyakorlatiasabb megközelítést választottak (*EC–SCB–Eurostat* [1999] 61–68. old.).

Bengt Swensson az Uppsalai Egyetem professzora egy 1998-ban végzett kutatásban nyolc vezető statisztikai hivatal minőségirányítási megoldását hasonlította össze (*EC–SCB–Eurostat* [1999] 80–86. old.). A kutatás célja nem a rangsor megállapítása volt, hanem az eltérő gyakorlatok bemutatása az egyes hivatalokban. Ezek között volt a kanadai, az ausztrál, az új-zélandi és a svéd statisztikai hivatal, valamint az Egyesült Államok négy statisztikai hivatala. A TQM megoldások választása érdekében érvényesített kormányzati követelmény országonként igen eltérő volt. Az Egyesült Államok szövetségi kormányzatának 1992-ben kezdett kampánya a TQM-alkalmazások bevezetésére azt eredményezte, hogy néhány évvel később a felmért szövetségi statisztikai intézmények már több mint fele elkezdte a TQM-rendszer megvalósítását. Más országokban nem volt ehhez hasonló kormányzati elvárás a közigazgatásban. A Kanadai Statisztikai Hivatal nem jelentett átfogó minőségirányítási megközelítést, de mind az 1985 óta működtetett minőségirányítási irányelvei, mind az 1986-ban bevezetett minőségbiztosítási keretrendszere megfelel a TQM alapelveinek (*UN Statistical Commission* [2009]). A Svéd Statisztikai Hivatal a felmérés idején éppen elkezdte a TQM-alapú minőségfejlesztési programját.

[5] Az ISO 9000 minőségirányítási rendszerek számára kialakított szabványok csoportja, melyet az Nemzetközi Szabványügyi Szervezet (International Organization for Standardization – ISO) határoz meg. Az ISO 9000 szabványcsalád célja olyan szervezeti teljesítmény biztosítása, amely időről időre képes a vevők minőséggel kapcsolatos követelményeit kielégítő termékeket és szolgáltatásokat nyújtani, függetlenül attól, hogy mit állít elő a szervezet, mekkora a mérete, a magánszektorban vagy a közszférában működik-e (*Magyar Szabványügyi Testület* [2009]: Minőségirányítási rendszerek. Követelmények (ISO 9001:2008), MSZ EN ISO 9001:2009 Beszerezhető: http://www.mszt.hu/tanusitas/mir.html).

1.2. Az EU minőségüggyel foglalkozó szakértői csoportjának (LEG) tevékenysége

A kilencvenes évtized végére az EU-n belül a svéd hivatal rendelkezett a legtöbb minőségirányítási tapasztalattal, így nem volt véletlen, hogy ők kezdeményezték egy ESR-en belüli munkacsoport felállítását. A svéd hivatal által vezetett LEG zárójelentésében kilenc témakört dolgozott ki, melyek közül az első a minőségügyi keretrendszer volt (*EC* [2002]). Ebben a pontban tárgyalták a minőség fogalmától, a termékminőség összetevőitől kezdve a folyamatminőség fejlesztését, a TQM fogalmát és jelentőségét. Az ajánlások listáján a 4. sz. ajánlás fogalmazza meg a minőségfejlesztésre vonatkozó rendszerszemléletű megközelítés követelményét.

„Az ESR minden szervezete alkalmazzon rendszerszemléletű megközelítést a minőség javítására. Az ESR tagjai fejlesztési munkájuk alapjaként használják az Európai Minőségfejlesztési Alapítvány (üzleti) kiválóság modelljét (European Foundation for Quality Management – EFQM, excellence model), hacsak már más, hasonló modellt nem alkalmaznak." (*KSH* [2005a])

1. ábra. Az EFQM Kiválóság Modell kritériumai és a hozzájuk tartozó súlyok és pontszámok (EFQM [2009])[6]
(a 2010. évi változat)

<table>
<tr><td colspan="3">Adottságok (50 százalék, 500 pont)</td><td colspan="2">Eredmények (50 százalék, 500 pont)</td></tr>
<tr><td rowspan="3">Vezetés
(10 százalék,
100 pont)</td><td>Munkatársak
(10 százalék,
100 pont)</td><td rowspan="3">Folyamatok, termékek és szolgáltatások
(10 százalék,
100 pont)</td><td>Munkatársakkal kapcsolatos eredményei
(10 százalék,
100 pont)</td><td rowspan="3">Kulcs-eredmények
(15 százalék,
150 pont)</td></tr>
<tr><td>Stratégia
(10 százalék,
100 pont)</td><td>Vevőkkel kapcsolatos eredmények
(15 százalék,
150 pont)</td></tr>
<tr><td>Partnerkapcsolatok és erőforrások
(10 százalék,
100 pont)</td><td>Társadalommal kapcsolatos eredmények
(10 százalék,
100 pont)</td></tr>
<tr><td colspan="5">Az eredmények hatása: tanulás, kreativitás és innováció</td></tr>
</table>

[6] A Gyakorlati kódexnek az EFQM Kiválóság Modellel való megfelelését tárgyaló anyag elérhető az alábbi linken: http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/code_of_practice/related_quality_initiatives (Elérés dátuma: 2010. június. 23.)

Az EFQM Kiválóság Modell olyan keretrendszer, amely 9 kritériumon alapul, ebből öt az „Adottságok" és négy az „Eredmények" részhez tartozik. Az „Adottságok" kritériumai fedik le azt, amit a szervezet végez. Az „Eredmények" kritériumai arra vonatkoznak, amit a szervezet elér. Az „Adottságokból következnek az „Eredmények", és az „Eredmények" visszajelzéseit használják az „Adottságok" fejlesztéséhez.

Az EFQM Kiválóság Modell több célra alkalmazható:

– az önértékelés eszközeként,
– más szervezetekkel való viszonyítási alapként és
– irányelvként a fejlesztendő területek azonosításához.

A LEG-ajánlások teljesülését 2002-ben és 2003-ban is felmérték, erről 2004-ben az Eurostat egy összefoglaló jelentést adott (*Eurostat* [2004]). A LEG-ajánlások teljesülését vizsgáló jelentés szerint nem teljesült a 4. sz. ajánlás. Tizenhat statisztikai hivatal válaszolt a felmérésre, ezek közül csak kilenc alkalmazott EFQM- vagy hasonló modellt, öt hivatal pedig tervezte ilyen modell alkalmazását 2005-ig.

Az Eurostat 2009. évi jelentésében javasolta, hogy az EFQM-modell alkalmazására vonatkozó LEG-ajánlást felül kellene vizsgálni, egyrészt a Gyakorlati kódex megvalósítási tapasztalatai alapján, valamint azért, mert, újabb fejlemények utalnak a Közös Értékelési Rendszer (Common Assessment Framework – CAF)[7] közigazgatáson belüli alkalmazásának vagy az ISO tanúsításnak a térnyerésére (*Eurostat* [2009]).

1.3. A Gyakorlati kódex elveinek teljesülése (2005–2009)

Az Európai Bizottság 2005. május 25-én elfogadott közleményében ajánlásként jelentetett meg egy szabályzatot, „Az európai statisztika gyakorlati kódexe" (a továbbiakban Gyakorlati kódex) címmel (*Európai Közösségek Bizottsága* [2005]). A Gyakorlati kódex előzménye az volt, hogy az EU gazdaságpolitikáját és a költségvetési politikáját irányító testületek egyre határozottabban fogalmazták meg igényüket a megbízható költségvetési és statisztikai adatok iránt. A 2004. évi görög költségvetési adatok eltitkolása és utólagos korrekciója felgyorsította az Európai Unió gazda-

[7] CAF teljes körű minőségirányítási eszköz a közigazgatás intézményei számára, amelyet az Európai Minőségirányítási Alapítvány (EFQM) Kiválóság Modellje és a németországi Speyer Közigazgatás-tudományi Egyetemének modellje alapján a maastrichti Európai Közigazgatási Intézeten (EIPA) belül hoztak létre a közigazgatásért felelős főigazgatók döntése nyomán. Az EFQM-modellhez hasonlóan 9 kritériumot, de csak 28 alkritériumot tartalmaz. A CAF kísérleti változatát 2000 májusában mutatták be, jelenleg a 2006. évi változat van érvényben, mely a MEH honlapjáról letölthető: https://caf.meh.hu/

ságpolitikai testületeinek döntését: EU-szintű követelmények felállítását a nemzeti statisztikai hivatalok függetlenségével, integritásával és elszámoltathatóságával kapcsolatban.

A Gyakorlati kódex az Eurostat, a nemzeti statisztikai hivatalok és az Európai Statisztikai Rendszerhez tartozó más intézmények számára határoz meg 15 követendő elvet. Az intézményi környezet, a statisztikai folyamatok, valamint a statisztikai adatok mint fő fejezetek köré csoportosított elvek a következők.

Intézményi környezet

1. Szakmai függetlenség
2. Felhatalmazás adatgyűjtésre
3. Megfelelő erőforrások
4. A minőség iránti elkötelezettség
5. A statisztikai adatok bizalmas kezelése (adatvédelem)
6. Pártatlanság és objektivitás

Statisztikai folyamatok

7. Megalapozott módszertan
8. Megfelelő statisztikai eljárások
9. Az adatszolgáltatók terheinek csökkentése
10. Költséghatékonyság

Statisztikai termékek

11. Relevancia
12. Pontosság és megbízhatóság
13. Gyorsaság/időszerűség és pontosság
14. Koherencia és összehasonlíthatóság
15. Hozzáférhetőség és érthetőség

Az egyes elvek tartalmát pár mondatos értelmező magyarázatban foglalták össze. Az elveken belül találjuk a teljesülésük igazolására szolgáló konkrét célokat, teljesítendő követelményeket, ezeket ismérveknek nevezték el a kódexben.[8] A Gyakorlati kódexben rögzített elvek teljesítésének ellenőrzésére, 2005 végére minden statisztikai hivatalnak önértékelést kellett elvégeznie egy – a kódex alapján kidolgozott – egységes, részletes kérdőív kitöltésével, melyben megismételték a minőségirányítási rendszer meglétére vonatkozó kérdést.

[8] A Gyakorlati kódexet elsőként *Szilágyi György* ismertette a *Statisztikai Szemlében* (*Szilágyi* [2005]).

2006 és 2008 között megtörtént a kódex elvei végrehajtásának külső ellenőrzése is ún. konzultatív szakértői vizsgálat (peer review) formájában. Ennek során az Eurostat és a tagállamok képviselőiből álló három fős szakértőcsoportok háromnapos látogatást tettek az egyes statisztikai hivataloknál. Az előzetesen megkapott hivatali anyagok megismerését követően találkoztak a különböző beosztású munkatársakkal, az adott statisztikai elvért felelős szakértőkkel, az adatszolgáltatók és az adatfelhasználók képviselőivel. Ezután értékelték az adott hivatal teljesítményét, rögzítették a legjobb megoldásokat és a fejlesztendő területekre vállalt akcióterveket. Az Eurostat rendszeresen jelentést kér a statisztikai hivataloktól a Gyakorlati kódex elveinek megvalósításáról és a fejlesztési tervek teljesítéséről, melyek alapján összefoglaló készül az Európai Parlament és a Tanács részére (*EC* [2008]).

A Gyakorlati kódex bevezetésével az EU statisztikai hivatalai egy egységes minőségértékelési keretrendszert kaptak, mely kifejezetten a hivatalos statisztika előállítói számára készült, elvei levezethetők az EFQM értékelési keretrendszerből, ugyanakkor az országok közötti összehasonlításokra is alkalmas eszköznek bizonyult (*Voineagu et al.* [2009]). A Gyakorlati kódexben az intézményi környezet keretében előírt elvek olyan követelményeket tartalmaznak, melyeket egy átlagos statisztikai hivatal nehezen tudott volna önmaga számára követelményként megfogalmazni (például a Szakmai függetlenség, a Felhatalmazás adatgyűjtésre, a Megfelelő erőforrások, a Pártatlanság és objektivitás). Nem véletlen, hogy az Ausztrál Statisztikai Hivatal kiegészítette saját Adatminőségi Keretrendszerét a Gyakorlati kódex intézményi környezetére vonatkozó követelményekkel, amit ebben a formában már más ausztrál kormányzati intézmények is alkalmaznak.

A konzultatív szakértői vizsgálat során azonosított jó gyakorlatok közé három ország statisztikai hivatalának minőségirányítási rendszerét választották ki (*Eurostat* [2008]).

Szlovákia. A Szlovák Statisztikai Hivatal az ISO 9001:2000 szabvány alapján fokozatosan vezette be a minőségirányítási rendszerét. Az értékteremtő folyamatok közé sorolták a nemzeti számlát és makrogazdasági statisztikát, az üzleti statisztikát, a társadalomstatisztikát és demográfiát, valamint a területi statisztikát. 2005 és 2007 között az összes folyamatot dokumentálták és auditálták a fejlesztendő területek azonosítása céljából. A statisztikai hivatal 2006 novemberében nyerte el az ISO 9001:2001 szabvány szerinti tanúsítást, amit 2009-ben megújítottak.

Portugália. A Portugál Statisztikai Hivatalban az ISO-szabványok szerinti minőségirányítási rendszert vezettek be 1996-ban. Létrehoztak egy minőségirányítási szervezeti egységet, ők koordinálják a hivatalon belül a minőségfejlesztési tevékenységeket, ők felügyelik a statisztikai termelési eljárások kézikönyvének felülvizsgálatát és a belső és külső auditok jelenlegi megújítását is.

Finnország. A Finn Statisztikai Hivatalban a minőség ISO-fogalmából indultak ki. A TQM-et a kilencvenes évek elején bevezették, a szervezeti szintű minőségirá-

nyítás az EFQM-modell és a Kiegyensúlyozott Mutatószámrendszer (Balanced Scorecard – BSC)[9] kombinált alkalmazásával valósul meg, ami 2005 óta kiegészül a Gyakorlati kódex alapján végzett önértékeléssel, a kódexhez kapcsolódó kérdőívvel, valamint a konzultatív szakértői vizsgálat eredményeit felhasználva más statisztikai hivatalokkal való összehasonlítással.

1.4. A statisztikai hivatalok jelenleg alkalmazott minőségirányítási rendszerei

Huszonhét európai statisztikai hivatalról összegyűjtött információk alapján a következő példákat találtuk a minőségirányítási rendszerek alkalmazásáról.

– A 2005-ben megjelent Gyakorlati kódex önként vállalt, de a hozzá csatolt kérdőívvel és az EU-s szakértői ellenőrzésekkel gyakorlatilag kötelezővé tett alkalmazása új minőségirányítási rendszerré vált az EU valamennyi statisztikai hivatala számára.

– A 2001. évi LEG-ajánlás hatására kilenc ország statisztikai hivatala alkalmazta az EFQM Kiválóság Modellt, a modell alkalmazását tervezik további 4 országban.

– Két statisztikai hivatal tervbe vette az EU közigazgatási intézményeire kidolgozott minőségirányítási rendszer: a Közös Értékelési Keretrendszer (CAF) bevezetését.

– Három ország az ISO 9001:2000-es Minőségirányítási szabvány alkalmazását választotta, az Egyesült Királyság statisztikai hivatala csak egy statisztikai területre próbálta ki.

– Hét ország (Dánia, Norvégia, Franciaország, Szlovénia, Egyesült Királyság, Hollandia, Magyarország) statisztikai hivatalai pragmatikus megoldást választottak, azaz stratégiai tervezésükbe építették be a minőségirányítás alapelveit és eszközeit, így egy formális modell választása nélkül is rendszerszemléletű minőségfejlesztést alkalmaznak.

– Hollandia az utóbbi években egy saját fejlesztésű minőségirányítási modell kísérleti alkalmazását kezdte el (*van Nederpelt* [2009]).

– A statisztikai hivatalok egy csoportja a Gyakorlati kódexen kívül többféle modell alkalmazását vállalta. (Lásd az 1. táblázatot.)

[9] A BSC egy vállalatirányítási rendszer, amely a szervezet jövőképét, illetve hosszú és középtávú kitekintő üzleti tervét célokra és mutatókra bontja négy nézőpontból. A négy nézőpont: a pénzügy (Mit várnak tőlünk a tulajdonosok?), a vevők (Mit várnak tőlünk a vevők?), a működés folyamatai (Milyen folyamatokban kell kiemelkedőt nyújtanunk?) és a tanulás (Hogyan őrizhetjük meg fejlődési képességünket?).

1. táblázat

Az egyes országok statisztikai hivatalaiban alkalmazott minőségirányítási rendszerek

Országok	Minőségirányítási rendszerek					
	Gyakorlati kódex	EFQM	CAF	ISO 9001:2000	BSC	Egyéb
Ausztria	X	X				
Belgium	X	Xt				
Bulgária	X					
Csehország	X	X				
Dánia	X					X
Egyesült Királyság	X			X		X
Eurostat	X	X				
Észtország	X	X				
Finnország	X	X			X	X
Franciaország	X					X
Hollandia	X					X
Írország	X	Xt				
Lengyelország	X	Xt				
Lettország	X					
Litvánia	X			X		
Luxemburg	X		Xt			
Németország	X	X				
Magyarország	X					X
Málta	Xt					
Norvégia	X					X
Portugália	X	X		X		
Románia	X	Xt				
Spanyolország	X	X				
Svájc	X		Xt			X
Svédország	X	X			X	X
Szlovákia	X			X		
Szlovénia	X					X
Összesen	*27*	*13*	*2*	*4*	*2*	*10*

Megjegyzés. Xt – tervezett.

Az EU-országok statisztikai hivatalainak minőségfejlesztési története azt bizonyítja, hogy egyrészt idő- és erőforrás-igényes feladat a minőség kultúrájának elterjesztése, másrészt erre alapozva többféle minőségügyi rendszer alkalmazása a jövő

útja. Ezt a fejlesztési folyamatot gyorsította az EU-n belül megszervezett információcsere, a legjobb gyakorlatok elterjesztése, együttműködés a kutatás és fejlesztés területén. Minden statisztikai hivatalnak hasonló feladatokat kell teljesíteni, amihez folyamatosan fejleszteniük kell a szervezetüket, a minőség komplex kezelése érdekében. Ez a nagy, decentralizált statisztikai hivataloknak más feladat, mint a kicsi, centralizált hivataloknak. Fontos tényező a minőségfejlesztési utak, módszerek kiválasztásában az adott ország statisztikai törvényi szabályozásának és a költségvetésének mindenkori helyzete, az infrastruktúra fejlettsége, az ország kulturális hagyományai. A legfontosabb azonban a hivatal vezetésének szándéka és elkötelezettsége.

2. Minőségügyi keretrendszer a magyar Központi Statisztikai Hivatalban

A KSH 2003-ban kezdett rendszerszerűen foglalkozni a hivatalos statisztika minőségének témájával, melyet a saját és más hivatalok minőséggel kapcsolatos tapasztalataira, az ENSZ és az Eurostat fejlesztéseire, valamint EU-s tréningek és projektek tapasztalataira alapozott. A folyamat során a KSH kialakította a minőséggel kapcsolatos jövőképét és bevezette *Michael Colledge* nemzetközi szakértő ajánlásait. 2005-ben elkészült a 2005 és 2008 közötti időszakra szóló, a Hivatal honlapján is elérhető középtávú stratégiai terv (*KSH* [2005e]). Ebben külön fejezet foglalkozik a KSH jóváhagyott minőségügyi koncepciójával, mely a három fő elemre bontható: minőség melletti elkötelezettség, a minőség definíciója és összetevői, statisztikai folyamatok és termékek minőségmérése, értékelése.

2.1. A minőség stratégiai megközelítése

A stratégia számos egyéb olyan célkitűzést, főirányt is kijelölt, melyek összhangban vannak a TQM alapelveivel. Ezek közé sorolható többek között a „Programtervezés és önértékelési rendszerek", a „Felhasználói kapcsolatok javítása" vagy a „Metainformációs rendszer" fejlesztése. A legfőbb minőséggel kapcsolatos főirány, a „Statisztikai termékek és folyamatok minőségbiztosítása"-nak célja egy belső minőségügyi keretrendszer kifejlesztése volt, a hivatal által szolgáltatott információk megbízhatóságának, minőségének javítása érdekében. A KSH-ban az elkészült statisztikai termékek minőségének értékelése, és a statisztikai adat-előállítási folyamat tartalmi értelemben vett minőségbiztosítása a korábbiakban is működött, ugyanakkor szervezeti egységenként jelentősen eltérő volt. A minőséggel foglalkozó főirány célja az volt, hogy a már mű-

ködő eljárásokra, módszerekre alapozva, kiindulva a meglévő követelményekből, egy szervezett, dokumentált, ellenőrzött minőségbiztosítási, minőségértékelési rendszert alakítson ki. A közzétett minőségjellemzők segítségével a felhasználók, az adatszolgáltatók és -előállítók világos és sokoldalú képet kaphatnak a statisztikák megbízhatóságáról. A főirány konkrét feladatai négy részből tevődtek össze.

1. A statisztikai termékek belső minőségellenőrzési rendszerének (kritériumok, mérési módszerek, értékelési eljárás, dokumentációs rendszer) kialakítása.

2. A fontosabb statisztikai munkafolyamatok minőségi kritériumai, standardok, dokumentálás.

3. A hivatal dolgozóinak minőségügyi képzése.

4. A minőségirányítás elemeinek összekapcsolása egy egységes rendszerbe, modellbe.

A minőséggel kapcsolatos projektek kivitelezésére a Statisztikai kutatási és módszertani főosztályt jelölték ki (*Szép–Mag–Vigh* [2006]). A Termékminőség-projekt keretében termékminőség-indikátorokat határoztunk meg, számítási módszerekkel és értékelési kritériumokkal, összeállítottunk egy standard minőségjelentés sémát és kialakítottuk ezek dokumentációját. A Folyamatminőség projekt során a statisztikai adat-előállítási folyamat egyes lépéseire minőségi irányelveket határoztunk meg, folyamatváltozókat fejlesztettünk ki az egyes folyamatszakaszokhoz és elkészítettük az önértékelés alapjául szolgáló DESAP (Önértékelés-fejlesztési projekt – Development of a Self Asessment Programme) önértékelő kérdőív[10] KSH-ra adaptált változatát.

A 2009 és 2012 közötti időszakra vonatkozó új középtávú stratégia szintén hangsúlyozza a KSH elkötelezettségét a minőség iránt, és ezt az elkötelezettséget a Hivatal egyik alapértékeként nyilvánítja ki a dokumentumban. A stratégiában a minőség témaköre több helyen is megjelenik, az intézmény irányítási és szervezetfejlesztési témájához kapcsolódóan a KSH célul tűzte ki egy intézményi szintű minőségirányítási rendszer bevezetését (*KSH* [2009b]).

2.2. Minőségpolitika

A Hivatal 2009 decemberében tette közzé stratégiájával összhangban levő minőségpolitikáját, mely magyarul és angolul is elérhető a honlapján (*KSH* [2009a]). Tartalmának legfőbb pontjai a következők.

[10] Az ESR adatfelvételeire kialakított egységes európai ellenőrzőlista, ami az adatfelvételek vezetői számára készült, azzal a céllal, hogy egységes alapokon értékelje a felvételek folyamatszakaszainak minőségét. Megtalálható az alábbi linken: http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/quality_reporting (Elérés dátuma: 2010. június 23.)

„A Központi Statisztikai Hivatal küldetése, hogy az Európai Statisztikai Rendszer részeként a társadalom, a gazdaság és a környezet állapotáról és változásairól a felhasználók igényeinek megfelelő, hiteles, jó minőségű statisztikai szolgáltatást nyújtson."

„A Központi Statisztikai Hivatal termékei a felhasználók számára átadott, publikált statisztikai adatok, statisztikai elemzések, osztályozások, regiszterek, módszerek, továbbá szolgáltatásként végzett egyéb statisztikai tevékenységek. A KSH által létrehozott statisztikai termékeknek – összhangban az ISO minőségdefiníciójával – felhasználásra alkalmasnak kell lenniük és meg kell felelniük a következő, egymással kapcsolatban levő minőség-összetevőknek:"

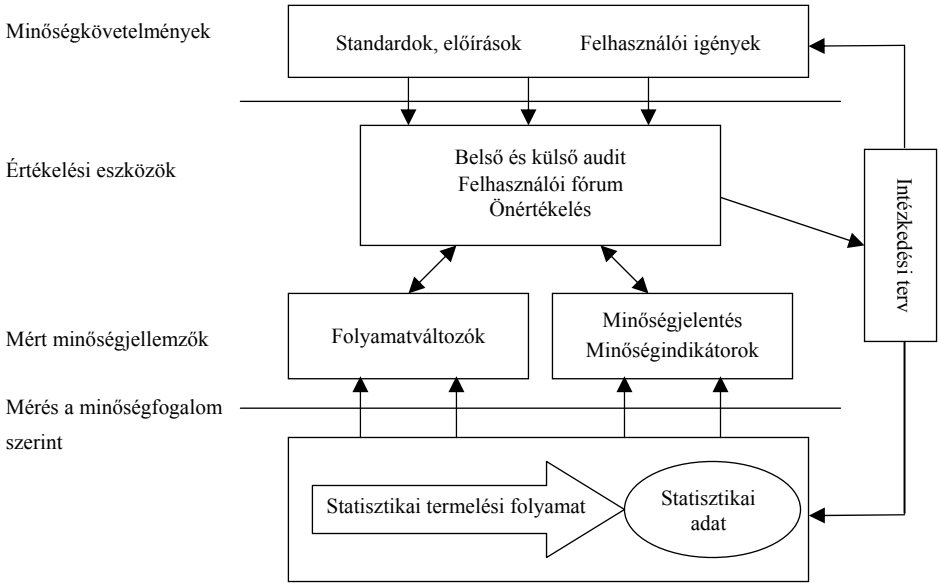
– relevancia;
– pontosság;
– időszerűség;
– időbeli pontosság;
– hozzáférhetőség;
– érthetőség;
– összehasonlíthatóság és koherencia.

Bár nem méri a minőséget, a statisztikák előállításának költsége és a válaszadói terhek szintén hatnak a minőségre. Az erőforráskorlátok ellenére, összességében, törekedni kell a legmagasabb szintű minőségre. „A KSH tevékenységét az Európai Statisztika Gyakorlati Kódexében foglalt alapelvekkel összhangban végzi, és elősegíti ezek érvényesülését a hivatalos statisztika más hazai intézményeinél is. Az európai szinten kialakított, egységes önértékelési rendszer keretében időről időre jelentést készít a kódex alapelveinek teljesüléséről. Az ezzel kapcsolatos dokumentumokat a KSH a honlapján magyar és angol nyelven közzéteszi." (*KSH* [2009a])

2.3. A KSH minőségügyi keretrendszerének elemei és működése

A statisztikai hivatalokra vonatkozó minőségügyi keretrendszer elemeivel, egységes koncepciójukkal, kialakításukkal az Eurostat Quality in Statistics elnevezésű munkacsoportja foglalkozik és valamennyi tagország részvételével évente egyszer üléseznek. A Grant-pályázat keretében német vezetéssel, a KSH részvételével nemzetközi munkacsoport dolgozta ki az adatminőség értékelésére ajánlott eszközöket és módszereket áttekintő kézikönyvet (*Bergdahl et al.* [2007]). A kézikönyvvel összhangban a hivatal minőségügyi keretrendszerének megvalósult, illetve tervezett elemeit a következő ábra szemlélteti.

2. ábra. Minőségmérés és -értékelés elemei és eszközei



2.3.1. Minőségkövetelmények

A minőségkritériumok több forrásból is származhatnak. Lehetnek nemzetközi (ENSZ- és EU-szintű) jogszabályok és ajánlások, hazai jogszabályok, KSH-szintű követelmények, de a felelősségi körébe tartozó szakstatisztikák esetében lehet(nek) valamely szakfőosztály saját vállalása(i) is. A szakfőosztályok a Tájékoztatási főosztállyal együttműködve felelősek az adott szakstatisztika felhasználói igényeinek megismeréséért, önállóan pedig a szakstatisztikával szembeni követelmények, az előző termelési ciklus tapasztalataiból adódó követelménymódosítások kitűzéséért.

Összhangban az Eurostat ajánlásaival, a stratégia keretében belső standardokat és irányelveket dolgoztunk ki. A „Statisztikai termékek és folyamatok minőségbiztosítása" főirány részeként a Folyamatminőség-projekt első eredményeként meghatároztuk a KSH statisztikai termelési folyamatát leíró folyamatmodellt, mely a minőség szempontjából meghatározó tizenkilenc szakaszra bontja az adatelőállítási folyamatot. Ennek alapján elkészült a folyamattal szembeni, folyamatszakaszonkénti legfőbb elvárásokat tartalmazó minőségi irányelvekről szóló dokumentum, mely széles körű belső konzultáció után, elnöki előírásként jelent meg 2007-ben. A kézikönyv tartalmazza az egyes folyamatszakaszok meghatározását, néhány rövid irányelvet ad azon alapelvekre, ajánlásokra és módszerekre, melyek figyelembe vétele nélkülözhetetlen a termelési folyamat végrehajtása során. A kézikönyv szerkezete követi a statisztikai termelési folyamatot.

Ennek megfelelően a következő fejezetekből áll.

– Regiszterek kiválasztása
– Felvételi keret meghatározása
– Célok, felhasználás és felhasználók meghatározása
– Fogalmak, definíciók és osztályozások meghatározása
– Felhasználható igazgatási adatok számbavétele, statisztikai hasznosításuk
– Mintavételi terv kialakítása
– Kérdőív és segédanyagainak tervezése
– Adatgyűjtés szervezése és adatgyűjtés
– Adatok előkészítése (rögzítés, editálás)
– Imputálás (pótlás)
– Súlyozás, becslés és mintavételi hiba számítása
– Indexszámok képzése
– Makrovalidálás (aggregált adatok validálása)
– Szezonális kiigazítás
– Elemzések készítése
– Adatok bizalmas kezelése és a felfedhetőség elleni védelem
– Tájékoztatás
– Archiválás
– Értékelés, felülvizsgálat, visszacsatolás

Az egyes fejezetek szerkezete megegyezik: leírások, alapelvek, minőségi irányelvek, vonatkozó legfontosabb jogszabályok, és módszertani dokumentumok. A Minőségi irányelveket 2009-ben felülvizsgáltuk, aminek eredményeként elnöki szabvány formájában elkészült a dokumentum második változata, mely a legújabb elvárásokat, a végrehajtott fejlesztések miatt szükségszerű változásokat is tartalmazza. A dokumentum mindenki számára elérhető a KSH honlapján (*KSH* [2010]).

2.3.2. Minőségmérés

A KSH minőségmérési eszközei a termékminőség-indikátorok, a minőségjelentés és a folyamatminőség-indikátorok. Ezeket az eszközöket a már említett stratégiai fejlesztési projektek keretében dolgoztuk ki. A KSH-ban elfogadott és alkalmazott minőségindikátorok számításával kapcsolatban jelenleg az a legfőbb célkitűzés, hogy megteremtsünk egy olyan informatikai rendszert, mely lehetővé teszi, hogy a folyamaton belül a minőségindikátorok és a minőségértékelés szempontjából fontos dokumentáció automatikusan előálljon. Ez nagymértékben csökkentené a szakfőosztályok terheit, és hatékonyabbá tenné a minőségértékelés folyamatát.

– Ennek érdekében 2008-ban KSH elnöki előírás formájában bevezettük *a termékminőség-indikátorok* rendszerét. A 24 indikátort tartalmazó indikátorrendszer összhangban az Eurostat ajánlásával, az elfogadott minőség-összetevőkhöz rendeli az indikátorokat. A belső szabályozás szerint az indikátorokat akkor kell számolni, amikor a termék, azaz a statisztikai adat elkészült. A publikáció gyakorisága szerint írásos kiegészítő magyarázatot is mellékelni kell. Az indikátorok számítása és az értelmezést segítő kiegészítő magyarázatok elkészítése a szakfőosztályok feladata.

– A KSH *Minőségjelentése* egy részletes, számszerű és szöveges információkat egyaránt tartalmazó, az egyes szakstatisztikákhoz tartozó statisztikai termékek minőségét bemutató jelentés. Évente egyszer kell véglegesíteni, a havi és a negyedéves adatfelvételek adatait a referenciaév elején kell megadni, ezután a jelentést frissíteni kell az év során, a publikáció gyakorisága szerint. A jelentés, az Eurostat ajánlásának megfelelően, a minőség-összetevők mentén van fejezetekre bontva.

A folyamatminőség-indikátorok a különböző termelési folyamatszakaszokkal kapcsolatban szolgáltatnak információt a minőségről. A termékminőség-indikátorokkal szemben, a folyamatváltozók legfőbb célja, hogy még az adat-előállítási folyamaton belül képet kapjunk a végrehajtott feladatok minőségéről, az adatok minőségére gyakorolt hatásukról, hogy szükség esetén a felelősök be tudjanak avatkozni, ki tudják javítani, vagy korrigálni tudják az előforduló hibákat. Mivel a különböző termelési folyamatok nagymértékben eltérhetnek egymástól, ezért a stratégiai fejlesztési projekt keretében elkészült katalógus egyrészt általános irányelveket tartalmaz az indikátorok kidolgozásához, másrészt minden egyes folyamatszakaszra meghatározza a leggyakrabban, általánosan használható folyamatváltozók körét. Az így meghatározott indikátorok felhasználás szempontjából két csoportba sorolhatók: az utólagos ellenőrzéshez, valamint a minőségbiztosítást támogató, folyamat közbeni ellenőrzéshez szükséges indikátorokra.

– *A felhasználói elégedettség* mérésére a Gyakorlati kódex elveivel összhangban a KSH egy átfogó, több különböző modult magában foglaló rendszert fejlesztett ki, melynek segítségével a felhasználói vélemények strukturált formában összegyűjthetők és elemezhetők. Az aktuális prioritásoknak megfelelően a KSH évenként több felmérést is végrehajt.

2.3.3. Minőségértékelés

A minőségértékelés legfontosabb feladatai a dokumentált minőség és az elvárások összehasonlítása, a megfelelés fokának meghatározása, a felmerülő hibák és hiányosságok feltárása, szükség esetén ajánlások vagy javaslat az intézkedési terv kidolgozására a kockázatok csökkentése és a követelmények vagy más elemek módosítása érdekében.

A minőségértékelés eszközei a következők lehetnek.

Önértékelés. A belső standard önértékelési kérdőívet a DESAP önértékelő kérdőív alapján, a KSH minőségügyi irányelveire alkalmazva fejlesztették ki. A kérdőív célja az adatfelvételek rendszeres és strukturált minőségértékelése, valamint a termékminőség értékelése.

Az önértékelő kérdőívet az adatfelvétel felelőse tölti ki, de más szakértők bevonása is szükséges a folyamatba. A kérdőív hasznos eszközként szolgál a szükséges intézkedések összegyűjtésére és a fejlesztési feladatok kijelölésére. A kérdőív kitöltése minden felvételre javasolt, de csak az auditok előtt, az új vagy jelentősen módosuló felvételek esetében kötelező. Megjegyezzük, hogy a KSH Minőségjelentése szintén tartalmaz a szakstatisztikák önértékelésével kapcsolatos kérdéseket, ezek azonban a statisztikai adat és nem a folyamat szempontjából értékelik a minőséget.

További értékelési eszközök a visszacsatolás a felhasználóktól és az auditok. Jelenleg ezeket a KSH nem alkalmazza, bár készültek tervek a későbbi bevezetésre, a rendelkezésre álló erőforrásoktól függően.

Visszacsatolás a felhasználóktól (feedback-talks). Célunk, hogy a fő felhasználókkal folytatott strukturált konzultációk révén információkat szerezzünk, elsőként a fő szakstatisztikákról, majd később szélesebb körben. A felhasználói vélemények megismerése történhet felméréssel vagy felhasználói fórum segítségével, de a legfontosabb felhasználókkal a folyamatos kapcsolattartás is elvárható.

Belső és külső minőségügyi audit. A belső és külső audit az értékelésbe vont szakértők, érdekeltek körének bővítésével több, mint az önértékelés. Sokat segíthet a szakstatisztika fejlesztésében egy külső szem, egyedi szempontok, ismeretek szakértői megjelenítésével. Belső és külső minőségügyi auditokat tervezünk előbb a főbb szakstatisztikákra, majd később igény szerint kiterjesztve. Az értékelési eljárás egy elfogadott ütemterv, értékelési terv szerint fog történni.

2.3.4. Intézkedési tervek

Az értékelési folyamatok (jelenleg önértékelések) eredményei alapján szükség esetén a szakfőosztályok intézkedési terveket készíthetnek. Ha a tervben foglaltak csak a főosztályt érintik, akkor ezeket be lehet építeni a rendszeres éves tervezési folyamatba. Ebben az esetben a szükséges források biztosítva vannak és az adott főosztály a felelős a tervezett intézkedések kivitelezéséért. Szélesebb kört érintő intézkedés esetén más főosztályok ajánlásait is figyelembe lehet venni, és szükség esetén projekt kezdeményezhető. Minthogy az intézkedési tervek az értékelések visszacsatolásai, így az értékelések tulajdonképpeni okát és célját adják.

Az összegzett jelentéseken keresztül átfogó intézkedési tervek készülhetnek intézményi szinten. A minőségügyi keretrendszert az MTA Statisztikai Bizottsága 2009. decemberi ülésén megvitatta, és egyetértéssel nyugtázta az elért eredményeket.

2.4. A minőségügyi keretrendszer jövőbeni fejlesztései

A minőségügyi keretrendszer továbbfejlesztése a KSH 2009 és 2012 közötti időszakra vonatkozó stratégiájában (*KSH* [2009b]) megfogalmazottakkal továbbá a keretrendszernek az egész hivatalos statisztikai szolgálatra való kiterjesztésével valósulhat meg. A statisztikai törvény aktuális felülvizsgálatánál a minőségügyi szempontokat is célszerű megjeleníteni.

*

A statisztikai adatok iránti igények növekedése, valamint az ugyanakkor szűkülő anyagi és emberi erőforrások szükségessé teszik a termelési folyamat hatékonyabbá tételét, a minőségbiztosítási eszközök fejlesztését, rendszerbe illesztett működtetését. A KSH-ban a minőségügyi keretrendszer fő elemeinek kialakítása az előző évek fejlesztései alapján befejeződött, lehetővé vált a rendszer működtetése.

A rendszer kialakításával párhuzamosan a fejlesztés európai szinten is folytatódott. A Gyakorlati kódexhez kapcsolódóan európai szinten jelentős előrelépés történt, amennyiben a 2009-ben elfogadott új európai statisztikai törvény pontos előírásokat tartalmaz az adatokkal egyidőben szolgáltatandó minőségjelentés tartalmáról. A KSH 2009-ben tizenhárom területen huszonhét minőségjelentést küldött az Eurostatnak, s ez a szám évről évre nő. A Gyakorlati kódex továbbfejlesztésére, az európai statisztikai rendszer függetlenségének, integritásának és elszámoltathatóságának felügyeletére 2008-ban létrehozták az európai statisztikairányítási tanácsadó testületet[11] (European Statistical Governance Advisory Board – ESGAB) (*Európai Unió Hivatalos Lapja* [2008]), és egy időszaki munkacsoportot.

A minőségügyi keretrendszerek témaköre az ENSZ Statisztikai Bizottsága napirendjén is szerepelt az ez évi bizottsági ülésen. A Kanadai Statisztikai Hivatal javasolta, hogy a Bizottság fogadjon el egy egységes minőségbiztosítási keretrendszersémát (*UN Statistical Commission* [2010]), amely lehetővé tenné az egyes országok speciális rendszereinek egységes elvek szerinti leírását, és így átláthatóságát is. A Bizottság a javaslatot támogatta, jelenleg egy munkacsoport dolgozik a 2011. évi ülésre[12] készülő előterjesztésen.

Irodalom

BERGDAHL, M. ET AL. [2007]: *Handbook on Data Quality Assessment Methods and Tools (DatQAM)*. European Commission. Wiesbaden. http://epp.eurostat.ec.europa.eu/portal/page/

[11] A testület honlapja elérhető: http://epp.eurostat.ec.europa.eu/portal/page/portal/esgab/introduction (Elérés dátuma: 2010. június 23.)

[12] A kanadai anyag és az országoktól, nemzetközi szervezetektől beérkezett javaslatok elérhetők az alábbi linken: http://unstats.un.org/unsd/dnss/nqaf.aspx (Elérés dátuma: 2010. június 23.)

portal/quality/documents/HANDBOOK%20ON%20DATA%20QUALITY%20ASSESSMENT%20METHODS%20AND%20TOOLS%20%20I.pdf (Elérés dátuma: 2010. június 23.)

CARLING, J. [2002]: Systematic Quality Work Experiences from Statistics Sweden and Other European Statistical Institutes. *The Survey Statistician*. 46. sz. 19–22. old. http://isi.cbs.nl/iass/N461part1.pdf (Elérés dátuma: 2010. június 23.)

CARSON, C. S. [2001]: *Toward a Framework for Assessing Data Quality.* IMF Working Paper, 01/25. http://www.imf.org/external/pubs/ft/wp/2001/wp0125.pdf (Elérés dátuma: 2010. június 23.)

COLLEDGE, M. – MARCH, M. [1991]: *Quality Management: Development of a Framework for a Statistical Agency.* Statistics Canada. Ottawa. http://www.amstat.org/sections/srms/Proceedings/papers/1991_125.pdf

COMMISSION OF THE EUROPEAN COMMUNITIES [2008a]: *Commission Staff Working Paper.* Brussels. 08/10/2008 SEC (2008) 2635 final. http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/documents/egreffe_adopted.pdf (Elérés dátuma: 2010. június 23.)

Czech Statistical Office [2008]: *TQM Activities at the CZSO in 2007*. Praha. http://www.czso.cz/eng/redakce.nsf/i/tqm_activities_of_the_czech_statistical_office_in_2007/$File/tqm2007eng.pdf (Elérés dátuma: 2010. június 23.)

EC (EUROPEAN COMMISSION) – SCB (STATISTICS SWEDEN) – EUROSTAT [1999]: *Quality Work and Quality Assurance within Statistics*. Stockholm. http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/documents/DGINS%20QUALITY%20Q98EN_0.pdf (Elérés dátuma: 2010. június 23.)

EC (EUROPEAN COMMISSION) [2002]: *Quality in the European Statistical System – The Way Forward.* Luxembourg. http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/documents/ESS_QUALITY_RECOMMENDATIONS_2002_EN_0_1.pdf (Elérés dátuma: 2010. június 23.)

EC (EUROPEAN COMMISSION) [2008]: *2008 Report from the Commission to the European Parliament and the Council on the Implementation of the Code of Practice.* Brussels. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2008:0621:FIN:EN:PDF (Elérés dátuma: 2010. június 23.)

EC (EUROPEAN COMMISSION) [2009]: *ESS Handbook for Quality Reports*. Metholodogies and working papers. Luxembourg. http://epp.eurostat.ec.europa.eu/portal/page/portal/ver-1/quality/documents/EHQR_FINAL.pdf (Elérés dátuma: 2010. június 23.)

EFQM [2009]: *EFQM Transition Guide How to upgrade to the EFQM Excellence Model 2010*. Brussels. www.efqm.org/en/PdfResources/Transition_Guide.pdf (Elérés dátuma: 2010. június 23.)

EURÓPAI KÖZÖSSÉGEK BIZOTTSÁGA [2005]: *A Bizottság közleménye az Európai Parlamentnek és a Tanácsnak a nemzeti és közösségi statisztikai hivatalok függetlenségéről, integritásáról és elszámoltathatóságáról. A Bizottság ajánlása a nemzeti és közösségi statisztikai hivatalok függetlenségéről, integritásáról és elszámoltathatóságáról, Brüsszel, 2005. 5. 25. COM (2005) 217 végleges* http://circa.europa.eu/Public/irc/dsis/coded/library?l=/2005_final_hupdf/_HU_1.0_&a=d (Elérés dátuma: 2010. június 23.)

EURÓPAI UNIÓ HIVATALOS LAPJA [2008]: *Az Európai Parlament és a Tanács 235/2008/EK határozata (2008. március 11.) az európai statisztikairányítási tanácsadó testület létrehozásáról (EGT-vonatkozású szöveg)* 2008.3.15. L73/17 http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2008:073:0017:0019:HU:PDF (Elérés dátuma: 2010. június 23.)

EURÓPAI UNIÓ HIVATALOS LAPJA [2009]: *A Európai Parlament és a Tanács 223/2009/EK Rendelete, (2009. március 11.) az európai statisztikákról és a titoktartási kötelezettség hatálya alá tartozó statisztikai adatoknak az Európai Közösségek Statisztikai Hivatala részére történő továbbításáról szóló 1101/2008/EK, Euratom európai parlamenti és tanácsi rendelet, a közösségi statisztikákról szóló 322/97/EK tanácsi rendelet és az Európai Közösségek statisztikai programbizottságának létrehozásáról szóló 89/382/ EGK, Euratom tanácsi határozat hatályon kívül helyezéséről. (EGT- és Svájc-vonatkozású szöveg)* 2009.3.31. L87/164 http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2009:087:0164:0173:HU:PDF (Elérés dátuma: 2010. június 23.)

EUROSTAT [2004]: *2004 LEG Implementation Status Report. The Second Interim Report of the LEG on Quality Implementation Group.* Meeting of Working Group 'Quality in Statistics', XVII. CIRCA Doc. ESTAT/02/Quality/2005/13.b/2004

EUROSTAT [2005]: *European Statistics Code of Practice for the National and Community Statistical Authorities.* http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/documents/VERSIONE_INGLESE_WEB%20new%20links.pdf (Elérés dátuma: 2010. június 23.)

EUROSTAT [2008]: *Summary of Good Practices Identified During the European Code of Practice Peer Reviews Carried Out During 2006–2008.* Version 1.0 of 10 June 2008. http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/documents/summary_good_practises.pdf (Elérés dátuma: 2010. június 23.)

EUROSTAT [2009]: *ESGAB's First Annual Report to the European Parliament and the Council on the Implementation of the European Statistics Code of Practice by Eurostat and the European Statistical System as a Whole.* http://epp.eurostat.ec.europa.eu/portal/page/portal/esgab/documents/ESGAB-2009-Annual-Report-EN.pdf (Elérés dátuma: 2010. június 23.)

FENWICK, D. – TIPPEN, G. [2003]: Quality Management Using ISO 9000 for Price Indices in the UK. *Journal of Official Statistics*. 19. évf. 4. sz. 365–382. old.

GILBERT, N. [2010]: *ABS Data Quality Framework: Linking Quality Assessment to Development of Performance Indicators.* http://q2010.stat.fi/media//presentations/session-4/gilbert_abs-data-quality-framework-linking-quality-assessment-to-development-of-performance-indicators_paper.doc (Elérés dátuma: 2010. június 23.)

HACKL, P. – HOLZER, W. [2008]: *Quality Management: A Pragmatic Approach. Experiences and Results*. Statistics Austria. Vienna. http://panda.hyperlink.cz/cestapdf/pdf08c2/hackl.pdf (Elérés dátuma: 2010. június 23.)

IMF (INTERNATIONAL MONETARY FOUND) [2003]: *Data Quality Assessment Framework – Generic Framework. (July)* http://dsbb.imf.org/images/pdfs/dqrs_Genframework.pdf (Elérés dátuma: 2010. június 23.)

JAPEC, L. ET AL. [2010]: *Striving for Business Excellence: Implementing the EFQM Excellence Model at Statistics Sweden.* http://q2010.stat.fi/media//presentations/session-3/japec-et-al_q2010-efqm_100501_paper.docx (Elérés dátuma: 2010. június 23.)

KÖVESI J. – TOPÁR J. (szerk.) [2009]: *A minőségmenedzsment alapjai*. Budapesti Műszaki és Gazdaságtudományi Egyetem. Typotex Elektronikus Kiadó Kft. Budapest.

KSH (KÖZPONTI STATISZTIKAI HIVATAL) [2004]: *ENSZ-alapelvek a hivatalos statisztikára.* http://portal.ksh.hu/pls/portal/docs/PAGE/KSHPORTAL/BEMUTATKOZAS/NEMZETKOZI_AJANLASOK/KSHSTRAT_2M.PDF (Elérés dátuma: 2010. június 23.)

KSH (KÖZPONTI STATISZTIKAI HIVATAL) [2005a]: *A Szakértői Csoport (LEG) minőséggel kapcsolatos ajánlásai.* http://portal.ksh.hu/pls/portal/docs/PAGE/KSHPORTAL/BEMUTATKOZAS/NEMZETKOZI_AJANLASOK/LEGD2.PDF (Elérés dátuma: 2010. június 23.)

KSH (KÖZPONTI STATISZTIKAI HIVATAL) [2005b]: *Az Európai Statisztikai Rendszer minőségügyi deklarációja.* http://portal.ksh.hu/pls/portal/docs/PAGE/KSHPORTAL/BEMUTATKOZAS/NEMZETKOZI_AJANLASOK/ESR.PDF (Elérés dátuma: 2010. június 23.)

KSH (KÖZPONTI STATISZTIKAI HIVATAL) [2005c]: *Az európai statisztika gyakorlati kódexe*. Budapest. http://portal.ksh.hu/pls/ksh/docs/bemutatkozas/hun/eu_stat_kodex.pdf (Elérés dátuma: 2010. június 23.)

KSH (KÖZPONTI STATISZTIKAI HIVATAL) [2005d]*: Az európai statisztika gyakorlati kódexe. Önértékelő kérdőív*. http://portal.ksh.hu/pls/portal/docs/PAGE/KSHPORTAL/BEMUTATKOZAS/NEMZETKOZI_AJANLASOK/KAPCSOLODO_DOKUMENTUMOK/ONERTEKELO_KERDOIV_KSH.PDF (Elérés dátuma: 2010. június 23.)

KSH (KÖZPONTI STATISZTIKAI HIVATAL) [2005e]: *KSH-stratégia, 2005–2008*. http://portal.ksh.hu/pls/ksh/docs/bemutatkozas/hun/STRATEGIA.pdf (Elérés dátuma: 2010. június 23.)

KSH (KÖZPONTI STATISZTIKAI HIVATAL) [2009a]: *A KSH minőségpolitikája.* http://portal.ksh.hu/pls/ksh/docs/bemutatkozas/hun/minpol_web_hu.pdf (Elérés dátuma: 2010. június 23.)

KSH (KÖZPONTI STATISZTIKAI HIVATAL) [2009b]: *KSH-stratégia, 2009–2012*. http://portal.ksh.hu/pls/ksh/docs/bemutatkozas/hun/strategia_2009_2012.pdf (Elérés dátuma: 2010. június 23.)

KSH (KÖZPONTI STATISZTIKAI HIVATAL) [2010]: *Minőségi irányelvek a Központi Statisztikai Hivatal statisztikai munkafolyamatainak egyes szakaszaira*. http://portal.ksh.hu/pls/ksh/docs/bemutatkozas/hun/minosegi_iranyelvek.pdf (Elérés dátuma: 2010. június 23.)

LISAI, D. [2007]: *Choosing and Implementing a Quality Management System at Statistics Sweden. Masters Thesis in Mathematical Statistics*. Linköping University. Stockholm. http://liu.diva-portal.org/smash/record.jsf?pid=diva2:17405 (Elérés dátuma: 2010. június 23.)

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.

MEH (MINISZTERELNÖKI HIVATAL) [2006]: *CAF 3.0. Európai Közös Értékelési Keretrendszer (CAF-modell) nemzeti változata.* https://caf.meh.hu/ (Elérés dátuma: 2010. június 23.)

VAN NEDERPELT, P. [2009]: *The Creation and Application of a New Quality Management Model. Statistics Netherlands.* The Hague. http://www.cbs.nl/NR/rdonlyres/D61D5537-2634-4F6F-A41C-748520564162/0/200940x10pub.pdf (Elérés dátuma: 2010. június 23.)

PRUAL, R. [2008]: *Implementation of EFQM Excellence Model in Statistics Estonia and Lessons Learned.* The European Conference on Quality in Official Statistics. July 8–11. Rome. http://q2008.istat.it/sessions/presentation/27/S2704Prual.ppt (Elérés dátuma: 2010. június 23.)

SÆBØ, H. V. – NÆS, P. [2010]: *Linking Management, Planning and Quality in Statistics Norway.* http://q2010.stat.fi/media/q2010_abstracts.pdf (Elérés dátuma: 2010. június 23.)

SORS (STATISTICAL OFFICE OF THE REPUBLIC OF SLOVENIA) [2008]: *Medium-Term Programme of Statistical Surveys, 2008–2012*. Ljubjlana. http://www.paris21.org/documents/3279.pdf (Elérés dátuma: 2010. június 23.)

STATISTICS CANADA [2002]: *Statistics Canada's Quality Assurance Framework*. Ottawa. http://www.statcan.gc.ca/pub/12-586-x/12-586-x2002001-eng.pdf (Elérés dátuma: 2010. június 23.)

STATISTICS CANADA [2009]: *Statistics Canada Quality Guidelines. Fifth Edition*. Ottawa. http://www.statcan.gc.ca/pub/12-539-x/12-539-x2009001-eng.pdf (Elérés dátuma: 2010. június 23.)

STATISTICS FINLAND [2007]: *Quality Guidelines for Official Statistics*. 2nd Revised Edition. Helsinki. http://www.stat.fi/meta/qg_2ed_en.pdf (Elérés dátuma: 2010. június 23.)

STATISTIKA CENTRALBYRAN [2001]: *The International Conference on Quality in Official Statistics*. Május 14–15. Stockholm. http://www.q2001.scb.se/abstracts.pdf (Elérés dátuma: 2010. június 23.)

SZÉP, K. – MAG, K. – VIGH, J. [2006]: Quality of Statistics in the Strategy of HCSO. Statistical Days. November 6–8. Radenci, Slovenia. http://www.stat.si/radenci/program_2006/C_Szep.doc (Elérés dátuma: 2010. június 23.)

SZÉP K. – VIGH J. [2004]: A minőség a hivatalos statisztikában. *Statisztikai Szemle*. 82. évf. 8. sz. 773–798. old. http://www.ksh.hu/statszemle_archive/2004/2004_08/2004_08_773.pdf

SZILÁGYI GY. [2005]: Gyakorlati kódex az európai statisztikában. *Statisztikai Szemle*. 83. évf. 10–11. sz. 911–918. old. http://www.ksh.hu/statszemle_archive/2005/2005_10-11/2005_10-11_911.pdf

TENNER, A. R. – DETORO, I. J. [1996]: *TQM – Teljes körű minőségmenedzsment.* Műszaki Könyvkiadó. Budapest.

UN STATISTICAL COMMISSION [2004]: *Implementation of the Fundamental Principles of Official Statistics*. http://unstats.un.org/unsd/statcom/doc04/2004-21e.pdf (Elérés dátuma: 2010. június 23.)

UN STATISTICAL COMMISSION [2009] *National Quality Assurance Frameworks*. Statistics Canada. http://unstats.un.org/unsd/dnss/qaf/qafreport.htm (Elérés dátuma: 2010. június 23.)

UN STATISTICAL COMMISSION [2010]: *Statistics Canada: National Quality Assurance Frameworks*. http://unstats.un.org/unsd/statcom/doc10/2010-2-NQAF-E.pdf (Elérés dátuma: 2010. június 23.)

VOINEAGU, V. ET AL. [2009]: „The European Code of Practice" – Comparisons Regarding the Implementation within the National Institute of Statistics from Romania and the Hungarian Central Statistical Office. *Revista Română de Statistică.* 10. sz. 37–50. old. http://www.revistadestatistica.ro/Revista/2009/sumar%2010_2009.pdf (Elérés dátuma: 2010. június 23.)

Summary

As a follow-up of the paper on quality concepts in official statistics (*Szép–Vigh* [2004]), the authors give an overview on the recent developments, on the implementation of quality frameworks in statistical offices, with special regard to the Hungarian Central Statistical Office (HCSO). In the first part of the study the Code of Practice and its implementation process are presented as a result of the preceding quality management related activities of different national statistical institutes, the Leadership Group on Quality, Eurostat and other international organisations. This chapter ends with an overview table on the quality frameworks implemented in European statistical offices. The

second part focuses on the quality framework of HCSO. The organization has published its quality policy on its website (http://portal.ksh.hu/pls/ksh/docs/bemutatkozas/eng/minpol_web_eng.pdf). The HCSO Quality Guidelines structure follows the statistical business process model, and contains the description, principles, quality guidelines, references to the related legal acts and methodological documents for each of the 19 steps of statistical workflow. A quality report scheme and a set of standard product quality indicators serve the measurement of output quality, a handbook on process variables offers possibilities to process quality monitoring. The adaptation of the DESAP-checklist to HCSO requirements has been developed to support survey self-assessment. Some key elements in the future plan are the adaptation of audits, the widening of the scope of the framework to the whole official statistical system.

A hibrid adatfelvétel módszertani kihívásai*

Pintér Róbert,
az Ipsos Zrt. online stratégiai igazgatója
E-mail: robert.pinter@ipsos.com

Kátay Bálint,
az Ipsos Zrt. operatív igazgatója
E-mail: balint.katay@ipsos.com

A tanulmány az elmúlt néhány évben egyre népszerűbb hibrid adatfelvétel módszertani kérdéseit vizsgálja a magyar Ipsos kutatási tapasztalatai alapján. A hibrid kutatás az online és a személyes adatfelvételi módszerek (elsősorban az ún. CAPI) ötvözése. A szerzők megvizsgálják, hogy mire alkalmas a hibrid kutatás, mi az oka ezen kutatások térhódításának (például a reprezentativitás biztosítása, a válaszadók jobb elérése, a válaszadási arány növelése, a költséghatékonyság). Ezt követően azt vizsgálják, hogy milyen területeken alkalmazható (például média, véleménykutatás, piackutatás, reklámkutatás stb.). Azt is bemutatják, hogy milyen szakmai, szervezési és módszertani nehézségek adódhatnak a hibrid kutatások készítése, többek között a munkaszervezés, a költséghatékonyság, a mintakészítés, a kérdőívkészítés, az adatfelvétel, az adatfeldolgozás során. A tanulmányhoz felhasználják az Ipsosban az elmúlt években végzett hibrid kutatások tapasztalatait.

TÁRGYSZÓ:
Statisztikai adatgyűjtés.
Interjú.
Internet.

* A szerzők ezúton mondanak köszönetet *Gábos Zsuzsának, Melles Katalinnak* és *Sterk Péternek* a tanulmány egy korábbi verziójához fűzött értékes megjegyzéseikért.

Mindenekelőtt tisztázzuk a hibrid kutatás[1] fogalmát, melyen összefoglalóan azokat a felméréseket értik, ahol nem csupán egyetlen adatfelvételi technikával veszik fel az adatokat.

1. A hibrid kutatás fogalma

A definícióból következik, hogy a hibrid kutatás nem napjaink „találmánya", hanem szerves része a (piac)kutatói kultúrának. Összetett projektekben évtizedek óta alkalmazott módszer, hogy a kutatások megbízhatóságának és az adatok érvényességének érdekében ötvözik az adatfelvételi technikákat: elsősorban a postai kérdőívek, a papíros kérdezés (PAPI), a telefonos felmérés (CATI), a számítógéppel segített személyes kérdezés (CAPI) alkalmazásával. Ahogy *Biemer* és *Lyberg* fogalmaz „az adatgyűjtő rendszerek általában nem egyetlen módot tartalmaznak, mivel a hibrid felmérések jelentik a normát napjainkban"[2] (*de Leeuw* [2008]).

Amiért a figyelem mégis megnőtt a hibrid kutatások iránt az ezredfordulót követően, az az online kutatások egyre szélesebb körű elterjedése és együttes alkalmazása a korábban már bevett személyes kérdezési technikákkal. Ahogy azt látni fogjuk, ennek több oka lehet, például a reprezentativitás javításának igénye, a költségek csökkentése, a minták méretének általában vett növelése vagy egyes részcsoportba tartozó tagok számának célzott megnövelése (ún. boost) stb. Az elsődleges ok azonban, hogy online módon még a legfejlettebb országokban sem érhető el a teljes lakosság, kézenfekvő tehát, hogy azoknál a kutatásoknál, ahol az online opció felmerül, de az alapsokaságban jelentős, online úton nem elérhető rétegek vannak (és a későbbiekben részletezett okokból mégis ragaszkodnak az online kutatási megoldások használatához), ötvözni kell a hagyományos és az új, online kutatási eszközöket.

Az online és offline adatfelvételi technikák ötvözésének többfajta módja lehetséges. Attól függően, hogy a kétfajta technikát hogyan fűzik egybe a vizsgálat során, többféle

[1] Angolul leginkább *mixed-mode*, *multi mode* vagy *switch mode* (*Bisschop* [2004]; *Macer* [2004], [2005]) vagyis „kevert", „többféle" vagy „váltott" (adatfelvételi) módú kutatások, de a legelterjedtebb, mára bevett összefoglaló elnevezés a mixed-mode (*de Leuuw* [2008]). Mivel azonban idehaza a *hibrid kutatás*, mint terminus technicus terjedt el és vált napjainkra az elfogadott összefoglaló megnevezéssé, ezért a tanulmányban mi is ezt a fogalmat alkalmazzuk a mixed-mode magyar megfelelőjeként.

[2] Saját fordítás. Bár *de Leuuw* megjegyzi, hogy bizonyos országokban (például Japánban) nem alkalmazzák a hibrid megközelítést, illetve nem minden kutatás hibrid kutatás, tehát normaként feltüntetni talán túlzás a hibrid módszertant, de tény, hogy egyre inkább terjedőben van.

adatfelvételi móddal készülő (multimodális) vagy szakaszos (szeriális) hibrid kutatásról beszélhetünk (*Bisschop* [2004]). Abban az esetben, ha a minta hibrid, többféle adatfelvételi móddal készülő hibrid kutatásról beszélünk, ilyen esetben egy válaszadó csak egyetlen módon, vagy online, vagy offline válaszolhat a kérdőívre, vagyis a minta egyik részét kizárólag online, másik részét kizárólag offline módon érjük el. A multimodális hibrid kutatásokban elsősorban a célcsoportok jobb elérése, reprezentálása a cél.

A szakaszos (szeriális*) vagy* párhuzamos (parallel) *hibrid kutatásnál* viszont nem a minta felől szerveződik a hibridizálás, hanem a kutatás (kérdőív) felől. A szakaszos *hibrid* kutatásban a kutatás több szakaszból épül fel, bizonyos kérdőívekre online, míg másokra offline módon kell válaszolni ugyanannak a személynek (tehát a kutató előre meghatározza, hogy melyik szakaszban milyen módon kell majd válaszolni a résztvevőknek). Például elképzelhető, hogy egy online szűrés után személyes kapcsolatot igénylő kutatásokban vesz részt a válaszadó, majd a benyomásait újra egy online kérdőív kérdéseire válaszolva mondja el (ilyen lehet például egy termék tesztelése). A szeriális hibrid kutatások egymásra épülő szakaszaiban sokszor toborzási, szűrési okokból váltogatják az adatfelvételi technikákat.

A párhuzamos (parallel) *hibrid kutatásban*[3] viszont a válaszadónak lehetősége van arra, hogy a kérdőív megválaszolása (kitöltése) közben változtasson az adatfelvétel módján, például telefonon vagy személyesen kezdjen megválaszolni egy kérdőívet, majd online fejezze be azt. Ilyen esetekben a hibrid mód inkább a válaszadási hajlandóság növelését, a válaszadó kényelmét növeli, hogy az ne hagyja abba a kérdőív megválaszolását vagy érzékeny témát érintő, intim kérdésekre is válaszoljon (elsősorban, amikor személyesen induló kutatást online módon fejeznek be).

Az Ipsosban az elmúlt években folyó hibrid kutatásainkban egyaránt alkalmaztuk a többféle adatfelvételi móddal készülő, valamint a szakaszos hibrid kutatási megoldásokat, illetve ezeknél ritkábban a párhuzamos metódust. Megítélésünk szerint a legnagyobb szakmai kihívást és a legígéretesebb lehetőséget a kutatás számára az online paneles adatfelvétel és az offline technikák (leginkább a CAPI, kisebb részben CATI[4]) többféle adatfelvételi módú ötvözése jelentheti.[5] Ennek megfelelően ebben a

[3] Ezt angolul sokszor switch mode-nak nevezik, mivel a válaszadó válthat a kérdőívek offline és online verziói között. Erről bővebben lásd *Macer* [2005].

[4] A Confirmit és a Meaning Ltd. 2004 óta folyó nem reprezentatív globális piackutatói felméréséből úgy tűnik, hogy a cégek leggyakrabban az online és a telefonos (CATI) adatfelvételi módokat alkalmazzák együtt következésképpen ritkábban fordul elő, hogy az online módszer mellett ne a telefonos adatfelvételi mód kerüljön elő (*Molloy–Macer* [2009]). Ennyiben tehát az Ipsos kutatóiként a mi tapasztalatunk némileg eltér a nemzetközi trendektől, aminek az oka elsősorban abban keresendő, hogy idehaza továbbra is a személyes kérdezés a leggyakoribb adatfelvételi technika.

[5] A szakaszos hibrid kutatásokban tulajdonképpen tisztán offline vagy tisztán online kutatási szakaszokról van szó, amelyek elkülönülnek egymástól, de egymásra épülnek. Így ezek, bár remek kutatási eszközt jelentenek, szakmailag nem jelentenek érdemben plusz kihívásokat az online vagy az offline kutatásokhoz képest, ezért ebben a cikkben külön nem foglalkozunk velük.

tanulmányban az ezzel kapcsolatos tapasztalataink fogalmazódnak meg, de természetesen a szakirodalom egyéb, témára vonatkozó releváns eredményeit is ismertetjük röviden. A következőkben tehát hibrid kutatás alatt elsősorban az online paneles és CAPI (CATI) kérdezéssel folyó felmérések multimodális ötvözését fogjuk érteni. Meglátásaink egy jelentős része azonban általánosítható az online és személyes adatfelvétel ötvözésével folyó bármely hibrid kutatásra.

2. Miért van szükség hibrid kutatásokra?

A válaszadási hajlandóság növelésének igénye az egyik legfontosabb indok, amely életre hívta a hibrid kutatásokat. Ennek megfelelően az elmúlt tíz évben a témához kapcsolódó szakirodalom érdeklődésének homlokterében a hibrid kutatások és a válaszadási hajlandóság viszonyának kutatása áll, amellyel majd mi is foglalkozunk röviden a következő részben.[6]

A válaszadási hajlandóság fokozatos csökkenése világszerte ismert probléma a kutatásban, ami természetesen nem kerülte el Magyarországot sem. A személyes megkérdezés napjainkban egyre nagyobb nehézségekbe ütközik. Az empirikus társadalomkutatások, piackutatások legelterjedtebb adatfelvételi eszközét, a survey-módszertant az elmúlt évtizedekben egyre szélesebb körben alkalmazták, és ez egyes lakossági és foglalkozási csoportok túlkérdezéséhez vezetett. A helyzetet súlyosbította a direkt marketing (DM) és direct sale (DS) alkalmazások robbanásszerű növekedése, amelyekben közvetlenül keresik meg a lakosságot marketing-, reklámüzenetekkel és vásárlási ajánlatokkal. A túlzott, illetve kéretlen megkeresések következtében csökkent a kutatások kapcsán a válaszadói hajlandóság, ami egyre nagyobb költségek mellett, egyre kisebb hatékonyságot eredményezett a személyes adatfelvételi módszerek esetében. Az Ipsos saját adatai szerint, míg 1995-ben jellemzően a kiinduló főcím mintába tartozók 50-60 százalékával készült sikeres személyes interjú, addig másfél évtizeddel később, 2009-ben ugyanez az arány már olykor a 20 százalékot sem érte el.[7] A válaszadási hajlandóság lassú erodálódása a kutatási módszertanok folyamatos megújítását teszik

[6] A témára vonatkozó korai dilemmákkal kapcsolatban lásd például *Dillman* [2001]; az utóbbi idők összetettebb problémafelvetései kapcsán pedig például *Heerwegh* [2009].

[7] Egy 1990-es évek eleje óta havonta lekérdezett személyes kutatásunkban, amely folyamatosan azonos témában, országos reprezentatív mintán folyt 1995-ben, három hónapot véletlenszerűen kiválasztva 57,83 százalékos volt a kiinduló főcím mintán a sikeres interjúk aránya, 2009-ben ugyanebben a három hónapban már csak átlagosan 17,8 százalékos arányt találtunk. Fontos azonban hangsúlyozni, hogy ez alacsonyabb, mint a tényleges válaszadási arány, mivel a főcím mintában benne vannak azok is, akik a kérdezés idején nem voltak otthon, vagy akiknek rossz volt a címük stb. Az arányra negatív hatással van az adatfelvételi idő lerövidülése is, ami szintén gyakori jelenség az elmúlt évtizedekben.

szükségessé, hogy a válaszadókat be lehessen vonni a kutatásokba. Részben ez a folyamat állt a telefonos módszer 1980-90-es évekbeli elterjedése vagy az online kutatások megjelenése és megerősödése mögött is az ezredfordulót követően, és ez hajtja legújabban az internetes közösségek vagy az ún. okos telefonok kutatási célú felhasználását, és nincs ez másképpen a hibrid kutatások esetében sem.

A Confirmit és a Meaning Ltd. 2008-ban piackutatók körében készült nem reprezentatív globális felmérése szerint is a válaszadási hajlandóság javítása áll a hibrid kutatások elterjedése mögött, de további meghatározó okok a minta méretének vagy a reprezentativitásnak a javítása, a válaszadó-barátság javítása (nyilván összefüggésben a válaszadási hajlandósággal), a költségek remélt csökkentése, illetve a kifejezett megrendelői igények.

*Miért csinálnak hibrid kutatásokat?**

Indok	Összes válasz	Elsődleges ok
	százalék	
A válaszadási hajlandóság javítása	68	34
A minta méretének vagy reprezentativitásának javítása	60	28
Válaszadó-barátság növelése	38	6
Adatfelvételi költségek csökkentése	35	10
A megrendelő hibrid módszertant kért	33	13
Az adatfelvételi erőforrások kapacitásának növelése	17	2
Az adatfelvételi idő csökkentése	15	0
Jól mutat az ajánlatokban	4	1
Egyéb	9	5

* Egy válaszadó több választ is megjelölhetett.
Forrás: *Macer–Molloy* [2009] 34. old.

Az Ipsosban a mi elsődleges motivációink is hasonlók voltak az előbb említettekhez. A reprezentativitás javítása, a válaszadók jobb elérése, a válaszadási arány növelése, a költséghatékonyság javítása és a módszer innovatív jellege miatt döntöttünk úgy, hogy 2008-tól elkezdünk foglalkozni a hibrid kutatási módszer fejlesztésével. Ennek első lépéseként ugyanebben az évben készült egy belső használatra szánt tanulmány (*Lakatos* [2008]), ami a hibrid módszer piackutatásban való alkalmazhatóságának bizonyos elméleti és módszertani kérdéseit vizsgálta. Ezt követően jött létre egy munkacsoport, hogy felmérje a módszer hatékony bevezetésével kapcsolatos gyakorlati teendőket, hozzáigazítva azt az Ipsos munkaszervezéséhez és alkalmazott megoldásaihoz, illetve javaslatot tegyen az adatfelvétellel kapcsolatos egyéb szükséges újításokra is. Ezen munka közben térképeztük fel, hogy milyen kihívásokat rejthet a hibrid módszertan üzemszerű integrálása egy piackutató cég napi működésébe.

3. A hibrid kutatásokkal kapcsolatos néhány kihívás

Ahhoz, hogy egy hibrid felmérés sikeres legyen, gondosan meg kell tervezni a kutatást, ellenkező esetben ugyanis jelentősen megnövekedhetnek a ráfordított költségek, az adatfelvételre fordított idő, a projekt kivitelezéséhez szükséges munkamennyiség, és ami még nagyobb gondot jelenthet, a kapott eredmények megbízhatóságára és érvényességére is hatással lehet. Számos olyan szempont van, amellyel érdemes előre számolni, mi ezek közül csak a legfontosabbakat emeljük ki és mutatunk rá igen röviden néhány figyelemre méltó tényezőre. A felsorolás a kutatás folyamatát veszi alapul a módszer kiválasztásától az adatok feldolgozásáig.

3.1. A módszertan kiválasztása és a megrendelő tájékoztatása

Hiába innovatív és vonzó az ajánlatokban a hibrid kutatás, mint minden újdonság esetében, itt is felmerülnek félelmek, kétségek, dilemmák. Fontos, hogy az adott projekt esetében a módszer előnyei mellett tisztában legyünk annak várható hátrányaival is, és amennyiben a kutatásnak külső megrendelője van, őt is pontosan tájékoztassuk a módszerről, alkalmazásának indokairól.

3.2. Munkaszervezés és költséghatékonyság

A legtöbb adatfelvétellel foglalkozó apparátus a tiszta adatfelvételi módokhoz szokott, amikor a kutatás adott szakasza egyetlen adatfelvételi módszerrel zajlik, tehát egyetlen kérdőívvel folyik a munka, és a kérdezést követően egyetlen adatbázis készül el. A munkaszervezet és ennek megfelelően a munka megszervezése is ezt a logikát követi. Hibrid kutatásokban azonban – különösen többféle adatfelvételi móddal készülő (multimodális) vagy a párhuzamos hibrid kutatásoknál –, ha a minta egyik része hagyományos, másik része online módon válaszol, akkor ugyanaz a kérdőív két változatban[8] is elkészül. Ennek megfelelően a két kérdőívvel két almintán zajlik az adatok felvétele, az eredményeket pedig össze kell illeszteni az elemzés előtt. Mindez megnöveli a projekthez szükséges munkamennyiséget.

Ezenkívül természetesen felmerülnek adminisztratív kérdések, például a hibrid projektek nyilvántartása, itt is megkettőződnek bizonyos dolgok. A hibrid kutatások esetében óhatatlanul megnövekednek a projekt felállításához kötődő költségek, valamint a módszer bevezetésének idején annak ismeretlensége miatt extra munka-

[8] Illusztrációként ezúttal tudatosan csak erről az egyszerűbb megoldásról beszélünk.

órákkal kell számolni, ami mindaddig magas marad, amíg az üzemszerű, bevett működésmód ki nem alakul. Emiatt a hibrid kutatás csak akkor lehet költséghatékony, ha más munkafolyamatokat leegyszerűsít, lerövidít vagy feleslegessé tesz.

3.3. Mintakészítés, mintakezelés és a reprezentativitás biztosítása

A hibrid kutatások megtervezésének az egyik legfontosabb szakasza és az egyik legnagyobb kihívása a mintakészítés, a minta kezelése a kutatás folyamán, és amennyiben a kutatás jellege megköveteli, a reprezentativitás biztosítása. A kihívás alapját az adja, hogy online nem lehetséges véletlen mintavétellel dolgozni[9], rendszerint paneles kvótás mintavétel folyik, ami a peremkvóták követése kapcsán az elfogadott szakmai konszenzus alapján tekinthető az internetezőkre nézve reprezentatívnak, miközben az offline kérdezés esetében a reprezentativitás a véletlen mintavétel következtében áll elő, ebben az esetben viszont az alacsony válaszadási hajlandóság miatt a kiinduló minta pótlására van szükség.

A hibrid minta kezelése során megoldás lehet, ha az offline mintából kizárjuk az internetezőket, ha úgy tervezzük, hogy azokat a kutatás online része reprezentálja majd. Szintén megoldás lehet, ha az offline rész önmagában is reprezentatív a teljes célcsoportra, az internetes részt pedig a minta méretének megnövelésére használjuk (ún. boost-minta), ami a részletekbe menő elemzést teszi lehetővé. Ez esetben a két mintafél együttes elemzése gondos módszertani mérlegelést igényel.

Szintén gyakori megoldás, hogy a mintát nem az internetezés, hanem egyéb változó mentén vágjuk ketté, ami lényeges hatással van az internetezésre (idehaza ilyen lehet például az életkor, a településtípus, vagy az iskolai végzettség). Ilyen esetekben például a fiatalokat, vagy a városiakat, vagy a magasabb végzettségűeket az internetes adatfelvétel tartalmazza, az időseket, falusiakat vagy alacsonyabb iskolai végzettségűeket pedig a hagyományos adatfelvétel hivatott reprezentálni.

Az a legszerencsésebb azonban, ha egy viszonylag „kemény” változó mentén van szétvágva a célpopuláció (például az említett internethasználat mentén). De gyakran kényszerülnek a kutatók arra, hogy az internetezés gyakoriságánál valamivel „lágyabb” változók mentén osszák fel a teljes mintát. A lényeg minden esetben az, hogy egyértelmű szempont szerint kell kettévágni a mintát.

Amennyiben online panelből származik a kutatás mintájának online része, akkor annak tulajdonságai is nagymértékben befolyásolják a minta összeállítását. Ilyenkor

[9] *Switch mode* esetén természetesen nem panelen folyik az online adatfelvétel, hanem az offline mintába kerülők számára ajánlják fel, hogy online válaszoljanak, ami növeli a válaszadási hajlandóságot, még akkor is, ha figyelembe vesszük, hogy sokan a kutatásban való részvétel nyílt visszautasítása helyett mondják azt, hogy inkább az online módszert választanák (*Macer* [2005]).

mérlegelni kell, hogy az online panel segítségével kiket lehet jól, érvényesen reprezentálni: általában az internethasználókat, vagy a rendszeres internetezőket, esetleg csak a naponta netezőket. Ez azért lényeges szempont, mert korlátozhatja a módszer alkalmazhatóságát, főként az alminták szétbontásában.

Minden esetben körültekintően meg kell azonban tervezni a mintát, hogy minden csoport megfelelően legyen reprezentálva (nincsen-e olyan csoport, ami kimarad, illetve alul- vagy felülreprezentálódik), különben nem jutunk megfelelő minőségű adatokhoz.

3.4. Kérdőívkészítés és -tesztelés

Csak akkor fogjunk hibrid kutatásba, ha megfelelő technikai platformmal rendelkezünk, amely kezelni tudja a különböző adatfelvételi módú kérdőíveket és adatfelvételt, sőt, akár a váltást is a két módszer között, különben kétszer kell ugyanazt a kérdőívet programozni és egyéb nehézségek is felmerülhetnek.

Ahhoz, hogy a kétféle módszerrel ne a kérdőívek különbözősége miatt kapjunk eltérő válaszokat, fontos odafigyelni a kérdések megjelenítésére, az ugratások egységes kezelésére és általában a kétféle kérdőív egyezésére. A hibrid módszer miatt egészen biztosan megnövekszik a tesztelés mennyisége és az ahhoz szükséges idő, miközben a javítások átvezetése összetettebb, és szintén időigényesebb. Ez különösen olyan esetekben okozhat gondot, amikor a kérdőív hosszas, több felet is bevonó egyeztetésen és folyamatos átalakuláson megy keresztül. Ilyenkor egyébként is érdemes először a kérdőívet véglegesíteni és azokat a változtatásokat is végigbeszélni, amik a kétféle módszer alkalmazásából fakadhatnak (például az önkitöltős kérdőívben az utasítások nem a kérdezőbiztosnak szólnak, emiatt a kérdések megfogalmazásán helyenként változtatni szükséges, előfordulhat tegezés, a válaszok lehetnek egyes szám első személyben stb.).

A kutatás megtervezésekor és a kérdőív elkészítésekor fontos azt a tényt is figyelembe venni, hogy az online adatfelvétel esetében a válaszadók 25-30 percnél hosszabb kérdőíveket már nem szívesen töltenek ki, jelentősen növekszik a félbehagyás valószínűsége, mivel nincsen kérdezőbiztos, aki a bevonódást fenntartaná és jelenlétével ösztönözné a válaszadót a részvételre.

Végül, a kutatás elindítása előtt szükséges meggyőződni arról, hogy a két kérdőív azonos-e, vagy legalábbis a közösen elemezni kívánt blokkok rendre megegyeznek-e.[10]

[10] A hibrid kutatások kérdőíveinek elkészítése nem egyszerű feladat, több dologra is oda kell figyelni, ezekkel kapcsolatban ad hasznos tanácsokat *Bäckström* és *Nilsson* [2002]. Fontosabb ajánlásaik a következőkre vonatkoznak: kérdőív-dizájn, a kitöltés előrehaladását mérő *status bar* alkalmazása, a kérdőív kérdéseinek egyértelmű sorszámozása, válaszadóbarát kérdőív, könnyen megérthető kérdések (ne legyen eltérő interpretálás az offline és az online kérdőív esetében), könnyen kitölthető kérdőívek (hogy a számítógépes ismeretek ne korlátozzák a kitöltést) stb.

3.5. A válaszadási helyzetekből fakadó eltérések

A bizonytalanságra okot adó alapvető kérdés a következő: ugyanaz a személy ugyanazt a választ adná, ha feltennék neki ugyanazt a kérdést a két különböző módszerrel? Saját méréseink is megerősítik azt a szakirodalomban közismert és sokszor vizsgált tényt, hogy bizonyos témák és kérdések eltérő válaszokat eredményezhetnek online és offline kutatások esetén. Megfigyelhető például, hogy az elégedettségi kérdésekre rendre magasabb értékeket adnak telefonos kutatásoknál, mint az online kutatásokban, és sokszor szélsőségesebbek is (*Macer* [2005]). Ugyanígy eltérés található a következő választáson való részvételi hajlandóság vagy a korábbi választásokon való részvételt firtató kérdésekre adott válaszok esetében is: telefonon többen jelzik, hogy elmennek szavazni, mint online (*Krosnick* [2009]).

A jelenséget alaposabban megvizsgálva azt találták, hogy az összefügg az érzékeny témákkal. Ha a válaszadóknak zavarba ejtő vagy kínos dolgokról kellene beszámolnia mások jelenlétében (értsd kérdezőbiztos), akkor gyakorta előfordul, hogy félrevezetően válaszolnak vagy eltitkolják az igazságot. *Tourangeau* és *Yan* szerint a torzítás mértéke több mindentől függ, például, hogy milyen a kérdőív felépítése, jelen van-e kérdezőbiztos, illetve mennyire hozná magát kínos helyzetbe a válaszadó, ha őszintén válaszolna (*Tourangeau–Yan* [2007]). A válaszadók akár akaratlanul is szeretnek pozitív benyomást tenni másokra, illetve megfelelni a ki nem mondott társadalmi elvárásoknak (például a választásokon való részvétel). Mindez torzíthatja a válaszokat.

A válaszadási helyzetből adódó eltérések sok témát érinthetnek (*Macer* [2005]), a teljesség igénye nélkül ilyenek például az egészségügy, a bűnözés, az étkezési szokások, az internethasználati szokások, az önkéntesség, a szelektív hulladékgyűjtés stb. Minden olyan témát befolyásolhat, ahol a válaszadó érték alapon fogalmazza meg a válaszait. Tourangeau és Yan felmérése szerint azonban a jelenség nem annyira általános és átfogó hatású, ahogy azt általában a kutatók feltételezik (*Tourangeau–Yan* [2007]), valójában a válaszadóknak csak egy része „kozmetikázza" a válaszait, azok akik úgy érzik, hogy az őszinteségük zavarba hozná őket.

Mindez persze nem jelenti azt, hogy nem lehet hibrid kutatásokat végezni, hanem inkább azt, hogy a kérdőív készítése közben folyamatosan meg kell kérdezni magunktól, hogy vajon az adott kérdés hallatán nem jönnek-e majd zavarba egyes válaszadók, képesek lesznek-e őszintén válaszolni. Ha úgy gondoljuk, hogy nem kapnánk őszinte válaszokat, akkor vagy egyetlen módszerrel kérdezzünk, vagy hibrid kutatás esetén mindenképpen biztosítsuk, hogy az adott kérdést a személyes adatfelvétel során is önkitöltős módon kérdezzük, hogy a kérdezőbiztos ne hallhassa-láthassa a válaszokat.

3.6. Az adatfelvétel összehangolása

A hibrid adatfelvételt igen gyakran választják, ha több országban folyik az adatfelvétel, ilyen esetben egy adott országban többnyire csak egyetlen adatfelvételi módot alkalmaznak, azonban az országok között már lehetnek eltérések, ilyenkor, az összehasonlíthatóság érdekében fontos az adatfelvétel összehangolása és az adatok körültekintő elemzése.

A legtöbb hibrid kutatás az adatfelvétel alapos felügyeletét igényli, főleg, ha az online és offline almintáknak együttesen kell biztosítaniuk a reprezentativitást. Folyamatosan érdemes monitorozni a kitöltött kérdőívszámot, a félbehagyást, online panel használata esetén a kvóták állását. Így hamarabb kiderülhet, ha az előzetes tervekhez és ütemezéshez képest eltérés tapasztalható; ez esetben a két adatfelvétel segíthet is egymásnak. Az adatfelvétel összehangolása azt jelenti, hogy a két apparátus – az online és az offline – folyamatosan összedolgozik és nem különálló részkutatásoknak tekintik a kutatást, ahol elegendő a saját részüket szakszerűen elvégezni.

Még nagyobb körültekintést igényel, ha párhuzamos hibrid módszerrel dolgoznak, vagyis a válaszadók megválaszthatják a válaszadás módját, ilyenkor érdemes előzetesen megbecsülni, hogy várhatóan mennyien fogják az egyik, illetve a másik módszert választani, hogyan lehet ezt nyomon követni, hogyan fog zajlani az emlékeztetés és mikortól tekintünk úgy a váltókra, hogy végül mégsem fognak válaszolni, vagyis pótolni kell őket a mintaméret megőrzése céljából. A váltási lehetőség ugyanis – bár növeli a válaszadási hajlandóságot – lehetővé teszi a bújtatott visszautasítást is, vagyis nem minden módszertant váltó fogja kitölteni a kérdőívünket. *Allison* és *O'Konis* egy 2002-es, pénzügyi szolgáltatásokról szóló online kutatásukban úgy találták, hogy a telefonon elért minta 88 százaléka döntött az online folytatás mellett, végül csak 54 százalékuk töltötte ki a kérdőívet online (*Allison–O'Konis* [2002]). Macer szerint más-más módszertanpárok hasonló eredményt hoznak a válaszadási hajlandóság tekintetében (*Macer* [2006]). Annak ellenére azonban, hogy a váltók közül nem mindenki tölti ki a kérdőívet, még így is nagyobb a válaszadási hajlandóság, mintha egyetlen módszert alkalmaztak volna.

3.7. Az adatok feldolgozása és elemzése

A kutatás kezdetén szükséges eldönteni, hogy az adatokat együtt, vagy különkülön szeretnénk feldolgozni és elemezni. Ha az adatokat együtt akarjuk elemezni, akkor a különböző adatfelvételi technikával készült részeredményeket egyetlen adatbázisban kell egyesíteni. Itt figyelni kell a rész adatfile-ok struktúrájának azonossá-

gára, illetve arra, hogy ha paneles online adatfelvételt alkalmaztunk, akkor érkeznek-e adatok a kérdőíven kívül magából a paneltagok adatbázisából.[11]

A körültekintő súlyozás szintén fontos, alaposan át kell gondolni, hogy kiket reprezentálnak a válaszadók, mire és hogyan érdemes súlyozni, milyen súlyok alkalmazhatók. Az adatok elemzésénél pedig erősen oda kell figyelni a megfelelő súlyok használatára.

További érdekes terület az adatok feldolgozásakor a nyitott kérdések kezelése. Az online módszertannak ugyanis sajátossága, hogy sokkal több és hosszabb válaszok születnek, mint a személyes adatfelvétel esetében, ráadásul ilyenkor, az önkitöltés miatt, a kérdezőbiztos nem végezhet akaratlanul is „előzetes kódolást", biztosan a válaszadó saját szavai olvashatók a válaszokban. Mindez összetettebbé teszi a nyitott kérdésre adott válaszok bekódolását.

Az adatok feldolgozásánál a legnehezebb annak a kérdésnek a megválaszolása, hogy a kapott eredményekre milyen hatással volt a módszertan. Nagyon nehéz megmondani, hogy az eltérések melyik részét okozta az alkalmazott adatfelvételi mód, a módszertanból fakadó eltérés, vagy a válaszmegtagadás eltérő szintje (*Voogt–Saris* [2005]), illetve a társadalmi elvárásoknak való megfelelés (*Heerwegh* [2009]). Ha azonban a fentebb taglalt szempontoknak megfelelően gondosan megterveztük a kutatást, az eredményeinkben ellenőrizhetővé és kiszűrhetővé válnak a módszertanból fakadó anomáliák.

Mindezen szempontok alapján, mintegy összegzésül elmondható, hogy a kutatás kezdetén gondos mérlegelést kíván, érdemes-e hibrid adatfelvételt csinálni, mivel az nem alkalmas minden esetben és minden téma vizsgálatára. Tesztkutatásokra és korábbi tapasztalatokra alapozott, gondos tervezésre van szükség, hogy a kutatás sikeres legyen. A következő részben pontosan azzal a kérdéssel foglalkozunk, hogy mikor érdemes hibrid kutatást végezni, vagyis mik az alkalmazás körülményei.

4. A hibrid kutatások alkalmazhatósága

Amikor kutatást tervezünk, akkor különböző paramétereket veszünk figyelembe (*de Leuuw* [2005]), például, hogy mik a kutatni kívánt kérdések, kiket kell megkérdezni (mi az alapsokaság és hogyan lehet belőle mintát venni), melyek a korlátozó

[11] A kérdőív rövidítése miatt a paneltagoktól gyakran nem kérdezik meg azokat a szocio-demográfiai jellemzőiket, amelyek tartósan azonosak (például nem, születési idő), vagy ritkán változnak (például településtípus). Utólag azonban ezeket a jellemzőket hozzá kell rendelni a válaszokhoz úgy, hogy közben a válaszadók anonimitása nem sérül (vagyis személyazonosságuk – például pontos lakcímük, nevük vagy e-mail címük – és a kérdőívre adott válaszuk nem kapcsolható össze a hatályos magyar jogszabályok alapján).

például időbeli, költségbeli, adatvédelmi stb.) körülmények, és a legalkalmasabb megoldást választjuk, amivel összegyűjthetők az adatok, tehát megbízható és érvényes eredményekhez juthatunk megfizethető áron. A hibrid kutatás kiválasztását is hasonló elemzésnek kell megelőznie.

Az elmúlt két év tapasztalatai alapján azt gondoljuk, hogy abban az esetben jó döntés hibrid kutatást csinálni, ha a csak online vagy csak offline kutatás elvégzése önmagában kedvezőtlenebb eredményhez vezetne, mintha ötvöznénk a két módszert. Tehát a hibrid kutatás akkor indokolt, ha például növeli a reprezentativitást, vagy képes a válaszadási hajlandóságon, illetve a korábbiakban felsorolt egyéb szempontokon (például költségek, az adatfelvétel kapacitása, a válaszadó-barátság stb.) javítani. A téma szakértői általában úgy fogalmaznak (*ESOMAR* [2007]), hogy valamely kommunikációs csatorna (például évtizedekkel korábban a telefon) akkor válik alkalmassá adatfelvételre, ha a vizsgálandó alapsokaság legalább 50 százaléka elérhető rajta keresztül. Magyarországon a felnőtt lakosság több mint 50 százaléka internetezik, és a kutatási célcsoportok jó részére igaz az 50 százalékos határ átlépése. Az online módszerek alkalmazása teljesen elterjedt. Bár az online adatfelvételt gyakran inkább az offline alternatívájaként emlegetik, a két módszer kombinációja a felhasználás körének bővítését teszi lehetővé. Bizonyos esetekben a hibrid megoldás egyértelműen jobb, mint az összes többi. Ez a helyzet akkor, amikor a nagy arányban internethasználó csoportok és a teljes, jelentős offline népességet tartalmazó alapsokaság reprezentációja egyaránt fontos.[12] Hasonlóan indokolt lehet a hibrid kutatás, amikor egy bizonyos csoport nem elérhető az online panelből, vagy mert az adott, szűk földrajzi régióhoz köthetően túl sok válaszadót kell megkérdezni, vagy mert eleve igen alacsony számú a csoport tagjainak előfordulása. Ilyen esetekben a hibrid adatfelvétel nem csupán reális, hanem a kizárólag offline vagy online kutatásoknál egyértelműen jobb minőségű mintát eredményező lehetőség, mivel egyetlen módszerrel az adott célcsoport nehezen elérhető.

Ha a célcsoport jellege és a vizsgálat témája lehetővé teszi a hibrid kutatást, akkor ennek a következő előnyei lehetnek.

– A vizsgált célcsoportok pontosabb, jobb elérése.
– A kutatási adatok minőségének javulása (ahhoz képest, mintha vagy csak online, vagy csak offline készülne el a felmérés).
– Rövidebb adatfelvételi idő (egy azonos mintaméretű kizárólag offline folyó kutatáshoz képest).
– Az online módszer használatából fakadó költséghatékonyság (feltéve, hogy a kutatás gondos tervezése folytán nem lesznek felesleges párhuzamosságok).

[12] Az Ipsos rendszeres kutatásai révén képes pontosan megítélni, hogy egy célcsoport mekkora hányada internethasználó, így körükben milyen sikerrel alkalmazható hibrid kutatás.

Ezek az előnyök azonban csak akkor jelentkeznek, ha kellő körültekintéssel tervezzük meg a kutatást. Egyértelműen és jól körülhatárolhatóan kell meghatároznunk, hogy az alapsokaságnak mely részét vizsgáljuk online és mely részét offline. Előre el kell dönteni, hogy az offline csatornán felvett mintában csak internetet nem használók szerepeljenek-e, esetleg offline is kérdezzünk-e internethasználókat, vagyis, hogy milyen típusú hibrid kutatást végezzünk. Bizonyos témák vizsgálatakor a kérdezőbiztos jelenléte feszélyezheti a válaszadót, aki emiatt „kozmetikázhatja" a válaszait, míg az online kutatásokban ez a hatás nem jellemző. Emiatt ugyanarra a kérdésre esetleg eltérő válaszokat kaphatunk (online kérdőívek esetében az önkitöltés miatt őszintébbek a válaszadók), amire fel kell készülni a kutatás tervezésekor. A költségcsökkentő hatás pedig csak akkor tud jelentkezni, ha a kétféle adatfelvétel nem eredményez felesleges párhuzamosságokat.

A hibrid kutatások számos területen alkalmazhatók a piac- és véleménykutatásban (*Lakatos* [2008]). Példaként említjük:

– Piackutatás: kiemelt részcsoportok (ún. niche csoportok) kutatása;
– Média: közönségkutatás, médiafogyasztás;
– Távközlési piac: szolgáltatások használata, megítélése;
– Reklámkutatás: hatásvizsgálatok, szponzorációs kutatások;
– Politikai tárgyú és véleménykutatások: nagy lakossági minták, szubkultúrák, fiatalok.

A következőkben Magyarország egyik legismertebb hibrid kutatását mutatjuk be röviden, amely a rádiós hallgatottság mérésével foglalkozik.

4.1. Egy konkrét példa a hibrid kutatásokra: a rádiós közönségmérés

Jelenleg Magyarországon az Ipsos és a GfK közösen végzi a legismertebb és legnagyobb online-offline hibrid kutatást, a hazai rádiók hallgatottságának és közönségének a mérését. A kutatás tervezése idején a céloknak megfelelő összetett módszertan kidolgozásán túl legalább ennyire fontos feladat volt a végül kiválasztani kívánt módszertan megismertetése és elfogadtatása a célpiaccal, hogy megértsék annak felépítését, alaposságát és az adott helyzetben a hibrid kutatás alkalmazásának indokait (*Melles* [2010]). Ez azért is különösen fontos volt, mivel a kutatás 2010 januárjától újult meg és váltotta fel az addigi mérést, ami közel egybeesett a két korábbi országos kereskedelmi adó (a Sláger és a Danubius) megszűnésével és az újak (a Neo és a Class) elindulásával. Emiatt a mérés külön figyelmet kapott, ami egyúttal lehetővé tette azt is, hogy a hibrid kutatási módszer ismertsége is nőjön Magyarországon.

A hibrid módszertan alkalmazása mellett több érv is szólt, így például a korábbi módszerek (naplós és önálló telefonos) hátrányainak kiküszöbölése, miközben azok előnyeit megőrizzük; a korábbi mérés két részének (országos és helyi) integrálása és a költségek racionalizálása a piac számára.[13] A kizárólag online kutatás azért nem volt kivitelezhető, mivel a rádióhallgatók jelentős része internetezik, de nem mindenki, ugyanakkor telefonon sem lehetséges a teljes célcsoport elérése, így kézenfekvőnek tűnt a hibrid megoldás választása.

A hibrid rádiós mérés lehetővé teszi, hogy azonos mintán, azonos időben, azonos adatfelvételi módszerekkel folyjon a kutatás az országos és a helyi rádióadók számára, illetve, hogy együtt, egységes adatbázisban dolgozzuk fel és tegyük közzé ezeket az adatokat. A két módszer azonos kérdőívvel dolgozik, így az adatok egyetlen adatbázisban kezelhetők.

A bevezetést alapos módszertani kutatások előzték meg. Ezek alapján alakult ki a különböző adatfelvételi módszerek egymás közötti aránya.

Végeredményben a hibrid rádiós mérés sikeres bevezetése bizonyítja, hogy lehetséges idehaza is jó hibrid kutatásokat végezni és megfelelő kommunikáció mellett a piac is fogékony erre az innovatív megoldásra.

5. Teret nyernek-e a hibrid kutatások?

Vajon a hibrid kutatásoké a jövő? Igen is és nem is. A Confirmit és a Meaning Ltd. már korábban említett nem reprezentatív globális felmérése szerint a résztvevő cégek közel fele végez hibrid kutatásokat, viszont ezek az összes kutatás mindössze 6 százalékát teszik ki (*Molloy–Macer* [2009]). Tehát a hibrid módszertan jelenleg nem tekinthető dominánsnak és ez feltehetően később sem fog érdemben változni.

A jövőben sem várható ugyanis, hogy minden kutatás hibrid módon készüljön el, hiszen ahogy az internet elterjedtsége nő és a legfontosabb célcsoportok kizárólag online is jól elérhetők, nem indokolt az összetettebb hibrid megoldás választása. Viszont minden olyan esetben érdemes hibrid kutatást készíteni, ahol azok jobb együttes reprezentálása miatt egyszerre szükséges online és offline elérhető célcsoportokon kutatást végezni.

Mivel Magyarországon az internet elterjedése elmarad a fejlettebb nyugati országokétól és alig haladja meg az 50 százalékot, itt sokkal gyakrabban fordulhat elő, hogy a hibrid kutatás reális választás, nem csak a divat, vagy az újdonságértéke miatt alkalmazzák azt. Ezért várható, hogy több területen is hibrid kutatások jelenjenek meg. Emiatt különösen fontos, hogy jobban megismerjük a hibrid kutatásokat, a ben-

[13] A korábbi mérésben az országos rádióadókat csak naplóval mérte a kutatás, a helyi rádiókat pedig csak telefonon, az új megoldásban mind a kettőt a hibrid módszer méri.

nük rejlő feladatokat és lehetőségeket. Tanulmányunk azt a célt kívánja szolgálni, hogy ezekből adjon egy kis ízelítőt. Aki ennek alapján kedvet érez a behatóbb vizsgálódásra, annak ajánljuk a felhasznált irodalomban hivatkozott, az interneten elérhető tanulmányokat, előadásokat, azok közül is elsősorban *de Leeuwe* 2005-ös, a *Journal of Official Statistics*-ban megjelent cikkét, amely remek kiindulópont a témával való további ismerkedéshez.

Irodalom

ALLISON, J. – O'KONIS, C. [2002]: If Given the Choice. *Quirk's Marketing Research Review*. 6–7. sz. 20. old.

BÄCKSTRÖM, C. – NILSSON, C. [2002]: *Mixed mode: Handling Method Differences Between Paper and Web Questionnaires.* Mid Sweden University. Östersund. http://www.modsurvey.org/text/MixedMode-MethodDiff.pdf (Elérés dátuma: 2010. június 30.)

BIEMER, P. P. – LYBERG, L. E. [2003]: *Introduction to Survey Quality.* John Wiley. NewYork.

BISSCHOP, A. [2004]: Switch Mode Research – Changing Non-response into On-Line Data. *Marketing Information Event.* 6. sz. http://www.niposoftware.com/Upload/documenten/Switch%20Mode%20paper.pdf (Elérés dátuma: 2010. június 30.)

DE LEEUW, E. D. [2005]: To Mix or Not to Mix Data Collection Modes in Surveys. *Journal of Official Statistics*. 21. évf. 2. sz. 233–255 old. http://www.irss.unc.edu/odum/content/pdf/deleeuw%20mixed%20mode%20jos%202005.pdf (Elérés dátuma: 2010. június 30.)

DE LEEUW, E. D. [2008]: *Question Design and Measurement in Mixed Mode Research*. Survey Measurement/ESRC Question Bank. London. http://surveynet.ac.uk/sqb/about/qbworkshop100408/deleeuw.ppt. (Elérés dátuma: 2010. június 30.)

DILLMAN, D A ET.AL. [2001]: *Response Rate Measurement Differences in Mixed Mode Surveys Using Mail, Telephone, Interactive Voice Response and the Internet.* AAPOR Annual Conference. Montreal.

ESOMAR [2007]: *Global Market Research 2007*. ESOMAR Industry Report.

HEERWEGH, D. [2009]: Mode Differences between Face-to-Face and Web Surveys: An Experimental Investigation of Data Quality and Social Desirability Effects. *International Journal of Public Opinion Research.* 21. évf. 1. sz. 111–121. old. http://ijpor.oxfordjournals.org/cgi/content/short/21/1/111 (Elérés dátuma: 2010. június 30.)

KROSNICK, J. A. [2009]: *Money for Surveys: What about Data-Quality?* 11th General Online Research Conference. Április. Bécs.

LAKATOS Z. [2008]: *„HIBRID" Offline és online adatfelvételek alkalmazása ugyanazon kutatásban*. (Kézirat.)

MACER, T. [2004]: Things to Look for in CATI/CAWI Software. *Quirk's Marketing Research Review.* Július/augusztus. http://www.meaning.uk.com (Elérés dátuma: 2010. június 30.)

MACER, T. [2005]: *Weaving, Not Drowning: Take-Up and Best Practices in Mixed Mode Research*. SPSS Directions User Conference. November 5. Las Vegas. http://www.meaning.uk.com/resources/articles_papers/files/spss_directions_2005.pps (Elérés dátuma: 2010. június 30.)

MACER, T. [2006]: *Think Global, Act Local. Taking a Hybrid Approach to Data Collection and Data Dissemination.* Pulse Train Users. Május 6. Barcelona. http://www.meaning.uk.com/resources/articles_papers/files/pulsetrain_2006.pps (Elérés dátuma: 2010. június 30.)

MELLES K. [2010]: *Új rádiós közönségmérés 2010.* Média Hungary 2010. Május 19. Balatonfüred.

MOLLOY, P. – MACER, T. [2009]: *Where We Are and Where We Might Be Going. Trends in Marketing Research Technology.* CASRO 14th Technology Conference. New York. Május 28–29. http://www.meaning.uk.com/resources/articles_papers/files/CasroTech09-Software-survey-presentation-Molloy-and-Macer.pdf (Elérés dátuma: 2010. június 30.)

TOURANGEAU, R. – YAN, T. [2007]: Sensitive Questions in Surveys. *Psychological Bulletin*. 133. évf. 5. sz. 859–883. old.

VOOGT, R. J. J. – SARIS, W. E. [2005]: Mixed Mode Designs: Finding the Balance Between Nonresponse Bias and Mode Effects. *Journal of Official Statistics*. 21. évf. 3. sz. 367–387. old.

Summary

The study discusses the methodological challenges of mixed-mode data collection – whose popularity has grown in the last few years – on the basis of Hungarian Ipsos' research experience. In this case, mixed-mode research is regarded as a compound of online and face-to-face methods (primarily CAPI). The authors examine the purpose of hybrid research and the reasons for its revaluation (for example ensuring representativeness, better availability of respondents, enhancement of the proportion of answering, cost efficiency). They also discuss its fields of application (media, opinion, market or advertisement research, etc.) and the different professional, organizational and methodological difficulties which may emerge during research (in organization of work, sampling, questionnaire design, data collection and processing). The authors use experience gained by Ipsos in the last few years.

Mintavételi módszerek ritka populációk esetén

Kapitány Balázs,
a KSH Népességtudományi Kutatóintézetének tudományos titkára
E-mail: kapitany@demografia.hu

A tanulmány áttekinti azokat a legfontosabb statisztikai mintavételi eljárásokat, melyek olyan esetekben alkalmazhatók, amikor nem áll rendelkezésre az alapsokaságról közvetlen mintavételi keret.

A bemutatott eljárások: mintanagyság növelése; minta aszimmetrikus rétegzése; minta szűrése; közbenső mintavételi egység szintjén történő szűrés; többszörös/dupla mintavételi keret; kapcsolathasznosító módszerek (hálózati, illetve adaptív csoportos mintavétel) és rejtőzködő populációk esetén alkalmazható módszerek (térben és időben meghatározott, illetve válaszadó által vezérelt mintavétel).

Ezek alkalmazása költséghatékony módon teszi lehetővé adatok gyűjtését olyan speciális társadalmi csoportokról, amelyekről a hagyományos mintavételi alapokon nyugvó adatgyűjtések nem, vagy csak nagyon költségesen tudnak megbízható információt szolgáltatni.

TÁRGYSZÓ:
Mintavétel.
Statisztikai módszertan.

A mintavétel módszertani kérdései közül az egyik legérdekesebb, legtöbbet vitatott probléma a ritka populációk (rare populations) körében végzett mintavétel. Ritka populációk alatt olyan közösségeket, társadalmi csoportokat értünk, amelyekről közvetlen mintavételi keret (tehát az adott populáció tagjairól egy csaknem teljes körű lista) nem áll rendelkezésünkre (nincs, vagy létezik, de nem elérhető). Ilyenek például a magas vérnyomásban szenvedők, egy adott vallási, etnikai csoporthoz tartozók, a 100 kilónál nehezebb személyek, a munkásokat feketén foglalkoztató vállalatok, a HIV-pozitívak és így tovább.

Minél ritkább e csoport, mint részpopuláció előfordulása az azt tartalmazó nagyobb populációhoz tartozó mintavételi kereten belül, annál nagyobb problémába ütközik a belőle történő reprezentatív, valószínűségi mintavétel. Ez utóbbi jellemzőit jelen tanulmányban úgy határozzuk meg (eltekintve a téma kapcsán létező viták ismertetésétől), hogy az adott populáció (csaknem) minden tagjának 0-nál nagyobb, meghatározható mértékű esélye van a mintába kerülésre, illetve a mintából mind az adott populáció ritkaságára, mind a ritka populációt jellemző belső arányokra statisztikai becslést lehet készíteni.

Általánosan elfogadott, hogy szélsőségesen ritka populáció (például 1/1000-es előfordulási gyakoriság) esetén az ilyen jellegű mintavételről lemondunk, és más módszerekhez (például hólabdatípusú vagy szakértői mintavételhez) folyamodunk. Az elmúlt évtizedekben azonban folyamatosan megfigyelhető módszertani törekvés arra, hogy minél alacsonyabb sűrűségű populáció esetén váljon lehetővé egy, az előző kritériumoknak megfelelő reprezentatív, valószínűségi mintavétel.

A következőkben először ez utóbbi esetben alkalmazható módszereket fogjuk bemutatni. Amellett érvelünk, hogy amennyiben egy ritka populáció sűrűsége eléri a 3–5 százalékot, megoldható feladat (igaz nehézségek árán) a reprezentatív mintavétel, még ha a ritka populáció eloszlása kis mértékben el is tér az egyenletestől. Az ismertetésben az egyszerűbb módszerektől haladunk az összetettebbek felé: *1*. mintanagyság növelése; *2*. minta aszimmetrikus rétegzése (disproportionate strati-fication); *3*. minta szűrése (screening); *4*. közbenső mintavételi egység szintjén történő szűrés; *5*. többszörös/dupla mintavételi keret (multiple/dual frame methods);[1] *6*. Kapcsolathasznosító módszerek (linkage exploitation methods): *6.1*. hálózati mintavétel (network sampling); *6.2*. adaptív csoportos mintavétel (adaptive cluster sampling); *7*. Rejtőzködő populációk (hidden populations) esetén alkalmazható módszerek: *7.1*. térben és időben

[1] Az angol kifejezések egy részének nincs még bevett magyar megfelelője. Az általunk javasoltak a magyar módszertani szakirodalomhoz illeszkednek, és nem feltétlenül az angol eredeti megnevezések szolgai fordításai.

meghatározott mintavétel (time-space sampling); *7.2.* válaszadó által vezérelt mintavétel (respondent driven sampling).

A különféle módszerek egymáshoz kapcsolódnak, egymást kiegészíthetik. Természetesen más jellegű csoportosítás, beosztás is lehetséges, ezen a téren a szakirodalom sem egységes. *Kalton* [2001] például 11 módszert különböztet meg, köztük azonban van nem reprezentatív jellegű is (hólabdatípusú mintavétel).

Tanulmányunk elsősorban a mintavételi eljárások ismertetésére összpontosító szakirodalmi áttekintés, tehát csak érintőlegesen foglalkozunk az egyes minták alkalmazása után, az adatbázis elemzésekor felmerülő becslési problémákkal.

Amikor arra lehetőség nyílik, a szakirodalmi példákon túl két visszatérő témán keresztül mutatjuk be e módszerek gyakorlati alkalmazásait. Ezek közül az egyik egy valóságos, az erdélyi magyar populációra készült reprezentatív kutatás („Életünk Fordulópontjai – Erdély" vizsgálat), ami a közelmúltban valósult meg (az alkalmazott mintavételi eljárásról részletesen lásd *Kiss–Kapitány* [2009]). [2]

A másik elméleti jellegű: egy olyan nem megvalósult vizsgálat példája, amely azt szeretné felmérni, hogy hányan, kik és milyen körülmények között végeznek ma hivatalosan, nem hivatalosan vagy félhivatalosan mezőgazdasági bérmunkát Magyarországon.

1. A mintanagyság növelése

Legegyszerűbb lehetőség a ritka populációk vizsgálatára a teljes mintanagyság növelése. Például az Egyesült Államokban, ha elemezhető (legalább 300 fős) fekete és spanyolajkú válaszadót (jelenleg hozzávetőleg 13, illetve 15 százalékos arányuk a felnőtt populációban) szeretnének a mintában a közvélemény-kutató cégek, egyszerűen 1 000 helyett 3 000 fős mintán kérdezik le a kérdőívet. Ha a mintavétel nincs is származás szerint kontrollálva, a 300 feletti elemszámokat a módszer nagy bizonyossággal garantálja.

Ritkább populáció és nagyobb esetszámigény esetén azonban ez az út nehezen járható: egy 2 500-as elemszámú mintához 5 százalékos sűrűség esetén akár 50 000 fős mintát kellene venni, ami 50 százalékos válaszadási hajlandósággal számolva 100 000 mintavételi egység (személy, cég stb.) megkeresését jelentené.

[2] Az adatfelvételre, melyben a KSH Népességtudományi Kutatóintézete (Budapest), a Nemzeti Kisebbségkutató Intézet (Kolozsvár) és a Max Weber Alapítvány (Kolozsvár) vett részt, 2007-ben került sor Erdély magyarul beszélő 20–45 éves lakossága körében. A mintanagyság 2 500 fő volt. A kutatás főbb eredményei magyarul elérhetők: *Spéder* [2009].

2. Aszimmetrikus mintarétegzés

Az aszimetrikus mintarétegzés akkor használható, ha a vizsgált ritka populáció egyes, a mintavétel során rétegzésként használható jellemzők (például nem, kor, lakóhely) szerint ismerten, de nem egyenletesen oszlik el. A mintavételnél a hasznos elemszám növelése érdekében azokat a rétegeket, ahol a ritka populáció nagyobb sűrűségét feltételezzük, felülreprezentáljuk. Ugyanakkor a feltehetően kisebb sűrűséggel rendelkező rétegeknek is hagyunk egy alacsonyabb, de ismert bekerülési valószínűséget. A terepmunka után súlyozással helyreállítjuk a helyes arányt, de ez a ritka populáció válaszadóinak a valós hasznos elemszámát már nem módosítja. (Márpedig ez meghatározó a mintavételi hiba, vagyis az adatok megbízhatósága szempontjából.)

Nézzünk példát egy ilyen jellegű mintavételre az erdélyi magyarok esetén. Első lépésként veszünk egy 3 000 fős reprezentatív mintát a tágan vett erdélyi megyék körében, megyénként rétegezve. Ebbe a mintába előreláthatólag kb. 600 (20%)[3] magyar anyanyelvű kerülne. Anyanyelvet nem figyelembe véve a 3 000 főből mintegy 500 válaszadó élne Hargita–Kovászna–Maros megyékben, utóbbiak közül hozzávetőleg 300 magyar anyanyelvű.

Eközben egy kiegészítő mintában még 1 500 válaszadót választunk ki e három megyéből, ezzel még háromszorosan felülreprezentálva azokat népességarányukhoz viszonyítva. Ezek közül hozzávetőleg 900 fő magyar anyanyelvű. Így összesen a 4 500 válaszadóból 1 500 lesz magyar. Ezen 1 500 ember között persze arányon felül vannak a székely megyékben élők (80 százalékos arányban a valós 50 százalék helyett), melyet utólag súlyozással helyre lehet állítani. A magyarokra vonatkozó adatok mintavételi hibája természetesen nagyobb lesz, mintha „normális" 1 500 fős mintavétel történt volna, de még mindig kezelhető és alacsonyabb, mint egy 600 fős véletlen minta esetén. Persze ilyenkor az elemzés során nagyon óvatosan kell kezelni az elemszámokat területi megoszlásokkal összefüggő tényezők esetén (a módszerről lásd például *Kalton* [2001]).

Természetesen ennek a módszernek is erőteljes korlátai vannak. Alkalmazásához egyfelől egyértelmű sűrűsödési pontokra (mint az erdélyi magyarok esetén a Székelyföld) van szükség, másfelől elengedhetetlen, hogy a vizsgálni szándékozott ritka populáció összességének minél nagyobb aránya legyen megtalálható ezeken a ritka populáció által sűrűn lakott területeken (*Waksberg–Judkins–Massey* [1997]). Ha például magyarországi romákra szeretnénk mintát venni, ahol a megyénkénti eloszlás nem ennyire aszimmetrikus, e módszerrel is legalább 10 000 fős minta kellene 1 000 roma válaszadóhoz.

[3] Az áttekinthetőség miatt kerekített számok.

3. A minta szűrése („szkríning”)

A következő megoldás az előzetes szűrés (screening, szkríning, szkríningelés). Ennek lényege, hogy nagyméretű mintát választunk ki egészen a mintavételi egységek (személyek, cégek stb.) szintjéig. Az utóbbiakkal (személyes, telefonos stb.) kapcsolatfelvételre is sor kerül, de a teljes adatgyűjtés csak ott történik meg, ahol a potenciális adatszolgáltató a keresett ritka populációba tartozik. A többi esetben csak egy rövid regisztrációs adatlapot töltünk ki. Ezzel a módszerrel a terepmunka költsége jelentősen (de a tapasztalatok szerint 50 százalékot csak ritkán meghaladó mértékben) csökkenthető, hiszen egy rövid szűrőkérdőív kitöltésének költsége lényegesen alacsonyabb a teljes kérdőívénél.

Nézzünk példát egy kombinált szkríningelésre azzal számolva, hogy egy szűrőkérdőív kitöltésének költsége 50 százaléka egy teljes kérdőív lekérdezésének. Vegyünk egy 20 000 fős, a munkavállalási korúakra reprezentatív mintát, azt feltételezve, hogy lesz benne 1 000 mezőgazdasági munkavállalásban érintett személy (5%). 2 000 személy – a 20 000 fő véletlen almintája – esetén azt az utasítást adjuk a kérdezőknek, hogy a válaszadótól a kérdőív adatait mindenképpen kérdezzék le, míg a többi esetben csak akkor, ha az illető a szűrőkérdőív alapján valóban érintett a mezőgazdasági munkavállalásban. Ezzel összesen 2 900 lekérdezett (1 900 nem érintett és 1 000 érintett) kérdőívet, valamint 17 100 szűrt, de nem kérdezett címet kapunk, utóbbit fél költségen. Így az adatfelvételi költség 20 000-ről 11 450 egységre csökkenthető. Ez a 43 százalékos megtakarítás ugyanakkor bizonyos szempontból látszólagos, hiszen összesen 2 900 kitöltött kérdőívért fizettük ki 11 450 árát.

A szűrésre jól bevett módszer más célú omnibusz adatfelvételek „szűrőkérdőívként” való használata. Havi 1 000 fős omnibusz adatfelvételekkel számolva egy év alatt 12 000 fő szűrése végezhető el. Ez 10 százalékos sűrűség esetén még az újbóli megkeresés által jelentett lemorzsolódással számolva is 1 000 fős mintát eredményezhet.

Ladányi és *Szelényi* [2001] a kelet-európai romákról szóló vizsgálatukban szintén ezt a módszert használták. Országonként eltérő számban, de mintegy 10–20 000 fős omnibuszos kutatás keretében készült interjúból „szűrték ki” azokat a roma válaszadókat, akikkel később a személyes beszélgetést lefolytatták.

A módszer hátulütője a korábbiakban bemutatottakkal szemben az, hogy nem eredményez a teljes alapsokaságra (tehát nem a ritka populációra) vonatkozó kontrolladatokat (hiszen a „nem kiszűrtekről” nem gyűjt információt). Márpedig ezek a kontrolladatok ritka populációkra történő mintavétel esetén minőségbiztosítási, validálási okokból lennének fontosak.[4] Ezért szokták azt a megoldást alkalmazni,

[4] Ha a teljes populációra vonatkozó adatok jelentősen eltérnek a korábbi felvételek hasonló értékeitől, valószínű, hogy a ritka populációhoz kapcsolódó adatokkal is baj van. Ha nincs információ a teljes populációra, ezt a kontrollt nem lehet elvégezni.

hogy a nagy minta egy kis részét (önmagában is reprezentatív almintáját) mindenképpen lekérdezik, többségét azonban csak szkríningelve. Az előzőkben említett omnibuszos jellegű szűrés erre a problémára is megoldást jelent.

A szűrés, mint módszer sok esetben kombinálható az aszimmetrikus rétegzéssel. Nézzük erre a korábbi erdélyi magyar példát. Először bontsuk szét a 3 000 fős „teljes erdélyi" mintát: 1 000 főt kérdezzünk le anyanyelvtől függetlenül, a másik 2 000-nél végezzünk előzetes szűrést. (Utóbbiak közül így 1 600 nem magyar nyelvű válaszadót csak szűrünk, s 400 magyart kérdezünk.) A székely megyék kiegészítő 1 500 fős mintája esetén ismét előzetes szűrést végzünk nyelv szerint (600 szűrés; 900 magyar teljes kérdezés). Így összességében – 2 200 fő kiszűrése után – kapunk 800 román és 1 500 magyar válaszadót. A szűréssel járó megtakarítás – 50 százalékos szűrési költséggel számolva – jelentős, hiszen egy 3 400 fős minta költségéből megtörtént a 4 500 embert lekérdező adatfelvétel.

Ha rendelkezésünkre áll olcsó szűrési módszer (például telefon), elvileg akár 1 százalékos sűrűségű populáció is elérhető ilyen kombinált mintavétellel. A 2001/2002. évi amerikai zsidó adatfelvétel (National Jewish Population Survey) esetén például zsidók által is lakott területeket felülreprezentálva több mint 170 000 háztartás telefonos szűrését végezték el a mintegy 5 000 fős minta kialakítása érdekében. (Eközben egy 4 000 fős nem zsidó kontrollmintát is felvettek.)

Komoly veszélyt jelent ugyanakkor maga a szűrési eljárás, amely újabb hibalehetőségeket hordoz magában. Kisebb baj, ha nem a keresett ritka populáció tagjai válaszolnak, hiszen az általuk szolgáltatott adatok utólag törölhetők. Nagyobb problémával jár azonban ennek a fordítottja, amikor a keresett ritka populáció tagjai valamilyen okból nem jutnak túl a szűrőkérdőíven. Az utóbbi téves besorolás tömegessége akár a teljes kutatást is ellehetetlenítheti, mivel emiatt alulbecsülhetjük a keresett ritka populáció sűrűségét.

4. Közbenső mintavételi egység szinten történő szűrés

Az egyszerű szűrésnél módszertanilag vitathatóbb, ugyanakkor költséghatékonyabb – és így kisebb sűrűségű populáció elérését teszi lehetővé –, ha a szűrés már a többlépcsős mintavétel valamelyik közbenső szintjén (is) megtörténik („screening with area sampling"). Ez azt jelenti, hogy a többlépcsős mintavétel során olyan mintavételi egységekben, ahol az adott ritka populációnak nincsenek, vagy alig vannak tagjai, csak elméletileg kerül sor mintakijelölésre, a gyakorlatban nem történik meg a válaszadók felkeresése. Ez a módszer értelemszerűen csak abban az esetben alkalmazható, ha a kutatni szándékozott ritka populáció eloszlása a közbenső

mintavételi szinten nem egyenletes. Így feltehető, hogy az adott népesség egyáltalán nincs jelen a közbenső mintavételi egységek (jellemzően települések) jelentős részénél.

Személyi szintű előzetes szűréssel kombinálva mi is ilyen módszert alkalmaztunk az „Életünk Fordulópontjai – Erdély" vizsgálat során az erdélyi magyar nyelvű populációra történő mintavételkor.

Első lépésben az erdélyi megyékre vonatkozóan kialakítottunk egy úgynevezett elméleti kontaktmintát. Itt – más vizsgálatokhoz hasonlóan – úgy jelöltük ki az (elméleti) kutatási pontokat, hogy az önreprezentáló települések mellett régió- és településméret szerinti rétegeket is létrehoztunk. A 10 000 fő alatti, nem önreprezentáló települések esetén az egyes települési rétegeken belül a reprezentativitáshoz szükséges elemszámot egyenletesen osztottuk el arra törekedve, hogy egy településen legalább 50 címet kijelöljünk. A települések meghatározásában nem volt szerepe azok nemzetiségi (vagy anyanyelvi) megoszlásának.

Ezután az elméleti kontaktminta és a tényleges terepmunka mintája közé egy köztes, településszintű szűrést ékeltünk. Azokra a településekre, ahol népszámlálási adatok alapján minimális volt az esély sikeres magyar nyelvű interjúra, nem küldtünk kérdezőbiztost.[5] A felkeresendő személyek száma így mintegy 45 százalékkal csökkent az elméleti kontaktmintához képest.

A kiválasztott mintavételi pontokon belül kijelöltük a reprezentativitáshoz szükséges interjúalanyok számát, magukat a megkérdezetteket pedig a települési névjegyzékből választottuk ki. Minden elérhető válaszadót felkerestünk, és a kérdezőbiztos közülük azokkal készített interjút, akik a szűrőkérdések alapján képesek voltak magyar nyelven válaszolni.

Az ilyen közbenső szintű szűrés esetén kritikus pont, vajon honnan és mennyire megbízhatóan tudjuk megállapítani, hogy az adott közbenső mintavételi egységben valóban nincs-e tagja a keresett ritka populációnak. Hiszen sok esetben erről nem vagy csak nagyon korlátozottan állnak külső információk a rendelkezésünkre.[6] Ennek a speciális, de nem ritka problémának a megoldását *Sudman* formalizálta először 1972. évi cikkében. Eszerint első lépésben a közbenső mintavételi szint minden tagjából kiválasztunk véletlenszerűen egy-egy elemet. Ha a kiválasztott válaszadó tagja a keresett ritka populációnak, folytatjuk a szűréssel kombinált lekérdezést az adott klaszterben. Ellenkező esetben a klaszter többi válaszadóját már nem keressük fel.

A módszer napjainkra lényegesen finomodott: a teljes minta kiválasztását követően kijelölünk egy újabb almintát oly módon, hogy az lehetőleg egyenlő számban tar-

[5] A határt az jelentette, hogy az adott településen a magyarok aránya a 2002. évi népszámláláson elérte-e a 8 százalékot. Ha nem, úgy vettük, mintha az adott településen egyetlen sikeres kérdezés sem valósult volna meg.

[6] Az előbb említett erdélyi példa nem ilyen volt, hiszen itt a településsoros népszámlálási adatok erre a célra elégséges információt adtak.

talmazzon elemeket minden egyes közbenső mintavételi egységből. A terepmunka első lépéseként ennek az almintának minden elemét felkeressük. Ezután azonban csak azokban a közbenső mintavételi egységekben folytatjuk a teljes mintán a lekérdezést, ahol az első alminta eredményes volt, vagyis találtunk a ritka populációhoz tartozó adatszolgáltatót.

További újítási lehetőség, hogy az első alminta esetén nem alkalmazunk szűrést, hanem mindenkit lekérdezünk. Ilyenkor kiküszöböljük a módszernek azt a hátulütőjét, hogy az nem eredményez a teljes populációra vonatkozó kontrolladatokat, hiszen a teljes kutatás eredményeképpen nemcsak a ritka populációt reprezentáló mintát kapjuk meg, hanem egy teljes populációra vonatkozó kontrollmintát is. (Ennek jelentőségéről már a szűrés kapcsán volt szó.) Ez esetben azonban az almintavételnek önmagában is reprezentatívnak kell lennie, ami sok esetben feloldhatatlan ellentmondásban van azzal az elvárással, hogy az almintának egyenlő számban kell tartalmaznia elemeket minden egyes közbenső mintavételi egységről.

5. Többszörös/dupla mintavételi keret alkalmazása

A többszörös/dupla mintavételi keret elnevezésű módszerek (multiple/dual frame methods) egy speciális mintavételi eljárási rendet képviselnek. Ennek lényege, hogy a mintavételkor több mintavételi keret együttes alkalmazására kerül sor. A már az 1950-es évek óta ismert, de tömegesen csak az 1980-as évek óta alkalmazott módszernek a gyakorlati megvalósítást tekintve rengeteg alfaja és ezekhez kapcsolódóan kiterjedt szakirodalma van (például *Skinner–Holmes–Holt* [1994], *Lohr–Rao* [2006]).

Mi a következőkben e módszernek először azzal a speciális esetével foglalkozunk, melyet jellemzően a ritka populációkból történő mintavételek során alkalmaznak. Eszerint a mintavételi keretek között van olyan, amely a teljes ritka populációt magába foglalja („A" keret), de drágán használható, és egy vagy több másik („B" keret), amely nem teljes körű, de a keresett ritka populáció tagjait nagy sűrűségben tartalmazza. (Például „B" keretnek nevezhetők az amerikai indián törzsek törzsi névjegyzékei. Értelemszerűen ezekben nem minden indián szerepel, de akik igen, azok valóban indiánok és könnyen elérhetők.) Tehát ebben az esetben elvileg lehetőség nyílna a korábban felsorolt módszerek alkalmazására – hiszen van egy, a teljes ritka populációt magába foglaló keret –, de a ritka populáció eloszlása mintavételi szempontból mégis olyan kedvezőtlen, hogy azok alkalmazása irreális költségekkel járna.

Az első módszertani kérdés az, hogy miképp lehet kombinálni ezt a két mintavételi keretet. A legegyszerűbb eljárás erre, hogy veszünk két független mintát. Először

az „A” keretből egy olyan nagyot, ami használható elemszámú (mondjuk 500 főnyi) válaszadót eredményez a keresett populációból. Ezután veszünk egy másikat (mondjuk 2 000 főst) a „B” keretből, amelyet azután hozzásúlyozunk a másik keretből kapott ritka populáció eloszlásához.[7] Ezzel a kapott mintanagyságot sikerült meglehetősen alacsony költségen megötszöröznünk. Természetesen ezután kritikus pont az eredmények pontosságának becslése (pont- és szórásbecslések) (*Lohr* [2007]).

Az előbb ismertetettnél lényegesen komolyabb mintavételi probléma, ha nem áll rendelkezésünkre a teljes ritka populációt magában foglaló mintavételi keret, hanem az egyes mintavételi keretek egymást átfedve léteznek, összességében lefedve a keresett ritka populáció minden tagját. Például, amennyiben a Magyarországon mezőgazdasági jellegű munkát végzőket szeretnénk tanulmányozni, szembe kell néznünk azzal, hogy a vizsgálandó populáció jelentős része nem magyar állampolgár, és magyarországi lakhellyel sem rendelkezik. Tehát a Közigazgatási és Elektronikus Közszolgáltatások Központi Hivatala (KEK KH) népesség-nyilvántartásából vett esetleges minta („A” keret) nem fedi le a teljes keresett populációt. Rendelkezésünkre áll egy „B” mintavételi keret is, mivel a külföldi mezőgazdasági munkavállalók (az ellenőrzésekre számítva még a feketén munkát vállalni szándékozók is) kiváltják az alkalmi munkavállalói kiskönyvet. Ez utóbbi a vizsgált populáció szempontjából lényegesen „sűrűbb”, ugyanakkor kisebb, és nem részhalmaza az „A” keretnek. Az „A” és a „B” együttesen viszont teljes lefedést biztosít.

6. Kapcsolathasznosító módszerek

A kapcsolathasznosító módszerek (linkage exploitation methods) abban az esetben alkalmazhatók, ha a keresett ritka populáció tagjai (fel)ismerik egymást, és vannak köztük kapcsolatok (*Kaslsbeek* 2000). (Így ez nem használható például allergiások közüli mintavételnél.)

E csoportba tartozó módszerek lényege, hogy kiválasztunk egy kiinduló (reprezentatív) mintát a keresett ritka népességből, majd ehhez kapcsolódva választunk, gyűjtünk újabb mintaelemeket. Így a mintanagyság olcsón és könnyen az eredeti kétszeresére vagy akár többszörösére növelhető.

A reprezentativitást legalább elvi szinten biztosító kapcsolathasznosító eljárások több fontos ponton megkülönböztethetők az első pillanatban hasonló „hólabdamódszertől”: esetükben *1*. a kiinduló mintának reprezentatívnak kell lennie; *2*. az ismerő-

[7] Vagy mások szerint inkább az „A” mintakeret azon alpopulációjához, amely tagja a „B”-nek is. Ehhez természetesen az „A”-beli adatgyűjtéskor meg kell tudni, hogy ki tagja szintén a „B” keretnek.

söktől már nem lépünk tovább az ismerősök ismerőseire, mivel minden további lépésnél hatványozódna az esetleges torzítás mértéke; *3*. a megadható kapcsolatok típusa és száma pontosan szabályozott.

A következőkben a kapcsolathasznosító mintavételi típusok két formájával, a hálózati (network sampling) és az adaptív mintavétellel (adaptive cluster sampling) foglalkozunk.

6.1. Hálózati mintavétel

A hálózati mintavétel, melynek névadója, „prófétája" *Monroe G. Sirken* (*Sirken* [1970], *Shimizu–Sirken* [2006], történeti áttekintést lásd *Sirken* [1998]), a nevével ellentétben nem mintavételi módszer, hanem csupán egy olyan eljárás, amely több esemény egy megfigyelési egységhez való társítását teszi lehetővé ritka populációk esetén. Erre példa speciális genetikai rendellenességgel élők keresése egy háztartási mintában. Nyilván sűrűn előfordul, hogy egy háztartásban több személy is szenved ilyen betegségben. „Normális" háztartási adatgyűjtés esetén egy család egy esetnek számít. A hálózati mintavétel módszere azonban megengedi, hogy egyrészt egy háztartás annyiszor számítson, ahányszor a rendellenesség benne előfordul, másrészt a válaszadó a háztartástagok külön élő testvéreinek genetikai rendellenességeiről is adjon információt.

Mindezekből következik az a továbblépés, hogy minden társított eseményhez külön adatgyűjtést, kérdőívet társítsunk.

A módszer kapcsán felmerülő két legkritikusabb kérdés a következő: *1*. hogyan jelöljük ki az „ismerősöket" (counting rules); *2*. utólag milyen módon súlyozzuk az adatokat (nyilván a kevés kapcsolatot megadók „értékesebbek"), és becsüljük a mintavételi hibát.

Az utóbbi problémával – hisz cikkünk elsősorban a mintavételi technikákkal foglalkozik – részletesen nem foglalkozunk. Azt azonban lényeges megjegyezni, hogy itt – egyedüliként az ismertetett mintavételi eljárások közül – valójában előbb volt ismeretes a becslési probléma, és utóbb született a mintavételi eljárás. Sirken ugyanis eredetileg épp azzal a kérdéssel foglalkozott, hogy milyen veszélyt jelent az adatok megbízhatóságára a többszörös kiválasztás (multiplicitás), amely nehézzé teszi a hálózati mintákra vonatkozó becslést. Így az alkalmazott becslési eljárások (multiplicity estimator; weighted multiplicity estimator stb.) már rögtön a módszer kialakulásakor rendelkezésre álltak, és az alkalmazási standard részévé váltak.[8]

[8] Nem merül tehát fel senkiben, hogy egy hálózati mintavétellel kapott adatbázisra éppoly módszerrel végezhetők becslések, mint egy hagyományos véletlen minta eredményeire. Más ismertetett módszerekről ez sajnos nem mondható el.

Az ismerősök kijelölésekor szigorúan és jól meghatározott kapcsolatok alkalmazására kerül sor. A mintákat általában a testvérekre, leszármazottakra, nagybácsikra, nagynénikre, közvetlen szomszédokra bővítjük ki. Módszertani minimum, hogy a kijelölt ismerősök száma minden válaszadó esetén ismert legyen, és az eredeti megkérdezettek tudják, ismerőseik vajon szintén a keresett ritka populáció tagjai közé tartoznak-e. Természetesen csak azon ismerősök bevonására kerül sor, akik szintén tagjai a keresett populációnak.

A módszer gyakorlati haszna nem elhanyagolható, egy szórványhelyzetű etnikai közösség esetében például két-háromszorosra lehetne növelni az elemszámot a mintába véletlenszerűen bekerült populációtagok (életben lévő) testvéreinek (a féltestvérek közül csak a kérdezettel egyneműek) felkeresésével.

6.2. Adaptív mintavétel

Az 1990-es évek közepétől egy új módszer került előtérbe, a *Steven K. Thompson* nevéhez köthető (rétegzett/inverz) adaptív mintavétel ((stratified/invers) adaptive cluster sampling) (*Thompson* [1990], *Thompson–Seber* [1996]). (Az elnevezésben szereplő adaptív jelző arra utal, hogy a mintavételi eljárás „adaptálódik", vagyis alkalmazkodik az adatgyűjtés folyamán talált adatokhoz.) Ezen eljárás kiindulópontja más volt, de a ritka populációk mintavételi kérdései felől közelítve lényegében hasonló eredményre jutott, mint a hálózati mintavétel. Eredetileg biostatisztikai célból fejlesztették ki (ritka állatfajok, például bálnák stb. vizsgálatára). (Többek között *Philippi* [2005]). A módszer – mivel az állatok és a növények nem tudnak beszélni, és nem képesek arra, hogy megjelöljék testvéreiket, legjobb barátjukat stb. – a kiinduló mintaelemek közelében automatizált algoritmus szerint keres újabb mintaelemeket, teljességgel megszüntetve az elsődleges mintaelemek ebben játszott – akármennyire is minimális, de – szubjektív szerepét.[9]

A módszer lényege, hogy első lépésben egy hagyományos, előre meghatározott elemszámú mintavételre kerül sor. Ezután azon elemek tekintetében, amelyek tagjai a ritka populációnak, megtörténik a „körülvevő" esetek adatfelvétele is. Ez utóbbiak közül a ritka populáció tagjainál (míg van ilyen) ismét mintába vonjuk a környező eseteket. Az ilyen mintába utólag bevont, de nem a ritka populáció részét képező elemeket peremelemeknek („edge units") nevezik.

A végső minta végül négyfajta elemet tartalmaz: a kiinduló minta ritka populációhoz tartozó és nem tartozó elemeit, az utólag bevont, ritka populációból származó mintaelemeket, illetve a peremelemeket.

[9] A hálózati mintavétel esetén például arra gondolhatunk, hogy a válaszadó „elfelejt" beszámolni a család „fekete bárányának" szerepét betöltő testvérről stb.

A különböző becslések és számítások elvégzésekor más-más eseteket vonunk az elemzésbe. A ritka populáció elterjedtségére vonatkozók készítésekor értelemszerűen a kiinduló minta tagjaiból indulunk ki. A ritka populáció belső arányaihoz kapcsolódó becslésnél a kiinduló minta ritka populációhoz tartozó elemei mellett az utólag bevont mintaelemeket is figyelembe vesszük. Az utóbbi esetben azonban az adatok előzetes belső átsúlyozására van szükség, hiszen a ritka populáció nagyobb sűrűsödési pontjainak mintába kerülésére nagyobb az esély, mint a szórványos elhelyezkedésűekéinek.[10]

A módszer egyszerű, érthető, frappáns, de sajnos csak nem életszerű mintavételi problémák esetén alkalmazható. A korrekt becslések előfeltételei ugyanis a következők: kölcsönösség (ha az A elem szomszédja B-nek, akkor a B elem is legyen szomszédja A-nak); egyenlő „szomszédszám" minden tag esetén; és a „szomszédos" elem egyértelmű definiálhatósága. A módszert így általában földrajzi alapú mintavétellel kombinálják, mivel a valós életben ritkák a szabályos, stabil mértani formákba rendezett válaszadók.

Tegyük fel, hogy egy bejelentésre nem kötelezett mezőgazdasági kultúra (például egy ritka fafajta) elterjedését és állapotát kívánjuk vizsgálni. Ekkor első lépésben szabályos (például 100×100 méteres) négyzetekre osztjuk a vizsgált területet. Majd ezek közül veszünk egy véletlen mintát előre meghatározva azt az egyedszámot vagy -sűrűséget, amelytől kezdve az adott területet „érintettnek" tekintjük. Végül e területi alapon lefolytatjuk az előzőkben leírt eljárást.

A mezőgazdasági munkavállalókra vonatkozó példa esetén a munkavégzés helye alapján történhetne a megközelítés. A véletlenszerűen kiválasztott munkavállalókat az adatgyűjtő elkíséri a megadott napon a munkavégzés helyszínére, és az ott lévő többi munkavállaló közül – egy előre megadott algoritmus alapján – bővíti a mintát.

7. Rejtőzködő populációk esetén alkalmazható módszerek

A ritka populációk egy speciális alcsoportját képezik a rejtőzködő populációk (hidden populations). Ezek tagjai nemcsak alacsony sűrűségben lelhetők fel, hanem még arra is hajlamosak, hogy a hagyományos megkeresési módok (például előzetes telefonos szűrés, kérdezőbiztosi felkeresés a kérdezett otthonában) esetén ne vállalják fel csoporttagságukat. A rejtőzködő populációkra tipikus példák: intravénás

[10] Merész hasonlat, de a közkedvelt torpedójátékban is az őrnaszádokhoz képest nagyobb az esély a több tagból álló anyahajók megtalálására. (http://www.logikaifeladatok.hu/torpedo/torpedo.html Elérés dátuma: 2010. május 26.)

droghasználók, örömlányok, homoszexuálisok csoportjai stb. Jelen tanulmányunkban két, ilyen populációkat elérő speciális módszert mutatunk be röviden.

7.1. Térben és időben meghatározott mintavétel

A térben és időben meghatározott mintavétel (time-space sampling)[11] arra alapoz, hogy a rejtőzködő ritka populációk tagjai is elérhetők bizonyos helyeken (klubokban, internetes chatszobákban, speciális nyilvános tereken stb.) és helyzetekben (csoportspecifikus felvonulásokon, ünnepeken stb.), amikor könnyebben felvállalják csoporttagságukat (*Stueve et al*. [2001], *Mansergh et al*. [2006], *Parsons–Grov–Kelly* [2008]). Az eljárás lényege, hogy mintavételi keretként ezek a helyszínek, események, illetve a helyszíneken belül az idő (nap és óra) számít. Tehát a mintavétel során az előzők közül véletlenszerűen kiválasztott helyszíneken és időpontokban az ott és akkor megjelenő összes személy közül szűrjük ki a célcsoporthoz tartozókat, s veszünk közülük mintát. Az eljárás így összességében három lépcsőből áll: először a helyszínek, majd a megfigyelési időszakok, végül a személyek közül választunk. Szokásos a mintavételi eljárás közbeni rétegzés és az aszimmetrikus felülreprezentálás is. (Például a helyszínek típusai közül rétegzünk, s felülreprezentáljuk azokat, ahol alacsony sűrűséget várunk (low yield venues).)

Statisztikai értelemben nyilvánvalóan sok probléma van a módszerrel. A legkomolyabb, hogy a minta tervezésekor nem lehet feltérképezni és felkeresni minden helyszínt, hiszen vannak kevésbé vagy egyáltalán nem nyilvánosak. A keresett ritka populáció bizonyos tagjai ezeken kívül máshol szinte soha nem tűnnek fel. Ezért a módszerrel elért populáció bizonyos szempontokból különbözhet a teljes rejtőzködő populáció jellemzőitől, ami – az egyébként szükséges – utólagos súlyozással nem kiszűrhető torzításokhoz vezethet. Emiatt a kutatók inkább a vizsgált populációt szűkítik le az ily módon elérhető sokaságokra. Ez okból készül például az „utcai szexmunkásokról" lényegesen több, empirikus adatgyűjtésen alapuló tanulmány az általában vett örömlányokhoz képest.

A hagyományos keresztmetszeti kutatásokkal való összevetések (többek között *Xia et al*. [2006]) arra utalnak, hogy nem árt fenntartással kezelni e mintavételi forma eredményeit. Vannak azonban olyan esetek, amikor a keresett populáció tagjai elkerülhetetlenül megjelennek bizonyos helyszíneken, és ez a szinte egyedüliként használható módszer. Így például, ha illegálisan foglalkoztatott, nem magyar állampolgárságú mezőgazdasági kampánymunkásokat vizsgálunk, nyugodtan állíthatjuk, hogy az alapsokaság tagjainak a meghatározásnál fogva meg kell jelenniük munkavégzésük helyén. Ezért ezek a helyszínek bizonyos időszakokban (szőlőmetszés,

[11] Használatos még a venue-time-space sampling, illetve a time-location sampling kifejezés is.

dinnyeszüret stb.) alkalmasak mintavételi egységnek. Vannak ezek mellett olyan helyek is, ahol ugyan a populáció nem minden tagja jelenik meg, de közöttük nagy a keresett populáció sűrűsége (például a mezőgazdásági települések „emberpiacai", a munkaügyi központok, ahol a külföldi munkavállalók szezonális alkalmi munkavállalói kiskönyveinek[12] kiváltása folyik stb.).

7.2. Válaszadó-vezérelt mintavétel

A rejtett populációk esetén az elmúlt évtizedben gyakran alkalmazott másik módszer a válaszadó-vezérelt (respondent driven) mintavétel (*Heckathorn* [1997], [2002]). Ez gyakorlatilag olyan minta, ahol a kiindulópontok a rejtőzködő populáció szakértő elemei, akik saját maguk kutatásban való részvételre felhívó „vócsereket" (részvételi jegyeket) osztanak ki a populáció általuk ismert más tagjai között. Ez utóbbiak, ha hajlandóak válaszolni, szintén kapnak ilyen vócsereket, melyeket szétoszthatnak, és így tovább, amíg csak el nem jutunk a kívánt mintanagyságig.

A vócserek kiosztási rendszerét és információtartalmát olyan módon kell megoldani, hogy ezek segítségével „visszafejthetők" legyenek az ajánlási hálózatok, illetve információt gyűjthessünk arról, hogy ki-ki hány másik tagot ajánlott (s közülük hány jelentkezett). Így a kutatás eredményeként kapott adatokat elsősorban nem személyi szintű, hanem hálózati jellegű adatbázisként kell felfogni, amely a hálózat jellemzőiről bír hasznos információkkal. Ezek már elégséges alapot nyújtanak a válaszadóktól kapott személyi szintű adatok olyan átsúlyozásához, hogy azokból valódi – a valószínűségi mintavételhez hasonlítható – becsléseket kapjunk. Ez nyilván felértékeli a kis zárt, és lebecsüli a nagy hálózatok tagjait (a levezetést és a becsléseket lásd *Salganik–Heckathorn* [2004]).

Kérdés azonban, hogy mi garantálja az elvi esélyt a populáció minden tagjának a mintába kerülésre. A módszer hívei azzal érvelnek, hogy a hálózati kutatások eredményei szerint szinte minden ember negyed- vagy ötödfokú ismerőse mindenkinek, tehát ha megfelelően hosszú ajánlási láncok működnek a válaszadók összegyűjtése során, akkor az az összes személynek esélyt nyújt a bekerülésre. Ez a logika azonban egyértelműen téves, hiszen e mintavétel esetén az „ismerős ismerősét" csak akkor tudjuk elérni, ha maga az „ismerős" is eleme a keresett ritka populációnak.

Ez tehát gyakorlatilag a hólabdatípusú módszer speciális, továbbfejlesztett és utólag súlyozott alfajának tekinthető, így megítélésünk szerint a meggyőző részeredmények ellenére sem valószínűségi mintavételi eljárás. Az alkalmazott súlyozási mód-

[12] Az alkalmi munkavállalói kiskönyvek e fajtája olyan speciális „intézmény" (volt), amelyet szinte kizárólag az illegális munkát vállalni szándékozók váltottak ki tevékenységük látszólagos lefedésére.

szerek pedig – az előző állítással szemben – nem képesek kompenzálni a mintavétel alapvető hiányosságait.[13]

Irodalom

CHRISTMAN, M. C. – LAN, F. [2001]: Inverse Adaptive Cluster Sampling. *Biometrics*. 57. évf. 4. sz. 1096–1105. old.

HECKATHORN, D. D. [1997]: Respondent-Driven Sampling: A New Approach to the Study of Hidden Populations. *Social Problems*. 44. évf. 2. sz. 74–99. old.

HECKATHORN, D. D. [2002]: Respondent-Driven Sampling II.: Deriving Valid Population Estimates from Chain-Referral Samples of Hidden Populations. *Social Problems*. 49. évf. 1. sz.11–34. old.

KISS T. – KAPITÁNY B. [2009]: Magyarok Erdélyben: A minta kialakítása és az adatfelvétel. In: *Spéder Zsolt* (szerk.): *Párhuzamok – Anyaországi és erdélyi magyarok a századfordulón*. KSH-NKI Kutatási jelentések. 86. Budapest. 31–54. old.

KASLSBEEK, W. D. [2000]: *Sampling Racial and Ethnic Minorities*. Summer Public Health Conference on Minority Health. Június 12–16. Chapel Hill, Észak-Karolina, Egyesült Államok. http://chsr.sph.unc.edu/Dissemination/MinHlth_2000.ppt (Elérés dátuma: 2010. május 26.)

KALTON, G. [2001]: *Practical Methods for Sampling Rare and Mobile Populations*. Proceedings of the Annual Meeting of the American Statistical Association. Augusztus 5–9. http://chsr.sph.unc.edu/Dissemination/asa_pres_2000_35mins.ppt (Elérés dátuma: 2010. május 26.)

LADÁNYI J. – SZELÉNYI I. [2001]: A roma etnicitás „társadalmi konstrukciója" Bulgáriában, Magyarországon és Romániában a piaci átmenet korszakában. *Szociológiai Szemle*. 11. évf. 4. sz. 85–95. old.

LOHR, S. [2007]: *Recent Developments in Multiple Frame Surveys*. http://www.amstat.org/sections/SRMS/proceedings/y2007/Files/JSM2007-000580.pdf (Elérés dátuma: 2010. május 26.)

LOHR, S. L. – RAO, J. N. K. [2006]: Estimation in Multiple Frame Surveys. *Journal of the American Statistical Association*. 101. évf. 405. sz. 1019–1030. old.

MANSERGH, G. ET AL. [2006]: Adaptation of Venue-Day-Time Sampling in Southeast Asia to Access Men Who Have Sex with Men for HIV Assessment in Bangkok. *Field Methods*. 18. évf. 2. sz.135–152. old.

PARSONS, J. T. – GROV, C. – KELLY, B. C. [2008]: Comparing the Effectiveness of Two Forms of Time-Space Sampling to Identify Club Drug-Using Young Adults. *Journal of Drug Issues*. 38. évf. 4. sz. 1061–1082. old.

PHILIPPI, T. [2005]: Adaptive Cluster Sampling for Estimation of Abundances within Local Populations of Low-Abundance Plants. *Ecology*. 86. évf. 5. sz. 1091–1100. old.

SALGANIK, M. J. – HECKATHORN, D. D. [2004]: Sampling and Estimation in Hidden Populations Using Respondent-Driven Sampling. *Sociological Methodology*. 34. köt. 193–239. old.

[13] A módszer híveinek saját honlapján (http://www.respondentdrivensampling.org/ Elérés dátuma: 2010. május 26.) erről további – bár némiképp elfogult – információk nyerhetők.

SHIMIZU, I. – SIRKEN, M. [2006]: *Network Sampling for Rare Trait Inference*. American Statistical Association Proceedings of the Survey Research Methods Section. 3664–3668. old. http://www.amstat.org/sections/srms/proceedings/y2006/Files/JSM2006-000397.pdf (Elérés dátuma: 2010. május 26.)

SIRKEN, M. G. [1970]: Household Surveys with Multiplicity. *Journal of the American Statistical Association*. 65. évf. 329. sz. 257–266. old.

SIRKEN, M. G. [1998]: *A Short History of Network Sampling*. American Statistical Association Proceedings of the Survey Research Methods Section. 1–6. old. http://www.amstat.org/sections/SRMS/proceedings/papers/1998_001.pdf (Elérés dátuma: 2010. május 26.)

SKINNER, C. J. – HOLMES, D. J. – HOLT, D. [1994]: Multiple Frame Sampling for Multivariate Stratification. *International Statistical Review*. 62. évf. 3. sz. 333–347 old.

SPÉDER Zs. (szerk.) [2009]: *Párhuzamok – Anyaországi és erdélyi magyarok a századfordulón*. KSH-NKI Kutatási jelentések. 86.

STUEVE, A. et al. [2001]: Time–Space Sampling in Minority Communities. *American Journal of Public Health*. 91. évf. 6. sz. 922–926. old.

SUDMAN, S. [1972]: On Sampling of Very Rare Human Populations. *Journal of the American Statistical Association*. 67. évf. 338. sz. 335–339. old.

THOMPSON, S. K. [1990]: Adaptive Cluster Sampling. *Journal of the American Statistical* Association. 85. évf. 412. sz. 1050–1059. old.

THOMPSON, S. K. – SEBER, G. A. F. [1996]: *Adaptive Sampling*. Wiley. New York.

WAKSBERG, J. – JUDKINS, D. – MASSEY, J. T. [1997]: Geographic-Based Oversampling in Demographic Surveys of the United States. *Survey Methodology*. 23. évf. 1. sz. 61–71. old.

XIA, Q ET AL. [2006]: The Effect of Venue Sampling on Estimates of HIV Prevalence and Sexual Risk Behaviors in Men Who Have Sex With Men. *Sexually Transmitted Diseases*. 33. évf. 9. sz. 545–550. old.

Summary

The present study addresses the most important sampling methods for rare populations.

The following methods are described: *1*. increase of the sample size; *2*. disproportionate stratification of the sample; *3*. screening; *4*. screening at the level of a primary sampling unit; *5*. multiple/dual frame methods; *6*. linkage exploitation methods: *6.1*. network sampling; *6.2*. adaptive cluster sampling; *7*. methods applicable in cases of hidden populations: *7.1*. time-space sampling; *7.2*. respondent driven sampling. The author values these methods and demonstrates cases in which it is worth applying them.

These methods allow the collection of data on special social subgroups in a cost-efficient manner, on which the traditional sampling methods cannot, or only very costly can provide reliable information.

A kiskereskedelmi forgalom havi megfigyelésének reprezentatív módszertana a 2000-es években*

Dr. Telegdi László,
a matematikai tudomány kandidátusa,
a KSH szakmai tanácsadója
E-mail: Laszlo.Telegdi@ksh.hu

A tanulmány ismerteti a kiskereskedelmi forgalom havi megfigyelésének reprezentatív módszertanát a 2000-es években; foglalkozik a megfigyelés jellemzőivel, a rétegzéssel, a rétegenkénti mintanagyság meghatározásával és a minta kiválasztásával, továbbá tárgyalja a hiányzó adatok pótlását, a felhasznált alternatív becslési módszereket és a becslések helyességének vizsgálatát.

TÁRGYSZÓ:
Kiskereskedelmi statisztika.
Statisztikai mintavétel.
Statisztikai módszertan.

* A szerző ezúton mond köszönetet *Csereháti Zoltánnak*, *Éltető Ödönnek*, *Horváth Józsefnének*, *Merczel Ágnesnek*, *Probáld Ákosnak*, *Süveges Évának* és *Szecsődi Ákosnénak*, akik értékes segítséget nyújtottak a módszertan kialakítása során. A tanulmányban előforduló esetleges hibákért kizárólag a szerzőt terheli felelősség.

Kiskereskedelmi forgalmon a kiskereskedelmi üzletek (boltok és vendéglátóhelyek) eladásainak összességét értjük. A havi kiskereskedelmi forgalom az egyik legfontosabb konjunktúramutató, amely nemcsak a kiskereskedelmi értékesítés, hanem – közvetve – a lakossági fogyasztás alakulását is jellemzi. A forgalom iránt megnyilvánuló érdeklődés ennek megfelelően jelentős: a gazdaság fejlődését kifejező információk köréből nem lehet nélkülözni ennek adatait. A havi kiskereskedelmi megfigyelés nagyszámú üzletre terjed ki, és 1991 óta mintavételen (más szóval mintakiválasztáson) alapul, reprezentatív. A reprezentatív megfigyelések ily módon a kiskereskedelmi statisztika lényeges elemévé váltak. A Központi Statisztikai Hivatal (KSH) a kiskereskedelmi forgalmat ugyanazon felvétel keretében figyeli meg, mint a vendéglátás és a gépjármű-kereskedelem forgalmát. Ezért amikor a dolgozatban kiskereskedelmi forgalomról beszélünk, ez utóbbiak forgalmát is beleértjük.

1. A megfigyelés jellemzői

A Kiskereskedelmi és Szálláshelyi Összeírás (KSZÖ), az önkormányzati adatgyűjtések, valamint az ezek alapján felállított és folyamatosan karbantartott Kiskereskedelmi Regiszter (KISREG) lehetővé tették, hogy a KSH 1998 januárjában a kiskereskedelem eladási forgalmának új megfigyelési rendszerét vezesse be (*Süveges* [2001]). Az Országos Statisztikai Adatgyűjtési Program (OSAP) 1045 nyilvántartási számú havonkénti felvételének kérdőíve a „Jelentés a kiskereskedelem és vendéglátás eladási forgalmáról". Ez – az üzletet üzemeltető vállalkozás egészére vonatkozó adatok mellett – a következőket tartalmazza:

– az üzlet havi eladási forgalma,
– az üzlet tevékenységében bekövetkezett változás kódja és
– a havi nyitvatartási napok száma.

Ezek közül az üzlet (havi eladási) forgalma a reprezentatív megfigyelés feldolgozásra kerülő mutatója. A megfigyelés – eddig teljesült – célja az egyes szakágazatok sokasági forgalmának megyénként történő becslése (volt) az összes üzletre, ezen belül 2004-ig a kiskereskedelemben külön a kiskereskedelmi vállalkozások üzleteire vonatkozóan oly módon, hogy a becslés régiónként jó legyen szakágazat-csopor-

tokra. A becslést 2007-ig a TEÁOR (Gazdasági tevékenységek egységes ágazati osztályozási rendszere) ’03, 2008-ban mind a TEÁOR’03, mind a TEÁOR’08, 2009 óta a TEÁOR’08 szerint végeztük, illetve végezzük.

A reprezentatív megfigyelés célsokasága a Magyarországon üzemelő, kiskereskedelmi és vendéglátó tevékenységet végző boltok, kiskereskedelmi telephelyek és vendéglátóhelyek, összefoglalóan üzletek összessége. A célsokaságra vonatkozó nyilvántartás a KISREG. Az ebben szereplő működő üzletek a megfigyelési és egyben mintavételi egységek. Ezek összessége a megfigyelés kerete, a mintavételi keret. Ennek nagysága, vagyis a megfigyelési egységek száma 2006-ig folyamatosan nőtt (akkor 216 ezer volt), azóta csökken; jelenleg 200 ezer. A reprezentatív adatgyűjtés értelemszerűen azokra a megfigyelési egységekre terjed ki, amelyeket a mintavétel során a mintavételi keretből kiválasztottunk. Az adatszolgáltató nem a megfigyelési egység, vagyis az üzlet, hanem az üzletet üzemeltető vállalkozás. Ezek száma, vagyis az adatszolgáltatói keret nagysága jelenleg 135 ezer (egy vállalkozásnak több üzlete is lehet). A beérkező adatokat – a hiányzókat a 4. részben leírtak szerint pótolva – teljeskörűsítjük.

A kiskereskedelem eladási forgalmának reprezentatív havi megfigyeléséhez a mintakiválasztást rétegzett mintavétellel hajtjuk végre. Ennek során a következő rétegeket (cellacsoportokat) képezzük.

1. A mintavételi keret alapján a kiskereskedelmi és a vendéglátó üzleteket megkülönböztetjük.

2. A kiskereskedelmen belül 2001-ig és 2007 óta 10 szakágazat-csoportot különböztettünk, illetve különböztetünk meg:

– gépjárművek és alkatrészeik kereskedelme, motorkerékpárok és alkatrészeik kereskedelme és javítása,
– üzemanyag-kiskereskedelem,
– élelmiszer jellegű vegyes kiskereskedelem,
– iparcikk jellegű vegyes kiskereskedelem,
– élelmiszer-, ital- és dohányáru-kiskereskedelem,
– gyógyszerek, gyógyászati termékek és illatszerek kiskereskedelme,
– textil-, ruházati, lábbeli- és bőráru-kiskereskedelem,
– bútorok és műszaki cikkek kiskereskedelme,
– kultúr- és egyéb cikkek kiskereskedelme,
– használtcikk-kiskereskedelem.

2002-től 2006-ig ezekhez még egy szakágazat-csoportot vettünk hozzá:

– fogyasztási cikk javítása.

A vendéglátáson belül a TEÁOR'03 szerint 2 szakágazat-csoportot különböztettünk meg:

– kereskedelmi vendéglátás,
– munkahelyi és közétkeztetés.

A TEÁOR'08 szerint a vendéglátás egy szakágazat-csoport.

3. Az egyes szakágazat-csoportokon belül a TEÁOR'03 szerint 2001-ig 24 kiskereskedelmi és 3 vendéglátó, összesen tehát 27 szakágazatot különböztettünk meg. 2002-től 2004-ig a megfigyelést 30 kiskereskedelmi és 3 vendéglátó, összesen tehát 33 ágazati egységben (29 szakágazat és 4 rész-szakágazat, a továbbiakban egységesen szakágazat) végeztük. A megfigyelt szakágazatok számát 2005-ben 1-gyel növeltük, 2007-ben 4-gyel csökkentettük.

Az egyes szakágazat-csoportokon belül a TEÁOR'08 szerint 29 (20 tényleges és 9 fiktív) kiskereskedelmi, valamint 3 vendéglátó, összesen tehát 32 szakágazatot különböztetünk meg.

4. A 7 régiót megkülönböztetjük.

5. A Közép-Magyarország régión belül Budapestet és Pest megyét megkülönböztetjük (ezen túlmenően azonban a mintakiválasztásnál a megyéket nem).

6. A kiskereskedelmi szakágazatokon belül 2004-ig a Gazdasági Szervezetek Regisztere (GSZR) alapján megkülönböztettük a kiskereskedelmi és az egyéb vállalkozások üzleteit. 1999-től 2004-ig a három vendéglátó szakágazaton belül – ugyancsak a GSZR alapján – megkülönböztettük a kiskereskedelmi és a vendéglátó vállalkozások üzleteit.

7. Mindezeken belül 1998-ban megkülönböztettük a KISREG-ben forgalmi adattal rendelkező és az ilyennel nem rendelkező üzleteket. Az utóbbi réteget reprezentatívan figyeltük meg. A forgalmi adattal rendelkező üzleteket nagyság szerint kategorizáltuk. Nagyságkategóriák képzésére a kiskereskedelmi árbevételnek a KISREG-ben található bázis értéke szolgált. Ez alapján egy teljes körűen és egy második reprezentatívan megfigyelt réteget alakítottunk ki. A mintába kiválasztott vagy teljes körűen megfigyelt üzleteket üzemeltető vállalkozások száma, vagyis az adatszolgáltatói kör nagysága azonban ily módon túl nagy volt. Az adatgyűjtés hatékonyságának növelése végett ezért a teljes körűen megfigyelt üzletek rétegeinek kialakítását 1999-ben új alapokra helyeztük. 1999-től 2004-ig a forgalmi adattal rendelkező és az ilyennel nem rendelkező üzletek nem lettek megkülönböztetve, és az előbbiek nem voltak nagyság szerint kategorizálva. Ezekben az években egy teljes körűen és egy reprezentatívan megfigyelt réteget alakítottunk ki. A TEÁOR'03 szerint 2005 óta 10 szakágazatban, a TEÁOR'08 szerint (2008 óta) 8 szakágazatban egy, a többi 24 szakágazatban két reprezentatívan megfigyelt réteget alakítottunk, illetve alakítunk ki.

Néhány kiegészítéssel a teljes körűen megfigyelt rétegbe soroltuk, illetve soroljuk

– 2006-ig a legalább 50 fős, legalább 2 üzletet üzemeltető és az 50 fő alatti, legalább 10 üzletet üzemeltető, 2007 óta a kiskereskedelemben a legalább 50 fős, legalább 6 üzletet üzemeltető és az 50 fő alatti, legalább 10 üzletet üzemeltető, a vendéglátásban a legalább 20 fős, legalább 4 üzletet üzemeltető vállalkozások üzleteit, valamint

– 2005 óta – a vendéglátó szakágazatok kivételével – azoknak a vállalkozásoknak az üzleteit, amelyek valamely üzlete az üzlet szakágazatához tartozó (nagyobb) küszöbértéket meghaladja.

Azokban a szakágazatokban, amelyekben két reprezentatívan megfigyelt réteget alakítunk ki, a reprezentatívan megfigyelt üzleteket az alapján soroljuk az egyik vagy másik rétegbe, hogy a kisebb küszöbértéket meghaladják (nagyobbak) vagy nem haladják meg (kisebbek). A küszöböt, illetve küszöböket a gyógyszer-kiskereskedelem szakágazatban a forgalom, a többi szakágazatban az alapterület alapján adjuk meg.

A mintakiválasztáshoz az előbbiek szerint kialakított reprezentatív rétegek száma tehát

1998-ban	$(24 \times 8 \times 2 \times 2) + (3 \times 8 \times 2)$	$= 816,$
1999-től 2001-ig	$27 \times 8 \times 2$	$= 432,$
2002-től 2004-ig	$33 \times 8 \times 2$	$= 528,$
2005-től 2006-ig	$(10 + 24 \times 2) \times 8$	$= 464,$
2007-től 2008-ig	$(7 + 23 \times 2) \times 8$	$= 424,$
2009 óta	$(8 + 24 \times 2) \times 8$	$= 448$

(volt). A teljes körű rétegek száma a gyógyszer-kiskereskedelemmel 2001-ig 448, 2002-től 2004-ig 544, 2005-től 2006-ig 272, 2007-től 2008-ig 240 volt, 2009 óta 256. A megyék cellákat képeznek a rétegeken belül, amelyek tehát cellacsoportoknak is tekinthetők.

2. A rétegenkénti mintanagyság meghatározása

A rétegenkénti mintaelemszám meghatározása 1998-ban az alábbi lépésekben történt. Az első lépésben a KSZÖ adatbázisából előállított táblázatok alapján a for-

galmi adattal rendelkező üzletek nagyság szerinti kategorizálásához rétegenként küszöböket adtunk meg; a KISREG-ben ennél nem kisebb forgalmi adattal rendelkező üzleteket teljes körűen, az egyéb üzleteket reprezentatívan figyeltük meg. A második lépésben e küszöböknek a figyelembevételével, a KISREG 1997. december 15-i állapota alapján meghatároztuk valamennyi réteg előzetes nagyságát és a forgalmi adattal rendelkező üzletek rétegeinek előzetes forgalmát. A harmadik lépésben ezen adatok figyelembevételével meghatároztuk a rétegenkénti előzetes mintanagyságot. Ezt szimulációs kísérletek segítségével végeztük, azzal a módszerrel, amelyről *Telegdi* [2004] számol be. A módszer lényege, 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. E szimulációs kísérletek alapján határoztuk meg az előzetes mintanagyságot az egyes szakágazatokban. A negyedik lépésben a teljes KISREG alapján meghatároztuk valamennyi réteg végleges nagyságát és a forgalmi adattal rendelkező üzletek rétegeinek végleges forgalmát. Az ötödik lépésben ezek alapján megadtuk a rétegenkénti végleges mintanagyságot.

1999 óta a rétegenkénti mintaelemszámot a következőképpen határozzuk meg. Az első lépésben 2007-ig az előző évi III. negyedéves, 2008 óta az előző évi II. negyedéves állapot szerint meghatároztuk, illetve meghatározzuk az egyes rétegek nagyságát és forgalmát, valamint 2000 óta relatív szórásukat és relatív hibahatárukat az – 1999-ben mind az 1998-ban használt, mind az új módon kialakított – teljes körűen, illetve reprezentatívan megfigyelt rétegek mellett, továbbá az – 1999-ben az 1998-ban használt módon kialakított – reprezentatívan megfigyelt rétegekhez tartozó előző évi mintaelemszámokat. A második lépésben ezek alapján, a bemutatott szimulációs módszerrel határozzuk meg a rétegenkénti tárgyévi mintanagyságot.

2000-ben – a becslés helyességének kívánatos javítása és az adatgyűjtési lehetőségek közötti reálisnak látszó kompromisszumként – lehetővé vált, hogy a mintanagyságot közel a kétszeresére, 3300-ról 6200-ra növeljük. Ugyanakkor a hibaszámítás eredménye nem tette szükségessé, hogy a minta összetételét alapvetően megváltoztassuk. Ezért a következőképpen jártunk el. Szakágazatonként az 1999-es mintanagyságot előbb a 2000-es és 1999-es sokaságnagyság – 1-hez közeli – hányadosának négyzetgyökével megszoroztuk, majd arányosan úgy növeltük, hogy az így meghatározott elemszámok összege 4100 legyen (6200 mintegy kétharmada; ezt 20 százalékos növeléssel értük el). Az ezek után maradt többlet elemszámot (2100) az említett szimulációs módszerrel osztottuk el. Az így meghatározott mintaelemszámok további elosztását az egyes szakágazatokon belül, kisebb módosításoktól eltekintve, arányosan végeztük.

Annak érdekében, hogy a TEÁOR'03 szerint kiválasztott minta a TEÁOR'08 szerint történő becslésre is jó legyen, 2008-ban a teljes mintanagyságot 10 százalék-

kal növeltük. A pontosság javítása érdekében 2009-ben a mintanagyságot további 9 százalékkal növeltük, 2010-ben viszont 2,5 százalékkal csökkenteni tudtuk. Jelenleg a mintanagyság 8,7 ezer (186 ezerből); 14,5 ezer üzletet teljes körűen figyelünk meg. Összesen tehát 23,2 (=8,7 + 14,5) ezer üzlet (az összes üzlet 11,6 százaléka) jelentett vagy pótolt adatából becsüljük a havi kiskereskedelmi forgalmat.

3. A minta kiválasztása

A megfigyelés sikerességéhez elengedhetetlen a kiválasztott minta karbantartása. Ennek fontos mozzanata a mintaelemek bizonyos idő utáni lecserélése, 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. (Az ismétlődő reprezentatív gazdaságstatisztikai felvételek során alkalmazott rotációt a KSH szabályozza; lásd például *Telegdi* [1999].) A kiskereskedelmi forgalom havi megfigyelése során az adatgyűjtés hatékonysága érdekében ezen túlmenően az adatszolgáltatókra, a vállalkozásokra is tekintettel kell lennünk.

Mindezeket figyelembe véve az egyes rétegekre a mintavételt a következőképpen végezzük. A mintavételi kerethez és az adott réteghez tartozó üzletek mindegyikéhez előállítunk egy a 0 és 1 között egyenletes eloszlású véletlen számot, vagy vesszük a korábban előállított ilyen számot. 2001-től 2003-ig abból a célból, hogy előnyben részesítsük azon vállalkozások üzleteit, amelyeknek egyetlen üzlete sem volt mintaelem 3 évvel korábban, ezen belül pedig azokat az üzleteket, amelyek már – bármely hónapban – az előző éves mintának is elemei voltak, a megfelelő véletlen számokat 2-vel, illetve további 1-gyel csökkentettük, majd az üzleteket az így módosított véletlen számok nagysága szerint növekvő sorba rendeztük. Ebben a sorban az üzletek tehát a következő rendben követték egymást.

1. Azoknak a vállalkozásoknak az előző éves mintához tartozó üzletei, amelyek egyetlen üzlete sem volt 3 évvel korábban mintaelem.

2. Azoknak a vállalkozásoknak az előző éves mintához nem tartozó üzletei, amelyek egyetlen üzlete sem volt 3 évvel korábban mintaelem.

3. Azoknak a vállalkozásoknak az üzletei (tekintet nélkül arra, hogy az előző éves mintához tartoztak-e), amelyek valamely üzlete mintaelem volt 3 évvel korábban.

Az ily módon véletlen sorba rendezett üzletek közül az elsőket választottuk (megfelelő számban) a mintába.

2001-ben a mintába az említettek szerint kiválasztott üzleteknek több mint a fele nem volt eleme a 2000-es mintának, ezért a mintavételt a módosított véletlen számok további módosításával megismételtük. Ennek során 3-mal csökkentettük az olyan üzletekhez tartozó véletlen számokat, amelyekre teljesültek a következők: *1*. az üzlet eleme volt a 2000-es mintának, *2*. a megfelelő vállalkozásnak ugyan volt olyan üzlete, amely eleme volt az 1998-as mintának, de egyetlen üzlete sem volt eleme az 1999-es mintának.

2002-ben a mintába a 2001 előtti években használt módon kiválasztott üzletek túl nagy része eleme volt a 2001-es mintának is, ezért a megfelelő mértékű rotáció biztosításáért a mintavételt módosítva megismételtük. Ennek során azok közül az eredetileg kiválasztott mintaelemek közül, amelyek az előző két évben mintaelemek voltak, lecseréltünk az előző éves mintához nem tartozó üzletekre annyit, hogy azoknak a mintaelemeknek a hányada, amelyek nem tartoztak az előző éves mintához, rétegenként elérje a 30 százalékot. 2003-ban a minta kiválasztása hasonlóan történt.

2004 óta abból a célból, hogy előnyben részesítsük azokat az üzleteket, amelyek már – bármely hónapban – az előző egy vagy két évben is mintaelemek voltak, a megfelelő véletlen számokat 2-vel, illetve 1-gyel csökkentjük. Az ily módon véletlen sorba rendezett üzletek közül az elsőket választjuk a mintába a mintanagyság 70 százalékáig. A többi mintaelemet elsősorban azon üzletek közül választjuk, amelyek 2002 óta egyetlen évben sem tartoztak a mintához.

A bemutatottak szerint elvégzett mintavétel a visszatevés nélküli rétegzett egyszerű véletlen kiválasztás módosított változata: az egyes rétegeket tulajdonképpen csoportokra osztjuk, és a csoportokból az üzleteket különböző valószínűséggel választjuk a mintába. A módosítás ellenére a teljeskörűsítésnél és a hibaszámításnál a mintát rétegzett egyszerű véletlen mintának tekintjük.

Az adatszolgáltatói kör nagysága jelenleg 9,4 ezer (az adatszolgáltatói keret 7 százaléka).

A megfigyelés során az adatszolgáltatók a kérdőíveket negyedévente postán kapják meg, és azokat havonta postán kell visszaküldeniük a KSH illetékes adatgyűjtő egységének.

Előfordulhat, hogy a mintához tartozó üzletek közül néhányat más rétegbe kell sorolni. Ezeket a módosításokat nemcsak a mintán, hanem a mintavételi kereten is elvégezzük.

4. A hiányzó adatok pótlása

Az adatgyűjtés eredményességét kedvezőtlenül befolyásolhatja a nemválaszolás. Ennek, vagyis a nem teljesített adatszolgáltatásnak okairól a KSH érkeztető rendszere nyújt információt az ún. MV19 kóddal (lásd például *Telegdi* [1999]). A hiányzó adatokat ennek felhasználásával a következőképpen pótoltuk, illetve pótoljuk (imputáljuk). Külön-külön a teljes körűen megfigyelt üzletekre, valamint 2004-ig az összes reprezentatívan megfigyelt (1998-ban együtt a megfelelő küszöbnél kisebb forgalmi adattal rendelkező és a forgalmi adattal nem rendelkező) üzletre, 2005 óta a reprezentatívan megfigyelt nagyobb és a kisebb üzletekre meghatároztuk, illetve meghatározzuk a válaszolók adatának $\overline{y_{0l}}$ szakágazati átlagát. 2000-ig azokban az esetekben, amikor a vállalkozás egyáltalán nem szolgáltatott adatot és a nem teljesített adatszolgáltatás okának MV19 kódja 101–105, 201–204 volt, vagy a vállalkozás az üzletre vonatkozóan nem szolgáltatott forgalmi adatot és az üzlet változáskódja azt jelezte, hogy az üzlet nem működik, a hiányzó adatot nem pótoltuk. Ellenkező esetben

– 1998-ban olyankor, amikor az üzletnek volt előző havi eredeti (nem pótolt) adata, akkor azzal, egyébként a

$$d_m \overline{y_{0l}}$$

mennyiséggel,

– 1999-ben és 2000-ben olyankor, amikor az üzletnek volt előző havi eredeti adata (ami tehát januárban még nem lehetett), akkor a hiányzó adatot ezen adat és a

$$d_m \frac{\overline{y_{0l}}}{\overline{y_{0l}^e}}$$

mennyiség szorzatával, egyébként a

$$d_m \overline{y_{0l}}$$

mennyiséggel

pótoltuk, ahol $\overline{y_{0l}^e}$ az előző havi szakágazati átlag, a d_m paraméter értékét a nyolc területi rétegre – a régiókra, a Közép-Magyarország régi-

ón belül pedig külön-külön Budapestre és Pest megyére – megadtuk. 2001 óta a következőképpen járunk el. Ha az üzletnek van előző havi eredeti, nem pótolt adata, akkor a hiányzót ennek, valamint a teljes körűen és reprezentatívan megfigyelt üzletek beérkezett adataiból számított átlagos dinamikának a szorzatával pótoljuk. Ellenkező esetben olyankor, amikor a vállalkozás egyáltalán nem szolgáltat adatot és a nem teljesített adatszolgáltatás okának MV19 kódja 101–105, 201–204, vagy a vállalkozás az üzletre vonatkozóan nem szolgáltat adatot és az üzlet változáskódja azt jelzi, hogy az üzlet nem működik, a hiányzó adatot nem, egyébként pedig a

$$d_m \overline{y_{0l}}$$

mennyiséggel pótoljuk.

5. A teljeskörűsítés

A feldolgozás során több, a forgalommal kapcsolatos mennyiség, paraméter teljeskörűsítését, sokasági értékének közelítő megállapítását, becslését is elvégezzük. Az egyes hónapokra háromszor becslünk: kétszer előzetesen – gyorsítottan már a havi adatok beérkezése közben, majd közvetlenül a beérkezés után – a KISREG és a GSZR akkori (az előző előtti negyedév utolsó hónapjának végét jellemző) állapota szerint, véglegesen az előző negyedévre vonatkozó önkormányzati adatok beérkezése után, a következő negyedév utolsó hónapjában.

A becslés folyamán a mintába kiválasztott és válaszoló üzletek adataiból vonunk le rétegenként következtetéseket az üzletek havi eladási forgalmáról, vagyis az ebben a mintában elvégzett megfigyeléseket teljeskörűsítjük. 2006 óta egyes reprezentatívan megfigyelt rétegeket két részre bontunk. Ennek során a réteg kiugró értékkel (outlierrel) rendelkező üzleteit – csak a szóban forgó hónapban – kiemeljük és áttesszük a teljes körűen megfigyelt megfelelő rétegbe.

A kiemelt üzleteket a következő módon határozzuk meg (részletesebben lásd *Csereháti* [2004]). A különböző rétegek összehasonlíthatóvá tétele céljából az üzletek adatait – a forgalom rétegátlagát levonva és a különbséget a forgalom rétegbeli szórásával osztva – standardizáljuk, majd az így kapott értékeket a réteg mintanagyságának függvényében módosítjuk (ezt a módosítást az teszi szükségessé, hogy kevesebb adathoz képest nagyobb valószínűséggel fordul elő nagy érték), és nagyság szerint csökkenő sorba rendezzük. Az ily módon sorba rendezett értékek közül a ma-

tematikai és tapasztalati megfontolások alapján megállapított küszöbnél nagyobb értékűeket tekintjük outliernek.

1999 óta a reprezentatívan megfigyelt üzletekre a teljeskörűsítést nem rétegenként, hanem rétegcsoportonként végezzük. 1999-ben a nyolc területi réteget összevontuk, 2000 óta egyrészt Budapest kivételével a területi rétegeket, másrészt 2004-ig mind Budapesten, mind a vidéken belül a kiskereskedelmi vállalkozások és az egyéb, illetve a vendéglátó vállalkozások üzleteit, 2002 óta pedig ezen túlmenően az élelmiszer-, ital- és dohányáru-kiskereskedelem szakágazat-csoport szakágazatait összevontuk, illetve összevonjuk (tehát szakágazatonként, ezen belül külön-külön Budapestre és a vidékre becslünk). Abból a feltételezésből indulunk ki, hogy a kiválasztott minta jól tükrözi, reprezentálja a célsokaságot (ezért mondjuk reprezentatívnak a megfigyelést). Ez azt jelenti, hogy minden egyes mintaelem a célsokaság bizonyos számú elemét (köztük természetesen saját magát is) reprezentálja. Ezt a – többnyire nem egész – számot a mintaelem súlyának nevezzük.

A teljeskörűsítés folyamán becsüljük a forgalom sokasági értékösszegét. E célból meghatározzuk az egyes üzletek súlyát. A bemutatottak szerint kialakított, reprezentatívan (nem teljes körűen) megfigyelt rétegeken (cellacsoportokon), illetve rétegcsoportokon belül az egyes mintaelemeknek ugyanaz a súlya. Ez 1998-ban a KISREG-ben forgalmi adattal rendelkező üzleteket tartalmazó rétegek esetén a KISREG és a GSZR szerint a teljeskörűsítéskor a réteghez (k) tartozó összes, N_k számú üzlet kiskereskedelmi bázis árbevétel X_k értékösszegének, valamint az ezek közül megfigyelt (a mintához tartozó és válaszoló, nemleges vagy pótolt), n_k számú üzlet kiskereskedelmi bázis árbevétel x_k értékösszegének

$$q_{ki} = q_k = \frac{X_k}{x_k}$$

hányadosa volt (vagyis a teljeskörűsítést ebben az esetben hányadosbecsléssel végeztük), 1998-ban a KISREG-ben forgalmi adattal nem rendelkező üzleteket tartalmazó rétegek, 1999-ben valamennyi rétegcsoport esetén a

$$q_{ki} = q_k = \frac{N_k}{n_k} \qquad /1/$$

hányados volt ($i = 1, 2, ..., n_k$).

2000 óta valamennyi rétegcsoport esetén kétféleképpen becsüljük a forgalom sokasági értékösszegét. Ennek során az üzletek súlyát két alternatív módszerrel határozzuk meg. Az első módszernél a súlyt az /1/ képlettel számítjuk. A második módszernél – összetett becslés (*composite estimation*, lásd például *Foreman* [1991], *Éltető–Mihályffy* [1997]) alkalmazásával – oly módon teljeskörűsítünk, hogy a tárgyhavi

mintaelemek közül fokozott mértékben vesszük figyelembe azoknak az üzleteknek az adatát, amelyek az előző év hasonló hónapjában mint bázishónapban is mintaelemek voltak. Nevezetesen az /1/ hányadost egy c_k korrekciós tényezővel szorozzuk:

$$q_{ki} = q_k = c_k \frac{N_k}{n_k}$$

($i = 1, 2, ..., n_k$). c_k értékét a következő módon határozzuk meg.

Mind a tárgyhónap, mind a bázishónap esetén kiszámítjuk az egyes rétegcsoportok összes (válaszoló, nemleges vagy pótolt) mintaelemének $\overline{y_k^{t1}}$ és $\overline{y_k^{b1}}$ átlagát, valamint ezek közül azoknak az üzleteknek az $\overline{y_k^{t2}}$ és $\overline{y_k^{b2}}$ átlagát, amelyek mindkét hónapban mintaelemek voltak. A *k*-adik réteghez tartozó c_k korrekciós tényezőt a

$$c_k = \left(\frac{\overline{y_k^{t2}} / \overline{y_k^{b2}}}{\overline{y_k^{t1}} / \overline{y_k^{b1}}} \right)^{\beta_k} \qquad /2/$$

képlettel határozzuk meg, amihez a β_k paraméterek 0 és 1 közé eső értékét úgy számítjuk ki, hogy a január és – februártól kezdődően – a tárgyhónap közötti hónapokra rétegcsoportonként meghatározzuk

– a rétegcsoport összes mintaelemének számát, valamint
– ezek közül azoknak a (párosodott) üzleteknek a számát, amelyek a bázishónapban is mintaelemek voltak,

és β_k értékéül az utóbbi és az előbbi hányadosának időbeli átlagát vesszük. (Ha a párosodott üzleteket nem vennénk fokozott mértékben figyelembe, akkor 0, ha a teljeskörűsítést csak ezek alapján végeznénk, akkor 1 értéket kellene adnunk β_k -nak.)

Bár a /2/ képletben szereplő $\overline{y_k^{t2}}$ és $\overline{y_k^{b2}}$ a *k*-adik rétegcsoport azon üzleteinek az átlaga, amelyek mindkét hónapban mintaelemek voltak, a c_k korrekciós tényező a rétegcsoport többi üzletére is vonatkozik, hiszen egy rétegcsoporton belül mindegyik mintaelemnek ugyanaz a súlya.

1998-ban egy-egy cellán (*j*) belül a forgalom Y_j sokasági értékösszegét úgy becsültük, hogy az egyes mintaelemekre vonatkozó y_{ji} tárgyhavi adatokat megszoroztuk a cellát tartalmazó rétegen, így a cellán belül is közös súllyal és összegeztük:

$$Y_j = \sum_{i=1}^{n_j} q_j y_{ji} = q_j \sum_{i=1}^{n_j} y_{ji}.$$

1999 óta az egyes rétegcsoportokra a forgalom Y_k sokasági értékösszegét úgy becsüljük, hogy az egyes mintaelemekre vonatkozó y_{ji} tárgyhavi adatokat megszorozzuk a rétegcsoporton belül közös súllyal és összegezzük:

$$Y_k = \sum_{i=1}^{n_k} q_k y_{ki} = q_k \sum_{i=1}^{n_k} y_{ki}.$$

Az egyes rétegcsoportokra vonatkozó, az előbbiek szerint meghatározott Y_k értékösszegeket megbontottuk, illetve megbontjuk a rétegcsoporthoz tartozó cellák, vagyis 2000-ig a megyék, 2001 óta

– külön Budapestre és külön a vidékre az élelmiszer-, ital- és dohányáru-kiskereskedelem szakágazat-csoport 7 szakágazata (2002-től), majd

– a vidék 19 megyéje, végül

– 2004-ig a kiskereskedelmi vállalkozások és az egyéb, illetve a vendéglátó vállalkozások üzletei

között. Ennek során a következőképpen járunk el. Jelölje y_{kl} és

$$y_k = \sum_{l=1}^{L} y_{kl}$$

a mintaelemek megfelelő értékösszegeit az egyes cellacsoportokra (cellákra), illetve az egész rétegcsoportra vonatkozóan, N_{kl} a megfelelő sokaságokhoz tartozó, n_{kl} pedig a mintába kiválasztott és válaszoló vagy nemlegesnek pótolt üzletek számát. L értéke időben, térben, illetve tevékenység szerint a következőképpen változott:

2000-ig			$L = 20$,
2001-ben		Budapestre	$L = 2$,
		a vidékre	$L = 19 \times 2 = 38$,
2002-től 2004-ig	az élelmiszer-, ital- és dohányáru-kiskereskedelem szakágazat-csoportban	Budapestre	$L = 7 \times 2 = 14$,
		a vidékre	$L = 7 \times 19 \times 2 = 266$,
	egyébként	Budapestre	$L = 2$,
		a vidékre	$L = 19 \times 2 = 38$,
2005 óta	az élelmiszer-, ital- és dohányáru-kiskereskedelem szakágazat-csoportban	Budapestre	$L = 7$,
		a vidékre	$L = 7 \times 19 = 133$,
	egyébként	Budapestre	$L = 1$,
		a vidékre	$L = 19$.

$L = 1$ esetén természetesen nincs szükség megbontásra.

Képezzük az

$$s_{kl} = \frac{N_{kl} - n_{kl}}{N_k - n_k}$$

súlyszámokat. A mintaelemekre vonatkozó értékösszegekkel csökkentett $(Y_k - y_k)$ sokasági értékösszegeket e súlyszámok alapján bontjuk meg a cellacsoportok (cellák) között:

$$Y_k - y_k = \sum_{l=1}^{L} W_{kl},$$

ahol

$$W_{kl} = s_{kl}(Y_k - y_k).$$

Az egyes cellacsoportok (cellák) adatát a

$$Y_{kl} = y_{kl} + W_{kl}$$

képlet segítségével határozzuk meg.

A teljes körűen megfigyelt cellák esetén (az imputálás után)

$$n_j = N_j, \qquad x_j = X_j \qquad \text{és} \qquad q_j = 1,$$

ezért ezekre a cellákra a sokasági értékösszeg

$$Y_j = \sum_{i=1}^{n_j} y_{ji} = \sum_{i=1}^{N_j} y_{ji}.$$

Cellákra együttesen a Y sokasági értékösszeget az egyes cellabecslések összegével becsüljük:

$$Y = \sum_j Y_j.$$

A sokasági értékösszegből becsüljük a sokasági átlagot mind az egyes cellákra, mind a cellákra együttesen:

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

(A reprezentatívan megfigyelt cellákra a sokasági átlag általában nem egyezik meg az $\overline{y_j} = \sum_{i=1}^{n_j} y_{ji} \Big/ n_j$ mintaátlaggal.)

6. A hibaszámítás

1998-ban azokra a reprezentatívan megfigyelt cellákra, amelyekre $n_j > 1$, a következőképpen jártunk el. A KISREG-ben forgalmi adattal rendelkező üzleteket tartalmazókra (ahol a teljeskörűsítést hányadosbecsléssel végeztük) meghatároztuk a

$$\sigma_j^2 = \frac{1}{n_j - 1}\left(\sum_{i=1}^{n_j} y_{ji}^2 + R_j^2 \sum_{i=1}^{n_j} x_{ji}^2 - 2R_j \sum_{i=1}^{n_j} y_{ji} x_{ji}\right)$$

mennyiséget, ahol

x_{ji} – a *j*-edik cella *i*-edik üzletének kiskereskedelmi bázis árbevétele,

$$R_j = \sum_{i=1}^{n_j} y_{ji} \Big/ \sum_{i=1}^{n_j} x_{ji} .$$

(σ_j^2 tehát a tárgyhavi adatnak és a kiskereskedelmi bázis árbevételnek a kovarianciája.)

A KISREG-ben forgalmi adattal nem rendelkező üzleteket tartalmazó cellákra kiszámítottuk a

$$\sigma_j^2 = \frac{1}{n_j - 1}\left(\sum_{i=1}^{n_j} y_{ji}^2 - n_j \overline{y_j}^2\right) = \frac{1}{n_j - 1}\left[\sum_{i=1}^{n_j} y_{ji}^2 - \frac{\left(\sum_{i=1}^{n_j} y_{ji}\right)^2}{n_j}\right]$$

(korrigált tapasztalati) szórásnégyzetet, valamint mindkét típusú cellára a

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

relatív becslési bizonytalanságot, amely a forgalmi adattal nem rendelkező üzleteket tartalmazó cellák esetén a relatív szórás ($n_j = 1$ esetén $\sigma_j = C_j = 0$). A reprezentatívan megfigyelt cellák mellett a *j*-edik cellának a mintához tartozó minden egyes üzletére vonatkozóan $N_j > n_j$ esetén a

$$\sigma_{y_{ji}} = \frac{N_j}{n_j}\sigma_j\sqrt{1-\frac{n_j}{N_j}}$$

képlet segítségével becsültük a sokasági értékösszegnek az üzletre jutó, a cellán belül közös

$$\sigma_{y_{ji}} = \sigma_{y_j}$$

szórását, más néven standard hibáját. ($N_j = n_j$ esetén $\sigma_{y_{ji}} = 0.$)

1999 óta azok közül a reprezentatívan megfigyelt rétegcsoportok közül, amelyekre $n_k > 1,$ minden rétegcsoportra meghatározzuk a

$$\sigma_k^2 = \frac{1}{n_k - 1}\left(\sum_{i=1}^{n_k} y_{ki}^2 - n_k\overline{y_k}^2\right) = \frac{1}{n_k - 1}\left[\sum_{i=1}^{n_k} y_{ki}^2 - \frac{\left(\sum_{i=1}^{n_k} y_{ki}\right)^2}{n_k}\right]$$

szórásnégyzetet és a

$$C_k = \frac{\sigma_k}{\overline{Y_k}}$$

relatív szórást. ($n_k = 1$ esetén $\sigma_k = C_k = 0.$) A reprezentatívan megfigyelt rétegcsoportok mellett a *k*-adik rétegcsoportnak a mintához tartozó minden egyes üzletére vonatkozóan $N_k > n_k$ esetén a

$$\sigma_{y_{ki}} = \frac{N_k}{n_k}\sigma_k\sqrt{1-\frac{n_k}{N_k}}$$

képlet segítségével becsüljük a sokasági értékösszegnek az üzletre jutó, a rétegcsoporton belül közös

$$\sigma_{y_{ki}} = \sigma_{y_k}$$

standard hibáját. ($N_k = n_k$ esetén $\sigma_{y_{ki}} = 0.$)

Cellákra a standard hiba a mintaelemek standard hibájának $\sqrt{n_j}$ -szerese:

$$\sigma_{Y_j} = \sqrt{n_j}\sigma_{y_j}.$$

A teljes körűen megfigyelt cellákra értelemszerűen

$$\sigma_{Y_j} = 0.$$

Cellákra együttesen a standard hiba a cellánkénti standard hibák négyzetösszegének négyzetgyöke:

$$\sigma_Y = \sqrt{\sum_j \sigma_{Y_j}^2}.$$

Az értékösszeg (pont)becslése köré konfidenciaintervallumot jelölünk ki. Nevezetesen a szokásos megbízhatósági követelménynek megfelelően (amikor is a valószínűségi szint 0,95) meghatározzuk azt a

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

abszolút hibahatárt, 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 meghatározzuk az értékösszeg

$$v_j = \frac{\Delta_j}{Y_j}, \quad \text{illetve} \quad v = \frac{\Delta}{Y}$$

relatív hibahatárát (amely egyben a relatív konfidenciaintervallum sugara). Az egyes teljes körűen megfigyelt cellákra értelemszerűen

$$\Delta_j = v_j = 0.$$

v értéke – az egész sokaság relatív hibájának 1,96-szorosa – 2009-ben a következők szerint alakult.

A kiskereskedelmi forgalom relatív konfidenciaintervallumának sugara 2009-ben
(százalék)

Hónap	1.	2.
	becslési módszer	
Január	2,20	2,09
Február	2,35	2,26
Március	2,35	2,25
Április	2,44	2,34
Május	2,55	2,43
Június	3,04	2,93
Július	2,87	2,76
Augusztus	2,90	2,82
Szeptember	2,94	2,86
Október	2,97	2,90
November	3,06	2,99
December	2,57	2,52

Irodalom

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

ÉLTETŐ, Ö. – MIHÁLYFFY, L. [1997]: Stability of Composite Estimators: Experiments with Hungarian LFS Data. *Hungarian Statistical Review.* 75. évf. Special number 1. 36–45. old.

FOREMAN, E. K. [1991] *Survey Sampling Principles.* Marcel Dekker. New York.

SÜVEGES É. [2001]: A kiskereskedelmi statisztikai rendszer, fejlesztési irányai, kapcsolata a nemzeti számlákkal. *Gazdaság és Statisztika.* 13. (52.) évf. 6. sz. 61–67. old.

TELEGDI L. [2004]: A kisszervezetek integrált reprezentatív évközi megfigyelése a 2000-es években. *Statisztikai Szemle.* 82. évf. 8. sz. 668–690. old.

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

Summary

The author reviews the sampling methodology of the monthly survey of retail trade in Hungary, in the recent decade. The paper deals with the general characteristics of the survey, stratification, the determination of the sample size by strata and the selection of the sample. Imputation of missing data, alternative methods of estimation and investigation into the correctness of the estimation are also discussed.

Sajátértékek a statisztikában

Dr. Hajdu Ottó,
a Budapesti Corvinus Egyetem
Statisztikai Tanszékének
tanszékvezetője
E-mail: hajduotto@uni-corvinus.hu

A tanulmány a statisztikai kapcsolatok mérési skála által meghatározott típusai – variancia, korreláció, asszociáció, látencia – mérésének sokváltozós mérőszámait tekinti át az elemzendő mátrixok sajátértékeinek tükrében. Kiemelkedően fontos alkalmazási területekre koncentrál.

TÁRGYSZÓ:
Statisztikai módszertan.
Korrelációszámítás.
Mátrixelmélet.

A többváltozós statisztikai kapcsolatok mérése nevezetes mátrixok sajátértékeinek meghatározására vezet. A kapcsolat jellemzése – jellegétől függetlenül – alapvetően a szóródás egy-, illetve kétváltozós mérésén alapul. Kézenfekvő a több változót egybesűríteni, vagy a kapcsolatot minden párosításban vizsgálni. E célt szolgálja a *szóródási mátrix*, összekapcsolva a kétféle megközelítést. A kapcsolat jellegétől függően – korreláció, diszkriminancia, asszociáció – a szóródási mátrix nevezetes formákat ölt, melyek sajátértékei nyújtják a megfelelő szóródási, illetve kapcsolatvizsgálati mértékeket. A tanulmány áttekinti az egyes kapcsolatok vonatkozó szóródási mátrixait és azok sajátértékeinek statisztikai tartalmát.

Lévén a többváltozós elemzések alapvető eszköze az ún. *szinguláris érték* felbontás, kiindulásként e módszert ismertetjük. Ezt követően tárgyaljuk a *variancia* tömörítését, majd a *korreláció–diszkriminancia–asszociáció* hármas többdimenziós kiterjesztését, végül a kapcsolatok mögött húzódó *latens változók* kérdését. A sajátérték-feladat és az egyes kapcsolattípusok többváltozós módszertani alapjainak ismeretét feltételezzük.

1. Az SVD-eljárás

Statisztikai változók komponensekre bontásának alapvető módja az *Eckart–Young-féle szinguláris érték felbontás* (SVD-eljárás) mely szerint bármely valós (*n,p*) rendű **X** mátrix felírható az alábbi multiplikatív formában:[1]

$$\mathbf{X} = \mathbf{FDV}^T, \qquad /1/$$

ahol **X** a *p* változókra végzett *n* megfigyelés értékeit tartalmazza, az ugyancsak *(n,p)* rendű **F** oszlopai az **X** bal oldali, a (*p,p*) rendű **V** mátrix oszlopai pedig az **X** jobb oldali szinguláris vektorait adják. A $\mathbf{D} = \langle \mu_1, \mu_2, ..., \mu_p \rangle$ diagonális mátrix diagonális elemei **X** (megfelelő) ún. szinguláris értékei. Másképpen fogalmazva **V** oszlopai a *p* dimenziós tér főtengelyeinek a bázisát, **F** oszlopai pedig a főtengelyekre vonatkozó koordinátákat jelentik.

[1] Singular Value Decomposition. A képletben szereplő „*T*" felső index transzponálást jelent.

Részletesebben felírva a modellt:

$$\mathbf{X} = \begin{bmatrix} \mathbf{f}_1 & \mathbf{f}_2 & \dots & \mathbf{f}_p \end{bmatrix} \begin{bmatrix} \mu_1 & & & \\ & \mu_2 & & \\ & & \ddots & \\ & & & \mu_p \end{bmatrix} \begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \dots & \mathbf{v}_p \end{bmatrix}^T.$$

Az SVD-feladat az $\mathbf{F}^T\mathbf{F} = \mathbf{I}$ és a $\mathbf{V}^T\mathbf{V}=\mathbf{I}$ ortonormáltsági feltételek mellett (ahol **I** a megfelelő rendű egységmátrixot jelöli) a (*p,p*) rendű $\Sigma = \mathbf{X}^T\mathbf{X}$ szóródási mátrix spektrális felbontásával oldandó meg, mivel a szóródási mátrix az SVD-szabály alkalmazásával az

$$\mathbf{X}^T\mathbf{X} = \mathbf{V}\mathbf{D}^2\mathbf{V}^T$$

sajátérték-sajátvektor feladatra vezet. Ekkor a szóródási mátrix $\mu_1^2 \geq \mu_2^2 \geq \dots \geq \mu_p^2$ sajátértékei a négyzetes szinguláris értékeket adják, miközben **V** oszlopai a megfelelő sajátvektorok. A szóródási mátrix főátló elemei a változónkénti szóródás, összegük pedig a totális szóródás mértéke. A saját értékek összege a spektrális felbontásból következően a totális szóródási mértékkel azonos:[2] $tr\left(\mathbf{X}^T\mathbf{X}\right) = tr\left(\mathbf{D}^2\right)$. Ezen összegen belül a rendre csökkenő sajátértékek feltételesen maximáltak. A szóródási mátrix pozitív (szemi-)definit, tehát minden sajátértéke nemnegatív, de empirikus adatokon alkalmazva gyakorlatilag szigorúan pozitív definit.

2. A variancia tömörítése

Közvetlenül megfigyelhető, *manifest* jellegű x_j (j=1,2,...,p) változók helyettesítését, illetve tömörítését főkomponensek szolgálják, melyek magukból a változókból képzett k_t (t=1,2,...,p) lineáris kombinációk, páronként korrelálatlan rendszert alkotva, és a manifest változókat maradék nélkül reprodukálják:

$$k_t = v_{1t}x_1 + v_{2t}x_2 + \dots + v_{jt}x_j + \dots + v_{pt}x_p, \qquad /2/$$

ahol

$$x_j = v_{j1}k_1 + v_{j2}k_2 + \dots + v_{jt}k_t + \dots + v_{jp}k_p. \qquad /3/$$

[2] tr(.) a mátrix nyomát jelenti, mely a főátló elemek összege.

A súlyok dupla alsó indexében az első (*j*) index az *x* változóra, a második (*t*) pedig a *k* főkomponensre utal. A v_{jt} súlyokat a **V** mátrixba foglalva, annak *t*. oszlopa az *x* változók súlyozására szolgál a k_t főkomponens számítása érdekében, *j*. sora pedig a *k* főkomponensek súlyozására az x_j változó kalkulálása céljából.

A feladat a manifest változók olyan *k* lineáris kombinációit megadni, melyek az *x* változók totális szóródásához rendre maximált hányadban járulnak hozzá.

A megoldás az SVD-**F** főkomponensek meghatározásával kezdődően:

$$\mathbf{F} = \mathbf{XVD}^{-1}, \qquad /4/$$

melyből átskálázással

$$\mathbf{K} = \mathbf{FD}. \qquad /5/$$

A skálázott *k* főkomponensek szóródási mátrixa:

$$\mathbf{K}^T\mathbf{K} = \mathbf{D}^T\left(\underbrace{\mathbf{F}^T\mathbf{F}}_{\mathbf{I}}\right)\mathbf{D} = \mathbf{D}^2. \qquad /6/$$

Lévén a változók szóródását a szóródási mátrix főátló elemei mérik, valamely főkomponens szóródásának mértékét a manifest változók szóródási mátrixának megfelelő sajátértékei adják.

Ekkor, ha az **X** változók szóródási mátrixa:

1. a **C** kovarianciamátrix, a főkomponens varianciája a kovarianciamátrix megfelelő sajátértéke:

$$Var\left(k_t\right) = \mu_t^{2(\mathbf{C})} \qquad (t = 1,2,...,p), \qquad /7/$$

2. az **R** korrelációs mátrix, a főkomponens varianciája a korrelációs mátrix megfelelő sajátértéke:

$$Var\left(k_t\right) = \mu_t^{2(\mathbf{R})} \qquad (t = 1,2,...,p). \qquad /8/$$

Ha a főkomponenseket az SVD-modellben transzformáljuk (rotáljuk) a (*p*,*p*) rendű **T** transzformációs mátrix alapján ($\mathbf{TT}^{-1}=\mathbf{I}$) akkor elfordulnak a főkomponensek a $\mathbf{K}^*=\mathbf{KT}=\mathbf{FDT}$ módon, és így a szóródási mátrix:

$$\mathbf{K}^{*T}\mathbf{K}^* = \mathbf{T}^T\mathbf{D}^T\left(\underbrace{\mathbf{F}^T\mathbf{F}}_{\mathbf{I}}\right)\mathbf{DT} \neq \mathbf{D}^2, \qquad /9/$$

tehát a manifest szóródási mátrix sajátértékei többé nem varianciatartalmúak.

3. Kategóriák diszkriminálása

A szóródás mérésének egyik feladata a g=1,2,...,m számú csoportokra bontott sokaság szóródásának többdimenziós mérése, tekintettel a csoporttagságokra is. Ekkor a szóródás kétféle hatás eredője: a csoportközi különbségeket jellemző külső és a csoporton belüli eltérésekben jelentkező belső szóródásé.

Célunk elhatárolni a totális szóródásban a külső és a belső faktoroknak tulajdonított hányadot. A megoldás alapja a kovariancia (mátrix) csoportközi felbontása:

$$\mathbf{C} = \mathbf{C}_K + \mathbf{C}_B, \qquad /10/$$

ahol $\mathbf{C}_K$ a csoportátlagokkal helyettesített sokaság kovarianciamátrixa, $\mathbf{C}_B$ pedig a súlyozott, átlagos csoporton belüli kovarianciamátrix.

A csoporton belüli homogenitás, illetve a csoportközi heterogenitás jellemzésére a Wilks-féle lambda mutatót használjuk, mely a belső általánosított varianciának a teljes általánosított varianciához való arányát fejezi ki:[3]

$$\Lambda = \frac{\det(\mathbf{C}_B)}{\det(\mathbf{C})}. \qquad /11/$$

Minél alacsonyabb ez a hányad, annál homogénebbek a csoportok, és annál inkább a csoportközi szóródás dominál a sokaság totális szóródásában.

A varianciahányados jellegű Wilks-lambda egyváltozós esetben a belső és a teljes variancia hányadosává egyszerűsödik. Többváltozós esetben kézenfekvő a külső és belső szóródás vizsgálatát visszavezetni egyváltozós esetre, a megfigyelt változók

$$z = b_1 x_1 + b_2 x_2 + ... + b_p x_p$$

lineáris kombinációját, a diszkriminanciaváltozót képezve, alkalmasan megválasztott b súlyok alkalmazásával. Ennek belső és külső varianciája:

$$Var(z) = Var_B(z) + Var_K(z),$$

mely kvadratikus formában (a b súlyokat a $\mathbf{b}$ vektorba foglalva):

$$Var(z) = \mathbf{b}^T \mathbf{C} \mathbf{b} = \mathbf{b}^T \left(\mathbf{C}_B + \mathbf{C}_K \right) \mathbf{b} = \mathbf{b}^T \mathbf{C}_B \mathbf{b} + \mathbf{b}^T \mathbf{C}_K \mathbf{b}. \qquad /12/$$

[3] A p-dimenziós tér általánosított varianciája a tér kovarianciamátrixának a determinánsa.

A diszkriminanciaváltozó egyváltozós Wilks-lambdája, illetve komplementere egységnyi belső varianciához normálva:

$$1-\Lambda(z)=\frac{Var_K(z)}{Var_B(z)+Var_K(z)}=\frac{Var_K(z)/Var_B(z)}{1+Var_K(z)/Var_B(z)}=\frac{\varphi}{1+\varphi}. \qquad /13/$$

Most a külső varianciát a belső varianciához viszonyító, értelemszerűen maximálandó diszkriminanciakritérium:

$$\varphi=\frac{Var_K(z)}{Var_B(z)}=\frac{\mathbf{b}^T\mathbf{C}_K\mathbf{b}}{\mathbf{b}^T\mathbf{C}_B\mathbf{b}}\rightarrow \max. \qquad /14/$$

A φ diszkriminanciakritérium **b** szerinti maximálása a

$$\frac{\partial\varphi}{\partial\mathbf{b}}=\frac{2\mathbf{C}_K\mathbf{b}\left(\mathbf{b}^T\mathbf{C}_B\mathbf{b}\right)-\left(\mathbf{b}^T\mathbf{C}_K\mathbf{b}\right)2\mathbf{C}_B\mathbf{b}}{\left(\mathbf{b}^T\mathbf{C}_B\mathbf{b}\right)^2}=\mathbf{0}$$

egyenlet megoldását igényli, mely a $\mathbf{b}^T\mathbf{C}_B\mathbf{b}$ skalárral való egyszerűsítés és keresztbeszorzás, majd φ /14/ definíciójának behelyettesítése után megfelelő átrendezéssel a

$$\left(\mathbf{C}_B^{-1}\mathbf{C}_K-\varphi\ \mathbf{I}\right)\mathbf{b}=\mathbf{0} \qquad /15/$$

sajátérték-sajátvektor feladatra vezet. Ez a

$$\left(\mathbf{C}_K-\varphi(\mathbf{C}-\mathbf{C}_K)\right)\mathbf{b}=\left((1+\varphi)\mathbf{C}_K-\varphi\mathbf{C}\right)\mathbf{b}=\mathbf{0}$$

átalakítással a

$$\left(\mathbf{C}^{-1}\mathbf{C}_K-\frac{\varphi}{1+\varphi}\mathbf{I}\right)\mathbf{b}=\mathbf{0}$$

sajátérték-sajátvektor feladat formában is megoldható. A súlyokat tartalmazó **b** sajátvektor mindkét feladatra közös.

A $\mathbf{C}^{-1}\mathbf{C}_K$ mátrixnak $min\{p,(m-1)\}=k$ számú pozitív sajátértéke van, melyek statisztikai tartalmuk szerint rendre egyváltozós Wilks-lambdák.

A $\mathbf{C}_B^{-1}\mathbf{C}_K$ nem szimmetrikus mátrix sajátértékei pedig statisztikai tartalmuk szerint rendre maximált diszkriminanciakritériumok.

Végül a több- és az egyváltozós Wilks-lambdák közötti kapcsolat:

$$\Lambda = \det(\mathbf{C}^{-1})\det(\mathbf{C}_B) = \det(\mathbf{C}^{-1}\mathbf{C}_B) = \det\left(\mathbf{C}^{-1}(\mathbf{C}-\mathbf{C}_K)\right) = \det(\mathbf{I}-\mathbf{C}^{-1}\mathbf{C}_K) = \quad /16/$$

$$= \prod_{j=1}^{k}\left(1-\frac{\varphi_j}{1+\varphi_j}\right) = \prod_{j=1}^{k}\left(\frac{1}{1+\varphi_j}\right). \quad /17/$$

4. Kanonikus korrelációk számítása

Többváltozós esetben a kétváltozós korreláció mérése kiterjeszthető két változócsoport közötti korreláció vizsgálatára, ha mindkét változócsoportot egy-egy lineáris kombinációval helyettesítjük. Tekintsük a standardizált változók $x_1,x_2,...,x_p$ magyarázó, és a velük oksági kapcsolatban lévő, eredmény jellegű, ugyancsak standardizált változók $y_1,y_2,...,y_q$ ($q \le p$) csoportját.

Képezzük az x magyarázóváltozók lineáris kombinációjaként az u, és az y eredményváltozók csoportjából a z lineáris kombinációk t=1,2,...,q párosait:

$$u_t = v_{1t}x_1 + v_{2t}x_2 + ... + v_{pt}x_p$$

$$z_t = w_{1t}y_1 + w_{2t}y_2 + ... + w_{qt}y_q\,,$$

ahol valamennyi változó standardizált, és $q \le p$. A v és w súlyokat úgy határozzuk meg, hogy az u_t és z_t kanonikus változók közötti lineáris korreláció maximált legyen, miközben a kanonikus változók bármilyen más párosításban korrelálatlanok. E követelményeket fogalmazza meg a *kanonikus változók korrelációs mátrixa* az alábbi partícionált formában:

$$\mathbf{R}_{uz} = \begin{array}{c|cccc|cccc} & u_1 & \cdots & u_q & & z_1 & \cdots & z_q \\ \hline u_1 & 1 & & 0 & & r_1 & & 0 \\ \vdots & & & & & & & \\ u_q & 0 & & 1 & & 0 & & r_q \\ \hline z_1 & r_1 & & 0 & & 1 & & 0 \\ \vdots & & & & & & & \\ z_q & 0 & & r_q & & 0 & & 1 \end{array}.$$

E korrelálatlansági feltételek mellett maximált $Cov(u_t,z_t)=r_t$ lineáris korrelációt a t. kanonikus korrelációnak, az (u_t,z_t) változópárost pedig a t. kanonikus változópárnak nevezzük.

A kanonikus korrelációk meghatározása érdekében particionáljuk a manifest változók ($q+p,q+p$) rendű korrelációs mátrixát az alábbiak szerint:

$$\mathbf{R} = \left[\begin{array}{c|c} \mathbf{R}_{yy} & \mathbf{R}_{yx} \\ \hline \mathbf{R}_{xy} & \mathbf{R}_{xx} \end{array}\right],$$

ahol az egyes mátrixok méretét az indexben szereplő változók számossága adja: például $\mathbf{R}_{yx}$ (q,p) rendű, vagyis nem négyzetes. Feladatunk az

$$r_{u,z} = r = \mathbf{v}^T\mathbf{R}_{xy}\mathbf{w} \rightarrow \max$$

korreláció maximálása a **v** és **w** súlyvektorok tekintetében, a

$$Var(u) = \mathbf{v}^T\mathbf{R}_{xx}\mathbf{v} = 1, \qquad Var(z) = \mathbf{w}^T\mathbf{R}_{yy}\mathbf{w} = 1$$

standardizáltsági megszorítások mellett. A Lagrange-féle multiplikátor-módszert alkalmazva, a keresett kanonikus korrelációt és a megfelelő súlyokat az

$$\mathbf{R}_{xy}\mathbf{w} = r\mathbf{R}_{xx}\mathbf{v}, \qquad \mathbf{R}_{yx}\mathbf{v} = r\mathbf{R}_{yy}\mathbf{w} \qquad /18/$$

egyenletrendszer megoldása szolgáltatja. Az első egyenletből kifejezve a **v** vektort, majd ezt a második egyenletbe helyettesítve, és végül az utóbbit átrendezve, az

$$\left(\mathbf{R}_{yy}^{-1}\mathbf{R}_{yx}\mathbf{R}_{xx}^{-1}\mathbf{R}_{xy} - r^2\mathbf{I}\right)\ \mathbf{w} = \mathbf{0}$$

sajátérték-sajátvektor feladatra jutunk, ahol a (q,q) rendű $\mathbf{R}_{yy}^{-1}\mathbf{R}_{yy}\mathbf{R}_{xx}^{-1}\mathbf{R}_{xy}$ mátrix sajátértékei a kanonikus korrelációk négyzeteit, a megfelelő sajátvektorok pedig az y (szűkebb körű) változókhoz tartozó súlyrendszereket nyújtják. A **w** súlyok ismeretében /18/ bármely egyenletéből a **v** súlyok is következnek.

5. Korrespondenciák feltárása

Jellegét tekintve az asszociáció a kategóriaskálán mért változók kimenetei közötti kapcsolat. Exploratív elemzési eszközeinek általános kerete a korrespondencia-

analízis (CA), mely a nagyméretű kontingenciatábla adatait hivatott áttekinthetővé tenni. Mivel itt a kapcsolatrendszer struktúrája szempontjából az egyes kategóriák előfordulásának nem az abszolút, hanem a relatív gyakorisága érdekes, a CA induló adatállományát – valamennyi empirikus f_{ij} gyakoriságot a gyakoriságok n összegével (a megfigyelések számával) osztva – a kontingenciatábla normált változata, az ún. korrespondenciamátrix alkotja. Ennek általános eleme $p_{ij} = f_{ij} / n$, az i sor és a j oszlop együttes bekövetkezésének relatív gyakorisága.

1. táblázat

Korrespondenciatábla

Kategória	Oszlop					*Sorösszesen*
	1.	...	*j.*	...	*J.*	
Sor *1.*	p_{11}		p_{1j}		p_{1J}	s_1
Sor *i.*	p_{i1}		$p_{ij} = f_{ij} / n$		p_{iJ}	s_i
Sor *I.*	p_{I1}		p_{Ij}		p_{IJ}	s_I
Oszlopösszesen	o_1		o_j		o_J	1

A sorok s_i és az oszlopok o_j összesen adatai peremgyakoriságként értelmezendők. A tábla sorainak, illetve oszlopainak belső szerkezeteit összehasonlítva a peremmel hozzuk egymással kapcsolatba azon (*i,j*) kategóriapárosításokat, melyek a sorok és az oszlopok szóródásához, illetve a közöttük lévő asszociációhoz a leginkább hozzájárulnak. Az egymást vonzó, illetve taszító (*i,j*) kategóriapárosítást a peremszerkezet alapján vártnál kiugróan magasabb vagy alacsonyabb p_{ij} gyakoriság jelzi.[4]

Matematikailag a korrespondenciaanalízis az asszociáció Pearson-féle χ^2 mértékét bontja komponensekre hasonló módon, mint azt a főkomponens-analízis a varianciával teszi. Az eljárás a sorokat (oszlopokat) a megoszlásaikból képzett, redukált dimenziójú, mesterséges térbe helyezi. Itt a tengelyeket úgy definiáljuk, hogy rendre csökkenő százalékos mértékben (sorrendben) járuljanak hozzá a χ^2 statisztikához.

A korrespondenciatábla kategóriái közötti asszociáció mértékét jellemző, egységnyi megfigyelésre jutó Pearson-féle χ^2 érték definíció szerint:[5]

[4] Az 1. táblázat „összesen" sorában és oszlopában foglalt relatív peremgyakoriságok szerkezete alapján várható gyakoriság: $p^*_{ij} = s_i \cdot o_j$.

[5] E tanulmányban *Pearson*-χ^2 alatt mindig az egységnyi megfigyelésre normált χ^2 értéket értjük.

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(p_{ij} - s_i o_j)^2}{s_i o_j} = \sum_{i=1}^{I}\sum_{j=1}^{J} g_{ij}^2,$$

ahol $s_i o_j$ az (i,j) cellának a peremmegoszlások alapján várt relatív gyakorisága az asszociáció teljes hiánya esetén. Ebből következően a

$$g_{ij} = \frac{p_{ij} - s_i o_j}{\sqrt{s_i o_j}}$$

standardizált korrespondenciagyakoriság zéró értéke az asszociáció hiányát, pozitív értéke pozitív, negatív értéke pedig negatív asszociációt jelez az i sor és a j oszlop között. Pozitív asszociáció esetén az i és j kategóriák gyakran következnek be együtt, vagyis vonzzák egymást, negatív asszociáció esetén pedig ritkán járnak közösen, tehát taszítják egymást. Az előzők alapján g_{ij}^2 az (i,j) cellának, $\Sigma_j g_{ij}^2$ az i sornak, $\Sigma_i g_{ij}^2$ pedig a j oszlopnak a hozzájárulását adja a χ^2 mértékhez.

Az oszlop- és sorprofilok ábrázolása nemcsak két, hanem kettőnél több szempont (változó) szerint kategorizáló táblák esetén is lehetséges. Az i sor és a j oszlop közötti kapcsolat vizsgálatát egyszerű korrespondenciaanalízisnek nevezzük. Ebből a szempontból érdektelen, hogy adott sor (oszlop) esetleg több változó kategóriáinak valamely együttes kombinációját definiálja. Többszörös korrespondenciaanalízist végzünk viszont akkor, ha a vizsgált változók számát kettőnél többre bővítve, az asszociáció vizsgálatát az előforduló kategóriák valamennyi párosítására kiterjesztjük.

5.1. Egyszerű korrespondenciaanalízis

Az egyszerű korrespondenciaanalízis a gyakorisági tábla sorait egy pontfelhő pontjaiként tekinti az oszlopok terében, oszlopait pedig egy másik pontfelhő pontjaiként a sorok terében. A pontfelhőket egy redukált, alacsony dimenziójú térben ábrázoljuk, és a pontok helyzetéből következtetünk arra, hogy a vizsgált változók mely kategóriái vonzzák, illetve taszítják egymást. A redukált tér dimenziója $K \leq \min\{I-1, J-1\}$, a sorok CA-koordinátáit az **X**, az oszlopokét pedig az **Y** mátrixok tartalmazzák.

Az asszociáció feltárása érdekében vegyük a sorok (majd az oszlopok) origóperemhez centrált szerkezeteit – profiljait –, melyeket általános jelölésekkel a 2. és 3. táblázatokba foglaltunk, ahol s_{ij} a j oszlop centrált részesedése az i sorban, míg o_{ij} az i sor centrált részesedése a j oszlopban.

2. táblázat

Centrált sorprofilok és helyettesítő korrespondenciakoordinátáik

Sorprofil	Centrált profil: **S** mátrix					Sor CA-koordináta: **X**				
1.	s_{11}	...	s_{1j}	...	s_{1J}	x_{11}	...	x_{1k}	...	x_{1K}
i.	s_{i1}		s_{ij}		s_{iJ}	x_{i1}		x_{ik}		x_{iK}
I.	s_{I1}		s_{Ij}		s_{IJ}	x_{I1}		x_{Ik}		x_{IK}
Centroid*	0		0		0	0		0		0

* A sorok az origó körül szóródnak.
Megjegyzés. $s_{ij} = p_{ij} / s_i - o_j$.

3. táblázat

Centrált oszlopprofilok és helyettesítő korrespondenciakoordinátáik

Oszlopprofil	Centrált profil: **O** mátrix					Oszlop CA-koordináta: **Y**				
1.	o_{1I}	...	o_{1i}	...	o_{1I}	y_{1I}	...	y_{1k}	...	y_{1K}
j.	o_{j1}		o_{ji}		o_{jI}	y_{j1}		y_{jk}		y_{jK}
J.	o_{J1}		o_{Ji}		o_{JI}	y_{J1}		y_{Jk}		y_{JK}
Centroid*	0		0		0	0		0		0

* Az oszlopok az origó körül szóródnak.
Megjegyzés. $o_{ji} = p_{ij} / o_j - s_i$.

A CA-koordináták súlyozott centroidja az origó:

$$\sum_{i=1}^{I} s_i x_{ik} = 0, \quad \sum_{j=1}^{J} o_j y_{jk} = 0.$$

Most a χ^2 mérőszám az előző jelölésekkel a következő formában is megfogalmazható:

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{s_i}{o_j}(s_{ij})^2 = \sum_{j=1}^{J}\sum_{i=1}^{I} \frac{o_j}{s_i}(o_{ij})^2 = INR. \qquad /19/$$

Ebben a formában a χ^2 mutatót *inerciamértéknek* nevezzük, mely láthatóan *a pontfelhő súlyozott, többdimenziós varianciája* egyidejűleg mind a sorok, mind az oszlopok azonos mértékű szóródását jellemezve saját peremeik körül. A centrált CA-koordinátákat (**X,Y**) úgy definiáljuk, hogy adott pontnak a saját centroidtól vett távolsága, és így a teljes inercia értéke változatlan maradjon:

$$INR = \sum_{i=1}^{I} s_i \sum_{k=1}^{K} x_{ik}^2 = \sum_{j=1}^{J} o_j \sum_{k=1}^{K} y_{jk}^2. \qquad /20/$$

A CA-koordináták meghatározása érdekében definiáljuk a $\mathbf{D}_s=<s_1,...,s_I>$, $\mathbf{D}_o=<o_1,...,o_J>$, $\mathbf{D}_\mu=<\mu_1,...,\mu_K>$ diagonális mátrixokat és a g_{ij} standardizált korrespondenciagyakoriságokat tartalmazó $\mathbf{G}_{(I,J)}$ mátrixot. Ekkor a **G** mátrix SVD-felbontása az alapja a teljes inercia CA-tengelyek közötti szétosztásának:

$$\mathbf{G} = \mathbf{D}_s^{1/2}\mathbf{S}\mathbf{D}_o^{-1/2} = \mathbf{D}_s^{-1/2}\mathbf{O}\mathbf{D}_o^{1/2} = \mathbf{U}\mathbf{D}_\mu\mathbf{V}^T. \qquad /21/$$

Az **U** mátrix oszlopai adják **G** oszlopfelhőjének főtengelyeit, míg a **V** oszlopai **G** sorfelhőjének főtengelyeit. A keresett **X** és **Y** CA-koordináták a főtengelyekre vonatkozó megfelelő főkoordinátákból származnak.

Látható, hogy a $\mu_1,\mu_2,...,\mu_K$ szinguláris értékek négyzetei a $\mathbf{G}^T\mathbf{G}$ és a $\mathbf{G}\mathbf{G}^T$ szóródási mátrixok közös sajátértékei, és egyben a CA-tengelyek maximált varianciái. Ekkor a teljes inercia:

$$INR = \mathrm{tr}(\mathbf{G}^T\mathbf{G}) = \mathrm{tr}(\mathbf{G}\mathbf{G}^T) = \sum_{k=1}^{K} \mu_k^2. \qquad /22/$$

5.2. Többszörös korrespondenciaanalízis

Kettőnél több kategóriaváltozót elemezve, célszerű a korrespondenciaanalízis többszörös változatát alkalmazni. Ez ekvivalens az indikátormátrix egyszerű analízisével. A $\mathbf{Z}_{(n,J)}$ indikátormátrix sorait az i=1,2,...,n megfigyelések, míg oszlopait a Q számú Z_q (q=1,2,...,Q) kategóriaváltozók kategóriái képezik, ahol a Z_q változónak J_q számú lehetséges kategóriája van. Így a mátrix oszlopainak száma $J=J_1+J_2+...+J_Q$, és az oszlopok a Q számú csoport valamelyikének a tagjai. Az indikátormátrix mindegyik sora Q számú „1” elemet tartalmaz attól függően, hogy az illető megfigyelés adott változó melyik kategóriájához tartozik. Egyébként a mátrix elemei zérók.

4. táblázat

Indikátormátrix

Megfigyelés	A **Z** indikátor mátrix oszlopai *(j=1,2,…,J)*														*Összesen*
	Z_1 kategóriái: $\mathbf{Z}_1$				...	Z_q kategóriái: $\mathbf{Z}_q$				...	Z_Q kategóriái: $\mathbf{Z}_Q$				
	1	2	…	J_1	…	1	2	…	J_q	…	1	2	…	J_Q	
1	1						1							1	Q
2		1					1					1			Q
⋮															
i				1		1						1			Q
⋮															
n		1							1		1				Q
Összesen (f_j)	f_1^1	f_2^1	…	$f_{J_1}^1$	…	f_1^q	f_2^q	…	$f_{J_q}^q$	…	f_1^Q	f_2^Q	…	$f_{J_Q}^Q$	nQ

A **Z** mátrix tehát nQ egyest tartalmaz, n darabot minden egyes $\mathbf{Z}_q$ almátrixban, $\mathbf{Z}_q$ bármely sorának összege 1, és **Z** bármely sorának összege Q. A többszörös CA eredményeinek értelmezése az indikátormátrix alábbi tulajdonságain alapul:

1. A $\mathbf{Z}_q$ mátrix $o_j = f_j / (nQ)$ peremprofiljainak az összege bármely q=1,2,…,Q esetén: 1/Q. Így bármely változó egyforma relatív súlyt kap, melyet szétoszt az 1,2,…,J_q kategóriái között, az f^q gyakoriságoknak megfelelően.

2. Az $O_{ij} = (1/f_j) = 1/(n \cdot Q \cdot o_j)$ oszlopmegoszlások centroidja bármely $\mathbf{Z}_q$ blokkon belül egybeesik az oszlopprofilok globális centroidjával. Adott sor relatív gyakorisága $s_i = Q/(n \cdot Q) = 1/n$ és megoszlása: 1/Q.

3. A $\mathbf{Z}_q$ változó valamennyi oszlopához tartozó teljes inercia:

$$INR(q) = \sum_{j_q=1}^{J_q} INR(j_q) = \frac{J_q}{Q} - \frac{1}{Q}.$$

4. Az oszlopok (sorok) totális inerciája:

$$INR = \sum_{q=1}^{Q} INR(q) = \frac{J}{Q} - 1.$$

5. A pozitív inerciával bíró, nem triviális dimenziók száma legfeljebb J–Q.

6. Az n számú sorprofil mindegyike $J_1,J_2,\ldots,J_Q$ számú egymástól különböző pont valamelyikével esik egybe.

7. A $\mathbf{B}_{(J,J)}=\mathbf{Z}^T\,\mathbf{Z}$ *Burt*-mátrix analízisének standardizált korrespondenciakoordinátái azonosak a $\mathbf{Z}$ indikátormátrix analízisében az oszlopok standardizált korrespondenciakoordinátáival. A *Burt*-mátrix az alábbi blokkstruktúrában is írható:

$$\mathbf{Z}^T\mathbf{Z}=\mathbf{B}=\begin{bmatrix}\mathbf{Z}_1^T\mathbf{Z}_1 & \mathbf{Z}_1^T\mathbf{Z}_2 & \cdots & \mathbf{Z}_1^T\mathbf{Z}_Q\\ \mathbf{Z}_2^T\mathbf{Z}_1 & \mathbf{Z}_2^T\mathbf{Z}_2 & & \mathbf{Z}_2^T\mathbf{Z}_Q\\ \vdots & & \ddots & \\ \mathbf{Z}_Q^T\mathbf{Z}_1 & \mathbf{Z}_Q^T\mathbf{Z}_2 & & \mathbf{Z}_Q^T\mathbf{Z}_Q\end{bmatrix}.$$

Mindegyik $\mathbf{Z}_q^T\mathbf{Z}_{q^*}$ ($q\neq q^*$) mátrix, mely $\mathbf{B}$ diagonálisán kívül esik, egyben egy kétváltozós kontingenciatábla, mely a q és q^* változók közötti asszociációt sűríti az n számú megfigyelés alapján. Ugyanakkor a $\mathbf{B}$ diagonálisán mindegyik $\mathbf{Z}_q^T\mathbf{Z}_q$ mátrix diagonális, és diagonálisán $\mathbf{Z}_q$ oszlopösszesen értékei szerepelnek.

A Burt-mátrix oszlopainak és sorainak analízise azonos CA-koordinátákat eredményez. Tehát az egyetlen különbség $\mathbf{B}$ és $\mathbf{Z}$ oszlopainak korrespondencia-analízise között a főinerciák értéke, mely érinti a főkoordináták skáláját. Ezért az indikátormátrix oszlopainak az analízise inkább tekinthető *páronkénti kétváltozós*, mint *tömörített többváltozós* elemzésnek.

A Burt-mátrix partícionált formában Q számú változó kovarianciamátrixának analógiája, ahol minden egyes $\mathbf{Z}_q^T\mathbf{Z}_{q^*}$ mátrix egy-egy kovarianciának felel meg.

6. Latens dimenziók feltevése

A latens modell szerint adott x_j manifest változó indikátorjellegű abban az értelemben, hogy értékei megfigyelésenként valamely latens – létező, de nem megfigyelhető – f_t faktorok mozgásainak megfelelően alakulnak, és az indikátort végül egy, csak hozzá tartozó *egyedi hibafaktor egészíti ki* teljessé:[6]

$$x_j=\lambda_{j1}f_1+\lambda_{j2}f_2+\ldots+\lambda_{jt}f_t+\ldots+\lambda_{jm}f_m+u_j\,. \qquad /23/$$

[6] A következőkben a mátrix zárójelben szereplő alsó indexe a mátrix rendjére utal.

Valamennyi (j=1,2,…,m) indikátor változót közös vektorba foglalva, mátrix formában írva:

$$\mathbf{x}_{(p,1)} = \Lambda_{(p,m)}\mathbf{f}_{(m,1)} + \mathbf{u}_{(p,1)}, \qquad /24/$$

ahol $\mathbf{x}=[x_1,x_2,...,x_p]^T$ tartalmazza a p indikátort, $\mathbf{f}=[f_1,f_2,...,f_m]^T$ az $m<p$ latens faktort és $\mathbf{u}=[u_1,u_2,...,u_p]^T$ a unique (egyedi) faktorokat.

A $\mathbf{\Lambda}$ súlymátrix elemei a λ_{jk} értékek. Minél magasabbak abszolút értelemben, annál fontosabb a faktor. Megfigyeléseket végezve, valamennyi indikátorra az SVD-modellel analóg, de lényegileg eltérő formula adódik:

$$\mathbf{X}_{(n,p)} = \mathbf{F}_{(n,m)}\Lambda^T{}_{(m,p)} + \mathbf{U}_{(n,p)}. \qquad /25/$$

A faktoranalízis hipotézise szerint az indikátorok körének korrelációs rendszerét mögöttes, latens változók okozati köre generálja.

A /24/ kifejezés alapján az indikátorok $\Sigma_{xx} = \mathbf{X}^T\mathbf{X}$ *szóródási* mátrixa:

$$\Sigma_{xx} = \Lambda\Sigma_{ff}\Lambda^T + \Sigma_{uu} + \Lambda\Sigma_{fu} + \Sigma_{uf}\Lambda^T, \qquad /26/$$

ahol $\Sigma_{fu} = \Sigma_{uf} = \mathbf{0}$. Korrelálatlansági megszorításokat téve az egyedi faktoroknak közös faktorokkal való kapcsolatára

$$\Sigma_{xx} = \Lambda\Sigma_{ff}\Lambda^T + \Sigma_{uu} \qquad /27/$$

adódik. Ha Σ_{uu} és Σ_{ff} *diagonálisak*, akkor a modellhez az

$$\Sigma_{xx} - \Sigma_{uu} = \Lambda\Sigma_{ff}\Lambda^T \qquad /28/$$

megoldására van szükség, mely csak akkor sajátérték-feladat, ha Σ_{ff} diagonális, és csak akkor végrehajtható, ha létezik az $\Sigma - \Sigma_{uu}$ redukált szóródási mátrix (vagy becslésének) spektrális felbontása. A megoldásra iteratív algoritmusok állnak rendelkezésre, figyelembe véve, hogy a redukált szóródási mátrix már nem pozitív definit.

Irodalom

HAJDU, O. [2002]: Category Selection and Classification Based on Correspondence Coordinates. *Hungarian Statistical Review*. 80. évf. 7. sz. 103–126. old.

HAJDU O. [2003]: *Többváltozós statisztikai számítások*. Központi Statisztikai Hivatal. Budapest.

HAJDU, O. [2004]: Diagnostics of the Error Factor Covariances. *Hungarian Statistical Review*. 82. évf. 9. sz. 68–94. old.

HUNYADI L. – VITA L. [2002]: *Statisztika közgazdászoknak*. Központi Statisztikai Hivatal. Budapest.

KERÉKGYÁRTÓ GY.-NÉ ET AL. [2008]: *Statisztikai módszerek és alkalmazásuk a gazdasági és társadalmi elemzésekben.* Aula Kiadó. Budapest.

Summary

The paper deals with the basic statistical relations – correlation, discrimination, association – in a multivariate approach with regard to the eigenvalues of the corresponding matricies to be analysed. The focus is mainly on the statistical meaning of the eigenvalues. A brief overview is presented.

A megfigyelési egységektől a makrogazdasági aggregátumokig – a mikroszimulációs modellezés néhány módszertani kérdése

Cserháti Ilona
PhD, az ECOSTAT GTI osztályvezetője
E-mail: ilona.cserhati@ecostat.hu

Keresztély Tibor,
az ECOSTAT GTI tudományos munkatársa
E-mail: tibor.keresztely@ecostat.hu

A gazdaságpolitika irányítói, a társadalmi-szociális kérdésekkel foglalkozó kutatók, sőt valójában valamennyi gazdasági szereplő számára fontos kérdés, hogy miként alakul bizonyos társadalmi csoportok jövedelmi helyzete, illetve ez hogyan változik az egyes kormányzati intézkedések hatására. E kérdések mikroszimulációs modellel válaszolhatók meg, mellyel lehetővé válik az elemző által kiválasztott változók – például régiók, életkor vagy családtípus – szerinti bontásban meghatározni a különböző jövedelmi tételeket. A számításokhoz alkalmazott modell a KSH által készített háztartások költségvetési felvételét (HKF) használja primer adatforrásként, a tényleges elemzéshez azonban meg kellett teremteni az ebből származó mikroadatok és az egyéb forrásokból vett makrogazdasági adatok konzisztenciáját. A szerzők cikkükben az ehhez a feladathoz kapcsolódó két legfontosabb módszertani problémát és az általuk alkalmazott megoldást mutatják be.

TÁRGYSZÓ:
Mikroszimulációs modell.
Makroökonómia.
Háztartásstatisztika.

Napjainkban egyre több országban válik alapvető elvárássá, hogy a gazdaságpolitika kialakításakor figyelembe vegyék annak jövedelmi rétegződésre gyakorolt hatásait is. Ahhoz, hogy nyomon tudjuk követni az egyes szabályozók, intézkedések réteghatásait, a szokásosan alkalmazott makromodellek mellett mindenképpen szükség van olyan módszerekre is, amelyek számszerűsíteni tudják az egyes mikroegységek valószínű reakcióit. Míg például az adózási rendszer és a jóléti juttatások szerkezetátalakításának vizsgálatakor egy makromodell csupán a teljes adóbevétel változását tudja számszerűsíteni, addig mikroszimuláció alkalmazásával az esetleges adópolitikai változások nyertesei és vesztesei, valamint a várható jövedelempolarizáció mértéke is elemezhető. Az eredmények lehetővé teszik, hogy a vizsgált sokaság különböző szegmenseit hasonlítsuk össze. A háztartások jövedelmi helyzetének változásait például érdemes különböző szempontok szerinti csoportosításban is vizsgálni. A rendelkezésre álló információs rendszertől függően célszerű az eredményeket például jövedelemdecilisenként, gyerekszám, korösszetétel, a háztartásfő aktivitása, a háztartásban keresők száma és település szerint, illetve regionális bontásban elemezni.

A mikroszimuláció egyik tipikus alkalmazási területe az adó- és a szociális rendszerbeli intézkedések jövedelmi hatásainak előrejelzése. Ennek különös jelentősége van napjainkban, mivel a Magyarországot is erősen sújtó világgazdasági válság káros hatásainak csökkentését, illetve az abból való kilábalást célzó intézkedések a társadalmi rétegződésre várhatóan jelentős hatással lesznek. A mikroszimulációs modellek további előnye, hogy segítségükkel nemcsak az azonnali változások, hanem a hosszabb távú hatások is elemezhetők. Amennyiben azonnali változásokat elemzünk, akkor a modell az egyik állapotból a mikroegységek feltételezett reakciói, válaszfüggvényei szerinti közvetlen átmenetet írja le úgy, hogy a mikroegységek sokaságának jellemzőit lényegében változatlannak feltételezi. Ilyenkor statikus modellről beszélünk, amely ugyan szigorú értelemben véve még nem tekinthető dinamikus szimulációnak, de jól alkalmazható a gazdasági környezet változásaiból adódó azonnali hatások kimutatására. A statikus modellek időhorizontja nem haladja meg az 1-2 évet. A dinamikus modellek ezzel szemben több időszakon, akár több generációt is átívelő időtartamon keresztül írják le a mikroegységek viselkedését úgy, hogy a megfigyelt sokaság időbeli változásait is figyelembe veszik. A sokaságban vagy a vizsgált mintában szereplő egyének magasabb képzettséget szereznek, belépnek a munkaerőpiacra, vagy kikerülnek onnan, házasodnak, gyerekeik lesznek stb. Nyilvánvaló, hogy bizonyos gazdaságpolitikai intézkedések esetében elkerülhetetlen a hosszú távú hatások vizsgálata. Ilyen például a nyugdíjrendszerrel kapcsolatos reformintézkedések elemzése, amelynek fenntarthatósága nagyban függ az alapsokaság időbeli változásától.

A mikroszimuláció a fejlett ipari országokban a múlt század nyolcvanas éveitől kezdve egyre fontosabb szerepet játszik a gazdaságpolitikai döntéshozatalban. Az Egyesült Államokban például ma már nem is hoznak adórendszerrel, illetve szociális ellátással kapcsolatos törvényeket előzetes mikroszimulációs elemzés nélkül (*Zaidi–Rake* [2001]). A dinamikus mikroszimulációs modellek ezen kívül alkalmasak ún. életciklus-vizsgálatokra is, amelyek az iskolai végzettség hatását mutatják az életpálya során megszerzett jövedelmekre. A módszer tehát a gazdaság- és társadalompolitikában egyaránt fontos döntéstámogató eszköz. Meg kell ugyanakkor jegyezni, hogy még egy egyszerű statikus modell kialakítása és működtetése sem könnyű feladat. A nehézségek zöme azonban nem a megfelelő reakciófüggvények kidolgozásában, hanem a számítások alapját képező megbízható és friss alapadatbázis létrehozásában, illetve annak aktualizálásában rejlik. Magyarországon a lakossági szektort érintő hatások modellezéséhez elsősorban a háztartások részletes jövedelmi és kiadási struktúráját feltáró háztartásstatisztikai adatrendszerre lehet támaszkodni. Alapvető probléma azonban, hogy ez a statisztika csak mintegy másfél éves késéssel áll rendelkezésre, tehát valamilyen, a reprezentativitást megtartó átvezetésre van szükség ahhoz, hogy leírjuk a jelenlegi aktuális állapotot.

A mikroszimulációs modellezéssel kapcsolatos módszertani kérdések tárgyalása előtt szeretnénk megemlíteni néhány, a nemzetközi és a hazai szakirodalomból ismert gyakorlati mikroszimulációs modellt, amelyek még árnyaltabban mutatják a módszer sokrétű használhatóságát.

Az elmúlt évtized második felében alakították ki az Unió statikus ún. Európai Adómegtérülési Modelljét (European Tax-Benefit Model – EUROMOD), amely az EU15 adó- és járulékfizetését szimulálta. Ezzel azt kívánták vizsgálni, hogy az egyes országok gazdaságpolitikái mennyiben járulnak hozzá a közös ügyekhez. Az EU15-ben már korábban is folytak változó színvonalú mikroszimulációs vizsgálatok, de az egységes modellel kapott eredmények közvetlenül összevethetőkké váltak. Az EUROMOD-dal elemezték a szociális és fiskális politikák szegénységre és jövedelemegyenlőtlenségre gyakorolt hatásait, és felhasználták azt a szükséges reformintézkedések költségeinek becslésére, a finanszírozási lehetőségek feltárására, illetve általában véve a reformok hatásainak kimutatására. A számítások során az Európai Közösségi Háztartási Panel és a nemzeti mintavételes felvételek adataira támaszkodtak.

Az elmúlt húsz évben az Egyesült Államokban is számos mikroszimulációs modellt fejlesztettek ki. Az egyik ilyen a Háztartási Transzferek Mikroelemzése (Micro Analysis of Transfers to Households – MATH) elnevezésű statikus modell, amellyel a tervezett társadalmi juttatások különféle csoportokra, rétegekre gyakorolt hatásait tudják számszerűsíteni és összehasonlítani. Szintén a kilencvenes években fejlesztették ki az Adó- és Gazdaságpolitikai Intézet adómodelljét (Institute on Taxation and Economic Policy Tax Model – ITEP-model), amely kifejezetten az adótörvények okozta következményeket vizsgálja különböző jövedelmi szintű adófizetői csoportok

tekintetében. Számszerűsíti az adótörvények módosítása esetén várható hozamokat, elemzi a jövedelem-, fogyasztás- és tulajdonalapú adók hatásait. A svéd FASIT-modellt (F-eloszlású analitikus statisztikai rendszer a bevételek és transzferek vizsgálatára – F-Distribution Analytical Statistical System for Incomes and Transfers) elsősorban az adó- és -visszatérítési rendszer módosítására vonatkozó elképzelések hatásvizsgálatára használják. Az Egyesült Államok gyakorlatához hasonlóan ezek a modellszámítások is beépültek a gazdaságpolitikai döntéshozatalba. A hazai modellek közül a TÁRKI Rt. által 2004-ben kifejlesztett TÁRSZIM elnevezésű modell említhető, amely az adók és a társadalmi juttatások hatásait szimulálja. A Központi Statisztikai Hivatalból (KSH), illetve az Adó- és Pénzügyi Ellenőrzési Hivatalból (APEH) származó adatokon alapuló modellel mikroszimuláció végezhető a jövedelemadó, az indirekt adók (áfa) és a központilag szabályozott pénzbeni juttatásokra vonatkozóan.

A KSH Háztartási Költségvetési Felvétele (HKF) a gazdaságelemzés egyik fontos adatforrása. A ma már rendkívül részletes, a jövedelmi és a kiadási, valamint az életkörülményekre vonatkozó információk felhasználhatók szociológiai kutatásokhoz (például az életszínvonal alakulásának elemzéséhez, jövedelemegyenlőtlenségi és rétegvizsgálatokhoz, létminimum-számításhoz, szegénységkutatáshoz) és nemzetközi összehasonlításokhoz. A felvételből származó információkat maga a KSH is többek között makrostatisztikai mutatók előállítására alkalmazza, például a GDP és a fogyasztói árindex számításához; az eredményeket pedig folyamatosan publikálja. A legfontosabb kiadvány a HKF adataiból számított táblázatokat és különböző bontású aggregátumokat tartalmazó Háztartás-statisztikai Évkönyv. A következőkben ismertetendő módszertan a HKF-adatbázis olyan átalakítását ismerteti, melynek köszönhetően az abban szereplő mikroszintű adatok – és így az azokból számított rétegenkénti átlagok – konzisztensek lesznek az egyéb forrásokból származó makroadatokkal. Az ezzel kapcsolatos munka még 2007-ben kezdődött, amikor a KSH és az ECOSTAT együttműködésében megkezdődött egy mikroszimulációs modell kidolgozása, ami során megtörtént a modell 2004-ről 2005-re történő átvezetése.[1] Az ECOSTAT Gazdaságmodellezési Műhelyében 2008-tól folytatódott a fejlesztés, jelenleg a 2007. évi állományból kiindulva készülnek jövedelembecslések a lakosság egyes rétegeire, a 2008–2009. évekre.[2]

Az HKF átalakítása során két fontos problémát kellett megoldani. Az egyik abból adódik, hogy a jövedelmeket a háztartások aluljelentik (ilyen probléma egyébként a fogyasztási oldalon is jelentkezik), amit a makroadatokkal való egybevetéssel lehet kimutatni. Amennyiben a személyijövedelemadó-bevallások adatai teljes körűen rendelkezésre állnak, ezeket az adatokat – rétegspecifikusan – célszerű innen

[1] Ennek eredményeit *Cserháti et al*. [2007] munkája tartalmazza.
[2] Lásd *Cserháti–Péter–Varga* [2009].

imputálni.[3] Természetesen vannak olyan jövedelmi adatok (például a nyugdíj) is, amelyeknél szintén jelentkezik az aluljelentés problémája, de a személyijövedelemadó-bevallásokból nem lehet rájuk vonatkozó információt szerezni. Ezekben az esetekben más külső adatforrásokat kell igénybe venni. A cikk első részében ez utóbbiak felhasználására teszünk javaslatot. A másik elvi probléma abból adódik, hogy a HKF súlyrendszere csak az adott évre vonatkozóan biztosítja a minta torzításmentességét. Ha a háztartási szintű adatokat tovább kívánjuk vezetni, ezeket a súlyokat is igazítani kell ahhoz, hogy a HKF-ből számítható adatok illeszkedjenek a makroszintűekhez. A második rész erre ad megoldási módszert.

1. Külső adatok imputálása

A HKF-adatbázis fő előnye az, hogy segítségével a jövedelmeket és fogyasztásokat tételenként, különféle lakossági csoportokra tudjuk meghatározni és időben követni. Az adatjavítás során tehát célszerű az imputálandó állományt is minél strukturáltabb módon figyelembe venni, ugyanakkor persze biztosítani kell azt, hogy a módosított HKF-ből számítható aggregált érték egyezzen a javításhoz felhasznált adatbázis aggregált adatával. Nem járható út tehát, ha a HKF minden egyes rekordját egyszerűen felszorozzuk az SZJA- és a HKF-adatok hányadosával, mert a legtöbb jövedelem megoszlása nem egyenletes az egyes háztartások között. Az adatállomány ezért az ilyen egyszerűsített megoldás alkalmazása esetében jelentősen torzulna.

Így azt a megoldást alkalmaztuk, hogy a háztartásokat a személyi jövedelemadó adatbázisából is kivehető ismérvek szerint csoportokba soroltuk, amelyekről már feltételezhető volt, hogy nagyjából homogének. Az SZJA-adatok imputálása esetében ilyen csoportképző ismérv volt az életkor, a lakhely, adott jövedelmi réteghez való tartozás, pontosabban ezek összes lehetséges kombinációja. Gyakorlati megfontolásokból az életkor alapján elég volt 5-8 csoportot kialakítani, a lakhelyet pedig régiónként – esetleg Budapestet vagy a nagyobb városokat kiemelve – megadni. A jövedelmi rétegeket (kvintiliseket vagy deciliseket) persze önmagában egyik adatbázis sem tartalmazza, azokat külön ki kellett számolni a számítógépes megvalósítás során (például SAS-környezetben erre standard eljárások állnak rendelkezésre). Az általunk alkalmazottnál részletesebb felosztás gyakorlatilag nem képzelhető el, mert a HKF mintája viszonylag kicsi (kb. 8500 mintaelemből áll), és torzítást okozhat, ha egy-egy csoportba csak néhány megfigyelés esik. (Ha például öt életkori csoportot veszünk, hét régiót és Buda-

[3] Megjegyezzük, hogy az összegyűjtött nyers háztartási adatokon maga a KSH is végez imputálásokat, de ezek a HKF torzításait és a hiányzó információkat hivatottak kiküszöbölni.

pestet, valamint a jövedelmi kvintiliseket, az már 200 csoportot határoz meg.) Az adathelyettesítés a csoportokon belül már természetesen a korábban említett módon történt, tehát a két adatbázisból számított átlagos érték hányadosát tekintettük minden rétegre, és az így kapott hányadosokkal szoroztuk át az eredeti adatokat.

A munkából származó jövedelmeken túl, szükség lehet a lakosság rendelkezésre álló jövedelméhez tartozó egyéb tételek korrekciójára is. A tulajdonosi jövedelmek közül a kamatok esetében például biztosan jelentkezik torzítás abból adódóan, hogy a HKF csak a felvett, nem pedig a képződött kamatokra kérdez rá. A munkajövedelmeken kívüli egyéb jövedelmek közül a legnagyobb tétel a nyugdíjellátás keretében szerzett jövedelmi rész, amelynél a tapasztalatok szerint szintén jellemző az aluljelentés, így itt is szükség van korrekcióra.

Az adatjavítás egyik eszköze az lehet, hogy az ismérvek által meghatározott csoportokra külső információk alapján eloszlást illesztünk. Általában vannak statisztikák az egyes jellemzők szerint, amelyekből meg lehet határozni külön-külön az empirikus peremeloszlásokat, majd ez utóbbiakból az együttes eloszlást. A nyugdíjak esetében például a peremek becsléséhez jól használhatók az Országos Nyugdíjbiztosító Főigazgatóság által publikált statisztikák (például *ONyF* [2010]). Ha az általunk kiválasztott ismérvek függetlennek tekinthetők, az együttes eloszlást a peremértékek, pontosabban a peremgyakoriságok egyszerű összeszorzásával képezhetjük, és az eredeti adatokra rábecsülendő többletet az így meghatározott együttes eloszlás alapján osztjuk szét. Annak ellenére, hogy ez a feltétel nem teljesül maradéktalanul, mivel az ismérvek nem tökéletesen függetlenek egymástól (például régiónként jelentősen különböző lehet az átlagos nyugdíjellátás), a módszer gyakorlatilag megfelelő közelítő eredményt ad. Az adatok korrekciójához egyéb, viszonylag egyszerű módszereket is lehet használni attól függően, hogy milyen külső információs forrásokat akarunk vagy tudunk igénybe venni. Ha például az illesztésnél az egy főre jutó nyugdíjak régiónkénti vagy megyénkénti alakulásáról rendelkezésünkre álló információkat vesszük figyelembe, akkor kiszámítjuk az aluljelentés mértékét, és az eredeti adatállomány megfelelő adatát felszorzással hozzáigazítjuk a külső információs forrás alapján megadott peremfeltételekhez.

1.1. A személyijövedelemadó-bevallás adatainak imputálása

A HKF-be olyan adatokat tudunk imputálni az SZJA-adatbázisból, amelyeknek megfelelő kategóriák egy az egyben megfeleltethetők az adóbevallásban is megtalálhatóknak. Ilyenek például a költségtérítés, az egyéni vállalkozásból, a mezőgazdaságból vagy a másodállásból származó jövedelmek és a végkielégítés.

A lakhely szerinti besorolás történhet régiónként. Az APEH adatai alapján megyei szintű bontás is lehetséges (az adatbázisban található kód lényegében a megyei

igazgatóságokat jelöli), így a régiók itt is azonosíthatók. Az imputáláshoz a jövedelmi rétegeket általában decilisenként célszerű meghatározni, hogy a legfelső, általában extrém értékekkel rendelkező rétegre vonatkozó számításokat külön végezzük el. Ettől csak akkor kell eltérni, ha a HKF adataiban az adott csoporthoz túlságosan kevés mintaelem tartozik. Az életkort a születési év alapján lehet meghatározni, itt a tízévenkénti bontás elegendő. Célszerű, ha az imputálás alapját (az eltérő létszámok miatt) immár az egyes jövedelemtípusok csoporton belüli átlagos értéke képezi. Először minden háztartásra (a HKF-adatállomány alapján), illetve adóbevallásra (az SZJA-adatállományból) meghatározzuk, hogy az adott háztartás melyik rétegbe tartozik a három rétegképző ismérv szerint, majd valamennyi vizsgált jövedelemre kiszámítjuk a rétegenkénti aggregált jövedelmet és létszámot. A kettő elosztásával jutunk az adott réteg átlagos jövedelméhez az egyes jövedelmi csoportokra vonatkozóan. A HKF-ből és az APEH-adatbázisból ilyen módon számolt értékek egymással történő elosztásával kapjuk az adott jövedelemcsoport adott rétegre vonatkozó korrekciós tényezőjét. Az utolsó lépésben visszatérünk a HKF-adatbázishoz, és – alapul véve a már kiszámított rétegtagságokat – minden háztartásra kiszámítjuk az egyes jövedelmeknek a háztartás rétegtagsága szerinti korrekciós tényezővel beszorzott (tehát imputált) értékét.

1.2. A tulajdonosi jövedelmek számítása

A tulajdonosi jövedelmek túlnyomó többségét a betétek után keletkezett kamatok jelentik. Itt a forrásadót közvetlenül vonják le a pénzintézetek, így az SZJA-bevallásban ez a típusú jövedelem nem jelenik meg; tehát ezt nem lehet a többihez hasonló módon onnan imputálni. A számításokhoz azonban külső adatforrásként részben a nemzeti számlák, részben pedig az MNB-statisztikák használhatók fel, ahol különféle betét- és hiteltípusonként állnak rendelkezésre az állományokra és a kamatlábakra vonatkozó adatok.

Első lépésben a háztartások tulajdonosi jövedelmére a nemzeti számlákban megadott értéket (ESA95 D.4.) bontjuk fel régiók szerint. Ezt a stADAT-rendszerben szereplő regionális GDP-értékek arányában javasoljuk elvégezni.[4] A második lépésben a regionális összesen adatokat jövedelmi decilisekre terítjük szét. Mivel az alsóbb jövedelmi rétegekben feltehetőleg nincs kamatjövedelem, itt nem az egyszerű jövedelemarányos szétosztást tartjuk célszerűnek, hanem a következő elosztási szabályt javasoljuk. Ha a regionális összesen értékeket S_i (i=1,…,7) jelöli, akkor

$$S_i = \alpha_i\, dec_{4i} + \alpha_i\, dec_{5i} + \alpha_i\, dec_{6i} + \alpha_i\, dec_{7i} + 2\,\alpha_i\, dec_{8i} + 2\,\alpha_i\, dec_{9i} + 3\,\alpha_i\, dec_{10i},$$

[4] Lásd http://www.statimpatika.hu/statimpatika01013169_nurofen_non-aqua_100mg_tabl_12x.html

ahol dec_{ji} a j. jövedelmi decilisbe (j = 4,…,10) tartozók nettó összjövedelme az i. régióban, és az egyenletből α értéke meghatározható. Ennek alkalmazásával tehát azt feltételezzük, hogy az alsó három decilisben egyáltalán nincs ilyen jellegű jövedelem, míg a 4.–7. decilisbe tartozóknál azonos a tulajdonosi jövedelmek aránya, a következő két decilisben viszont ehhez képest kétszer, a legfelsőben pedig háromszor magasabb. Ezután a megfelelő α-val szorozva kapjuk a háztartások tulajdonosi jövedelmét.

A nemzeti számlák csak 2007-ig állnak jelenleg rendelkezésre, így a 2008–2010. évi felbontandó összesen értékeket az előzőkben említett MNB-statisztikákból becsüljük.[5] A háztartások nettó pénzügyi vagyona a pénzügyi eszközök és a kötelezettségek különbségeként adódik. Ez az adat három hónapos késéssel érhető el.

A betéti és hitelkamatok alakulásáról havi bontású táblázatok állnak rendelkezésre az alábbi bontásban: látra szóló és folyószámlabetét; lekötött betétek (éven belül, illetve éven túl lekötött, ez utóbbin belül maximum 2 éves vagy azt meghaladó); repoügyletekből származó kötelezettség; folyószámlahitel; fogyasztási hitel (változó kamatozás vagy legfeljebb 1 éves kamatfixálás, 1 éven túli kamatfixálás, ezen belül 5 éven belüli vagy túli); lakáscélú hitel (változó kamatozás vagy legfeljebb 1 éves kamatfixálás, legalább 1, de legfeljebb 5 éves kamatfixálás, 5 éven túli kamatfixálás); egyéb hitel.

Az előző adatok alapján meghatározhatók a betétek és hitelek teljes volumene, melyek egyenlegét tekintjük az összes szétosztandó tulajdonosi jövedelemnek a 2007 utáni években/időszakokban. Mivel a további szétosztáshoz újabb adatok nem állnak rendelkezésre, 2008-tól a korábbi arányokat vesszük alapul. Ez gyakorlatilag azt jelenti, hogy például 2008-ra a 2008. és a 2007. évi összesen értékek hányadosát tekintjük, és ezzel szorozzuk át az egy évvel korábbi jövedelmeket.

2. Átsúlyozás

A következőkben egy olyan módszert ismertetünk, melynek segítségével a HKF-hez tartozó mintaelemsúlyok úgy módosíthatók, hogy az azokkal „makrósított” jövedelmi mutatók adott célértékekhez minél közelebb kerüljenek.[6] A probléma azért vetődik fel, mert a HKF adatai csak mintegy másfél éves késéssel állnak rendelkezésre. Ha ezeket aktualizálni akarjuk, sőt, előre is szeretnénk jelezni, a régi súlyokkal való

[5] Az adatok regionális megbontásához azonban még ez is kevés, ahhoz ugyanis a – meglehetősen nagy késéssel rendelkezésre álló – regionális GDP értékeit is fel kell használni.

[6] Az itt alkalmazott módszer alapja a KSH eljárása, ami a HKF elemeihez rendelt súlyrendszer meghatározására szolgál.

átvezetés torzított eredményt adhat, a velük való felszorzás és összegzés pedig nem nyújt megfelelő makroszintű értékeket. Az általunk kidolgozott módszer viszont lehetőséget biztosít arra, hogy a megfigyelt háztartási mutatók tetszőleges kombinációira (azaz a háztartások tetszőleges csoportjaira) megadott célértékekhez tudjuk akár negyedévenkénti gyakorisággal is a mintát igazítani. Ily módon lehetővé válik a HKF jövedelmi adatainak átvezetése, előrejelzése, és ez alapján különféle jövedelmi mutatók számítása az egyes háztartási szegmensekre.

2.1. Az eredeti súlyrendszer kialakítása

A HKF adatállományának publikálásakor a KSH minden háztartáshoz megad egy súlyt, amely az adott mintaelem által képviselt háztartások számát mutatja. A szimulált alapsokaság tehát úgy áll elő, hogy minden háztartást meg kell szorozni az adott súllyal, és a háztartások így kapott összessége reprezentálja a teljes háztartási szektort. (Megjegyezzük, hogy az alapsokaságot a Magyarországon magánháztartásban élő magyar állampolgárok adják. Nem kerülnek tehát a megfigyelésbe a határon belül lakó külföldiek, de a határon kívüli magyar állampolgárok sem. Szintén nem terjed ki a felvétel az úgynevezett intézeti háztartások, például gyermekotthonok, szociális otthonok stb. lakóira.) A HKF egyébként nemcsak a háztartások szintjére, hanem az adott háztartásban élő személyekre is tartalmaz adatokat. A közölt súlyok ún. integrált súlyok, ami azt jelenti, hogy az állomány bármely háztartásában az oda tartozó személyek súlya megegyezik a háztartás súlyával.

Magát a mintát a települések szerint alakítják ki, és ehhez meghatározzák az induló súlyokat (*Éltető–Mihályffy* [2002]). A válaszmegtagadások következtében a felvételben közreműködő háztartásokban élő személyek különféle szempont szerint vett megoszlása eltérhet az alapsokaságtól, ezért az általános statisztikai gyakorlatnak megfelelően az esetleges alul- és felülreprezentáltságot a súlyok további kalibrálásával küszöbölik ki (lásd még *Molnár* [2005]). A kalibrálás alapja a makroszintű adatokhoz való illesztés. A gyakorlatban ehhez a KSH demográfiai adatokat használ. Ilyenek a régió és a településtípus szerinti elhelyezkedésen kívül a nemek és életkorcsoportok szerinti megoszlás, az iskolai végzettség, a gazdasági aktivitás[7] és a gyerekszám szerinti megoszlás. A kalibrálásnál az ún. Darroch–Ratcliff-féle iteratív skálázás módosított változatát alkalmazzák (*Darroch–Ratcliff* [1972]). Maga az eljárás egy adott valószínűség-eloszlás meghatározására irányul lineáris feltételek mellett, amellyel egyébként biztosítható a súlyok nemnegativitása, és az is, hogy az egy háztartásban élő személyek azonos súlyt kapjanak. Az eljárás a következő matematikai programozási feladat megoldását igényli:

[7] Gazdaságilag aktív az a személy, aki dolgozik vagy munkanélküli.

$$\sum_{j=1}^{n}\left[w_j \left(\log w_j - \log w_j^0\right) - \left(w_j - w_j^0\right)\right] \to \min$$

$$\sum_{j=1}^{n} w_j q_{ij} = c_i, \quad i = 1,...,m$$

$$l \le w_j \le u, \qquad j = 1,...,n,$$

ahol n a háztartások száma a mintában, w_j^0 és w_j az eredeti és a kalibrált mintasúlyok, q_{ij} az i-edik kontrollváltozó értéke a j-edik háztartásban, c_i az i-edik kontrollváltozó értéke a teljes népességben, l és u a kalibrált súlyokra adott alsó és felső korlát. A célfüggvény az ún. információdivergencia. Ennek minimalizálása azt célozza, hogy a kalibrálás során lehetőleg minimális mértékben változzanak meg az eredeti súlyok.

2.2. Az átsúlyozás algoritmusa

A HKF-beli értékek továbbvezetésénél a korábban megadott szempontok mellett a tulajdonosi jövedelmek alakulása alapján súlyozzuk át a mintát. A jövedelmi adatokat természetesen bármilyen más csoportképző ismérvvel kombinálhatjuk, így megkaphatjuk ezeket a lakossági szféra bármely szegmensére.

Jelölje m a mintabeli háztartások számát, n pedig az átsúlyozásnál figyelembe veendő csoportképző ismérvekét. Az utóbbi – mivel az átsúlyozásnál az eredeti szempontokat továbbra is tekintetbe kívánjuk venni – magában foglalja az előző ismérvek szerint kialakult csoportokat is. A HKF kiindulási mintája alapján így felépíthető egy $m \times n$ méretű $\mathbf{X}$ mátrix, amelynek általános eleme legyen x_{ij}. A HKF-ben közölt eredeti súlyok m elemű vektorát jelölje $\mathbf{w}$, az elérni kívánt célértékeket tartalmazó n elemű kontrollvektort pedig $\mathbf{k}$. Mivel az algoritmus akkor áll le, ha előre kívánt pontossággal megközelítettük a célértékeket, vagy ha az iterációk lépésszáma elért egy megadott értéket, rögzítjük a pontosságot mérő ε pozitív konstanst, továbbá a lehetséges maximális iterációszámot, amelyet N-nel ($N>1$) jelölünk. Az iterációs algoritmus lépései a következők.

Legyen $\mathbf{w_k}=\mathbf{w}$, $\mathbf{u} = \underline{\mathbf{1}} \in \mathbf{R^m}$, ahol $\underline{\mathbf{1}}$ az összegzővektor, valamint $h=1$, és képezzünk egy $\mathbf{u_1}$ vektort, amelynek i. eleme az $\mathbf{X}$ mátrix sorösszesen értéke:

$$\mathbf{u_1} = \left(\sum_{j=1}^{n} x_{1j} \quad \cdots \quad \sum_{j=1}^{n} x_{mj} \right)^T \in \mathbf{R^m},$$

ahol T a transzponált jele. Módosítsuk ezek után a $\mathbf{w_k}$ korrigált súlyvektort, amelynek i. eleme $w_k(i) = w(i)u(i)$, $i=1,...,m$.

Ha $h=N$, az algoritmus leáll, és $\mathbf{w_k}$ az új súlyrendszer. Következő lépésként ellenőrizzük, hogy az új súlyok az eredetihez képest nem mozdultak-e el jelentős mértékben. Ha $w_k(i) < 0,5w(i)$, akkor $w_k(i) = 0,5w(i)$, ha $w_k(i) > 2w(i)$, $w_k(i) = 2w(i)$, $i=1,...,m$.

Képezzük a c vektort a következő módon:

$$\mathbf{c} = \left(\sum_{i=1}^{m} x_{i1} w_k(i) \quad \cdots \quad \sum_{i=1}^{m} x_{in} w_k(i) \right)^T \in \mathbf{R^n}.$$

Ha teljesül, hogy

$$\max_i \left| \mathbf{k}(i) - \mathbf{c}(i) \right| \leq \varepsilon,$$

akkor az algoritmus leáll, és w_k az új súlyrendszer. Ha nem teljesül, a következő lépésre kerül sor. Legyen $r(i)=k(i)/c(i)$, $i=1,...,n$. Állítsuk elő az **S** segédmátrixot az

$$\mathbf{S} = \begin{pmatrix} x_{11} r(1) & \cdots & x_{1n} r(n) \\ \vdots & \ddots & \vdots \\ x_{m1} r(1) & \cdots & x_{mn} r(n) \end{pmatrix} \in \mathbf{R^{m \times n}}$$

definíció alapján, és jelölje ennek általános elemét s_{ij}.

Legyen

$$\mathbf{u_2} = \left(\sum_{j=1}^{n} s_{j1} \quad \cdots \quad \sum_{j=1}^{n} s_{jm} \right)^T \in \mathbf{R^m},$$

azaz a vektor i-edik eleme az **S** mátrix $i.$ sorösszege.

Képezzük az új **u** vektort oly módon, hogy annak i. eleme ($i=1,...,s$) a következő legyen: $u(i)=u_2(i)/u_1(i)$, azaz itt az **u** vektort elemenként elosztjuk az eredeti mátrixunk megfelelő sorösszegeivel. Ez lesz az az új vektor, amely segítségével a súlyrendszert módosítjuk. Legyen $h=h+1$, és térjünk vissza a 2. lépésre.

Az algoritmus futtatása nem biztosítja azt, hogy a súlyvektor elemei egészek legyenek, így azt utólag kerekíteni kell.

3. A számítások eredményei

Most röviden be szeretnénk mutatni, hogy az imputálásnak köszönhetően mennyivel kerültek közelebb a HKF-ből számított makroadatok a valós értékekhez. A

táblázatban összefoglaltuk az egyes jövedelemtételekre vonatkozó számításaink részeredményeit. A vizsgált tételek megtalálhatók a nemzeti számlákban, így az ott található számokat tekintettük a tényleges makroadatoknak. A vastagított számokkal jelölt értékek nem szerepeltek a HKF-ben, így azokat a nemzeti számlákból vettük át.

A lakossági jövedelmek egyes tételeinek összevetése a nemzeti számlák adataival, 2007
(folyó áron, millió forint)

Jövedelemkategória	Nemzeti számlák	Eredeti HKF	Imputált HKF	Eltérés a nemzeti számlák és az imputált HKF között
Főállású munkajövedelem	6 774 959	5 766 482	6 740 075	34 884
Egyéb munkajövedelem	2 536 379	1 959 450	1 993 303	543 076
Bérek és keresetek (D11)	*9 311 338*	*7 725 932*	*8 733 378*	*577 960*
Munkaadók TB hozzájárulásai (D12)	2 660 995	**2 660 995**	**2 660 995**	–
Munkavállalói jövedelem (D1)	*11 972 333*	*10 386 927*	*11 394 373*	–
Működési eredmény, nettó (B2n)	626 423	**626 423**	**626 423**	–
Vegyes jövedelem, nettó (B3n)	2 252 785	**2 252 785**	**2 252 785**	–
Tulajdonosi jövedelem (D4)	719 399	10 085	720 969	–1 570
Elsődleges jövedelem (B5n)	*15 570 940*	*13 276 220*	*14 994 550*	*576 390*
Jövedelemadó (D5)	1 939 165	1 249 822	1 700 238	238 927
Nyugdíjjárulék	1 497 626	1 172 231	1 329 774	167 852
Társadalombiztosítási hozzájárulások (D61)	4 158 621	3 833 226	3 990 769	167 852
Nettó elsődleges jövedelem	9 473 154	8 193 172	9 303 543	169 611
Pénzbeni társadalmi juttatások (D62)	4 109 185	3 549 875	3 545 942	563 243
Egyéb folyó jövedelemátutalások (D7)	2 220	178 403	161 613	–159 393
Rendelkezésre álló jövedelem, nettó (B6n)	*13 584 559*	*11 921 450*	*13 011 098*	*573 461*

Látható, hogy amennyiben a rendelkezésre álló jövedelmet az eredeti HKF-ből becsüljük, mintegy 1 663 milliárd forinttal térünk el a tényleges értéktől, ami nagyjából 12 százalékos differenciát jelent. Az imputált adatok viszont érezhetően pontosabban közelítik a valós összeget, körülbelül egyharmadára csökken az eltérés. A rendelkezésre álló jövedelem tételeit megvizsgálva azt tapasztaljuk, hogy ez a hiba-

csökkenés nem egyenletes. A legfontosabb résztételnek tekinthető főállású munkajövedelem esetében csaknem a teljes különbözet eltűnik. Ez annak köszönhető, hogy ebben az esetben megfelelő mennyiségben álltak rendelkezésre olyan pótlólagos információk, amelyek segítségével végrehajthattuk az imputálást. Hasonló a szituáció a tulajdonosi jövedelmeknél, mint korábban leírtuk, ennek a tételnek a javítására külön módszert használtunk. Alig lett jobb viszont a helyzet az egyéb munkajövedelem és a pénzbeni társadalmi juttatások kapcsán, előbbinél elsősorban a mezőgazdasági jövedelmek becslése okozott gondot. A nyugdíjjárulékok és a jövedelemadók számítását az imputált adatok használata tette pontosabbá, de így is megmaradt az eredeti hiba fele-harmada. Összességében elmondható, hogy az imputálás látványos eredményeket hozott a lakossági jövedelmek becslésében, ám nem minden tételt sikerült érdemben javítani.

*

Cikkünkben a mikroszimulációs modellezés két fontos módszertani problémáját mutattuk be, és vázoltuk ezekre az általunk alkalmazott megoldást. Mindkét probléma a felhasznált adatbázissal kapcsolatos, kezelésük alapvető feltétele a mikroszimulációs módszertan megfelelő alkalmazásának. A válaszadók hibás adatközléseit imputálással lehet – részlegesen – korrigálni. Ennek során olyan külső adatforrást használtunk fel, amely ugyan nem tartalmaz egyedi adatokat, de az abban szereplő makroadatok megbízhatóbbak az eredeti adatbázisunk egyedi adataiból számítható makroaggregátumoknál. Rámutattunk arra is, hogy nem minden jövedelemtétel imputálható, és a különböző tételekhez más-más módszerekre lehet szükség. Adatbázisunk időbeli továbbvezetése vetette fel a másik problémát, miszerint egy adott évi súlyrendszer nem lesz érvényes a későbbiekben. Ezért az adatok átsúlyozására van szükség, ami egy összetett szélsőérték-feladat iteratív megoldásával valósítható meg. A cikk utolsó részében egy táblázat segítségével illusztráltuk, hogy az eredeti adatok imputálásával milyen mértékben sikerült javítani a lakosság makroszintű jövedelemtételeinek becslésén.

Irodalom

DARROCH, J. N. – RATCLIFF, D. [1972]: Generalized Iterative Scaling for Log-Linear Models. *The Annals of Mathematical Statistics*. 43. évf. 5. sz. 1470–1480. old.

CSERHÁTI I. ET AL. [2007]: A háztartások jövedelemalakulásának elemzése mikroszimulációs modellel. *A gazdaságelemzés módszerei*. II. sz. ECOSTAT – Központi Statisztikai Hivatal. Budapest.

CSERHÁTI I. – PÉTER I. – VARGA ZS. [2009]: A lakosság jövedelmi rétegződésének tendenciái 2008–2009-ben. *Fejlesztés és Finanszírozás*. 3. sz. 70–78.old.

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

ATKINSON, A. B. ET AL. [1999]: *An Introduction to EUROMOD*. EUROMOD Working Paper No. EM0/99. http://ideas.repec.org/p/ese/emodwp/em0-99.html (Elérés dátuma: 2010. június 11.)

MOLNÁR GY. [2005]: Az adatállomány és a rotációs panel. In: *Kapitány Zs.–Molnár Gy.– Virág I.*: *Háztartások a tudás- és munkapiacon*. KTI Könyvek. MTA Közgazdaságtudományi Intézet. Budapest. 141–147. old.

ONYF (ORSZÁGOS NYUGDÍJBIZTOSÍTÓ FŐIGAZGATÓSÁG) [2010]: *Nyugdíjban, nyugdíjszerű ellátásban részesülők állománystatisztikai adatai*. http://www.onyf.hu/index.php?module=news&fname=onyf_left_menu_kiadvany&root=ONYF (Elérés dátuma: 2010. június 11.)

ZAIDI, A. – RAKE, K. [2001]: *Dynamic Microsimulation Models*: *A Review and Some Lessons for SAGE*. ESRC SAGE Research Group's Discussion Paper No. 2. London School of Economics. London.

Summary

It is an important question for economic policy decision makers, for researchers working on social issues and for all actors in the economy how incomes of certain social groups change in response to government measures. This can be answered with the help of micro-simulation modelling which makes it possible to separate different types of incomes according to criteria (such as region, age or family type) chosen by analysts. For calculations, the model uses the Household Budget Survey prepared by the Hungarian Central Statistical Office as a primary source. However, for the actual analysis, consistency had to be created among micro data collected from the survey and macro data taken from other sources. The authors present the two most important methodological issues related to this task and their solution to the problem.

Nemnormális, parametrizált eloszlású valószínűségi változók*

Kotosz Balázs

PhD, a Budapesti Corvinus Egyetem adjunktusa

E-mail: balazs.kotosz@uni-corvinus.hu

Ferenci Tamás

MSc, a Budapesti Corvinus Egyetem demonstrátora

E-mail: tamas.ferenci@medstat.hu

Szimulációs vizsgálatok során gyakran szükségessé válik adott jellemzőkkel rendelkező eloszlásból származó véletlen számok generálása. Amennyiben valamilyen jellegzetes, közismert eloszlásról van szó, a szükséges műveletek könnyen elvégezhetők, illetve a megfelelő programcsomagok ezeket tartalmazzák. Ha azonban bizonyos paraméternek tekintett tulajdonságokkal, például adott értékű momentumokkal rendelkező eloszlásokra van szükségünk, komoly akadályokba ütközhetünk. A szerzők tanulmányukban bemutatnak és megvizsgálnak néhány megoldási lehetőséget (Pearson-, Johnson-eloszláscsaládok, általánosított λ-eloszlás, Burr XII, Tukey-féle „g-and-h" és Fleishman transzformációs módszer), azok alkalmazhatósági korlátaival együtt, részletesen tárgyalva az illesztéssel kapcsolatos témákat is.

TÁRGYSZÓ:
Statisztikai módszertan.
Valószínűség-eloszlás.
Momentumok.

* Itt szeretnénk köszönetet mondani a lektornak értékes észrevételeiért. Természetesen a tanulmányban előforduló esetleges hibákért kizárólag a szerzőket terheli felelősség.

Eloszlásillesztés alatt első közelítésben azt a statisztikai feladatot értjük, melynek során valamilyen empirikus adatsorhoz (mintához) olyan elméleti eloszlást keresünk, hogy az empirikus adatsor eloszlása és az elméleti eloszlás a leghasonlóbb legyen (a hasonlóság valamilyen mértéke szerint). E dolgozatban kizárólag az egyváltozós statisztika területén fogunk mozogni.

Ahhoz, hogy a feladatot végre tudjuk hajtani, két részfeladat megoldására van szükség: először a mintánkból az eloszlására vonatkozó információkat szükséges kinyerni, majd ezeket kell felhasználni az elméleti eloszlás meghatározásakor.

Ez utóbbi – figyelembe véve, hogy a gyakorlatban paraméterek által befolyásolt eloszlásokkal találkozunk – ismét csak két részfeladatot jelent: a felhasznált eloszlás megválasztását, majd, miután ezt rögzítettünk, az optimális paraméterezés megállapítását. Az első feladatot, az eloszlás mellett történő elköteleződést számos tényező befolyásolhatja (a modellező implicit ismeretei a szóba jövő eloszlások szakmai tartalmáról, előzetes várakozások stb.), ezért kevésbé algoritmizálható. (Valamilyen illeszkedésvizsgáló próbát használva azonban arra is mód van természetesen, hogy a modellező több, önmagában optimálisan paraméterezett eloszlásnak az empirikus adatokkal vett illeszkedése alapján válasszon.)

Az eloszlás kiválasztása után következő feladat az optimális paraméterek meghatározása. Mivel itt a mintából kell következtetni a sokasági paraméterre, jól láthatóan egy becslési feladatot kaptunk, amelyre számos közismert eljárás, például a népszerű maximum likelihood-elv (ML) használható.

Ha azonban történetileg visszatekintünk erre a kérdésre, azt látjuk, hogy a XX. század elején – bár az eloszlásillesztések ekkor szinte fénykorukat élték – még nem volt, legalábbis mai formájában, széles körű használatban az ML-elv. (Noha *Pearson* már a századforduló környékén megsejtette, és bizonyos értelemben használta is e módszert.) A kor statisztikusai tehát más elvre támaszkodtak, az egyik legnépszerűbb a momentumok alapján történő illesztés, az ún. momentumok módszere (MM) volt.

Ennek során meghatározták az empirikus adatsor első néhány (tipikusan négy) momentumát, majd azt tekintették optimális paraméterkombinációnak, mely ugyanilyen momentumokkal rendelkező elméleti eloszlást adott. Amíg az ML-elv minden mintaelemet közvetlenül felhasznál, addig a momentumok alapján történő illesztés 4 számra redukálja az adatbázist – ami nyilvánvalóan rontja az illeszkedést. Hatalmas előnye viszont, hogy a legtöbb eloszlás esetén az elméleti eloszlás paramétereinek függvényében felírt, és az empirikus adatokkal egyenlővé tett momentumok, mint egyenletek alkotta egyenletrendszer analitikusan megoldható volt.

Későbbiekben, az elméleti és számítástechnikai fejlődésnek köszönhetően a momentumok módszerének ilyen alkalmazása kikerült a napi gyakorlatból. Érdekes viszont, hogy az utóbbi időben, egészen más indíttatásokból, ismét előtérbe kerültek e módszerek. Ezen okok egyike[1] a Monte-Carlo-szimulációs módszerek széles körű elterjedése, melyekkel végzett bizonyos vizsgálatoknál szükségessé válik adott értékű momentumokkal rendelkező eloszlásokból származó véletlenszámok generálása. Mivel jelen dolgozatunkat egy ilyen alkalmazás inspirálta, e kérdést röviden bemutatjuk. (Részletesebben lásd például *Ferenci* [2009]-et.)

Tegyük fel, hogy egy statisztikai próba valamilyen eloszlási (tipikusan normalitási) feltevéssel él a sokaságokra vonatkozóan, melyből a mintái származnak (mint a közismert Student-féle *t*-próba), és vizsgálni kívánjuk, hogy a próba mennyire robusztus e feltevés megsértésére nézve. Ennek egyike lehetősége a Monte-Carlo-módszer, melynek során a feltevést irányítottan megsértő (adott mértékben nemnormális) sokaságból származó véletlenszámok tömegét generáljuk, és – ezeken végrehajtva a vizsgált tesztet – megfigyeljük, hogy az empirikus elsőfajú hibaarány konvergál-e a szignifikanciaszinthez. Ehhez szükséges, hogy képesek legyünk adott mértékben nemnormális sokaságból származó véletlenszámok generálására; ez tipikusan adott (nemnormális) ferdeséget/csúcsosságot jelent. Fontos megjegyezni, hogy e feladat jó minőségű megoldása azt is igényli, hogy olyan eloszlást válasszunk, melyből a lehető legtöbb ferdeség/csúcsosság értékhez generálható véletlenszám, tehát a lehető legszélesebben, legtöbb ferdeség/csúcsosság eléréséhez paraméterezhető (hiszen a megoldás során majd a ferdeség/csúcsosság síkon akarunk végigiterálni).

Vegyük észre, hogy ez a probléma eltér a momentumok módszerének alapfeladatától, hiszen itt a momentumok nem egy empirikus adatsorból számolhatók, hanem előre, a modellező által meghatározottak.[2] Ez a tény (tehát hogy az említett két feladatrész közül csak a másodikat szükséges megoldani: az elméleti eloszlást úgy megválasztani és paraméterezni, hogy momentumai meghatározott értékek legyenek) egy lényeges módosulást mégis jelent: a feladat innentől nem statisztikai értelemben vett becslés (így például a becsléselméleti tulajdonságait sem lehet vizsgálni, szemben a szó hagyományos értelmében vett momentumok módszerével, ahol ez központi kérdés). Mi e különbségtétel hangsúlyozása végett használjuk az „eloszlásillesztés” kifejezést.

Ennek vizsgálata arra motivál minket, hogy fellapozzuk a momentum módszer és az eloszlásillesztés klasszikus irodalmát, és újra áttekintsük a korábban még egészen más okból vizsgált feladatot.

[1] Egy másik fontos és aktuális téma, melyre itt csak utalni tudunk, a momentum módszer egy általánosítása, a GMM. Erről lásd például *Hall* [2005]-öt.

[2] Egy másik terület, ahol – teljesen más indíttatásból – de épp ugyanerre szükség lehet, és melyet ismét csak utalás szintjén tudunk megemlíteni, a bayes-i statisztika (*Lee* [2009]). Itt ugyanis a priorok létrehozásához használt külső információ gyakran épp momentumok (vagy épp kvantilisek) formájában áll rendelkezésre.

Annál is inkább szükség van erre, mert a legtöbb jól ismert, alapozó statisztikai kurzuson is oktatott eloszlás (például normális, t, χ^2, F, exponenciális, lognormális) nem alkalmas 4 momentum alapján történő illesztésre (sem); a szóba jövő eloszlások pedig még egyetemi szinten is újszerűek lehetnek. Ezek áttekintését kíséreljük meg most.

Ezt az alapfeladatot kiegészítjük azon kérdés vizsgálatával, hogy hogyan lehetséges egy eloszlást (momentumai helyett) kvantiliseivel illeszteni. (Bár ennek megoldására csak egy, az eloszlásoknak még az előzőnél is szűkebb köre képes.) A momentumok kapcsán eddig elmondott legtöbb megjegyzés változatlanul érvényes kvantilisek alapján történő illesztésre is.

A dolgozat első részében az eloszlásillesztés két módszerét, a momentumokon és a kvantiliseken alapuló illesztést tekintjük át, különös tekintettel a fogalmak és a jelölések egységes definiálására. A második részben a céloknak megfelelő eloszlásokat, illetve eloszláscsaládokat mutatjuk be, így rendre a Pearson-, a Burr-, a Johnson-, az általánosított λ, a g-and-h, végül a Fleishman-eloszlást. Egyes levezetések és eredmények – bonyolultságuk, hosszuk miatt – az internetes Mellékletben kaptak helyet (www.ksh.hu/statszemle).

1. Az eloszlásillesztés két módszere

Ebben a részben definiáljuk pontosan, hogy mit értünk a momentumok, illetve kvantilisek alapján történő illesztésen. A dolgozat egészében alkalmazott eloszlásfüggetlen jelölésrendszert is itt vezetjük be. Ez már csak azért is fontos, mert a forrásmunkák majdnem egy évszázadot fognak át, amely idő alatt igen jelentősen változtak bizonyos statisztikai jelölésekkel kapcsolatos szokások; így most egyúttal arra is kísérletet teszünk, hogy ezeket egységes keretben mutassuk be.

1.1. Momentumok alapján történő illesztés

Egy valószínűségi változó[3] n-edik nyers momentumának a

$$\mu'_n = \int_{-\infty}^{+\infty} x^n \cdot f(x)\,\mathrm{d}x \qquad n = 0,1,...$$

integrált nevezzük, ha az konvergens (*Kendall–Stuart* [1977]). Egy változó nulladik nyers momentuma szükségképp 0, az első nyers momentuma a várható értéke.

[3] A továbbiakban sokszor – némileg hanyagul – erre úgy is fogunk hivatkozni, mint egy „eloszlás momentumra", tudva természetesen, hogy itt precízen egy valószínűségi változó momentumáról van szó.

Az *n*-edik centrális momentumának a

$$\mu_n = \int_{-\infty}^{+\infty} \left(x - \mu_1'\right)^n \cdot f\left(x\right) \mathrm{d}x \qquad\qquad n = 0,1,...$$

integrált nevezzük, ha az konvergens. (Könnyen belátható, hogy ha egy valószínűségi változónak létezik *n*-edik nyers momentuma, akkor létezik *n*-edik centrális momentuma is.) Mint látható, a centrális momentumot a várható érték körül értelmeztük. Ez nem szükségszerű, de mi a mostani tárgyalásunkban ezt fogadjuk el definíciónak. A nulladik centrális momentuma értelemszerűen minden eloszlásnak 1, az első centrális momentum értéke 0, míg a második a szórásnégyzet.

Végül bevezethető a standardizált centrális momentum fogalma is. Mivel ennek a három, és annál magasabb momentumok esetén van igazán értelme, így a jelölés indexe is a harmadik ilyen momentumnál vesz fel 1 értéket:

$$\gamma_{n-2} = \frac{\mu_n}{\left(\mu_2^{1/2}\right)^n}.$$

Így γ_1 a ferdeség, γ_2 pedig a csúcsosság mutatója lesz.[4] Példának okáért, a normális eloszlás ferdesége e mutatókkal $\gamma_1 = 0$, csúcsossága[5] $\gamma_2 = 3$.

Mindezek alapján – a Cauchy–Bunyakovszkij–Schwarz-egyenlőtlenség felhasználásával – belátható, hogy szükségképp minden eloszlásra teljesül a

$$\gamma_2 \geq \gamma_1^2 + 1$$

összefüggés, mely a ferdeség függvényében határoz meg egy minimális csúcsosságot. (Kissé leegyszerűsítve azt mondja ki, hogy nem léteznek nagyon ferde, és mégis lapult eloszlások.) Ebből következik, hogy a ferdeség/csúcsság síkon létezik egy – parabolikus görbe által kijelölt – „lehetetlen tartomány", ahol nem létezhet eloszlás.

A ferdeség és csúcsosság specifikált értékeit rendre g_1-gyel és g_2-vel fogjuk jelölni.

[4] A mutatók jelölésének tekintetében az irodalom megosztott. Az α, a β, a γ és a δ különböző sorszámai egyaránt felbukkannak különböző írásokban, sokszor hasonló, vagy éppen négyzetes tartalommal. Az általunk választott jelölésekben a mutatókat *Pearson* nyomán – bár vele nem teljesen azonosan – definiáltuk. Ezzel kapcsolatban megjegyezzük, hogy a ferdeség/csúcsosság síkot sokszor $\gamma_1^2 - \gamma_2$ tengelyeken ábrázolták, ráadásul a függőleges tengelyt fejjel lefelé fordítva. Mi konzisztensen a szokott állású $\gamma_1 - \gamma_2$ síkot fogjuk használni.

[5] Pontosan ez utóbbi az oka annak, hogy több helyen az általunk definiált mutató 3-mal csökkentett értékét nevezik csúcsosságnak („excess kurtosis", „többlet csúcsosság"), hiszen így a normális eloszlásra mindkét mutató 0 értékű lesz. Mivel a mi álláspontunk szerint e megoldás ad hoc, valószínűségelméletileg nem tiszta, nem követjük ezt a szisztémát, és az említett módon definiált mutatót fogjuk dolgozatunkban használni.

Ezek szerint a momentumok alapján történő illesztés feladata a következőként határozható meg. Adott egy $f\left(\underline{\Theta}\right)$ eloszlás, ahol a $\underline{\Theta}$ paramétervektor az eloszlás jellemzőit határozza meg. Keressük azt a $\underline{\Theta}^*$ paramétervektort, melyre teljesül, hogy

$$\mu_n\left(\underline{\Theta}^*\right) = m_n$$

valamely $m_n, n = 1,2,...H$ számsorozatra. Itt tipikusan $H = 4$.

1.2. Kvantilisek alapján történő illesztés

A feladat igen hasonló az előbbihez, egyedül a p-ed rendű ($p \in (0,1)$) kvantilis

$$\rho_p : \int_{-\infty}^{\rho_p} f(x)\mathrm{d}x = p$$

egyenlettel definiált fogalmát kell bevezetnünk.

Ekkor a kvantilisek alapján történő illesztés feladata így fogalmazható meg. Adott egy $f\left(\underline{\Theta}\right)$ eloszlás, ahol a $\underline{\Theta}$ paramétervektor az eloszlás jellemzőit határozza meg. Keressük azt a $\underline{\Theta}^*$ paramétervektort, melyre teljesül, hogy

$$\rho_{q_n}\left(\underline{\Theta}^*\right) = \rho_n$$

valamely $\{\rho_n, q_n\}$ (n=1,2,...H) párokból álló sorozatra.

2. Eloszláscsaládok

A következőkben bemutatjuk a legfontosabb olyan eloszláscsaládokat, melyek gyakorlati problémák esetén lehetővé teszik a momentumok és/vagy kvantilisek alapján történő illesztést. (Természetesen mindenhol megadjuk ennek korlátait is.) A leírások során bemutatjuk az eloszlásokat, és külön kitérünk az illesztés elvégzésének statisztikai hátterére.

2.1. Pearson-eloszláscsalád

A modern statisztika egyik legnagyobb alakja, *Karl Pearson* a XX. század első évtizedeiben vezette be azt az eloszláscsaládot, mely mai napig az ő nevét viseli. Az

ezzel kapcsolatos ismereteket cikkek egész sorában közölte (*Pearson* [1893], [1895], [1901], [1916]), melyek közül az első 1893-ben, az utolsó 1916-ban jelent meg. A sokszor a saját korát is megelőző közlés 12, római számmal azonosított eloszlás meglehetősen kusza rendszerét eredményezte, melyek számát Pearson folyamatosan növelte a cikkek során, de időközben bizonyos eloszlásokat át is definiált, míg másokról megállapította, hogy az előzők speciális esetei.

Tovább bonyolítja a helyzetet, hogy ezen eloszlások egy része – ahogy a valószínűségelmélet fejlődésével feltárultak az összefüggések – más nevet kapott a későbbiekben. Így fordulhat elő, hogy a Pearson-rendszerben vannak olyan eloszlások, amelyek – bár rájuk csak egy rejtélyes római szám utal – valójában teljesen triviálisak, míg más eloszlásoknak oly kevés a gyakorlati jelentősége, hogy azóta szinte feledésbe mentek.

Pearson eredeti célja az volt, hogy – biostatisztikai indíttatásból – olyan eloszlásrendszert alkosson, mely lehetővé teszi az illesztést a legkülönfélébb ferdeségű és csúcsosságú empirikus adatokra; még pontosabban, hogy olyan eloszlásrendszert adjon meg, mellyel minden esetben elvégezhető az illesztés az első négy momentum alapján – épp, amire nekünk is szükségünk van a korábban már vázolt okokból.

Pearson alapötlete az volt, hogy az eloszlásokat a sűrűségfüggvényükkel adta meg, de nem közvetlenül ($f(x)$ alakban), hanem egy rá vonatkozó differenciálegyenlettel ($\frac{\mathrm{d}f(x)}{\mathrm{d}x}$ alakban):

$$\frac{\mathrm{d}f}{\mathrm{d}x} = \frac{x}{b_0 + b_1 x + b_2 x^2} \cdot f. \qquad /1/$$

Bár első ránézésre igen szokatlan megadása ez egy sűrűségfüggvénynek, bizonyos tulajdonságok mégis kényelmesen leolvashatók. Egyrészt, ha egy ennek megfelelő eloszlás az egész $x \in \mathsf{R}$ számegyenesen értelmezett, akkor egy és csakis egy helyen, az $x = 0$ pontban vesz fel szélsőértéket. (Mivel itt eleve adott a sűrűségfüggvény deriváltja, ez közvetlenül leolvasható.) Az is könnyen belátható (lásd a Mellékletben), hogy ez a szélsőérték maximum lesz, tehát levonhatjuk azt a következtetést, hogy az ilyen (egész számegyenesen értelmezett) Pearson-eloszlások unimodálisak, módusszal a 0 pontban. Természetesen ez nem szükségképp teljesül azokra az eloszlásokra, melyek nem értelmezettek a teljes valós számegyenesen: ezeknél szélsőérték (módusz) lehet továbbá az értelmezési határokban is. Ezen felül az is észrevehető, hogy az $x \to \pm\infty$ határátmenetben a $\frac{\mathrm{d}f}{\mathrm{d}x}$ szintén nullába tart, tehát az eloszlás mindkét végén elenyészik. (Ezek teljesülése Pearsont is motiválta az eloszláscsalád kialakításakor.)

Az /1/ egyenlet átrendezésével és kiintegrálásával megkaphatjuk a Pearson-rendszer nyers momentumaira érvényes következő összefüggést:

$$nb_0\mu'_{n-1} + (n+1)b_1\mu'_n + \left[(n+2)b_2 + 1\right]\mu'_{n+1} = 0.$$

Ez jól láthatóan egy rekurzív összefüggés, mely lehetővé teszi, hogy $\mu'_0 (=1)$ és μ'_1 (tehát valójában csak μ'_1) ismeretében meghatározzuk az összes momentumot. (Ismerve természetesen az eloszlást leíró 3 paramétert, b_0 -t, b_1 -t és b_2 -t.) Belátható, hogy e momentumok közül az első négy egyértelműen meghatározza az eloszlás 3 paraméterét, így annak sincs akadálya, hogy ezeket a (centrális) momentumok függvényében írjuk fel (lásd a Mellékletet).

Mivel a fenti együtthatók még semmit nem mondanak magukról az eloszlásokról, így a problémát tovább kell vizsgálnunk: meg kell oldalunk a bemutatott differenciálegyenletet. Ennek menete terjedelmi okokból a Mellékletben található, mi most a végeredményre koncentrálunk.

2.1.1. A három alapvető Pearson-eloszlás

Bár (amint azt a 2.1. bevezetésében is említettük) Pearson 12 eloszlást definiált, ezek közül csak 3 van, ami nemnulla területet fed le a ferdeség/csúcsosság síkból, a többi – ebből is következően – átmeneti, illetve elfajuló eset. Mi a következőkben – összhangban eredeti céljainkkal – erre a 3 eloszlásra, a Pearson I, IV és VI eloszlásokra fogunk koncentrálni. Ezek az eloszlások a bemutatott általános esetből származtathatók, bizonyos sajátosságokat ($b_0 + b_1 x + b_2 x^2$ -nak van-e valós gyöke, illetve ha igen, akkor azok hogy helyezkednek el) is figyelembe vevő paraméterezéssel. (Ennek oka a Mellékletben közölt differenciálegyenlet-megoldásból válik világossá.)

Pearson IV. Amennyiben a $b_0 + b_1 x + b_2 x^2$ -nek nincs valós gyöke, a következő alakot érdemes (lásd a Mellékletet) használni:

$$f(x) = k\left(1 + \frac{x^2}{\alpha^2}\right)^m \cdot \exp\left[\nu \cdot \operatorname{arctg}\left(\frac{x}{\alpha}\right)\right]. \qquad /2/$$

Mivel a /2/ eloszlásfüggvény minden $x \in \mathsf{R}$ esetében értelmezett és valós, így az eloszlás tartója a teljes számegyenes. Összevetve ezt az áttekintésben mondottakkal, egyből adódik, hogy az eloszlás unimodális és harang[6] alakú.

[6]Az eloszlások alakja kapcsán a harang közismert jelentésű, az U-alakkal arra utalunk, hogy a sűrűségfüggvény egy lokális maximum szélsőértéktől indulva csökken, majd a globális minimum után növekszik, és egy lokális maximum az értelmezési tartomány felső határa. (A két lokális maximum közül bármelyik lehet globális maximum.) Az L- és a J-alakú eloszlások egymás tükörképei, így az L-alakú eloszlásoknál a módusz az értelmezési tartomány alsó, míg a J-alakúaknál a felső határa.

Az integrációs konstansból adódó – és eddig kötetlen – k értéke azon peremfeltétellel határozható meg, hogy a sűrűségfüggvény integrálja a teljes számegyenesen egységnyi. A részletek mellőzésével (lásd például *Heinrich* [2004]-et) ennek értéke:

$$k = \frac{\Gamma(m)}{\sqrt{\pi}\alpha\Gamma(m-1/2)}\left|\frac{\Gamma(m+i\nu/2)}{\Gamma(m)}\right|^2 = \frac{\left|\frac{\Gamma(m+i\nu/2)}{\Gamma(m)}\right|^2}{\alpha B(m-1/2,1/2)}.$$

Ebből a felírásból az utolsó szükséges információ is kiolvasható: az eloszlás akkor normálható, azaz akkor létezik, ha $m > 1/2$.

Pearson I és VI. Ha a $b_0 + b_1 x + b_2 x^2$-nek van valós gyöke, akkor (lásd a Mellékletet):

$$f(x) = k \cdot (x - a_1)^{\frac{\sqrt{b_1^2 - 4b_0b_2} + b_1}{2b_2\sqrt{b_1^2 - 4b_0b_2}}} \cdot (a_2 - x)^{\frac{\sqrt{b_1^2 - 4b_0b_2} - b_1}{2b_2\sqrt{b_1^2 - 4b_0b_2}}}.$$

A *Pearson I eloszlás* esetén az így meghatározott sűrűségfüggvény az $x \in (a_1, a_2)$ tartományon vesz fel valós értéket, csak ott értelmezett. Ez az eloszlás tehát mindkét irányból korlátos tartón, egy véges intervallumon értelmezett csak. Az eloszlás harang-, U- és J-alakú is lehet.

Vezessük be az $m_1 = \frac{\sqrt{b_1^2 - 4b_0b_2} + b_1}{2b_2\sqrt{b_1^2 - 4b_0b_2}}$ és az $m_2 = \frac{\sqrt{b_1^2 - 4b_0b_2} - b_1}{2b_2\sqrt{b_1^2 - 4b_0b_2}}$ jelöléseket, továbbá transzformáljuk a számegyenest úgy, hogy az origót a_1-be toljuk, az egységet pedig $(a_2 - a_1)$-nek választjuk. Ekkor a sűrűségfüggvény így írható:

$$f(x) = kx^{m_1}(1-x)^{m_2}.$$

Ekkor a normalizációs konstans beláthatóan $1/B(m_1+1, m_2+1)$ lesz, így (az immár nyilvánvalóan a $(0;1)$ intervallumon értelmezett) sűrűségfüggvény:

$$f(x) = \frac{1}{B(m_1+1, m_2+1)} x^{m_1}(1-x)^{m_2}.$$

Végül pedig, ebből az eloszlás létezésének feltétele is leolvasható: $m_1, m_2 > -1$.

Pearson VI eloszlásnál az előző eset előjeleinek megfordításával, alkalmas transzformációval a következő sűrűségfüggvényhez jutunk:

$$f(x) = \frac{1}{B(m_1+1, m_2+1)} x^{m_1}(1+x)^{-m_1-m_2-2}. \qquad /3/$$

Ez azért új lényegileg, mert nem egy tartományon (két gyök között), hanem azon kívül értelmezett; a /3/ speciális esetben például az $x \in (0, \infty)$-n. Ezen eloszlás tartója tehát mindig egy félegyenes. Alakja harang vagy J.

A /3/ sűrűségfüggvényből közvetlenül látható, hogy a létezés feltétele, hogy $m_2 > -1, m_1 + m_2 < -1$.

2.1.2. Az alapvető Pearson-eloszlások illesztése momentumok alapján

A következőkben megadjuk az eloszlások illesztéséhez szükséges összefüggéseket.

Pearson IV. Vezessük be az $r = 2m - 2$ jelölést. Ezzel a nyers momentumok számítási módszere:

$$\begin{aligned}
\mu_1' &= -\frac{a\nu}{r}, \\
\mu_2' &= \frac{a^2}{r(r-1)}\left(r + \nu^2\right), \\
\mu_n' &= \frac{a}{r-n+1}\left[(n-1)a\mu_{n-2}' - \nu\mu_{n-1}'\right].
\end{aligned}$$

Ezekből a standardizált momentumok és a ferdeség/csúcsosság ($\gamma_1 - \gamma_2$) mutatói számíthatók. Ha ez utóbbit megtesszük, és az eredményeket egyenlővé tesszük a specifikált g_1 és g_2 értékekkel, majd a kapott egyenletrendszert megoldjuk, akkor a következőket kapjuk:

$$r = 2(m-1) = \frac{6\left(g_2 - g_1^2 - 1\right)}{2g_2 - 3g_1^2 - 6},$$

$$\nu = \frac{r(r-2)g_1}{\sqrt{16(r-1) - g_1(r-2)^2}},$$

$$a = \frac{\sqrt{\mu_2\left[16(r-1) - g_1(r-2)^2\right]}}{4}.$$

Pearson I. Pearson itt is megadta a nyers és standardizált momentumokat, ízelítőül az első két nyers momentum (ehhez legyen $b = a_1 + a_2$):

$$\mu_1' = \frac{b(m_1+1)}{m_1 + m_2 + 2}, \quad \mu_2' = \frac{b^2(m_1+2)(m_1+1)}{(m_1 + m_2 + 3)(m_1 + m_2 + 2)}.$$

(A további tagok számítására rendelkezésre áll egy (igaz meglehetősen összetett) rekurzív formula.)

A paraméterek számítása specifikált momentumok alapján:

$$r = 2\frac{6\left(g_2 - g_1^2 - 1\right)}{3g_1^2 - 2g_2 + 6}, \quad \varepsilon = \frac{r^2}{4 + \frac{1}{4}g_1^2\left(r+1\right)^2 \big/ \left(r+1\right)}.$$

Ezek meghatározása után a két ismeretlen paraméter az

$$\left(m+1\right)^2 - r\left(m+1\right) + \varepsilon = 0$$

másodfokú egyenlet két gyökeként kapható meg.

Pearson VI. A nyers momentumok számítására meglehetősen bonyolult (de explicit alakú) formula áll rendelkezésre; alakjukat tekintve a Pearson I-gyel lesznek analógak. Ebből is következően, a számítás menete egyezik Pearson I-gyel.

2.1.3. A Pearson-eloszláscsalád lefedési tartománya

Az eloszlásokat bemutató részben minden esetben megadtuk az eloszlás létezésének feltételét. (Tipikusan származtatott paraméterek alapján, ám végeredményben az eredeti differenciálegyenlet paramétereit használva.) Nincsen akadálya tehát annak, hogy ezeket a feltételeket átírjuk a minta tulajdonságaira; ez minden esetben megtehető lesz pusztán a ferdeség és csúcsosság felhasználásával.

Pearson IV és VI létezésének a feltétele $2g_2 - 3g_1^2 - 6 > 0 \Rightarrow g_2 > \frac{3}{2}g_1^2 + 3$, míg a Pearson I-é $6 + 3g_1^2 - 2g_2 > 0 \Rightarrow g_2 < \frac{3}{2}g_1^2 + 3$.

A Pearson IV-et és VI-ot a $g_2 = \frac{3\left(2\sqrt{g_1^6 + 12g_1^4 + 48g_1^2 + 64} + 13g_1^2 + 16\right)}{32 - g_1^2}$ egyenletű görbe különíti el egymástól.

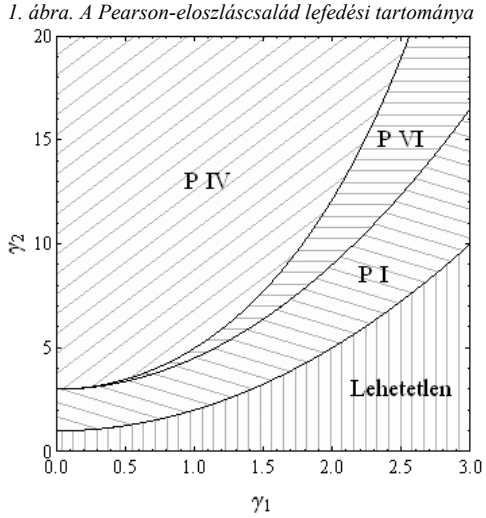
Ezen görbe alatt a Pearson VI, fölötte (ad infinitum) a Pearson IV eloszlás található. (Lásd az 1. ábrát.)

A Pearson-eloszláscsalád legfontosabb, minden más jó tulajdonságánál fontosabb előnye, hogy minden ferdeség/csúcsosság pontra illeszthető; lefedi az egész ferdeség/csúcsosság síkot. Még a most bemutatott igen általános eloszlások közül is csak kevés bír ehhez fogható lefedéssel.

Hátránya viszont, hogy az eloszlás, illetve kvantilisfüggvénye nem adható meg explicit alakban, így nincs egyszerű, általános módszer Pearson-eloszlásból származó

véletlenszámok generálására. Nem is beszélve arról, hogy ha iterálunk a ferdeség/csúcsosság síkon, akkor még az algebrai formák között is váltogatnunk kell, ami szintén impraktikus számítástechnikai szempontból.

1. ábra. A Pearson-eloszláscsalád lefedési tartománya



2.2. Burr-eloszláscsalád

Irwing W. Burr amerikai statisztikus 1942-ben publikált cikke (*Burr* [1942]) tekinthető az első közlésnek a témában. Burr 12 eloszlást adott meg írásában (mindegyiket eloszlásfüggvényével), melyeket gyakorlati szempontból fontosnak nevezett. Ezen eloszlások közül azonban egyetlen, a XII-es kapott nagy figyelmet a későbbiekben.

Már maga Burr is kiemelte ezt az eloszlást az idézett cikkében, és példaként tárgyalta az empirikus adatokhoz való illesztésének módszerét. Helyesen mutatott ugyanis rá, hogy az eloszlás paraméterei révén változatos γ_1 ferdeség és γ_2 csúcsosság mutatókat tud felvenni.

Hatke [1949] már azt is vizsgálta, hogy milyen ferdeség/csúcsosság értékekre végezhető el az illesztés, ám a ma szokásos $\gamma_1 - \gamma_2$ sík helyett a $\gamma_1^2 - \delta$ síkot használta a lefedettség megadásához, amely δ mérték ma már nincs[7] használatban, ráadásul később sikerült igazolni, hogy adatai részben tévesek: az eloszlás nagyobb területen illeszthető, mint a cikk megadta.

[7] $\delta = \dfrac{2\gamma_2 - 3\gamma_1^2 - 6}{\gamma_2 + 3}$ definíció szerint; használatát *Craig* [1936] javasolta, alapvetően azért, mert bevezetésével a Pearson-eloszlások sokkal egyszerűbb formát öltenek bizonyos számításokban.

Hosszú szünet után, 1968-ban Burr új cikkekkel jelentkezett (*Burr–Cislak* [1968]), melyben elsődlegesen a Burr-eloszlású sokaságokból vett minták becsléselméleti tulajdonságaival foglalkozott, ám emellett rámutatott *Hatke* [1949] előbb említett hibájára, és frissítette a lefedettséget mutató ábrát. Nem sokkal később (*Burr* [1973]) rövid közleményben bemutatta azokat az immár elektronikus számítógéppel, nagy pontossággal számított táblázatait, melyekkel finoman lehet illeszteni az eloszlásokat. Az eredményeket még mindig a $\gamma_1^2 - \delta$ síkon adta meg grafikusan, és továbbra sem foglalkozott a határok analitikus felírásának kérdésével.

Ebből a szempontból *Rodriguez* [1977] jelentett óriási előrelépést. A szerző egyrészt a ma szokásos $\gamma_1 - \gamma_2$ síkon adta meg az eloszlás lefedését (megmutatva, hogy számos, gyakorlatban fontos eloszlás illeszthető a Burr XII-vel), másrészt a lefedettséget illetően nem numerikus számításokon alapuló, analitikus eredményeket is elért.

Az utolsó fontos elméleti fejlemény *Tadikamalla* [1980] cikke, melyben tisztázta a kapcsolatot a Burr XII és pár egyéb fontos eloszlás között, egyúttal felhívta a figyelmet a Burr III eloszlás XII-höz hasonló kedvező illesztési tulajdonságaira.

2.2.1. A Burr XII eloszlás származtatása és definíciója

Burr eredeti cikkében (*Burr* [1949]) azt tűzte ki feladatul, hogy a gyakorlatban előforduló adatokhoz történő illesztésre alkalmas eloszlásokat adjon meg eloszlásfüggvénnyel[8]. A korszak legnépszerűbb, empirikus adatok illesztésére szolgáló rendszere, a Pearson-eloszláscsalád nem felel meg ennek a szempontnak, hiszen az eloszlások sűrűségfüggvényét ragadja meg (ahogy azt mi is tárgyaltuk a 2.1. pontban) egy, a sűrűségfüggvényre felírt differenciálegyenlet segítségével.

Burr úgy látott neki a feladatnak, hogy megalkotta Pearson differenciálegyenletének analógiáját eloszlásfüggvényre felírva:

$$\frac{\mathrm{d}F(x)}{\mathrm{d}x} = F(x)\left[1 - F(x)\right]g(x).$$

Az analógia nyilvánvaló, ha a $g(x) = \frac{1}{a + bx + cx^2}$-et tekintjük, és figyelembe vesszük, hogy a nevezőben itt csak $F(x)$ szerepelhet ($x \cdot F(x)$ nem), hogy az minden $x \in \mathsf{R}$-re nemnegatív legyen. (Ellenkező esetben sérülne az eloszlásfüggvény nemcsökkenő tulajdonsága.)

[8] Ez azért hangsúlyos, mert a korszak korlátozott számítástechnikai lehetőségei mellett komoly előnyökkel bírt empirikus adatok illesztésénél az eloszlásfüggvény használata (hiszen az intervallumok valószínűsége integrálás helyett egyszerű kivonással kapható meg) szemben az egyébként szokásosabb sűrűségfüggvényekkel. Hasonlóképp könnyebben ragadhatók meg a kvantilisértékek is.

Rögtön látható, hogy ez egy szétválasztható változójú differenciálegyenlet, amit szeparálva, majd az integrálást parciális törtekre bontással elvégezve, kapjuk, hogy:

$$F(x) = \frac{1}{e^{-G(x)} + 1}.$$

Burr a cikkében 12, általa fontosnak vélt konkrét $F(x)$ eloszlásfüggvényt ad meg. Ezek közül az utolsó, továbbiakban a Burr XII:

$$F(x) = 1 - \frac{1}{\left(1 + x^c\right)^k},$$

amely csak az $x \in (0, \infty)$ tartományon értelmezett, és $0 < c, k \in \mathbb{R}$. (A Mellékletben megmutatjuk, hogy az általános formulából hogyan kapható meg a Burr XII.)

2.2.2. A Burr XII tulajdonságai

Sűrűségfüggvény. A Burr XII sűrűségfüggvénye egyszerű deriválással adódik:

$$f(x) = F'(x) = \frac{kcx^{c-1}}{\left(1 + x^c\right)^{(k+1)}},$$

ahol továbbra is $x > 0$.

A Mellékletben részletesebben is elemezzük a sűrűségfüggvény jellegét. Ebből ki fog derülni, hogy $c > 1$ esetben az eloszlás unimodális, $\sqrt[c]{\frac{c-1}{kc+1}}$ módusszal, $c \leq 1$ esetben L-alakú.

Kvantilisfüggvény. A Burr XII kvantilisfüggvénye (tehát az eloszlásfüggvényének az inverze) egyszerű algebrai átalakításokkal megkapható az eloszlásfüggvényből:

$$F^{-1}(p) = Q(p) = \left[\frac{1}{(1-p)^{1/k}} - 1\right]^{1/c}.$$

Ezzel kapcsolatban kiemeljük, hogy lehetséges a kvantilisfüggvényt zárt alakban előállítani, ami nagy számítástechnikai egyszerűsítést jelent, ha ilyen eloszlást követő véletlenszámokat kell generálnunk.

2.2.3. A Burr XII momentumai analitikusan

A momentumokon alapuló illesztés kulcsfeladata az elméleti eloszlás momentumainak felírása általánosságban, az eloszlás ismeretlenjeinek segítségével.

Átlag és szórás illesztése. A Burr XII-nek csak két paramétere van, így világos, hogy az első 4 momentumot – célkitűzésünk szerint – bizonyosan nem fogjuk tudni tetszőlegesen megszabni. Egyik megoldás, hogy a ferdeség/csúcsosság beállítása után meghatározzuk az – így már adódó – átlagot és szórást, majd az eloszlás x változójához hozzáadjuk az elméleti és az empirikus átlag különbségét, illetve szorozzuk azt az elméleti és empirikus szórás hányadosával. Ez a művelet könnyen belefoglalható az eloszlásfüggvénybe is, például:

$$F(x) = 1 - \frac{1}{\left[1 + \left(\frac{x-\mu}{\sigma}\right)^{c}\right]^{k}},$$

ahol választható például $\mu = \mu' - \bar{\mu}(c,k)$ és $\sigma = \frac{\sigma'}{\bar{\sigma}(c,k)}$. (Itt $\bar{\mu}(c,k)$, illetve $\bar{\sigma}(c,k)$ jelenti a harmadik és negyedik momentumhoz (az első kettőtől függetlenül) illesztett eloszlás első két momentumát, μ' és σ' pedig a kívánt várható értéket és a szórást.) Ilyen módon a 2 paraméteres eloszlásunk könnyedén 4 paraméteressé alakítható. (Ez a megoldás látható például *Hönschová* [2008]-ban is.)

Amennyiben nem definiált momentumokhoz, hanem empirikus adatokhoz illesztünk, akkor egy másik (kézenfekvő, és az előző által is sugallt) lehetőség, hogy az eloszlásból csak az adódó várható értéket vonjuk ki, illetve szórásával osztjuk le, majd ehhez az empirikus adatok standardizáltját illesztjük. (Ez kézi számításnál praktikus, hiszen táblázatba célszerű volt eleve a standardizált értékeket foglalni.)

Mivel az említettek semmilyen lényegi módosítást nem jelentenek, így a továbbiakban az eredeti eloszlást használjuk, nem törődve a várható értékkel és szórással, tudva, hogy azokat tetszőlegesen beállíthatjuk anélkül, hogy az bármiben módosítaná a most következő tárgyalást.

Ferdeség és csúcsosság. A következőkben az eloszlás momentumait, centrális momentumait és standardizált centrális momentumait származtatjuk, hogy így megkapjuk a ferdeség és csúcsosság már bemutatott γ_1 és γ_2 mutatószámait (a levezetések terjedelmi okokból a Mellékletben kaptak helyet). Az így kapott formulákat használhatjuk később az illesztéshez.

$$\gamma_1 = \frac{\mu_3}{\sigma^3} = \frac{\mu_3}{\mu_2^{3/2}} = \frac{\Gamma^{-3}\left[2\lambda_{c,k}^3(1) - 3\Gamma(k)\lambda_{c,k}(1)\lambda_{c,k}(2) + \Gamma^2(k)\lambda_{c,k}(3)\right]}{\left\{\Gamma^{-2}(k)\left[\Gamma(k)\lambda_{c,k}(2) - \lambda_{c,k}^2(1)\right]\right\}^{3/2}} = \frac{2\lambda_{c,k}^3(1) - 3\Gamma(k)\lambda_{c,k}(1)\lambda_{c,k}(2) + \Gamma^2(k)\lambda_{c,k}(3)}{\left[\Gamma(k)\lambda_{c,k}(2) - \lambda_{c,k}^2(1)\right]^{3/2}}. \quad /4/$$

$$\gamma_2 = \frac{\mu_4}{\sigma^4} = \frac{\mu_4}{\mu_2^2} =$$

$$= \frac{\Gamma^{-4}\left[-3\lambda_{c,k}^4(1) + 6\Gamma(k)\lambda_{c,k}^2(1)\lambda_{c,k}(2) - 4\Gamma^2(k)\lambda_{c,k}(1)\lambda_{c,k}(3) + \Gamma^3(k)\lambda_{c,k}(4)\right]}{\left\{\Gamma^{-2}(k)\left[\Gamma(k)\lambda_{c,k}(2) - \lambda_{c,k}^2(1)\right]\right\}^2} = \quad /5/$$

$$= \frac{-3\lambda_{c,k}^4(1) + 6\Gamma(k)\lambda_{c,k}^2(1)\lambda_{c,k}(2) - 4\Gamma^2(k)\lambda_{c,k}(1)\lambda_{c,k}(3) + \Gamma^3(k)\lambda_{c,k}(4)}{\left(\Gamma(k)\lambda_{c,k}(2) - \lambda_{c,k}^2(1)\right)^2}.$$

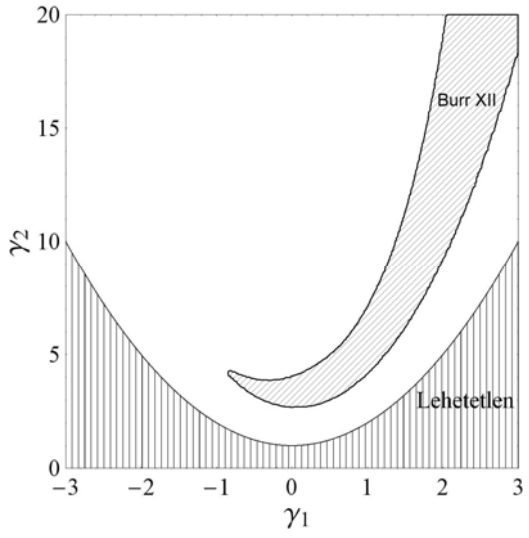
Ezeket a kifejezéseket (melyek c és k függvényei) egyenlővé kell tenni a specifikált g_1 és g_2 értékekkel, majd a kapott egyenletrendszert meg kell oldani c-re és k-ra. Ezt természetesen csak numerikusan tudjuk megtenni, ráadásul a megoldás még így is számos problémát felvet(het): numerikus instabilitás (például kerekítésekből adódó hibák), konvergencia kérdése stb. Egyszóval, bár itt ezt a kérdést egyáltalán nem tárgyaljuk, fontos jelezni, hogy a megoldás ettől még nem feltétlenül triviális.

A várható érték és a szórás illesztésének kérdését már tárgyaltuk, így az illesztés az eddigiek ismeretében teljeskörűen elvégezhető a Burr XII lefedési tartományában.

2.2.4. A Burr XII lefedési tartománya

A /4/ és /5/ egyenletek c és k argumentumait végigfuttatva lehetséges tartományukon, könnyen meghatározhatjuk – legalábbis empirikusan – a lefedési tartományt. (Lásd a 2. ábrát.)

2. ábra. A Burr XII eloszlás lefedési tartománya



A tartományt határoló görbékre *Rodriguez* [1977] analitikus egyenleteket is ad, ezekkel most – részben matematikai bonyolultságuk miatt – nem foglakozunk.

A Burr XII eloszlás, bár első ránézésre a lehetséges ferdeség/csúcsosság sík kis részét fedi, valójában igen praktikus, hiszen e „kis" rész számos, nagy gyakorlati jelentőségű eloszlást tartalmaz (többek között részeket mindhárom alapvető Pearson-típusból, a normális és logisztikus eloszlást, részeket mind a Johnson-féle S_U, mind az S_B eloszlásokból, részeket a Weibull- és a gamma-eloszlásokból stb.). Ennek következtében a kis fedés ellenére igen sok gyakorlati alkalmazásban jön szóba a használata, amit számos publikáció mutat az elmúlt évtizedekből.

Ezzel kapcsolatban az is előnyként jegyzendő meg, hogy a lefedett rész egyetlen algebrai alakú eloszlással érhető el (szemben például a Johnson- vagy Pearson-eloszlásokkal), így nem szükséges tartományonként eltérő eszköztár használata.

A Burr XII további előnye, hogy az eloszlásfüggvénye, illetve – ami ebből a szempontból még fontosabb – annak inverze (a kvantilisfüggvény) is megadható zárt alakban. Ez – figyelembe véve a közismert valószínűségszámítási tételt – azt jelenti, hogy a Burr XII eloszlást követő véletlen számok generálása igen egyszerűen, mindössze egy egyenletes véletlenszám-generátorral megvalósítható. Ez igen komoly előny akkor, ha számítógépes szimulációkhoz van szükség nagy mennyiségű Burr XII eloszlású véletlenszámra.

2.3. A Johnson-eloszláscsalád

Norman L. Johnson (*Karl Pearson* fiának témavezetése alatt készített) PhD-dolgozatában mutatta be a később róla elnevezett eloszláscsaládot. 1949-es munkájában, Pearson megoldásához hasonlóan, *Johnson* [1949] is eloszláscsaládot definiált, vagyis nem egyetlen formula paraméterezésével, hanem a $\gamma_1 - \gamma_2$ sík különböző területeire különböző függvényeket definiálva érte el célját.

Az eloszlásokat sűrűségfüggvényükön keresztül határozta meg, implicite:

$$z = \gamma + \delta \cdot \log f\left(\frac{x-\mu}{\lambda}\right),$$

ahol *z* standard normális eloszlás, míg az *f* függvény háromféle lehet:

– lognormális $S_L : f(u) = u$,
– korlátozatlan $S_U : f(u) = u + \sqrt{1+u^2}$,
– korlátozott $S_B : f(u) = u/(1-u)$.

A háromféle eloszlás együttesen teljesen lefedi a lehetséges $\gamma_1 - \gamma_2$ síkot.

Az S_L eloszlás ennek a síknak egyetlen egyenesét fedi le, így itt a két mutató egymást egyértelműen meghatározza. Az így kapott eloszlások egy oldalon korlátozottak, míg végtelenek a másik oldalon.

Az S_U eloszlások a $\gamma_1 - \gamma_2$ sík S_L vonal feletti területet fedik le, magukban foglalva a Pearson IV, V, VII eloszlásokat, illetve bizonyos VI-os eloszlásokat. Az így kapott eloszlások mindkét oldalukon végtelenek.

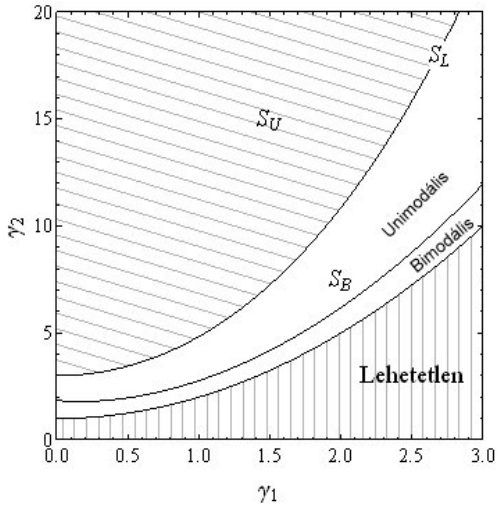
Az S_B eloszlások az S_L vonal és az eloszlások létezésének – korábban ismertetett – alsó határa közti területet fedik le, azaz ide értendők Pearson I, II, III és bizonyos VI-os eloszlásai. Ezek az eloszlások mindkét oldalon korlátozottak (*Draper* [1952]).

Johnson [1949] azt is megmutatta, hogy az S_B eloszlások bimodalitásának szükséges és elégséges feltétele, hogy

$$\delta < 2^{-1/2} \qquad\qquad |\gamma| < \frac{\sqrt{1-2\delta^2} - 2\delta^2 \tanh^{-1}\sqrt{1-2\delta^2}}{\delta}.$$

Ez a feltétel tágabb területet fed le, mint az a terület, ahol minden létező eloszlás szükségszerűen kétmódusú (*Draper* [1952]). Az eloszláscsalád lefedési tartományát a 3. ábra szemlélteti.

3. ábra. A Johnson-eloszláscsalád lefedési tartománya



Az eloszlás kiterjesztéseként *Johnson* [1954], illetve *Tadikamalla* és *Johnson* ([1980], [1982]) a Laplace-, illetve a logisztikus eloszlást (L-eloszláscsalád) használta a normális helyett. Utóbbi az eloszlásfüggvény és az inverz eloszlásfüggvény egyszerűbb kifejezhetősége miatt könnyebb illesztést tesz lehetővé.

2.3.1. A Johnson-eloszláscsalád illesztése momentumok alapján

A három eloszlás illesztését különválasztva kell kezelni. A szétválasztáshoz először a kívánt g_1 érték alapján az $\omega = e^{\delta^{-2}}$ helyettesítéssel a

$$(\omega - 1)(\omega + 2)^2 = g_1$$

egyenletet kell megoldani, majd ebből a

$$\gamma_2 = \omega^4 + 2\omega^3 + 3\omega^2 - 3$$

kifejezést kell értékelni. Ha $\gamma_2 > g_2$, akkor az S_B, egyébként az S_U eloszlás illesztése szükséges (*Hill et al.* [1976]).

Az S_U görbe esetén a momentumok zárt alakban kifejezhetők, az illesztés így ezek alapján numerikusan elvégezhető (az egyes formulák a függelékben megtalálhatók). Az illesztés megkönnyítésére Johnson több alkalommal is (*Johnson* [1965], [1974]) publikált táblázatokat, amelyek a $\gamma_1 - \gamma_2$ értékekhez tartozó γ és δ értékeket tartalmazzák. Helyettesítések sorozatával a kifejezések egyszerűsíthetők, egy negyedfokú, kétismeretlenes egyenletrendszerre, amiből az eredeti paraméterek visszaszámíthatók (*Tuenter* [2001]).

Az S_B görbék momentumai nem fejezhetők ki zárt alakban, így az illesztés még nehezebb. A megfelelő közelítő táblázatokat *Pearson* és *Hartley* [1972] közölte.

A momentumok alapján történő illesztés esetén tehát – hasonlóan Pearson eloszláscsaládjához – először meg kell találni, hogy melyik a megfelelő eloszlás, és a paraméterek becslése csak ezután végezhető el.

Amint az a 3. ábrából is kitűnik, a Johnson-eloszláscsalád a teljes ferdeség/csúcsosság síkot lefedi, ez nagyon fontos elméleti előnye.

2.3.2. A Johnson-eloszláscsalád illesztése kvantilisek alapján

Az illesztéshez több tanulmány szerint is szimmetrikus percentiliseket célszerű választani (*Bukac* [1972], *Mage* [1980], *Slifker–Shapiro* [1980]). Belátható, hogy ebben az esetben helyettesítések sorozatával a probléma egy másodfokú egyenletrendszer megoldásához vezet, ami a jelenlegi számítástechnikai háttér mellett általában nem okoz nehézséget.

A kvantilisen alapuló meghatározás szimulációs vizsgálatok alapján (különösen a korlátozott függvényre) nem csak egyszerűbb, de kisebb négyzetes hibával (MSE) is rendelkezik, mint a momentumokon alapuló (*Wheeler* [1980]).

2.4. Az általánosított λ-eloszlás

Az általánosított λ-eloszlás (GLD) ötlete eredetileg *Tukey-tól* származik (*Tukey* [1960]). Az eloszlásnak mindössze egyetlen szabadon állítható paramétere van, így a

normálistól több szempont szerint adott módon eltérő eloszlások előállítására nem alkalmas (lásd a Mellékletet).

A helyzet és a terjedelem kezelését biztosító technikák ismeretében – szimmetrikus eloszlásokat eredményező – triviális általánosítást adott meg *Ramberg* és *Schmeiser* [1972].

Két évvel később került sor (*Ramberg–Schmeiser* [1974]) a formula további általánosítására, a továbbiakban erre RS-eloszlásként hivatkozunk:

$$Q(u) = \lambda_1 + \lambda_2^{-1}\left[u^{\lambda_3} - (1-u)^{\lambda_4}\right],$$

ahol λ_1 a helyzetért, λ_2 a szóródásért, λ_3 és λ_4 az eloszlás alakjáért felelősek. Összhangban az 1972-es eredményekkel, a $\lambda_3 = \lambda_4$ eset szimmetrikus eloszlásokat ad.

Ramberg és szerzőtársai (*Ramberg et al.* [1979]) megmutatták, hogy bizonyos paraméter-kombinációkra kapott eredmények nem lehetnek az eloszlás kvantilisei (adott λ_2 mellett a λ_3–λ_4 tér bizonyos kombinációi nem érvényesek). A létezés feltétele a sűrűségfüggvény nemnegativitásával, azaz a

$$\frac{\lambda_2}{\lambda_3 u^{\lambda_3 - 1} + \lambda_4 (1-u)^{\lambda_4 - 1}} \geq 0$$

feltétellel egyenértékű (*Su* [2005]).

Az elérhető eloszlások túlnyomó része egymóduszú; azonban korlátozott formában, de U-alakú és nyesett (L-alakú) eloszlások is előállíthatók. A $\lambda_3 \geq 1, \lambda_4 \leq 2$ paraméterezés U-alakú, míg a $\lambda_3 = 0$ paraméter L-alakú eloszlásokhoz vezet (*Ramberg et al.* [1979]).

A λ_3–λ_4 sík teljes lefedettségének biztosításra is találtak megoldást (*Freimer et al.* [1988]) a következő paraméterezésen keresztül (FMKL):

$$Q(u) = \lambda_1 + \lambda_2^{-1}\left[\frac{u^{\lambda_3} - 1}{\lambda_3} - \frac{(1-u)^{\lambda_4}}{\lambda_4}\right].$$

Az FMKL-eloszlás már a teljes λ_3–λ_4 térben definiált, az illesztés egyetlen feltétele, hogy $\lambda_2 > 0$ legyen. Az eloszlás *k*-adik momentuma – hasonlóan az RS-eloszláshoz – akkor véges, ha $\min(\lambda_3, \lambda_4) > -1/k$. Az illesztéshez szükséges alapszámításokra vonatkozó irodalom ugyanakkor meglehetősen hiányos.

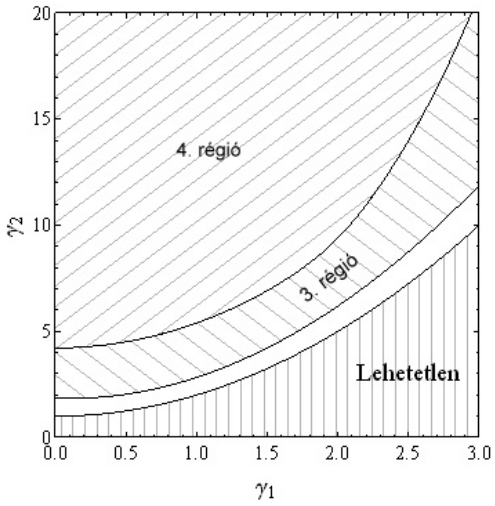
2.4.1. Az általánosított λ-eloszlás illesztése momentumok alapján

Az RS-eloszlás momentumokon alapuló illesztéséhez az eloszlás sűrűségfüggvényéből tudunk kiindulni. A momentumok definíciói alapján a paraméterek függvé-

nyében kifejezhetők a szükséges centrális momentumok, illetve a ferdeség és csúcsosság mutatói is. A Mellékletben található formulákból látható, hogy a γ_1 és γ_2 értékek csak a λ_3 és λ_4 függvényei, így az eloszlás ferdeségének és csúcsosságának meghatározása után a várható érték és a variancia külön paraméterezhető. Problémát okoz, hogy a λ_3- és a λ_4-értékek zárt alakban nem fejezhetők ki, így az egyenletrendszer megoldása erősen számításigényes, különösen a formulákban található béta függvények számítása miatt.

Karian és *Dudewicz* arra is felhívja a figyelmet, hogy az RS-eloszlás csak az $1{,}8\left(\gamma_1^2+1\right)\le\gamma_2$ teret tudja lefedni, így $\gamma_1-\gamma_2$ tér egy szűk sávjában nem lehetséges eloszlások generálása. Ez éppen az a sáv, ahol a lehetséges minimális csúcsosságnál csak kissé nagyobb csúcsosságértékek találhatók. (*Karian–Dudewicz* [2000]), amint a 4. ábra is mutatja. Valóban látható, hogy az általánosított λ-eloszlás egy szűk, közvetlenül a lehetetlen tartomány fölötti sáv kivételével lefedi a ferdeség/csúcsosság síkot. Az ábrán azt is megadtuk, hogy melyik *Karian–Dudewicz* [2000] szerinti régióból kikerülő paraméterekkel végezhető el a lefedés. (További régiók a nagyobb ferdeségeknél kaphatnak szerepet; csak többféle lefedésre adnak módot, a lefedés teljességét nem befolyásolják.)

*4. ábra. Az általánosított λ-eloszlás lefedési tartománya**



* A 3. régióban $\lambda_3, \lambda_4 > 0$, míg a 4. régióban $\lambda_3, \lambda_4 < 0$.

Az egyes paraméter-kombináció intervallumokban javasolt kezdőértékekről jó áttekintést ad (*Karian–Dudewicz* [2000]), míg *Lakhany–Mausser* [2000]-nél további illesztési módszerek értékelését is megtaláljuk.

2.4.2. Az általánosított λ-eloszlás illesztése kvantilisek alapján

A GLD-eloszlás valamennyi változata kvantilis függvényével adott, így kézenfekvőnek látszik a kvantiliseken alapuló illesztés. A 4 paraméteres változatok (RS és FMKL) esetén 4 kvantilis érték megadásával meghatározhatók a paraméterek. A kvantilis függvény formájából adódik, hogy az egyenletrendszerből gyorsan kiejthető a λ_1 és a λ_2 paraméter, így egy kétegyenletes, kétismeretlenes nemlineáris egyenletrendszert kell megoldani. Az egyenletrendszer csak hatványfüggvényeket tartalmaz, megoldása tehát nagyságrendekkel gyorsabban elvégezhető, mint a momentumokon alapuló illesztés (*Su* [2005]).

A problémát ebben az esetben az okozza, hogy 4 kvantilis közvetlenül nem tudja jól leírni az eloszlás alakját. A kvantiliseken alapuló csúcsosságmutatók is általában 4 kvantilist használnak (*Kim–White* [2004]), amik azonban éppen az eloszlás helyzetét nem határozzák meg. A probléma áthidalására *Karian* és *Dudewicz* 4 kvantilisen alapuló mutatót javasolt. Az első mutató, a medián szolgál az eloszlás helyzetének kiváltására (az első nyers momentum párjaként az eloszlás helyzetéért felel). A második mutató valamilyen interpercentilis mutató lehet, az adatok középső tartományának terjedelmét mutatja, $0 < u < 0{,}25$ (a szóródás mutatója). A harmadik mutató a ferdeséget írja le, míg a negyedik a csúcsosság egy lehetséges mérőszáma.

$$\rho_1 = F^{-1}(0{,}5) = \lambda_1 + \frac{\left(\frac{1}{2}\right)^{\lambda_3} - \left(\frac{1}{2}\right)^{\lambda_4}}{\lambda_2}$$

$$\rho_2 = F^{-1}(1-u) - F^{-1}(u) = \frac{(1-u)^{\lambda_3} - u^{\lambda_4} + (1-u)^{\lambda_4} - u^{\lambda_3}}{\lambda_2}$$

$$\rho_3 = \frac{F^{-1}(0{,}5) - F^{-1}(u)}{F^{-1}(1-u) - F^{-1}(0{,}5)} = \frac{(1-u)^{\lambda_4} - u^{\lambda_3} + \left(\frac{1}{2}\right)^{\lambda_3} - \left(\frac{1}{2}\right)^{\lambda_4}}{(1-u)^{\lambda_3} - u^{\lambda_4} + \left(\frac{1}{2}\right)^{\lambda_4} - \left(\frac{1}{2}\right)^{\lambda_3}}$$

$$\rho_4 = \frac{F^{-1}(0{,}75) - F^{-1}(0{,}25)}{\rho_2} = \frac{\left(\frac{3}{4}\right)^{\lambda_3} - \left(\frac{1}{4}\right)^{\lambda_4} + \left(\frac{3}{4}\right)^{\lambda_4} - \left(\frac{1}{4}\right)^{\lambda_3}}{(1-u)^{\lambda_3} - u^{\lambda_4} + (1-u)^{\lambda_4} - u^{\lambda_3}}.$$

A harmadik és a negyedik mutató csak λ_3 és λ_4 függvénye, így ebben az esetben is alkalmazható a rekurzív megoldás, először λ_3 és λ_4 meghatározása, majd abból λ_2, végül λ_1 kalibrálása.

2.5. A g-and-h-eloszlás

A g-and-h-eloszlás, az eredeti λ-eloszláshoz hasonlóan, *John Wilder Tukey* nevéhez fűződik. Az eloszlás egy 1977-es konferencia-előadásban (*Tukey* [1977]) került ismertetésre, amelyből tanulmány nem készült.

A g-and-h-eloszlást kvantilisfüggvényével (inverz eloszlásfüggvényével) definiáljuk, a standard normális eloszlásból (z) kiindulva, az alábbi transzformációval:

$$q(z) = g^{-1}\left(\mathrm{e}^{gz} - 1\right)\mathrm{e}^{hz^2/2} \quad g \neq 0 \quad h > 0\,.$$

A paraméterek közvetlenül alakítják az eloszlás alakját, így g felel a ferdeségért (irányban és nagyságban), h pedig a csúcsosságért (a kurtózissal pozitívan korrelál).

A két paraméter szerinti határeloszlások ($g \to 0$, illetve $h \to 0$) is meghatározhatók, illetve könnyen belátható, hogy a $g \to 0, h \to 0$ paraméterezés szerinti határeloszlás éppen a standard normális eloszlást adná vissza.

A g-and-h-eloszlás sűrűségfüggvénye az a következő alakban írható fel:

$$f_{q(Z)}\left(q(z)\right) = f_{q(Z)}\left(q(z), \frac{f_Z(z)}{q'(z)}\right).$$

Ahogy *Headrick* (*Headrick et al.* [2008]) megmutatja, a $q(z)$ transzformáció szigorú monotonitása miatt a sűrűségfüggvény unikális, globális maximumponttal rendelkezik, vagyis a kapott eloszlások egymóduszúak. Az eloszlások helyzetének jellemzésére az inverz eloszlásfüggvénnyel való megadásnak megfelelően a medián mutatkozik a legegyszerűbb középértéknek. Belátható, hogy a medián a $q(z=0)=0$ helyen lesz, ahogy a kiindulásul szolgáló standard normális eloszlásra is igaz.

2.5.1. A g-and-h eloszlás illesztése momentumok alapján

A sűrűségfüggvény felhasználásával az eloszlás momentumai definiálhatók:

$$E\left[q(z)^k\right] = \int_{-\infty}^{+\infty} q(z)^k \, f_Z(z)\,\mathrm{d}z\,.$$

A k-ad rendű momentum létezésének feltétele, hogy $0 \le h < 1/k$ teljesüljön. Ebből az első négy nyers momentum viszonylag egyszerűen származtatható. A pontos formulák a Mellékletben találhatók.

A harmadik- és negyedik centrális momentumokon alapuló ferdeség és csúcsosság mutatók (γ_1 és γ_2) a g és h függvényében megadhatók (a formulák a Melléklet-

ben találhatók). Ahogy *Rayner–MacGillivray* [2002] jelzi, valamennyi ferdeség elérhető, ugyanakkor a 3 alatti csúcsosságok (lapult eloszlások) nem képezhetők.

A momentumokon alapuló illesztéshez – még akkor is, ha az csupán a ferdeséghez és a csúcsossághoz való igazodást jelenti – megoldandó egyenletrendszer numerikusan is erősen számításigényes.

Az eloszlásnak két paramétere van, így γ_1 és γ_2 alapján *g* és *h* értékéből az első két momentum egyértelműen következik. Ha tetszőleges – vagy standard – első két momentummal rendelkező eloszlást szeretnénk illeszteni, akkor további általánosítás szükséges, ami az inverz eloszlásfüggvénnyel való megadás miatt analitikusan nehezen kezelhető, ismereteink szerint jelenleg megoldatlan probléma.

2.5.2. A g-and-h-eloszlás illesztése kvantilisek alapján

A kvantilisfüggvénnyel való definiálás sugallja a kvantiliseken alapuló illesztést. Figyelembe véve, hogy az eloszlás mediánja mindig 0, a triviális kvantilisen túl két kvantilis megadása egyértelműen meghatározza az eloszlást. Az 1.2. pontban leírtak alapján ez előny, ha célunk adott empirikus eloszláshoz elméleti eloszlás illesztése, ugyanakkor a normálistól adott mértékben eltérő eloszlás paraméterezése nem oldható meg. A kvantiliseken alapuló illesztés előnyei a többdimenziós eloszlások esetén jelennek meg (*Field–Genton* [2006]).

2.6. Fleishman-eloszlás

Bár „eloszlás” néven említjük, a Fleishman-eloszlás sokkal inkább egy nevezetes eloszlás-transzformációs eljárás. *Allen Fleishman* 1978-ben publikálta (*Fleishman* [1978]); már eredetileg is Monte-Carlo-szimulációkra gondolva.

Annak ellenére, hogy a módszer igen elegáns, és számítástechnikai szempontból is rendkívül jól kezelhető, mintegy 2 évtizedig szinte a feledés homályába merült. Jelentős elméleti fejlemény egészen a 2000-es évekig nem történt, és az alkalmazások többsége (*Headrick–Sheng–Hodis* [2007]) is az 1990-es évekre esik.

2002-ben Headrick kiterjesztette a módszert (*Headrick* [2002]), hogy az 6 momentumig alkalmas legyen illesztésre, majd ugyanő 2007-ben megoldotta (*Headrick–Kovalchuk* [2007]) a legfontosabb nyitott kérdést: megadta analitikus alakban a Fleishman-eloszlás sűrűség- és eloszlásfüggvényét.

2.6.1. A Fleishman-eloszlás áttekintése, illesztés

Fleishman módszere mindössze egy standard normális eloszlású véletlenszámot igényel. Alapötlete: készítsük el az így generált véletlen változó egy transzformáltját.

A transzformáló függvény konkrét alakja, paraméterei nyilván hatással lesznek a transzformált változó eloszlására, így momentumaira is. Ha a transzformáló függvénynek egyetlen paramétere van, akkor azzal nyilván egyszerre állítjuk az összes momentumot (legjobb esetben is), így általánosságban nincs arra remény, hogy el tudjuk érni, hogy a transzformált változó több (például négy) momentuma előírás szerinti legyen. Viszont ahogy növeljük paramétereinek számát, úgy válhat kellően „testreszabhatóvá" a transzformált függvény. A polinomiális transzformáló függvény alkalmas arra, hogy ezt megvalósítsa, hiszen kényelmesen állítható a szabad paraméterek száma. Ha például 4 paraméterre van szükségünk, mert a transzformált változó négy momentumát szeretnénk előírás szerintire állítani, akkor válasszuk az

$$Y = G(Z) = a + bZ + cZ^2 + dZ^3 \qquad /6/$$

transzformáló függvényt, ahol tehát $Z \sim N(0,1)$.

Nincs más dolgunk, mint meghatározni a paramétereket. Ehhez azt kell tudnunk, hogy a transzformált változó momentumai hogyan írhatók fel a transzformációs függvény ismeretében. Például, az első momentum, a várható érték esetén nincs nehéz dolgunk, hiszen az ismert tétel szerint, ha képezzük egy X valószínűségi változó $V = g(X)$ transzformáltját (ahol $g : \mathsf{R} \to \mathsf{R}$), akkor $E(V) = \int_{-\infty}^{\infty} g(x) \cdot f_X(x)\mathrm{d}x$ folytonos esetben. Azaz pusztán a transzformáló függvény és az eredeti sűrűségeloszlás ismeretében (egyetlen integrálással) megadható a transzformált változó várható értéke. Maradva a /6/ transzformáló függvénynél, azt kapjuk, hogy

$$E(Y) = \int_{-\infty}^{\infty} \left(a + bZ + cZ^2 + dZ^3\right) \cdot \varphi(x)\mathrm{d}x = a + c.$$

Ha tehát azt szeretnénk, hogy generált eloszlásunk várható értéke μ legyen, nincs más dolgunk, mint kielégíteni a $a + c = \mu$ egyenletet.

A további momentumokra hasonló kifejezések kaphatók, bár a számítások, ha nem is bonyolultabbak, de mindenestre jóval munkaigényesebbek lesznek (a hosszas polinomszorzások miatt), ezért csak a végeredményt közöljük. Felírva tehát ugyanezt a következő három momentumra is, három további egyenletet kapunk, így már megoldható lesz az négy ismeretlenes és immár négy egyenletes egyenletrendszerünk. Az eredményül kapott egyenletrendszer tehát, standardizált ($\mu = 0$, $\sigma = 1$) esetben:

$$a + c = 0$$

$$b^2 + 6bd + 15d^2 + 2c^2 = 1$$

$$2c\left(b^2 + 24bd + 105d^2 + 2\right) = g_1$$

$$24\left(bd + c^2\left(1 + b^2 + 28bd\right) + d^2\left(12 + 48bd + 141c^2 + 255d^2\right)\right) = g_2.$$

Ennek megoldásával a keresett transzformációs együtthatókat kapjuk. (Természetesen teljesen nyilvánvaló, hogy a megoldást numerikus úton kell végeznünk.)

Ha megvannak a transzformációs együtthatók, nincs más dolgunk, mint a standard normális eloszlású véletlenszámokat generálni, majd a felparaméterezett polinomiális függvénnyel transzformálni őket.

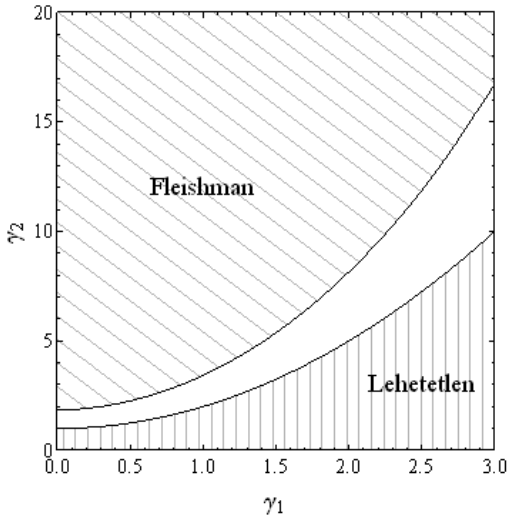
Ahogy említettük, *Headrick* [2002] megadja a szükséges formulákat az első 6 momentumhoz történő illesztésre is.

2.6.2. A Fleishman-eloszlás lefedési tartománya

A Fleishman-eloszlás nagyjából lefedi a teljes ferdeség/csúcsosság síkot, ám a módszer egy, az elméleti minimumnál kicsit magasabban húzódó minimum-csúcsossággal rendelkezik (adott ferdeségre). Ebből következően a lapult eloszlások generálása problémába ütközhet. Példának okáért, *Headrick–Sawilowsky* [2002] megmutatta, hogy szimmetrikus esetben a legkisebb elérhető csúcsosság $\gamma_1 = 1{,}85$ (1 helyett).

Saját számítási eredményeinket[9] e tekintetben az 5. ábra közli, melyen jól látható a nemgenerálható tartomány elhelyezkedése.

5. ábra. A Fleishman-transzformációs módszer lefedési tartománya



A Fleishman-módszer legnagyobb előnye akkor jelentkezik, ha nagy mennyiségű adott eloszlást követő véletlenszám-generálás szükséges. A módszer tervezéséből

[9] Ezen ábránál (és a többinél is) a számításokat és a rajzolást végző .nb Mathematica munkafüzet elérhető a szerzőknél.

adódóan ehhez mindössze egy standard normális eloszlású véletlenszám-generátor szükséges, ennek birtokában néhány szorzással és összeadással, azaz számítástechnikailag igen egyszerűen előállíthatók a kívánt véletlenszámok.

A módszer további jellemzője, hogy a ferdeség/csúcsosság síkon a lefedett terület igen nagy, ám nem a teljes elméletileg lehetséges terület.

*

A tanulmányban áttekintettük azokat az eloszlásokat, illetve eloszláscsaládokat, amelyek alkalmasak lehetnek változatos alakú – lehetőség szerint a ferdeség/csúcsosság síkot minél jobban lefedő – eloszlásból származó minták generálásához. Fontos szempontnak tartottuk, hogy eloszláscsaládok esetén a megfelelő eloszlás kiválasztása minél egyszerűbb legyen, az eloszlások paramétereinek megválasztása a kívánt empirikus eloszláshoz minél egyszerűbben megtörténhessen. Az irodalomban erre a célra fellelhető hat eloszlás(család) legfontosabb jellemzőit a táblázatban foglaljuk össze.

Az eloszlások legfontosabb jellemzői

Eloszlás(család)	Az eloszlás megadása	Momentumokon alapuló illesztés	Kvantiliseken alapuló illesztés	Ferdeség/csúcsosság lefedés
Pearson	sűrűségfüggvény (közvetetten)	lehetséges	nem ajánlott	Teljes; 3 algebrai alakkal
Burr	eloszlásfüggvény	lehetséges	nem ajánlott	Részleges, túl kis és túl nagy minimum feletti csúcsosságok egyaránt lehetetlenek; 1 algebrai alakkal
Johnson	sűrűségfüggvény (közvetetten)	lehetséges	ajánlott	Teljes; 2 algebrai alakkal
GLD	inverz eloszlásfüggvény	lehetséges, de nem ajánlott	ajánlott	Szinte teljes; kis minimum feletti csúcsosságok lehetetlenek; 1 algebrai alakkal
g-and-h	inverz eloszlásfüggvény	lehetséges	nem kivitelezhető	Szinte teljes; kis minimum feletti csúcsosságok lehetetlenek; 1 algebrai alakkal
Fleishman	sűrűségfüggvény	ajánlott	nem megoldott	Szinte teljes; kis minimum feletti csúcsosságok lehetetlenek; 1 algebrai alakkal

Irodalom

BUKAC, J. L. [1972]: Fitting S_B Curves Using Symmetrical Percentile Points. *Biometrika*. 59. évf. 688–690. old.

BURR, I. W. [1942]: Cumulative Frequency Functions. *The Annals of Mathematical Statistics.* 13. évf. 2. sz. 215–232. old.

BURR, I. W – CISLAK, P. J. [1968]: On a General System of Distributions: I. Its Curve-Shape Characteristics; II. The Sample Median.; III. The Sample Range. *Journal of the American Statistical Association*. 63. évf. 322. sz. 627–643. old.

BURR, I. W. [1973]: Parameters for a General System of Distributions to Match a Grid of α_3 and α_4 . *Communications in Statistics*. 2. évf. 1. sz. 1–21. old.

CRAIG, C. C. [1936]: A New Exposition and Chart for the Pearson System of Frequency Curves. *Annals of Mathematical Statistics*. 7. évf. 1. sz. 16–28. old.

DRAPER, J. [1952]: Properties of Distributions Resulting from Certain Simple Transformations of the Normal Distribution. *Biometrika*. 39. évf. 3–4. sz. 290–301. old.

FERENCI, T. [2009]: *Using Massively Parallel Processing in the Testing of the Robustness of Statistical Tests with Monte Carlo Simulation*. Challenges for Analysis of the Economy, the Businesses, and Social Progress International Scientific Conference. November 19–21. Szeged.

FIELD, C. – GENTON, M. G. [2006]: The Multivariate g-and-h Distribution. *Technometrics*. 48. évf. 1. sz. 104–111. old.

FLEISHMAN, A. I. [1978]: A Method for Simulating Non-normal Distributions. *Psychometrika*. 43. évf. 521–532. old.

FREIMER, M. ET AL. [1988]: A Study of the Generalized Tukey Lambda Family. Communications in Statistics. *Theory and Methods*. 17. évf. 10. sz. 3547–3567. old.

HALL, A. R. [2005]: *Generalized Method of Moments*. Oxford University Press. Oxford.

HATKE, M. A. [1949]: A Certain Cumulative Probability Function. *Annals of Mathematical Statistics*. 20. évf. 3. sz. 461–463. old.

HEADRICK, T. C. [2002]: Fast Fifth-Order Polynomial Transforms for Generating Univariate and Multivariate Non-normal Distributions. *Computational Statistics & Data Analysis*. 40. évf. 685–711. old.

HEADRICK, T. C. – SAWILOWSKY, S. S. [2000]: Weighted Simplex Procedures for Determining Boundary Points and Constants for the Univariate and Multivariate Power Methods. *Journal of Educational Behavioral Statistics.* 25. évf. 417–436. old.

HEADRICK, T. C. – SHENG Y. – HODIS F. A. [2007]: Numerical Computing and Graphics for the Power Method Transformation Using Mathematica. *Journal of Statistical Software*. 19. évf. 3. sz. 1–17. old.

HEADRICK, T. C. – KOWALCHUK, R. K. [2007]: The Power Method Transformation: Its Probability

Density Function, Distribution Function, and Its Further Use for Fitting Data. *Journal of Statistical Computation and Simulation*. 77. évf. 229–249. old.

HEADRICK, T. C. – KOWALCHUK, R. K. – SHENG, Y. [2008]: Parametric Probability Densities and Distribution Functions for Tukey g-and-h Transformations and Their Use for Fitting Data. *Applied Mathematical Sciences*. 2. évf. 9. sz. 449–462. old.

HEINRICH, J. [2004]: *A Guide to the Pearson Type IV Distribution*. http://www-cdf.fnal.gov/publications/cdf6820_pearson4.pdf (Elérés dátuma: 2010. május 15.)

HILL, I. D. – HILL, R. – HOLDER, R. L. [1976]: Fitting Johnson Curves by Moments. *Journal of the Royal Statistical Society*. Series C (Applied Statistics). 25. évf. 2. sz. 180–189. old.

HÖNSCHHOVÁ, E. [2008]: *Estimation of the Scale Parameter in Burr Distribution*. ROBUST 2008 Poster Section. Szeptember 8–12. Prbylina.

JEFFREYS, H. [1948]: *Theory of Probability.* Oxford University Press. Oxford.

JOHNSON, N. L. [1949]: Systems of Frequency Curves Generated by Methods of Translation. *Biometrika*. 36. évf. 1–2. sz. 149–176. old.

JOHNSON, N. L. [1954]: Systems of Frequency Curves Derived from the First Law of Laplace. *Trabajos de Estadistica.* 5. évf. 283–291. old.

JOHNSON, N. L. [1965]: Tables to Facilitate Fitting S_U Frequency Curves. *Biometrika*. 52. évf. 3–4. sz. 547–558. old.

JOHNSON, N. L. [1974]: Extensions and Corrections to 'Tables to Facilitate Fitting S_U Frequency Curves'. *Biometrika*. 61. évf. 1. sz. 203–205. old.

KARIAN, Z. – DUDEWICZ, E. [2000]: *Fitting Statistical Distributions: The Generalized Lambda Distribution and Generalized Bootstrap Methods*. CRC Press. Boca Raton.

KENDALL, M. G. – STUART, A. [1977]: The Advanced Theory of Statistics. *Distribution Theory*. Vol. 1. Charles Friffin & Company. London.

KIM, T-H. – WHITE, H. [2004]: On More Robust Estimation of Skewness and Kurtosis. *Finance Research Letters*. 1. évf. 56–73. old.

LAKHANY, A. – MAUSSER, H. [2000]: Estimating the Parameters of the Generalized Lambda Distribution. *Algo Research Quarterly*. 3. évf. 3. sz. 47–58. old.

LEE, P. M. [2009]: *Bayesian Statistics: An Introduction*. Wiley. New York.

MAGE, D. T. [1980]: An Explicit Solution for S_B Parameters Using Four Percentile Points. *Technometrics*. 22. évf. 247–251. old.

PEARSON, K. [1893]: Contributions to the Mathematical Theory of Evolution. *Proceedings of the Royal Society of London.* 54. köt. 329–333. old.

PEARSON, K. [1895]: Contributions to the Mathematical Theory of Evolution, II: Skew Variation in Homogeneous Material. *Philosophical Transactions of the Royal Society of London*. 186. köt. 343–414. old.

PEARSON, K. [1901]: Mathematical Contributions to the Theory of Evolution, X: Supplement to a Memoir on Skew Variation. *Philosophical Transactions of the Royal Society of London*. Series A. Containing Papers of a Mathematical or Physical Character. 197. köt. 443–459. old.

PEARSON, K. [1916]: Mathematical Contributions to the Theory of Evolution, XIX: Second Supplement to a Memoir on Skew Variation. *Philosophical Transactions of the Royal Society of London*. Series A. Containing Papers of a Mathematical or Physical Character. 216. köt. 429–457. old.

PEARSON, E. S. – HARTLEY, H. O. [1972]: *Biometrika Tables for Statisticians*. Vol. 2. University Press. Cambridge.

RAMBERG, J. S. ET AL. [1979]: A Probability Distribution and Its Uses in Fitting Data. *Technometrics*. 21. évf. 2. sz. 201–214. old.

RAMBERG, J. S. – SCHMEISER, B. W. [1972]: An Approximate Method for Generating Symmetric Random Variables. *Communications of the ACM*. 15. évf. 11. sz. 987–990. old.

RAMBERG, J. S. – SCHMEISER, B. W. [1974]: An Approximate Method for Generating Asymmetric Random Variables. *Communications of the ACM*. 17. évf. 2. sz. 78–82. old.

RAYNER, G. D. – MACGILLIVRAY, H. L. [2002]: Numerical Maximum Likelihood Estimation for the g-and-k and the Generalized g-and-h Distributions. *Statistics and Computing*. 12. évf. 57–75. old.

RODRIGUEZ, R. N. [1977]: A Guide to the Burr Type XII Distributions. *Biometrika*. 64. évf. 1. sz. 129–134. old.

SLIFKER, B. K. – SHAPIRO, S. S. [1980]: The Johnson System: Selection and Parameter Estimation. *Technometrics*. 22. évf. 239–246. old.

SU, S. [2005]: A Discretized Approach to Flexibly Fit Generalized Lambda Distributions to Data. J*ournal of Modern Applied Statistical Methods*. 4. évf. 2. sz. 408–424. old.

TADIKAMALLA, P. R. – JOHNSON, N. L. [1980]: *Systems of Frequency Curves Generated by Transformations of Logistic Variables.* Kézirat.

TADIKAMALLA, P. R. [1980]: A Look at the Burr and Related Distributions. *International Statistical Review / Revue Internationale de Statistique*. 48. évf. 3. sz. 337–344. old.

TADIKAMALLA, P. R. – JOHNSON, N. L. [1982]: Systems of Frequency Curves Generated by Transformations of Logistic Variables. *Biometrika*. 69. évf. 2. sz. 461–465. old.

TUENTER, H. J. H. [2001]: An Algorithm to Determine the Parameters of S_U-curves in the Johnson System of Probability Distributions by Moment Matching. *Journal of Statistical Computation and Simulation*. 70. évf. 4. sz. 325–347. old.

TUKEY, J. W. [1960]: *The Practical Relationship Between the Common Transformations of Percentages of Counts and of Amounts.* Technical Report 36. Statistical Techniques Reasearch Group. Princeton University. Princeton.

TUKEY, J. W. [1977]: *Modern Techniques in Data Analysis.* Regional Research Conference. Június 13–17. North Dartmouth, MA.

WHEELER, R. E. [1980]: Quantile Estimators of Johnson Curve Parameters. *Biometrika*. 67. évf. 3. sz. 725–728. old.

Summary

In simulational studies, it is often necessary to generate random numbers coming from distributions that have specified properties. If a well-known, typical distribution is used, the necessary steps can be performed easily, and they are included in statistical program packages. However, if we need distributions that have properties considered to be parameters, such as arbitrarily set moments, we might face problems. Now we present and examine a few solutions for this problem (Pearson-, Johnson-distribution families, Generalized λ-distribution, Burr XII, Tukey "g-and-h" and Fleishman transformation), with their limits of application, and an analysis of the questions that arise when fitting them.

Regressziós modellek becslése és tesztelése Excel-parancsfájl segítségével (szoftverismertetés)*

Kehl Dániel,
a Pécsi Tudományegyetem egyetemi tanársegéde
E-mail: kehld@ktk.pte.hu

Dr. Sipos Béla,
a Pécsi Tudományegyetem egyetemi tanára
E-mail: sipos@ktk.pte.hu

A regressziószámítás az egyik legegyszerűbb és leggyakrabban alkalmazott ökonometriai módszer. A szerzők a lineáris regressziós analízis metódusához kidolgoztak egy Excel-környezetű parancsfájt, amelynek részletes használati-értelemezési útmutatója e cikk. Az Excel-parancsfájl felhasználási lehetőségét példák is illusztrálják. A programcsomag a statisztikai elemzés és modellezés graduális oktatásának hasznos eszköze lehet.

A parancsfájl letölthető a PTE-KTK honlapjáról (Excel-parancsfájlok felhasználása statisztikai elemzésekhez, kézikönyv és a BSC.zip és MSC.zip Excel-parancsfájlok, szám szerint 35, köztük az ismertetésre kerülő regresszio.xls).[1]

TÁRGYSZÓ:
Regressziószámítás.
Statisztikai módszertan.
Excel.

* A tanulmányban előforduló esetleges hibákért kizárólag a szerzőket terheli felelősség.
[1] http://www.gmi.ktk.pte.hu/index.php?mid=33#SiposB.

A valós méretű statisztikai modellek, ezen belül a többváltozós regressziós feladatok megoldása kézi számításokkal általában nem, vagy csak nehezen végezhető el. A számítógépes feldolgozás lehetősége azonban új utakat nyitott meg a statisztika tudományában is. Napjainkban a számolási igény – a személyi számítógépek megjelenése és elterjedése miatt – már nem jelent különösebb akadályt, a számítások megkönnyítésére matematikai-statisztikai és ökonometriai szoftverek léteznek.

A jelenleg legnépszerűbb irodai programcsomag, a Microsoft Office változata 1990-ben jelent meg. A Microsoft Office (*Baczoni* [2007], *Bártfai* [2002]) és ezen belül az MS Excel (továbbiakban Excel) világviszonylatban és Magyarországon is széleskörűen alkalmazott szoftver. E program sok statisztikai műveletet képes elvégezni, és az alapfunkciókon túl, függvények segítségével felépíthetők a bonyolultabb statisztikai és ökonometriai módszerek is. További előny, hogy a módszerek, a felhasznált képletek alakíthatók, az adott feladat megoldásához testre szabhatók, láthatóvá és követhetővé válnak a részeredmények és a mellékszámítások. Az Excel – a speciális statisztikai szoftverekhez hasonlóan, de messze nem olyan részletességgel – a statisztika módszertanának nagy részét felöleli beépített modulja (Analysis ToolPak) segítségével, de több apróbb hiba (például rossz vagy félreérthető magyarra fordítás) és hiányosság is a sajátja. Az említett félrefordításoknál nagyobb hibák is megfigyelhetők, melyek a program korábbi verzióiban csakúgy megtalálhatók, mint a legújabbakban: a következtetéses statisztikában oly fontos eloszlások esetén némely speciális esetben hibás, félrevezető értékeket szolgáltat. A témakör bőséges irodalommal rendelkezik, tanulmányunkban csak utalunk *Knüsel* ([1998], [2002], [2005]), *McCullough* és *Wilson* ([1999], [2002]) és *Yalta* [2008] munkáira, melyekből kimerítő „hibalista" meríthető. Az említett, több éve ismert hiányosságokat a jelentős tudományos kritika ellenére sem javították még ki. Ugyanakkor az alkalmazás kétségtelen és messze legfontosabb előnye, hogy az Office-csomag elterjedése miatt szinte mindenhol megtalálható.

Megemlítjük továbbá azt a fontos tényt, hogy a statisztika oktatásában Magyarországon az egyetemeken és főiskolákon az Excel, mint táblázatkezelő szoftver elterjedt, főként könnyű elérhetősége okán (lásd e témában *Rappai* [2001]. Ismereteink szerint csak az Excel alapszolgáltatásainak használata terjedt el az oktatásban és az üzleti életben Magyarországon (*Balázsné Mócsai–Csetényi* [2003], *Jánosa* [2005]), pedig a program ennél többre képes, lehet batch file-okat, kötegelt parancsállományokat (a továbbiakban parancsfájlokat, illetve programokat) készíteni.

Tanulmányunkkal kapcsolódni kívánunk a *Rappai Gábor* által indított szakmai beszélgetéshez, ami a statisztikaoktatás átalakulásával, átalakításával foglalkozik. Az informatikai támogatottsággal és az Excel felhasználásával kapcsolatban Rappai a

következőket mondja: „meggyőződésem szerint a legszélesebb körben rendelkezésre álló támogatóeszköz használata a legindokoltabb" (*Rappai* [2008] 840. old.). A modernizáció jelentőségére hívja fel a figyelmet *Kovács Péter* [2008b] tanulmánya is, aki a Szegedi Tudományegyetemen bevezetett tanterven keresztül mutatja be a szegedi modellt, ami szintén erősen támaszkodik az Excelre. Úgy gondoljuk, hogy az általunk felvázolt Excel-alapú oktatás – melynek az egyik szelete a bemutatandó parancsfájl – az egyik, természetesen nem kizárólagos irány lehet a jövőben. A szakmai közösség tagjait továbbra is bíztatjuk tapasztalataik és javaslataik megtételére.

Rátérve az alkalmazási lehetőségekre, véleményünk szerint az adatelemzés öt szintje oldható meg az Excellel.

Az *első szint* az, amikor a Függvény beszúrása varázslót (ikont) használjuk, tehát beépített statisztikai, matematikai és trigonometriai, mátrix, adatbázis stb. függvényeket alkalmazunk. A *második szint*, amikor az Eszközök/Adatelemzés[2] menüpont szolgáltatásait (például korrelációanalízis, regresszió stb.) használjuk. A *harmadik szint*, amikor magunk írunk konkrét adatsorhoz vagy adatsorokhoz képleteket, mivel nem minden feladathoz áll rendelkezésre beépített függvény. A *negyedik szint* az, amikor parancsfájlokat készítünk – vagyis a harmadik szintet általánosítjuk –, amelyek segítségével az általunk megadott adatbázis terjedelméig új adatbázisok felhasználásával korlátlan számban végezhetünk számításokat a programozott képletekkel, illetve függvényekkel. Gyakran igen sok számítást kell elvégezni. Eben az esetben az idővel való takarékos gazdálkodás a cél, mert a harmadik szintnél egy feladatsor számításainak elvégzése sokszor több óra vagy nap is lehet, amit a parancsfájlok felhasználásával egy perc alatt el lehet végezni. Az *ötödik szint* az, amikor a feladat a hagyományos módon nem oldható meg. Erre példa a CES (constant elasticity of substitution – konstans helyettesítési rugalmasságú függvény) termelési függvény, ahol a változók száma több mint a rendelkezésre álló egyenletek száma. A feladat a legjobban illeszkedő függvény paramétereinek megkeresése.[3] A logisztikus és egyéb speciális trendfüggvények esetében a függvényeket nem lehet lineárisra transzformálni, a cél megkeresni azokat a paramétereket, amelyek mellett az illesztés a legpontosabb.[4] A logisztikus regressziós függvények sem linearizálhatók, de iterációs eljárással a paraméterek becsülhetők, meghatározható olyan függvény, ahol a többszörös determinációs együttható elégségesen nagy. Az Excel a Visual Basic for Applications (VBA) felhasználásával programozható, így ezek a feladatok megoldhatók.

A szoftverek alkalmazásának egyik legnagyobb problémáját abban látjuk, hogy a számítási lépések nem követhetők, a felhasználó nem minden esetben érzékeli, hogy az adatok és azok kismértékű változásai hogyan hatnak az eredményre. Az általunk

[2] Az Eszközök/Bővítménykezelő/Data Analysis Toolpak hozzáadása után.

[3] Lásd a CES-függvény becslését, ha három normálegyenlet áll rendelkezésre és a becsült paraméterek száma öt. A parancsfájl: ces1.xls

[4] *Kehl–Sipos* [2009] és logisztikusregresszio.xls.

elkészített parancsfájlok[5] – véleményünk szerint – kiküszöbölik ezt a hiányosságot: egyetlen cella, vagy vezérlőelem (Checkbox, legördülő menü stb.) módosításakor nyomon követhetjük az eredmények változását.

A munkalapokat egységes szerkezetben építettük fel. A változtatható, illetve megadható vagy megadandó adatokat sárga mezők jelölik, az eredményeket pedig egységes struktúrában jelenítettük meg. A végeredmények és az egyes cellák számításához használt képletek valamennyi esetben láthatók.

1. A regressziós Excel-parancsfájl működésének bemutatása

A regressziós modell készítésének (*Hajdu et al.* [1994–1995] 110–111. old.) első lépése a specifikáció, ami alatt a jelenséget leíró, modellben szereplő eredmény- és magyarázóváltozók kiválasztását, valamint a függvény konkrét formájának meghatározását értjük. Fontos szerepet játszik a specifikáció szakaszában az adatbázis, amelynek minősége, szerkezete nagymértékben befolyásolja a folyamat eredményességét. A gyakorlati munkában idősoros és keresztmetszeti adatokkal dolgozhatunk, ennek a modell feltételrendszerének ellenőrzésekor lesz jelentősége. Panel adatbázisokkal jelen anyagunkban nem foglalkozunk.

A specifikáció munkafázisának lezárása után a számításokat a regresszio.xls parancsfájllal lehet elvégezni. Ennek fontosabb lépései a következők:

1. A regressziós paraméterek becslése a klasszikus legkisebb négyzetek módszerével, melynek feltételei:

a) a magyarázóváltozók nem sztochasztikusak, tehát mérési hibát nem tartalmaznak és lineárisan függetlenek (multikollinearitás hiánya),

b) a hibatényezők (hibatagok, reziduumok) várható értéke 0, varianciájuk konstans, normális eloszlásúak és nem autokorreláltak.

2. A modell feltételrendszerének ellenőrzése. Ez a munkafázis visszahat mind a specifikációra, mind a paraméterbecslésre. Ebben a munkaszakaszban a modellező megállapítja, hogy adott szignifikanciaszint mellett mennyire fogadható el a modell. A fontosabb hipotézisellenőrzések: a regressziós modell paramétereinek globális és parciális tesztelése (a paraméterbecslés pontosságának vizsgálata, a paraméterek standard hibája, konfidencia intervalluma stb.), valamint a reziduumok vizsgálata: az autokorreláció és a homoszkedaszticitás tesztje, a ma-

[5] Meg kívánjuk jegyezni, hogy elsősorban oktatási célból, de a gyakorlati alkalmazásokat is segítendő, 36 Excel parancsfájlt dolgoztunk ki, ezek egy része az alapképzésben használható fel, más része az ökonometriai jellegű tárgyakban használható. Az internetes hozzáférést biztosítottuk.

gyarázóváltozók közötti kapcsolat szorossága, a multikollinearitás ellenőrzése. A próbákkal nyert információk alapján döntést lehet hozni a modell esetleges megváltoztatásáról vagy a becslési módszer módosításáról. Ezek a döntések természetesen visszahatnak a specifikációra és indokolt esetben az egész eljárás (specifikáció, becslés, hipotézisellenőrzés) megismétlését igényelhetik.

3. A regressziós modell felhasználása elemzésre és előrejelzésre.

4. A verifikálás, aminek során a modellt szembesítjük a valósággal.

A program a bemutatásra kerülő regressziószámítást maximum 16 magyarázóváltozó és 2000 megfigyelés esetében végzi el.[6] A programban a munkalapokon megjelenő színeknek jelentése van. A halványsárga cellák változtathatók, itt történik meg az adatok bevitele, a kívánt szignifikanciaszint beállítása, valamint a becslés/előrejelzés alapadatainak megadása. A tesztek végeredményei színes számokkal jelennek meg a fájlban. A modell ellenőrzésénél háromféle színt alkalmaztunk, a zavaró eredmények piros, a megfelelők zöld, a nem egyértelmű kimenetelek kék színnel jelennek meg.

Tanulmányunkban az elméleti háttér részletes ismertetésétől eltekintünk (kivéve a homoszkedaszticitás tesztjeit, ahol a felsőoktatásban ritkábban alkalmazott teszteket ismertetjük), mert az az Irodalom részben felsorolt szak-, illetve tankönyvekben, tanulmányokban megtalálható, célunk csupán a szoftver bemutatása, gyakorlati, oktatási célokra való közreadása.

2. Munkalapok

A továbbiakban a munkalapok tartalmát ismertetjük, és mivel a program képes az autokorreláció és a homoszkedaszticitás tesztelésére is, ezért két példán keresztül szemléltetjük a számításokat. Az első példa idősoros adatállomány, a második pedig keresztmetszeti, az adatállományokat elhelyeztük a regresszio.xls parancsfájlban.

2.1. Az Adat munkalap

Az Adat munkalap két nagyobb egységből áll. A bal oldali, sárgával jelölt terület az adatok bevitelére szolgál, itt kell rögzíteni az aktuális adatállományt. Új adatok bevitele előtt a megjelenő mintafeladat adatállományát az Adatok törlése gombra való kattintással törölhetjük. Az új adatokat kell beilleszteni annak érdekében, hogy a

[6] A magyarázóváltozókra érvényes korlát az Excel sajátja. A megfigyelésekre vonatkozó korlát igény szerint bővíthető, a korlátozás oka a gyors számítási sebesség megtartása.

parancsfájl formátuma megmaradjon. A jobb oldali egység a regressziós modell alapstatisztikáit közli:

– Regressziós statisztika: R – többszörös korrelációs együttható; R^2 – többszörös determinációs együttható; $\tilde{R}^2$ – korrigált determinációs együttható; s – modell standard hibája; n – megfigyelések száma.

– Varianciaanalízis: a többváltozós regressziós modell varianciaanalízis táblája.

– Regressziós együtthatók: együtthatók értékei és standard hibái, t-értékei, p-értékei, valamint konfidencia intervallumai (tetszőleges megbízhatósági szinten); változók bevonásáról/kihagyásáról döntő jelölőnégyzetek.

A formátum követi az Excel adatelemző menüpontja által használtat, azzal a különbséggel, hogy az egyes cellák függvényeket tartalmaznak, így az adatok megváltozásának hatása azonnal nyomon követhető az eredményeken. Szintén eltérés a beépített funkcióhoz képest, hogy az eredeti adatok meghagyása mellett is kihagyhatunk, illetve újra bevonhatunk változókat a paraméterek soraiban található jelölőnégyzetek segítségével.

A varianciaanalízis tábla segítségével a modell globális próbáját végezhetjük el. A hipotézisrendszerről való döntés – didaktikai okokból – két módon is elvégezhető: a tetszőlegesen beállítható szignifikanciaszinthez tartozó kritikus érték, valamint a p-érték alapján.

A gyors parciális tesztelés lehetőséget biztosít a backward eliminációs módszer alkalmazására. A módszer lényege, hogy az első lépésben olyan regressziós függvényt határozunk meg, amely az összes megfigyelt magyarázóváltozót tartalmazza, majd lépésenként kihagyjuk azokat a változókat, amelyek nem járulnak hozzá szignifikánsan a reziduális négyzetösszeg csökkentéséhez. A változók szelektálásához a p-értékeket használjuk: ha ennek értéke magasabb, mint amit megengedünk (például 0,05), akkor elfogadjuk a nullhipotézist, a regressziós paraméter nem különbözik szignifikánsan nullától. Amennyiben több változó p-értéke is magasabb a kívántnál, úgy a legmagasabb értékkel rendelkező változót hagyjuk ki elsőként. Az eliminációt addig folytatjuk, míg valamennyi bevont paraméter szignifikáns nem lesz.

A változók szelektálását természetesen elvégezhetjük a multikollinearitás vagy a homoszkedaszticitás parciális tesztjei alapján is.

2.2. A Mátrix munkalap

A Mátrix munkalapon a többváltozós regressziószámítással kapcsolatos mátrixok, valamint az ezekhez tartozó statisztikák találhatók meg. A mátrixok maximális

mérete a magyarázóváltozók maximális számával van összhangban. A munkalapon megjelenő mátrixok a következők:

– A „C" oszloptól kezdődően rendre: teljes korrelációs mátrix (valamennyi változóra); bevont korrelációs mátrix (a meghagyott magyarázó változókra, ha valamennyi magyarázóváltozó szerepel a végleges modellben, akkor megegyezik az előző mátrix tartalmával); bevont korrelációs mátrix inverze; determinációs együtthatók a teljes adatmátrixra; determinációs együtthatók a bevont adatmátrixra; bevont változók parciális korrelációit tartalmazó háromszögmátrix; $\mathbf{X}^T\mathbf{X}$ mátrix a teljes adathalmazra; $\mathbf{X}^T\mathbf{X}$ mátrix a bevont változókra; $(\mathbf{X}^T\mathbf{X})^{-1}$ a bevont változókra.

– A „V" oszloptól kezdődően rendre: teljes korrelációs mátrixhoz tartozó *t*-értékek; bevont korrelációs mátrixhoz tartozó *t*-értékek; bevont magyarázóváltozók korrelációs mátrixának inverze; bevont változók parciális korrelációihoz tartozó *t*-érték.

– Az „AP" oszloptól kezdődően rendre: teljes adatmátrixra a sajátértékek és sajátvektorok; bevont adatmátrixra a sajátértékek és sajátvektorok; a sajátértékek megoszlási és kumulált megoszlási viszonyszámai; főkomponenssúly-mátrix; főkomponenssúlyok négyzete.

A felsorolt mátrixok közül több önmagában is fontos információkat hordoz a regresszióval kapcsolatban, néhány kiszámítása pedig a további vizsgálatok miatt szükséges. Didaktikai okokból mindegyik mátrix bemutatását szükségesnek tartottuk.

2.3. A Maradék munkalap

A Maradék munkalapon az aktuális modell empirikus maradékaiból képzett oszlopvektorok találhatók meg, valamint lehetőség van becslés, előrejelzés elvégzésére is. A munkalapon található oszlopvektorok a következők:

– $\mathbf{y}$ – a vizsgált eredményváltozó értékeinek vektora;

– $\hat{\mathbf{y}}$ – az eredményváltozó értékeinek becsült vektora, a bevont magyarázó változókkal történő pontbecslés;

– $\hat{\mathbf{y}}^2$ – az eredményváltozó becsült értékeinek négyzete;

– e – empirikus reziduumok (hibatényezők, hibatagok, maradékok) $\left(e = \mathbf{y} - \hat{\mathbf{y}}\right)$;

– e_{t-p} – az empirikus maradék p-vel (p = 1,2,…,12) késleltetett értékei (p nagyságát az Autokorreláció munkalapon lehet megadni, jellemzően $p = 1$).

Előrejelzést (idősorok esetén), illetve pontbecslést (keresztmetszeti adatbázisok esetén) a „H" oszloptól kezdődően készíthetünk, a sárga mezőkbe a magyarázóváltozók értékeit kell beírni. Technikai okokból valamennyi (bevont és be nem vont) változóhoz értékeket kell megadni, ezekből csak azokat fogja a program figyelembe venni, amelyek a bevont változókhoz tartoznak. Egyszerre maximum 20 becslés, illetve előrejelzés hajtható végre. A helyesen kitöltött magyarázóváltozó-értékekhez tartozó becsült eredményváltozó-érték a „H" oszlopban olvasható le.

2.4. A Multikollinearitás munkalap

A Multikollinearitás munkalap a magyarázóváltozók összefüggésének problémáját vizsgálja.[7] Az elvégezhető tesztek közül nem építettük be a programba valamennyit, csupán az oktatásban gyakran alkalmazott, általánosan elterjedt próbákat. A beépített tesztek és módszerek a következők:

– a multikollinearitás globális tesztelése: χ^2 – próba; kondícióindexek és kondíciószám (gyökös formula); Petres-féle RED-mutató (*Kovács–Petres–Tóth* [2004, 2005]);

– a multikollinearitás lokalizálása: parciális korrelációs együtthatók tesztelése; *F*-próba; VIF-mutató (variancia infláló faktor); tolerancia-mutató; a multikollinearitás kiküszöbölése: főkomponens regresszió.[8]

2.5. Az Autokorreláció munkalap

Az autokorreláció (*Kovács I.* [1977] 605. old.) mértékét a reziduális autokorrelációs együtthatóval mérhetjük. A *p*-ed rendű (*p* időegységgel késleltetett, a parancsfájl esetében $p = 1, 2, \ldots, 12$) elméleti autokorrelációs együtthatót az egymástól *p* időegységnyi távolságra álló maradéktagok korrelációs együtthatójaként becsülhetjük.

A gyakorlatban az elsőrendű autokorrelációs együtthatót $(p = 1)$ szoktuk tesztelni. A fájlban közöltünk több késleltetésre vonatkozó adatot is, amire például szezonalitást mutató adatsorok esetén lehet szükség. Az Excel-parancsfájlban az eredeti, és nem a közelítő *p*-ed rendű reziduális autokorrelációs együtthatóval számoltunk, majd azt Student-féle *t*-próba felhasználásával teszteltük. A program kiszámítja a Durbin–Watson-mutató közelítő értékét és teszteli is azt. A kritikus érté-

[7] A multikollinearitás témaköre jelentős irodalommal rendelkezik, a regressziós modellek becslése és alkalmazása során jelentkező probléma legfrissebb magyar nyelvű összefoglalóját *Kovács Péter* [2008a] írása adja.

[8] A számítások elvégzéséhez szükség van a mátrix.xls parancsfájlra.

keket az Excel nem szolgáltatja, így azokat 1 és 5 százalékos szignifikanciaszintekre táblázatból keresi ki a program.

Általánosságban elmondhatjuk, hogy az autokorreláció jelenléte mellett készített paraméter- és pontbecslések ugyan torzítatlanok maradnak, de nem lesznek hatásosak. Különösen óvatosan kell kezelni az autokorrelált modellt, ha segítségével előrejelzéseket kívánunk készíteni. Autokorrelált modellek esetében az együtthatók standard hibái torzítottak, így sem a standard hibákhoz kapcsolódó próbák, sem az előrejelzésekhez kapcsolódó konfidenciaintervallumok nem használhatók fel.

A program ábrázolja e_{t-p} függvényében az e_t alakulását. Az ábra alapján vizuálisan is következtethetünk az autokorreláció létére, illetve hiányára.

2.6. A Homoszkedaszticitás munkalap

A keresztmetszeti adatok esetében a hibatényező varianciájának állandóságát tesztjük. Ha konstans a hibatényező varianciájának várható értéke, akkor:

$$E\left(\varepsilon_i^2\right) = \sigma^2 \quad i = 1, 2, \ldots, n .$$

Keresztmetszeti adatok esetén homoszkedaszticitás szempontjából is tesztelnünk kell a modelleket, hiszen elméleti feltétel, hogy a hibatényező varianciája állandó legyen (*Pintér* [1991] 18. old.).

A nullhipotézis:

$$H_0 : \sigma_1^{\,2} = \ldots = \sigma_n^{\,2} .$$

Az alternatív hipotézis:

$$H_1 : \sigma_l^2 \neq \sigma_m^2 ,$$

ahol $l, m = 1, 2, \ldots, n \quad (l \neq m)$.

A nullhipotézis azt fogalmazza meg, hogy a hibatényező szórásnégyzetei (varianciái) állandók. A nullhipotézis teljesülése egyben azt is jelenti, hogy a modell homoszkedasztikus, az alternatív hipotézis a heteroszkedaszticitás feltételezését szimbolizálja. A heteroszkedaszticitás jelensége esetén a regressziós együtthatók becslése torzítatlan, ugyanis továbbra is feltesszük, hogy a hibatényező várható értéke nulla. Ugyanakkor a paraméterek varianciájára vonatkozó becslés nem lesz hatásos,[9] a paraméterek standard hibái torzítottak, használatuk megkérdőjelezhető, a se-

[9] Ez azt jelenti, hogy a klasszikus legkisebb négyzetek módszere (KLNM, Ordinary Least Squares – OLS) alkalmazása esetén a becslések ebben az esetben nem lesznek hatásosak, vagyis található egy másik torzítatlan lineáris becslés, aminek kisebb a varianciája, mint az KLNM (OLS)-becslésnek (*Ramanathan* [2003] 365–366. és 397–398. old.)

gítségükkel elvégzett próbák (például *t*- és *F*-próbák) és becslések félreinformálhatnak.

2.6.1. Globális (csoportos) BPG, Glejser és KB-próba

A Breusch–Pagan–Godfrey (BPG) és a Glejser-próba esetében a nullhipotézis megegyezik az előzőkben leírtakkal, a hipotézisrendszer általánosabb formában (*Glejser* [1969] 316–323. old., *Godfrey* [1978] 227–236. old., *Breusch–Pagan* [1979] 1287–1294. old., továbbá *Ramanathan* [2003] 367–369. old, *Pintér* [1991] 21–24. old., *Gujarati* [2003] 411–412. old., Maddala [2004] 244–246. old.):

$$H_0 : \sigma_1^2 = ... = \sigma_n^2$$

$$H_1 : E\left[f\left(\varepsilon_i\right)\right] = \sigma^2 \left[h(\mathbf{Z}\boldsymbol{\alpha} + \mathbf{v})\right],$$

ahol

f – az eredeti reziduumok függvénye (például abszolút értéke, négyzete, logaritmusa);

h – a magyarázóváltozók függvénye (a függvény alakja lineáris, hatványkitevős, exponenciális);

$\mathbf{Z}$ – a heteroszkedaszticitást magyarázó változók $n \times \left(k+1\right)$ típusú mátrixa;

$\boldsymbol{\alpha}$ – a véletlent becslő modell $\left(k+1\right) \times 1$ típusú paramétervektora;

$\mathbf{v}$ – $n \times 1$ típusú, véletlen elemeket tartalmazó vektor.

F-próbával teszteljük a nullhipotézist, aminek elfogadása esetén a modell homoszkedasztikus, elutasítása esetén pedig heteroszkedasztikus.

A globális próbák a következők:

– Glejser-próba:

$$\left|e_i\right| = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + ... + \alpha_k x_k + v_i \, .$$

A regresszio.xls fájlban a pótlólagos regresszió többszörös determinációs együtthatója: $R^2\left(\left|e\right|; x\right)$.

– Breusch–Pagan–Godfrey (BPG)-próba:

$$e_i^2 = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + ... + \alpha_k x_k + v_i \, .$$

A regresszio.xls fájlban a pótlólagos regresszió többszörös determinációs együtthatója: $R^2\left(e^2;x\right)$.

– Koenker–Bassett (KB)-próba (*Gujarati* [2003]):

$$e_i^2 = \alpha_0 + \alpha_1 \hat{y}_i^2 + v_i .$$

A regresszio.xls fájlban a pótlólagos regresszió többszörös determinációs együtthatója $R^2\left(e^2;\hat{y}^2\right)$.

A képletek jelölései:

k – az eredeti regressziós függvényben a magyarázóváltozók száma;
$i = 1,2,\ldots,n$: a megfigyelések száma;
$|e_i|$ – az eredeti modell reziduális változójának abszolút értéke;
e_i^2 – az eredeti modell reziduális változójának négyzete;
$\hat{y}_i^2$ – az eredeti függvénnyel becsült eredményváltozó négyzete;
α_j – a becsült paraméterek $\left(j = 0,1,2,\ldots,k\right)$;
v_i – a pótlólagos regresszió reziduális változója.

A regresszió paramétereinek együttes szignifikanciája a globális *F*-próba segítségével mindegyik bemutatott teszt esetében vizsgálható.

2.6.2. A homo- és heteroszkedaszticitás vizsgálata

A Glejser- és a BPG-próba lehetővé teszi a heteroszkedaszticitás lokalizálását. Amennyiben feltételezzük, hogy a magyarázóváltozók lineáris függvényei a reziduális változók abszolút értékei vagy a négyzetei, akkor felírható magyarázóváltozónként egy-egy pótlólagos regressziós egyenlet.

A pótlólagos, *j*-edik magyarázóváltozóra vonatkozó regressziós egyenletek a következők:

– Glejser-próba esetén:

$$|e_i| = \alpha_0 + \alpha_1 x_{ji} + v_i ,$$

– BPG-próba esetén:

$$e_i^2 = \alpha_0 + \alpha_1 x_{ji} + v_i ,$$

ahol x_{ji} a *j*-edik magyarázóváltozó *i*-edik értéke.

A regressziós együtthatót (meghatározó szerepe az α_1 együtthatónak van) a Student-féle *t*-próbával teszteljük.

3. Gyakorlati alkalmazások bemutatása idősoros és keresztmetszeti adatok alapján

A regresszio.xls parancsfájl minden esetben közli az Autokorreláció és a Homoszkedaszticitás munkalapokon a számításokat. Az autokorreláció idősoros adatoknál jelentkezik, ahol az adatok sorrendje kötött. A keresztmetszeti adatok sorrendje változtatható, ebben az esetben a homoszkedaszticitást szoktuk vizsgálni. Megjegyezzük, hogy keresztmetszeti adatoknál is előfordul, hogy a szomszédos hibatagok korrelálnak egymással, amit térbeli korrelációnak neveznek. Az autokorreláció vizsgálatánál az ökonometriai szakirodalomban ettől eltekintenek, kizárólag az idősorok hibatagjainak vizsgálata tartozik e témakörbe (*Maddala* [2004] 273–274. old., *Ramanathan* [2003] 361–363. és 399–400. old.; *Gujarati* [2003] 401–403. és 441–443. old.) A maradékváltozó (reziduális változó) vizsgálatánál tehát lényeges kérdés, hogy idősoros vagy keresztmetszeti adatokkal dolgozunk-e. Idősoros adatbázis esetén az autokorrelációt, míg keresztmetszeti adatoknál a homoszkedaszticitást teszteljük. Ennek megfelelően két példát mutatunk be, mindkettő valós magyarországi adatokat tartalmaz.

1. Idősoros példa

Az 1985 és 2008 közötti magyarországi cementtermelést és az azt befolyásoló tényezőket vizsgálatuk.[10]A regressziós modell változói: y – cementtermelés (ezer tonna), x_1 – GDP volumenindexe (1985 = 100 százalék); x_2 – épített lakások száma (darab) – x_3 – építőanyag-ipar volumenindexe (1985 = 100 százalék); x_4 – népesség száma (ezer fő).

A rendszerváltozás idején a hazai cement-előállítás megközelítette az évi négymillió tonnát, ezt követőn azonban drasztikusan visszaesett, és 2000-ig közel egymillió tonnával alatta maradt a csúcsévek termelésének, majd 2001-től emelkedett ugyan a kibocsátás, de 2008-ban is közel félmillió tonnával maradt el az 1990-es szinthez képest.

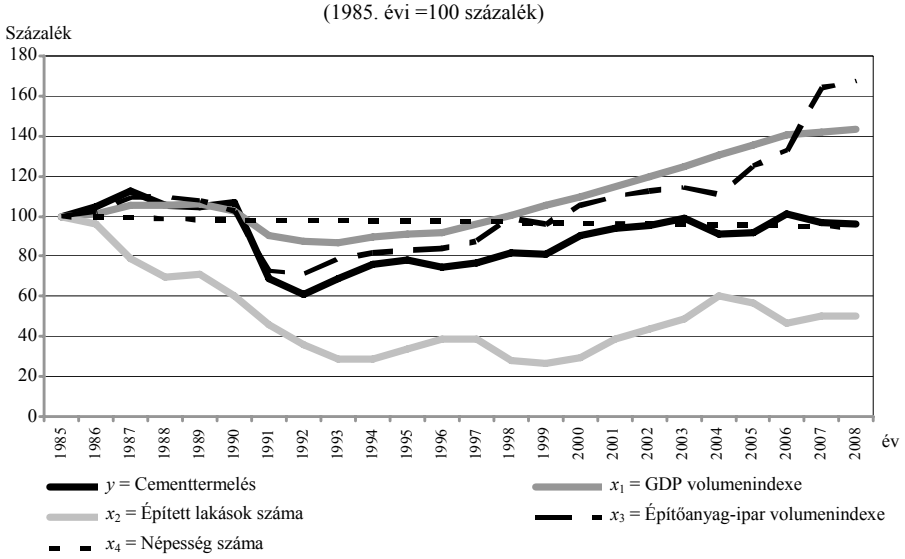
A számítások megkezdése előtt célszerű az adatokat ábrázolni, hogy feltárjuk azok tendenciáit. A cementtermelés és a vizsgált magyarázóváltozók alakulását a következő ábra mutatja. Az ábrakészítés során a vizsgált mutatók arányosságának biztosítása érdekében mindegyik mutatót (tehát a cementtermelést, az épített lakások számát és a népességszámot is) 1985-ös bázison számítottuk.

Az ábra alapján látható, hogy a cementtermelés és a vizsgált magyarázóváltozók sok tekintetben hasonlóan mozognak. A termelés mélypontját a rendszerváltást köve-

[10] Az adatok forrásai: *Polt* [2005] 996. old.; *Hunyadi–Vita* [2008] II. köt. 204. old., CD-melléklet: Adatok8.xls; *KSH* [1985–2005]; *KSH* [1985–2008].

tő években érte el. A magyarázóváltozók közül a népességszám eltérően alakult: Magyarországon a vizsgált időszakban a népességszám folyamatosan csökkent, aminek mértéke a huszonnégy év alatt –5,2 százalék volt. Eltérést mutat az épített lakások számának alakulása is, amely 1985 óta csökkenő tendenciát mutat, kivéve az 1995 és 1997, valamint a 2000 és 2003 közötti időszakot.

1. ábra. A cementtermelés és az azt befolyásoló tényezők alakulása Magyarországon 1985 és 2008 között
(1985. évi =100 százalék)



Vizsgálhatjuk a ciklusok fordulópontjait is, az átlagos periódushossz[11] a cementtermelésnél 3, a GDP volumenindexénél 10, az épített lakások számánál 6, az építőanyag-ipar volumenindexénél 5 év. A népességszám esetében nem voltak fordulópontok.

A termelés elemzése és előrejelzése a regressziószámítás felhasználásával a cementipar esetében arra épült (*Polt* [2005] 996–1000. old.), hogy az építőanyagok és ezen belül a cement termelése, szorosan követi a GDP változását, valamint függhet az épített lakások számának, az építőanyag-ipar teljesítményének és a népesség számának alakulásától is. A népességszám változása és az épített lakások száma közötti kapcsolatot az Egyesült Államok adatbázisán először Kuznets modellezte, kidolgozva a róla elnevezett 15–25 éves építési ciklus elméletét (*Kuznets* [1930]). Az építőanyagok és ezen belül a cement felhasználását az elmúlt években elsődlegesen az építési piac alakulása, pontosabban az infrastruktúra- (autópályák) és a lakásépítés befolyásolta.

[11] A ciklusfordulópontok számítása Excel-parancsfájl felhasználásával.

A számítások eredményei

Varianciaanalízis:

Összetevő	*df*	*SS*	*MS*	*F*-érték	*p*-érték
Regresszió	4	4845009,2	1211252,3	17,1	0,000004
Maradék	19	1347194,2	70905,0		
Összesen	*23*	*6192203,4*			

A varianciaanalízis tábla alapján a nullhipotézist elutasítjuk, tehát van legalább egy olyan magyarázóváltozó, amely szignifikáns hatással rendelkezik, létezik legalább egy nullától eltérő értékű regressziós paraméter.

Regressziós együtthatók:

Együttható	Érték	Standard hiba	*t*-érték	*p*-érték	Alsó 95%	Felső 95%
b_0	–54913,67	22223,78	–2,47	0,0231	–101428,57	–8398,77
b_1	49,15	20,68	2,38	0,0281	5,88	92,42
b_2	–0,01	0,01	–0,89	0,3820	–0,04	0,02
b_3	1,23	6,91	0,18	0,8606	–13,24	15,70
b_4	5,16	2,04	2,54	0,0201	0,90	9,43

A regressziós paraméterek parciális tesztelése: a backward eliminációs módszer alkalmazása alapján először mind a négy magyarázóváltozót bevontuk a modellbe, majd az így meghatározott regressziófüggvényből szelektáltuk azokat a változókat, amelyek nem járulnak hozzá szignifikánsan a reziduális négyzetösszeg csökkenéséhez (*Mundruczó* [1981] 117–118. old.). A változók szelektálásához a *p*-értékeket használtuk. Ennek alapján először az x_3 változót, majd az x_2 magyarázóváltozót hagytuk ki a modellből. Meg kívánjuk jegyezni, hogy szakmailag indokolt lenne a modellben szerepeltetni a két kihagyott változót.

Regressziós együtthatók:

Együttható	Érték	Standard hiba	*t*-érték	*p*-érték	Alsó 95%	Felső 95%
b_0	–37518,27	5884,69	–6,38	0,0000	–49756,15	–25280,39
b_1	38,65	4,62	8,36	0,0000	29,03	48,27
b_4	3,56	0,53	6,67	0,0000	2,45	4,67

A regressziófüggvény tehát:

$$\hat{y} = -37518,27 + 38,65x_1 + 3,56x_4 .$$

A multikollinearitás tesztjei:

A χ^2 globális próba alapján 5 százalékos szignifikanciaszinten van multikollinearitás $\left(\chi^2 = 8{,}18;\ \mathrm{d}f = 1;\ p = 0{,}0042\right)$.

A parciális korrelációs együtthatók alapján számított *t*-statisztika értéke –11,66, a kritikus érték pedig 2,08, a két magyarázóváltozó között van multikollinearitás.

A *p*-értékek is a multikollinearitás létét igazolják. A VIF-mutató értéke 2–5 között van, tehát erős, zavaró a multikollinearitás mértéke.

	y	R^2	*F*-érték	*p*-érték	VIF_j	T_j
R^2x	x_1	0,584	30,86	0,0000	2,40	0,42
R^2x	x_4	0,584	30,86	0,0000	2,40	0,42

A kondícióindexek (*CI* – condition index) a magyarázóváltozók korrelációs mátrixának legnagyobb $\left(\lambda_{\max}\right)$ és *j*-edik $\left(\lambda_j\right)$ $j = 1, 2, \ldots, k$ sajátértékei alapján határozhatók meg (*Kotz et al.* [2006] 1239–1240. old.):

$$CI = \sqrt{\frac{\lambda_{\max}}{\lambda_j}}\ .$$

Ha a legkisebb sajátértéket $\lambda_{\min}$-nel jelöljük, akkor a kondíciószám (CN – condition number):

$$CN = \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}}\ .$$

Ha a magyarázóváltozók lineárisan függetlenek, valamennyi sajátérték egy, akkor a *CN*-mutató értéke is eggyel egyenlő. Minél nagyobb a mutató, annál erősebb a multikollinearitás mértéke. A multikollinearitás mértéke gyenge, ha a $1 < CN < 5$, zavaró, ha $5 < CN < 10$, igen zavaró, ha $CN > 10$.

Esetünkben a kondíciószám 2,734, azaz a mutató szerint gyenge multikollinearitást tapasztalunk a két magyarázóváltozó között.

A Petres-féle RED-mutatót is számszerűsítettük, 76,4 százalékos eredményt kaptunk, a kritikus érték pedig 100 százalék. Ha minden sajátérték egy, akkor $\mathrm{RED}(\%) = 0\%$. Ez azt jelenti, hogy a sajátértékek szorzata, vagyis a magyarázóváltozók korrelációs mátrixának a determinánsa eggyel egyenlő. Ebben az esetben a mátrix ortogonális, nincs multikollinearitás, a magyarázóváltozók függetlenek egymástól. Amennyiben a sajátértékek távolodnak ettől az esettől, akkor a RED-mutató értéke növekszik. A maximális redundancia esetén a mutató értéke 100 százalék.

Ha a számított érték a kritikusnál kisebb, akkor a lineáris regressziós modell illesztése után kapott becsült paraméterek szórásnégyzeteinek az összege, illetve átlaga biztosan véges. Ellenkező esetben a lineáris regressziós modell illesztése után kapott becsült paraméterek szórásnégyzeteinek az összege, illetve átlaga nem biztos, hogy véges, az adatállomány redundáns.

Esetünkben az adatállomány a Petres-féle RED-mutató alapján nem redundáns.

Az autokorreláció tesztelése:

Az elsőrendű reziduális autokorrelációs együttható alapján nincs szignifikáns autokorreláció a modellben:

Autokorreláció rendje	ρ	t	t_{krit}	p-érték
1	0,342	1,707	2,074	0,1018

A népesség a vizsgált időszakban végig csökkent, a GDP volumenindexe pedig – a rendszerváltást követő éveket leszámítva – növekvő trendet mutatott, ezért a két magyarázóváltozó együttes alkalmazása multikollinearitást okozott. Az optimális regressziós egyenes meghatározásához ezért más megoldást kellett keresnünk.

Szakmai indokok alapján új modellt építettünk, és azt kaptuk, hogy a modell globálisan és parciálisan is elfogadható, ha az x_2 és x_3 változókat vonjuk be a modellbe. Nyilvánvaló, hogy az épített lakások számának és az építőanyag-ipar volumenindexének változása (növekedése vagy csökkenése) a cementfelhasználást, és így a termelést is jelentősen befolyásolja. Természetesen befolyásoló tényező a cement export- és importvolumene, de ennek vizsgálatától eltekintettünk. Megállapítható továbbá, hogy a multikollinearitás mértéke és az autokorreláció nem zavaró.

A varianciaanalízis F-próbájához tartozó p-érték ebben az esetben 0,000002, tehát a nullhipotézist elutasíthatjuk.

Regressziós együtthatók:

Együttható	Érték	Standard hiba	t-érték	p-érték	Alsó 95%	Felső 95%
b_0	1476,00	285,95	5,16	0,0000	881,33	2070,67
b_2	0,0204	0,00	4,86	0,0001	0,01	0,03
b_3	10,2928	2,58	4,00	0,0007	4,94	15,65

$$\hat{y} = 1476 + 0,0204x_2 + 10,2928x_3 .$$

A multikollinearitás próbái:

A χ^2 globális próba alapján 5 százalékos szignifikanciaszinten nincs multikollinearitás $\left(\chi^2 = 0,49;\ \mathrm{d}f = 1;\ p = 0,4821\right)$.

A parciális korrelációs együtthatók alapján számított *t*-statisztika –1,77. A kritikus érték 5 százalékos szignifikanciaszinten 2,08, tehát a két magyarázóváltozó között nincs multikollinearitás. A *p*-értékek is a multikollinearitás hiányát igazolják. A VIF-mutató értéke 1 és 2 között van, tehát nem zavaró a hatás.

y	R^2	F-érték	p-érték	VIF_j	T_j
x_2	0,052	1,20	0,2860	1,05	0,95
x_3	0,052	1,20	0,2860	1,05	0,95

A kondíciószám esetünkben 1,26, ami gyenge multikollinearitásra utal.

A Petres-féle RED-mutató:

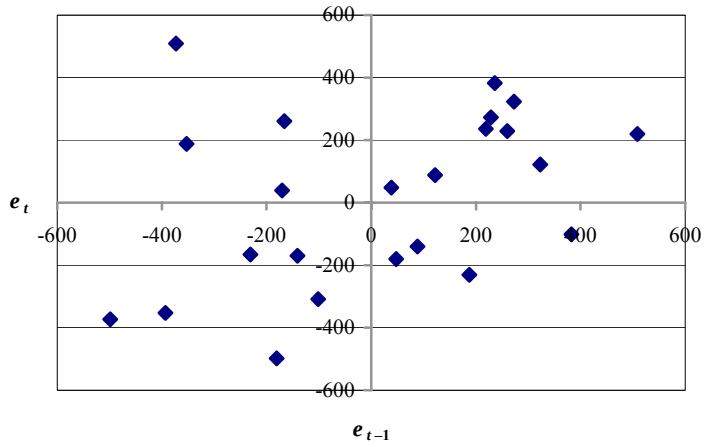
A modell nem redundáns. $\mathrm{RED}(\%) = 22,7\%$, ami azt jelenti, hogy az adott méretű és minimális redundanciájú adatállományhoz képest a hasznos tartalmat hordozó adatok aránya 77,3 százalék, azaz az adatok átlagos együttmozgásának a maximálishoz viszonyított mértéke 22,7 százalék.

Az autokorreláció tesztelése:

Az elsőrendű reziduális autokorrelációs együttható alapján nincs szignifikáns autokorreláció a modellben:

Autokorreláció rendje	ρ	t	t_{krit}	p-érték
1	0,365	1,837	2,074	0,0797

2. ábra. Reziduumok ábrája



A Durbin–Watson-féle teszt eredménye: 1,27, ami a bizonytalansági tartományba esik mindkét kérhető szignifikanciaszinten.

A kiválasztott modell az elméleti feltételeknek megfelel, elemzésre és előrejelzésre felhasználható.

2. Keresztmetszeti adatokon alapuló példa

A keresztmetszeti adatok alapján történő regressziószámítást egy tapasztalatiárindex-modellen keresztül mutatjuk be.

Az ökonometriai modellek egyik speciális fajtája a tapasztalati (hedonikus) árindex-modell (*Ramanathan* [2003] 23. old.), amelyben egy árucikk ára a jellemzőitől függ, példa erre a gépkocsi ára és tulajdonságai közötti összefüggés. A vizsgálatba a 10 millió forintnál olcsóbb, hazai forgalmazású autókat vontuk be. A gépkocsik árát nemcsak mérhető tulajdonságai befolyásolják, hanem minőségi tényezők is, mint például a márka, a biztonság, garancia stb.

A mintafeladatban 119 autó adatait vizsgáltuk 2008. évi áron (forrás: http://www.auto2.hu/. A modell változói: y – a termék, az új autók alapárai (ezer forint); x_j – a termék, az új autók tulajdonságai, az autók árát befolyásoló tényezők.

A magyarázóváltozók a következők: x_1 – KÖBCM hengerűrtartalom (cm^3); x_2 – TELJ teljesítmény (LE); x_3 – NYOM maximális nyomaték (Nm); x_4 – GYORS 0-ról 100 km/h-ra gyorsulás ideje (sec); x_5 – VMAX végsebesség (km/h); x_6 – TÖMEG satját tömeg (kg); x_7 MTÖMEG megengedett össztömeg (kg); x_8 – HOSZZ hosszúság (mm); x_9 – SZÉLES szélesség (mm); x_{10} MAGAS magasság (mm); x_{11} FOGYV fogyasztás városban (liter/100 km); x_{12} FOGYVK fogyasztás városon kívül (liter/100 km).

Az autóárak és az autóárakat befolyásoló 12 magyarázóváltozó közötti regressziós kapcsolat vizsgálata alapján a következő fontosabb megállapításokat tehettük:

– A modell minden számított teszt alapján homoszkedasztikus.

– A modellben minden számított teszt alapján káros mértékű a multikollinearitás. Ennek oka, hogy az autók tulajdonságai közül a teljesítmény erőteljesen befolyásolja a többi magyarázóváltozót (a sebességet, a fogyasztást, a gyorsulást, a végsebességet, a tömeget stb).

– A multikollinearitás miatt a regressziós paraméterek standard hibái nagyobbak (a VIF-mutató például 10 magyarázóváltozó esetében a kritikus értéknél nagyobb), és csak a b_0 és b_3 regressziós paraméter különbözik a t-próba alapján 5 százalékos szignifikanciaszinten nullától.

– Figyelembe véve, hogy mind a 12 magyarázóváltozónak a modellben való megtartása indokolt, célszerű a főkomponens-elemzést (PCA – Principal Components Analysis) elvégezni.

A regresszio.xls program közli a bevont változókra vonatkozó, a számításokhoz szükséges sajátértékeket és sajátvektorokat, továbbá a sajátértékek megoszlási és kumulált megoszlási viszonyszámait.

A főkomponensanalízis-számítások részletei megtalálhatók a http://www.gmi.ktk.pte.hu/index.php?mid=33#SiposB oldalon letölthető kézikönyv 124–126. oldalán. A transzformált paramétereket, a számítások végeredményét az alábbi táblázatban mutatjuk be.

Változók	Transzformált paraméterek
x_1	2,099
x_2	2,615
x_3	0,597
x_4	–3,513
x_5	3,204
x_6	0,290
x_7	–0,196
x_8	1,151
x_9	0,382
x_{10}	–3,290
x_{11}	2,894
x_{12}	2,218

4. Összefoglalás

A regresszio.xls program felhasználása nagymértékben segíti a regressziós modellezést, valamint annak oktatását. A különböző magyarázóváltozók kombinálásával kialakítható modellek gyors értékelésére ad módot, az adatállomány tetszőleges változtatására az eredmények minden esetben reagálnak. A magyarázóváltozók számának növekedésével a lehetséges modellvariánsok száma megtöbbszöröződik.

Nemcsak a modell globális és parciális tesztelésének az eredményét látjuk azonnal, hanem idősorok esetén az autokorreláció, keresztmetszeti adatoknál pedig a homoszkedaszticitás tesztjeit is értékelhetjük, valamint a reziduum ábrákat elemezhetjük. A magyarázóváltozók összefüggésének vizsgálatára több teszt is lehetőséget ad.

Tanulmányunkban két példán keresztül mutattuk be a kifejlesztett alkalmazást: az idősoros példa alkalmas volt a backward regresszió, valamint a szakmai ismeretek alapján történő modell felállítására is. Keresztmetszeti adatokon a főkomponensregresszió alkalmazhatóságát mutattuk be. A modellezés során a modell feltételeinek különböző tesztjeit is minden esetben figyelembe kell vennünk.

Irodalom

ACZEL, A. D. [2002]: *Complete Business Statistics*. McGraw-Hill/Irwin. Boston.

BACZONI P. [2007]: *Egyszerűen Microsoft Office Excel 2003*. Panem Kiadó. Budapest.

BALÁZSNÉ MÓCSAI A. – CSETÉNYI A. [2003]: *Kvantitatív technikák, II*. Zsigmond Király Főiskola. Budapest.

BALOGH M. [2000]: *Statisztikai ismeretek*. Perfekt Kiadó. Budapest.

BÁRTFAI B. [2002]: *Office XP. World 2002. Exel 2002. Power Point 2002. Outlook Access 2002*. BBS-Info Kft. Budapest.

BEDŐ, ZS. – RAPPAI, G. [2006]: Is There Causal Relationship Between the Value of the News and Stock Returns? *Hungarian Statistical Review.* Special Number 10. 81–99. old. http://www.ksh.hu/statszemle_archive/2006/2006_K10/2006_K10_081.pdf (Elérés dátuma: 2010. május 18.)

BELSLEY, D. A. – KUH, E. – WELSCH, R. E. [1982]: *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. John Wiley and Sons. New York.

BERENSON, M. L. – LEVINE, D. M. – KREHBIEL, T. C. [2006]: *Basic Business Statistics: Concepts and Applications*. Pearson/Prentice Hall. New Jersey.

BESENYEI L. – GIDAI E. – NOVÁKY E. [1977]: *Jövőkutatás, előrejelzés a gyakorlatban*. Közgazdasági és Jogi Könyvkiadó. Budapest.

BREUSCH, T. S. – PAGAN, A. R. [1979]: Simple Test for Heteroscedasticity and Random Coefficient Variation. *Econometrica* (Econometric Society). 47. évf. 5. sz. 1287–1294. old.

EVANS, J. R. [2007]: *Statistics, Data Analysis, and Decision Modeling.* Pearson-Prentice Hall. New Jersey.

FARRAR, D. E. – GLAUBER, R. R. [1967]: Multicollinearity in Regression Analysis: The Problem Revisited. *Review of Economics and Statistics*. 49. évf. 1. sz. 92–107. old.

GLEJSER, H. [1969]: A New Test for Heteroscedasticity. *Journal of the American Statistical Association*. 64. évf. 325. sz. 316–323. old.

GODFREY, L. [1978]: Testing for Multiplicative Heteroscedasticity. *Journal of the American Statistical Association.* 8. évf. 2. sz. 227–236. old.

GOLDFELD, S. M. – QUANDT, R. E. [1965]: Some Tests for Homoscedasticity. *Journal of the American Statistical Association.* 60. évf. 310. sz. 539–547. old.

GREENE, W. H. [2003]: *Econometric Analysis.* Pearson Education International. Upper Saddle River. Prentice Hall. New Jersey.

GUJARATI, D. N. [2003]: *Basic Econometrics.* McGraw-Hill Higher Education.

HAJDU O. – HUNYADI L. [1995]: Varianciafelbontás: előfeltevések és következtetések. *Szigma*. 1–2. sz. 1–18. old.

HAJDU O. ET AL. [1987]: *Ökonometriai alapvetés.* Tankönyvkiadó. Budapest.

HAJDU O. ET AL.[1994–1995]: *Statisztika I-II*. JPTE Kiadó. Pécs.

HAJDU O. [2003]: *Többváltozós statisztikai számítások*. Központi Statisztikai Hivatal. Budapest.

HARRISON, M. J. [1982]: Tables of Critical Values for a Beta Approximation to Szroeter's Statistic for Testing for Heteroscedasticity. *Oxford Bulletin of Economics and Statistics*. 44. évf. 2 sz. 159–167. old.

HARVEY, A. C. [1976]: Estimating Regression Models with Multiplicative Heteroscedasticity. *Econometrica.* 44. évf. 3. sz. 461–466. old.

HARVEY, G. [2000]: *Excel 2000 for Windows for Dummies*. Kossuth Kiadó. Budapest.

HILL, R. C. – GRIFFITHS, W. E. – LIM, G. C. [2008]: *Principles of Econometrics*. John Wiley and Sons. New York.

HUNYADI L. – VITA L. [2002]: *Statisztika közgazdászoknak*. Központi Statisztikai Hivatal. Budapest.

HUNYADI L. – VITA L. [2008]: *Statisztika I-II.* Aula Kiadó. Budapest.

HUNYADI L. [2001]: *Statisztikai következtetéselmélet közgazdászoknak.* Központi Statisztikai Hivatal. Budapest.

HUNYADI L. [2006]: A heteroszkedaszticitásról egyszerűbben. *Statisztikai Szemle*. 84. évf. 1. sz. 75–82. old. http://www.ksh.hu/statszemle_archive/2006/2006_01/2006_01_075.pdf (Elérés dátuma: 2010. május 18.)

JÁNOSA A. [2005]: *Adatelemzés számítógéppel*. Perfekt Kiadó. Budapest.

KÁDAS K. [1944]: Az emberi munka termelékenységének statisztikai vizsgálata a magyar gyáriparban. (A Cobb-Douglas féle statisztikai törvény kiegészítése.) *Magyar Statisztikai Szemle.* 22. évf. 7–8. sz. 270–318. old. http://www.ksh.hu/statszemle_archive/viewer.html?ev=1944&szam=07-08&old=3&lap=46 (Elérés dátuma: 2010. május 18.)

KEHL D. – SIPOS B. [2009]: A telítődési, a logisztikus és életgörbe alakú trendfüggvények becslése Excel parancsfájl segítségével. *Statisztikai Szemle*. 87. évf. 4. sz. 381–411. old. http://www.ksh.hu/statszemle_archive/2009/2009_04/2009_04_381.pdf (Elérés dátuma: 2010. május 18.)

KERÉKGYÁRTÓ GY. – MUNDRUCZÓ GY. – SUGÁR A. [2001]: *Statisztikai módszerek és alkalmazásuk a gazdasági, üzleti elemzésekben*. Aula Kiadó. Budapest.

KERÉKGYÁRTÓ GY. – MUNDRUCZÓ GY. [1995]: *Statisztikai módszerek a gazdasági elemzésben*. Aula Kiadó. Budapest.

KING M. L. [1981]: A Note on Szroeter's Bound Test. *Oxford Bulletin of Economics and Statistics.* 43. évf. 3. sz. 315–322. old.

KNÜSEL L. [1998]: On the Accuracy of Statistical Distributions in Microsoft Excel 97. *Computational Statistics and Data Analysis.* 26. évf. 3. sz. 375–377. old.

KNÜSEL L. [2002]: On the Reliability of Microsoft Excel XP for Statistical Purposes. *Computational Statistics and Data Analysis*. 39. évf. 1. sz. 109–115. old.

KNÜSEL L. [2005]: On the Accuracy of Statistical Distributions in Microsoft Excel 2003. *Computational Statistics and Data Analysis.* 48. évf. 3. sz. 445–449. old.

KOOP, G. [2008]: *Közgazdasági adatok elemzése*. Osiris Kiadó. Budapest.

KÖRÖSI G. – MÁTYÁS L. – SZÉKELY I. [1990]: *Gyakorlati ökonometria*. Közgazdasági és Jogi Könyvkiadó. Budapest.

KOTZ, S. ET AL. [2006]: *Encyclopedia of Statistical Sciences*. 16 Volume Set. John Wiley and Sons. New York.

KOVÁCS I. [1977]: Az autokorreláció vizsgálata regressziós modellekben. *Statisztikai Szemle*. 552. évf. 6. sz. 603–621. old. http://www.ksh.hu/statszemle_archive/viewer.html?ev=1977&szam=06&old=45&lap=19 (Elérés dátuma: 2010. május 18.)

KOVÁCS P. – PETRES T. – TÓTH L. [2004]: Adatállományok redundanciájának mérése. *Statisztikai Szemle*. 82. évf. 6–7. sz. 595–604. old.

http://www.ksh.hu/statszemle_archive/2004/2004_06-07/2004_06-07_595.pdf (Elérés dátuma: 2010. május 18.)

KOVÁCS P. [2008a]: A multikollinearitás vizsgálata lineáris regressziós modellekben. *Statisztikai szemle*. 86. évf. 1. sz. 38–67. old.
http://www.ksh.hu/statszemle_archive/2008/2008_01/2008_01_038.pdf (Elérés dátuma: 2010. május 18.)

KOVÁCS P. [2008b]: A statisztikaoktatás módszertanának modernizálása? *Statisztikai Szemle*. 86. évf. 12. sz. 1143–1157. old.
http://www.ksh.hu/statszemle_archive/2008/2008_12/2008_12_1143.pdf (Elérés dátuma: 2010. május 18.)

KOVACS, P. – PETRES, T. – TOTH, L. [2005]: A New Measure of Multicollinearity in Linear Regression Models. *International Statistical Review*. 73. évf. 3. sz. 405–412. old.

KOVALCSIK G. [2000]: *Excel 2000*. ComputerBooks. Budapest.

KREKÓ B. [1966]: *Mátrixszámítás*. Közgazdasági és Jogi Könyvkiadó. Budapest.

KSH (KÖZPONTI STATISZTIKAI HIVATAL) [1985–2005]: *Ipari és építőipari statisztikai évkönyv*. Budapest.

KSH (KÖZPONTI STATISZTIKAI HIVATAL) [1985–2088]: *Magyar statisztikai évkönyv*. Budapest.

KUZNETS, S. S. [1930]: *Secular Movements in Production and Prices: Their Nature and Bearing upon Cyclical Fluctuations*. Houghton-Mifflin. Boston.

LÉNÁRT I. – RAPPAI G. [2001]: Néhány gondolat a varianciabecslés hibahatáráról. *Statisztikai Szemle*. 79. évf. 7. sz. 613–621. old.
http://www.ksh.hu/statszemle_archive/2001/2001_07/2001_07_613.pdf (Elérés dátuma: 2010. május 18.)

MADDALA, G. S. [2004]: *Bevezetés az ökonometriába*. Nemzeti Tankönyvkiadó. Budapest.

MCCULLOUGH B. D. – WILSON B. [2002]: On the Accuracy of Statistical Procedures in Microsoft Excel 2000 and Excel XP. *Computational Statistics and Data Analysis*. 40. évf. 4. sz.713–721. old.

MCCULLOUGH, B. D. – WILSON, B. [1999]: On the Accuracy of Statistical Procedures in Microsoft EXCEL 97. *Computational Statistics and Data Analysis*. 31. évf. 1. sz. 27–37. old

MUNDRUCZÓ GY. [1981]: *Alkalmazott regressziószámítás*. Akadémiai Kiadó. Budapest.

MUNDRUCZÓ GY. [1998]: *Útmutató a statisztikai modellezéshez*. Aula Kiadó. Budapest.

NASH, J. C. [2008]: Teaching Statistics with Excel 2007 and Other Spreadsheets. *Computational Statistics and Data Analysis*. 52. évf. 10. sz. 4602–4606. old.

NYITRAI F.-NÉ – RÉDEY K. [1974]: *Statisztika III*. (Korszerű statisztikai módszerek és alkalmazásuk a gyakorlati közgazdasági munkában). Tankönyvkiadó. Budapest.

PARK, R. E. [1966]: Estimation with Heteroscedastic Error Terms. *Econometrica*. 34. évf. 4. sz. 888. old.

PAWLOWSKI Z. [1970]: *Ökonometria*. Közgazdasági és Jogi Könyvkiadó. Budapest.

PÉTERY K. [2003]: *Táblázatkezelés Excel 2002*. Kossuth Kiadó. Budapest.

PINTER J. – RAPPAI G. (szerk.) [2007]: *Statisztika*. Pécsi Tudományegyetem, Közgazdaságtudományi Kar. Pécs.

PINTER J. [1991]: A heteroszkedaszticitás diagnosztizálása. *Statisztikai Szemle*. 69. évf. 1. sz. 16–36. old.

http://www.ksh.hu/statszemle_archive/viewer.html?ev=1991&szam=01&old=18&lap=21 (Elérés dátuma: 2010. május 18.)

PINTÉR J. [2000]: *Bevezetés a statisztika módszereibe*. Pécsi Tudományegyetem. Pécs.

POLT R. [2005]: Levegőkereskedelem – a Nemzeti Kiosztási Terv kialakítása. *Statisztikai Szemle*. 83. évf. 10–11. sz. 990–1009. old. http://www.ksh.hu/statszemle_archive/2005/2005_10-11/2005_10-11_990.pdf (Elérés dátuma: 2010. május 18.)

RAMANATHAN, R. [2003]: *Bevezetés az ökonometriába alkalmazásokkal.* Panem Kiadó. Budapest.

RAPPAI G. [2001]: *Üzleti statisztika Excellel.* Központi Statisztikai Hivatal. Budapest.

RAPPAI G. [2008]: Gondolatok a gazdaságtudományi képzési területen folyó statisztikaoktatásról. *Statisztikai Szemle*. 86. évf. 9. sz. 829–849. old.

SAVIN, N. E. – WHITE, K. J. [1977]: The Durbin-Watson Test for Serial Correlation with Extreme Sample Sizes or Many Regressors. *Econometrica*. 45. évf. 8. sz. 1989–1996. old.

SIPOS B. [1982]: *Termelési függvények – vállalati prognózisok.* Közgazdasági és Jogi Könyvkiadó. Budapest.

SIPOS, B. – KISS, T. [1998]: REGAL: Expert System for Multiple Linear Regression Analysis. *Hungarian Statistical Review.* Special Number 2. 35– 49. old. http://www.ksh.hu/statszemle_archive/1998/1998_K2/1998_K2_035.pdf (Elérés dátuma: 2010. május 18.)

SPIEGEL M. R. [1995]: *Statisztika – elmélet és gyakorlat (SI mértékegységekkel)*. Panem–McGraw-Hill. Budapest.

STUART, A. – ORD, J. K. [1994]: *Kendall's Advanced Theory of Statistics*. Vol. 1. Distribution Theory. Edward Arnold. London.

SYDSAETER, K. – HAMMOND, P. I. [2006]: *Matematika közgazdászoknak*. Aula Kiadó. Budapest.

SZROETER, J. [1978]: A Class of Parametric Tests for Heteroscedasticity in Linear Econometric Models. *Econometrica.* 46. évf. 6. sz. 1311–1327. old.

THEISS E. (szerk.) [1958]: *Korreláció és trendszámítás.* Közgazdasági és Jogi Könyvkiadó. Budapest.

YALTA, A. T. [2008]: The Accuracy of Statistical Distributions in Microsoft® Excel 2007. *Computational Statistics and Data Analysis.* 52. évf. 10. sz. 4579–4586. old.

Summary

The Excel environment of the regression file developed by the authors is accessible for almost anyone, which helps the proliferation of further applications and its usability in higher education. Models built by the manipulation of explanatory variables, as well as the dataset itself can be evaluated rapidly. The authors have developed numerous Excel files also in the fields of special regression analysis, primarily in time series decomposition model applications. All files and the manual are accessible online.

Szezonális kiigazítás a gazdasági válságban – adatelőállító szemmel*

Bánhegyi Péter,
az MNB vezető statisztikai elemzője
E-mail: banhegyip@mnb.hu

Horváth Beáta,
a KSH tanácsosa
E-mail: Beata.Horvath@ksh.hu

Lénárt Imre,
az MNB statisztikai elemzője
E-mail: lenarti@mnb.hu

Urr Beáta,
a KSH tanácsosa
E-mail: Beata.Urr@ksh.hu

Az idősorokat klasszikusan három,[1] külön-külön közvetlenül nem megfigyelhető, egymástól független komponensre szokás felbontani, melyek a trend,[2] a szezonális ingadozások és a véletlen zaj. Mivel a szezonális (és naptári) ingadozások a konjunktúravizsgálat szempontjából irreleváns információk, ezért a szezonális kiigazítás során elsődleges cél ezek becslése és kiszűrése az idősorból.

A 2008-ban kirobbant nemzetközi gazdasági és pénzügyi válság megjelenését követően a társadalmi és gazdasági folyamatok mutatói közül továbbra is a szezonálisan

* A szerzők ezúton mondanak köszönetet *Anwar Klárának*, *Földesi Erikának*, *Montvai Beátának*, *Némethné Székely Edinának* és *Vályi Istvánnak*, akik számos hasznos javaslattal és megjegyzéssel segítették a tanulmány elkészítését.

[1] A gyakorlatban a klasszikus három összetevőn túl külön komponenst feleltetnek meg a naptárhatás különböző tényezőinek (például ünnepnaphatás), a kiugró értékeknek (például szinteltolódás) és egyéb az idősort rövid távon befolyásoló hatásoknak, melyek kiszűrése a szezonális komponenssel együtt történik, így nem szerepelnek a szezonálisan kiigazított adatokban.

[2] Trend alatt a trendciklus együttesét értjük, azaz az idősor alapirányzata a hosszabb távon megfigyelhető ingadozásokkal, hullámzásokkal kerül kifejezésre.

kiigazított adatok számítanak az egyik legfontosabb információforrásnak, amelyek többek között lehetőséget biztosítanak különböző jelenségek, folyamatok időbeli lefolyásának elemzésére, tendenciák, fordulópontok megragadására, idősorok közötti kapcsolatok vizsgálatára, összehasonlítások készítésére. A gazdaságstatisztikai gyakorlatban is nagy szerepet kapnak a tendenciát érzékeltető idősorok, azaz a szezonálisan kiigazított, illetve trendadatok, mivel a gazdasági folyamatokat, a konjunktúra alakulását az eredeti idősor alapján nehéz érzékelni. A szezonális kiigazítás célja egyrészről, hogy a különböző szezonális hatások kiszűrésével tisztább és pontosabb információt kapjunk az idősor lefolyásának és tendenciájának alakulásáról; másrészről, hogy térben (például nemzetközileg) és időben összehasonlítható adatok álljanak rendelkezésre.[3]

A hivatalos statisztikai szolgálatokon belül végzett szezonális kiigazítási eljárások összetett becslési, modellezési eljáráson alapulnak, amelyek még stabil körülmények, állapotok között is hordoznak magukban némi bizonytalanságot. Jelenleg a gazdasági folyamatokat a válság nyomán különböző hatások, gyors reakciók jellemzik, amelyek strukturális változásokat eredményezve átalakíthatják a korábbi szerkezeteket, megoszlásokat. Ezek a hatások nagymértékben megváltoztatják az idősorok viselkedését, tovább növelve az adatok hordozta szokásos bizonytalanságot.

Recesszió időszakában ugyanakkor különös figyelmet kapnak a szezonálisan kiigazított rövid távú indexek, hiszen ebből a mutatóból hamarabb azonosíthatók a gazdasági, társadalmi folyamatok fordulópontjai, illetve növekedési szakaszai. Azonban változó körülmények között az idősor végén különösen nehézkes az egyértelmű trendforduló megállapítása, mivel a szezonális kiigazítás eredményei a szokásosnál több bizonytalanságot hordoznak, melyhez nagyobb mértékű revízió is társulhat.[4] Abban az esetben, ha az idősor végén kiugró érték (outlier) jelenik meg, annak léte és típusa matematikai-statisztikai szempontból nem egyértelmű, hiszen csak a későbbi adatok ismeretében derül ki, hogy a hatás – amennyiben a későbbiekben is kimutatható – egy (additív kiugró érték), néhány (csillapodó jellegű törés) vagy az összes további adatra (szinteltolódás) vonatkozik-e. Új adatok beérkezésével a kiugró érték típusa változhat, akár eliminálódhat is, ami az utolsó kiigazított adatok jelentős változását is maga után vonhatja. Ennek elkerülésére hasznos, ha megfelelően alátámasztott szakmai információ áll rendelkezésre az idősorra gyakorolt hatásokról, és ennek megfelelően a kiugró érték típusa – természetesen matematikai-statisztikai szempontból is tesztelve – rögzíthető.

[3] A szezonális kiigazításról a módszertani füzetben (*KSH* [2005]) olvashat részletesebben.

[4] A szezonálisan kiigazított adatsor minden egyes új tényadat bevonásával a teljes idősor hosszára vonatkozóan módosulhat. Azaz a jelennel együtt a múlt is megváltozik, és a módosulás különösen nagymértékű lehet az idősor végén. Ennek oka egyrészről a modellezési eljárásban keresendő, mivel új adatok beérkezésével az újbóli becslés (szezonális kiigazítás) során revideálódhatnak a korábbi modell paraméterei, esetleg maga a modell is (*Ferenczi–Jakab* [2002]). Másrészről a revízió mértékét jelentősen meghatározza az új tényadat értéke és annak információtartalma.

Az idősorok viselkedésének megváltozása során bekövetkező revíziók természetesen problematikusak lehetnek a felhasználók, az elemzők szemszögéből. A revízió ugyanakkor önmagában természetes dolog, így amennyiben az adatok ezt indokolják, a különböző hatásokat célszerű bemutatni a felhasználók számára. Annak érdekében, hogy a publikált idősorok mind az adatelőállítók szempontjainak, mind a felhasználók igényeinek eleget tegyenek, különös figyelmet kell fordítani a szezonális kiigazítási eljárás során használt eszközökre.

A tanulmány néhány idősor példáján keresztül mutatja be, hogy a gazdasági válság milyen nehézségek elé állította a Központi Statisztikai Hivatalt (KSH) és a Magyar Nemzeti Bankot (MNB), mint a szezonálisan kiigazított adatok előállítóit.

1. Nemzetközi ajánlások és a hazai gyakorlatok

A nemzetközi intézmények a szezonális kiigazításra vonatkozóan – egységes jogszabály hiányában – különböző ajánlásokat fogalmaznak meg az adatelőállítók részére. Ezen ajánlásokkal az Eurostatnak (Európai Unió Statisztikai Hivatala), illetve az Európai Központi Banknak nemcsak a statisztikai adatok európai szintű hasznosítása, hanem azok nemzetközi szintű összehasonlíthatóságának és minőségének biztosítása is alapvető célja. 2008-ig az Eurostat a statisztikai hivatalok számára szakstatisztikánként különböző részletezettségű ajánlásokat fogalmazott meg. 2008-ban az SPC[5] és a CMFB[6] által is elfogadott, az Európai Statisztikai Rendszerre érvényes, egységes – szakstatisztikáktól független – szezonális kiigazítási irányelvek (*Eurostat* [2009]) léptek életbe, melyek implementálása a tagországokra vonatkozóan jelenleg is folyamatban van.

1.1. Az Eurostat szezonális kiigazításra vonatkozó irányelvei

Az irányelvek célja a szezonális kiigazításra vonatkozó ún. „legjobb gyakorlatok" kidolgozásának az elősegítése, mellyel a nemzeti gyakorlatok harmonizáltabbá válnak, és hozzájárulnak a robusztusabb aggregált EU-statisztikák kialakításához továbbá mind elméleti, mind gyakorlati szempontból tartalmazzák a végrehajtáshoz szük-

[5] Statistical Programme Committee (Statisztikai Programbizottság): 2009. március 31-e óta Európai Statisztikai Rendszerbizottság (ESSC – European Statistical System Committee). A bizottság tagjai a nemzeti statisztikai hivatalok vezetői, elnöke az Eurostat főigazgatója, fő feladataként szakmai iránymutatást nyújt az európai statisztikai rendszer számára az európai statisztikák fejlesztéséhez, előállításához és terjesztéséhez.

[6] Committee on Monetary, Financial and Balance of Payment (Monetáris, Pénzügyi- és Fizetésimérleg-statisztikákkal Foglalkozó Bizottság): a nemzeti statisztikai hivatalok, az Eurostat, a nemzeti központi bankok és az Európai Központi Bank statisztikusait fogja össze. A bizottság sokféle statisztikai kérdésben, de legfőképpen a túlzott deficiteljárással kapcsolatban ad tanácsot az Európai Bizottságnak, valamint biztosítja az európai szintű együttműködést és a statisztikai munka összehangolását.

séges információkat úgy, hogy azok a legszélesebb felhasználói kört kiszolgálják. További hangsúlyos eleme az irányelveknek, hogy az átláthatóság és a megbízhatóság érdekében az egész szezonális kiigazítási folyamat adekvát és igény szerint hozzáférhető módon legyen dokumentálva az adatok összehasonlíthatóságának érdekében. A hazai gyakorlat vonatkozásában és alkalmazásában kiemelkedően fontosak a következő szempontok.

– Új tényadat, illetve az alapadat revíziója esetén a modellek részleges újrabecslése ajánlott, mindemellett fontos, hogy a revízió nagysága lehetőleg minimális legyen. A modell és komponensei (autoregresszív tagok száma, differenciálás foka, a naptárhatásokra és kiugró értékekre vonatkozó regresszorok) lehetőleg évente egyszer kerüljenek felülvizsgálatra, emellett az új adatok beérkezésekor az egyes paraméterek újrabecsülhetők. Amennyiben szükséges (például jelentős és hosszabb visszamenőleges revízió esetén), a modellt és annak komponenseit is újra kell becsülni.

– A szezonálisan kiigazított adatok revideált idősorait legalább arra az időintervallumra közölni kell, amelyben a kiigazítatlan adatok is változtak (ebben az esetben a szezonálisan kiigazított idősor nem homogén), de a revideált idősorok ennél hosszabb visszamenőleges időszakra történő publikációja is (ez utóbbi azonban esetleg sok plusz irreleváns információt is tartalmazhat) lehetséges.

– Kiugró értékek vizsgálatánál három típus kerüljön tesztelésre: additív kiugró érték (additive outlier), csillapodó jellegű törés (transitory change) és szinteltolódás (level shift).

– Az aggregátumok lehetőleg a direkt igazítás módszerével kerüljenek kiigazításra, vagyis az alágazatok mellett az aggregátumok is külön idősorként kezelendők, amennyiben az aggregátumok és a komponenseik szezonális mintázata hasonló. Az alágazatok és az aggregátum közötti additivitás így nem teljesül automatikusan, azonban ez az eljárás általában kezelhetőbb és pontosabb eredményt ad, mint az indirekt kiigazítás, továbbá ez utóbbi reziduális szezonalitást is tartalmazhat. Az indirekt kiigazítás akkor javasolt elsősorban, ha az aggregált idősor egyes komponensei eltérő szezonalitást követnek.

– A szezonális kiigazítás minősége érdekében valamennyi, a program által felkínált diagnosztikát, grafikus ellenőrzési eszközt figyelembe kell venni, és a transzparencia érdekében az eredményeket egy egységes séma alapján célszerű dokumentálni, beleértve a revíziókat is.

A szezonális kiigazításra vonatkozó irányelvek az egységes alkalmazás előnye mellett kitérnek a hátrányokra, veszélyekre is, többek között arra, hogy a szezonális

kiigazítás nem precízen definiált fogalom, és az eredmény nagyban függ a választott, különböző hipotéziseken alapuló modelltől. Fontos szempont, hogy a szezonálisan kiigazított adat minőségét nagy mértékben befolyásolja az alapadat minősége. A nem megfelelő szezonális kiigazítás megtévesztő eredményekhez vezethet, növelve a statisztikai adatok iránti bizalmatlanságot.

Az irányelvek alkalmazásának új szoftvere[7] a Demetra+, mely nagyban hozzájárul a tagállamok eltérő gyakorlatának közeledéséhez, megfelelő teret hagyván az egyes országok sajátosságainak kezelésére.

1.2. Hazai gyakorlatok

A KSH-ban 2004 áprilisában lépett életbe a szezonális kiigazítás gyakorlatáról szóló szabályzat, mely alapján a Hivatal szakstatisztikai területektől függetlenül, egységes elvek alapján, valamint az Eurostat ajánlásaival összhangban végzi idősorainak kiigazítását. Kialakításra kerültek továbbá a statisztikai termelési folyamat egyes szakaszaira vonatkozó minőségi irányelvek, melynek a szezonális kiigazítás is részét képezi (*KSH* [2010]). Az MNB-ben szintén folyamatban van az Eurostat ajánlásainak megvalósítása.

A két intézmény a szezonális kiigazítást a különböző felhasználói igények kielégítésére, a kiigazított adatsorokhoz való hozzájutás biztosítása érdekében végzi. Az adatok felhasználói köre igen sokrétű: az újságíróktól kezdve, különböző kutatóintézetek munkatársain keresztül, a döntéshozókig. A legtöbb esetben a publikált szezonálisan kiigazított adatokat különböző elemzések, modellezések, előrejelzések készítésére használják.

A szezonális kiigazításhoz mindkét intézmény a TRAMO/SEATS-módszert[8] alkalmazza a Demetra program felhasználásával. Az alkalmazott módszer és a kialakított gyakorlat összhangban van az Európai Statisztikai Rendszer új, szezonális kiigazításra vonatkozó irányelveivel is.

A kis mértékű revízió, valamint az eredmények minőségének egyidejű elérése érdekében a KSH-ban alapszabályként évi egyszeri modell- és paraméterrögzítés került kialakításra. A homogenitás érdekében a publikálások során a szezonálisan kiigazított adatok az idősor teljes hosszára vonatkozóan közlésre kerülnek. Az MNB – az Eurostat ajánlásaival összhangban – alapszabályként szintén évente egyszer vált – szükség esetén – modellt, és a revízió időszakára újraközli az új szezonálisan kiigazított adatokat is.

[7] A szoftver fejlesztése jelenleg a tesztelési fázisban tart, a hivatalos verzió legkorábban 2010 őszén lesz elérhető, tényleges gyakorlati alkalmazása a következő időszak feladata közé tartozik.

[8] A TRAMO (idősorregresszió ARIMA-zajjal, hiányzó megfigyelésekkel és kiugró értékekkel) és a SEATS (jelkinyerés ARIMA idősorokban) programot *Augustin Maravall* és *Victor Gomez* fejlesztették ki 1996-ban a Spanyol Nemzeti Bankban.

Abban az esetben, ha a magyar naptárhatás, azaz a munkanap-, ünnep- és húsvéthatások meghatározók és magyarázhatók az idősorban, akkor általános elv ezen hatások kiszűrése, mely nagyban függ az idősor hosszától. A KSH a szezonálisan kiigazított adatok mellett naptárhatással kiigazított adatokat is publikál úgy, hogy az eredeti idősorból csak a naptárhatást szűri ki, a szezonalitást nem. Az MNB naptárhatás kiszűrést szezonális kiigazítás nélkül nem végez.

2. A gazdasági válság hatása a KSH idősoraira

A gazdasági válságból adódó rendkívüli helyzet kialakulásával a KSH arra az álláspontra jutott, hogy egy esetlegesen több területet is érintő változás kezelésére egységes eljárást kell alkalmazni. Mivel a jelenlegi bizonytalan gazdasági helyzetben az eddigi gyakorlat nem minden esetben vezetett megfelelő eredményre, a kialakított módszertan a nemzetközi ajánlásokkal összhangban pontosításra került a válság jelenségeinek egységes kezelése érdekében, különös tekintettel a modellválasztás és az idősorok végén megjelenő kiugró értékekre vonatkozóan.

A szezonális kiigazítás során továbbra is alapvető cél, hogy az eddigi gyakorlatnak megfelelően a megalapozott szakmai érvek mellett a matematikai-statisztikai szempontok is maximálisan figyelembe legyenek véve. Az idősorokon minden egyes időszakban – a rögzített modell- és paraméterértékekkel történő kiigazítás mellett – különböző automatikus modell- és paraméterbecsléseket is végre kell hajtani annak érdekében, hogy a lehető legtöbb rendelkezésre álló információ felhasználásra kerüljön a becslési eljárás során. Amennyiben jelentős eltérés adódik a rögzített beállításokhoz képest, a szakstatisztikus és a módszertani szakértő közösen döntenek az alkalmazandó eljárásról.

A különböző modellek futtatása az idősorok végén megjelenő kiugró értékek kezelése szempontjából is fontosak. Abban az esetben, ha az automatikus tesztelés eddig nem kezelt, újonnan megjelenő kiugró értéke(ke)t von be az idősorba, azok figyelembe vételéről a szakstatisztikus a rendelkezésére álló, kellően alátámasztott szakmai információk alapján dönthet, figyelembe véve a kapcsolódó idősorokban már (akár más szakstatisztikai területen) alkalmazott eljárást, kezelést és a matematikai-statisztikai, valamint minőségi szempontokat.

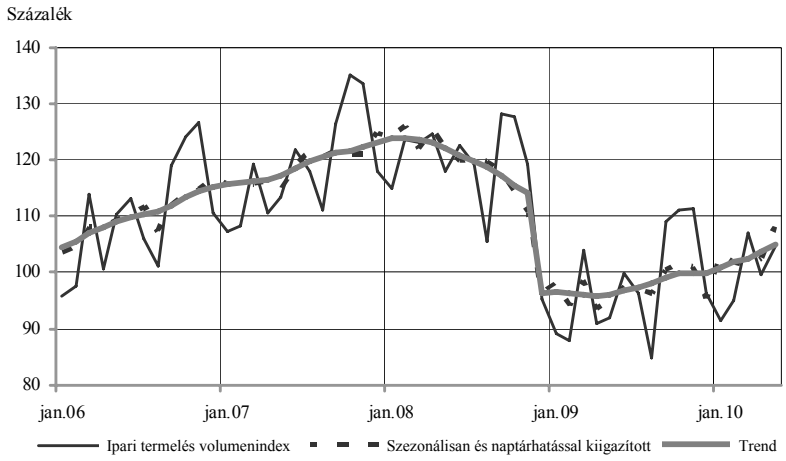
Amennyiben az automatikus tesztelés nem von be kiugró értékeket, azonban a szakértő kellő időben rendelkezésre álló, megfelelően alátámasztott szakmai információkkal rendelkezik a kiugró értékek helyéről, típusáról, akkor kérheti azok figyelembe vételét a szezonális kiigazítás során, ha az eredmények konzisztensek maradnak a kapcsolódó idősorokkal, valamint az eredmények megfelelnek valamennyi matematikai-statisztikai szempontnak és minőségi elvárásnak.

2.1. A gazdasági válság hatása a havi idősorokra

A 2008 második felében kibontakozó gazdasági válság a legtöbb idősorra erőteljes hatást gyakorolt. Ezek közül a havi gyakoriságú idősoroknál a Demetra program 2008 decemberére a KSH-ban alkalmazott gyakorlatnak megfelelően az automatikus tesztelés eredményeként szinteltolódással járó kiugró értéket illesztett az idősorok jelentős részére. Ez a fajta hatás jellemzi például a kulcsszerepet játszó külkereskedelmi és ipari termelés idősorait is. A fő aggregátumok mellett a legtöbb alágazatban is hasonlóan megjelent ez a jellegű változás. Mivel a havi gyakoriságú idősorok tendenciájában bekövetkező erőteljes törés és annak hosszú távon elhúzódó normalizálódása a rendelkezésre álló információk alapján alátámaszthatónak bizonyult, így a gazdasági válság hatása számos havi idősor esetén szinteltolódás típusú kiugró értékként jelentkezik.

Az ipari termelés volumenindexének szezonálisan kiigazított adata, például az előző időszakhoz viszonyítva, 2008 decemberében közel 20 százalékponttal csökkent, mely az alapadatból is jól látható. (Lásd az 1. ábrát.) Az automatikus tesztelés az idő előrehaladtával továbbra is erőteljes szignifikáns törést mutat 2008 decemberére. A gazdasági válság megjelenése óta a szinteltolódás folyamatosan jelen van az említett idősorokban.

1. ábra. Ipari termelés volumenindexe, 2006. január–2010. május
(2005. év havi átlaga=100 százalék)



Forrás: KSH.

Természetesen vannak olyan havi gyakoriságú idősorok is, amelyekre a válság nem volt akkora hatással (például már a válság előtt is visszaesést mutató építőipar), így szakmai megfontolások alapján az a döntés született, hogy kiugró értékek bevonása nem indokolt az idősorba.

2.2. A gazdasági válság hatása a negyedéves idősorokra

A havi idősorokkal ellentétben a negyedéves gyakoriságúaknál már nem volt tapasztalható ekkora mértékű változás. Ez részben azzal magyarázható, hogy a havi idősorok sokkal dinamikusabban tudnak reagálni a változásokra, míg negyedéves vonatkozásban a különböző hatások csökkenthetik, akár ki is olthatják egymást. Ezért természetes módon felvetődött a negyedéves idősorok esetén a válság kiugró értékekkel vagy azok nélküli kezelésének kérdése.

A makrogazdasági mutatók közül különösen érzékeny adatnak számítanak a GDP szezonálisan kiigazított negyedéves indexei. A pontosított gyakorlatnak megfelelően minden negyedévben az automatikus futtatás végrehajtásán túl számos vizsgálat, elemzés készül a különböző modellek, hipotézisek eredményei alapján.

A válság kirobbanását követő időszakokban a kiugró értékek jelenléte bizonytalannak mutatkozott, melyben nagymértékben szerepet játszott az említett kisebb gyakoriságú idősorok sajátossága. Néhány modell esetében megjelentek kiugró értékek az idősorban, azonban számos esetben, többek között az Airline[9] modell esetében nem. Ugyanakkor a GDP alágazatainak vizsgálata során sem lehetett egyértelműen kimutatni a kiugró értékek jelenlétét.

További kérdéseket vetett fel a vizsgálandó kiugró érték típusa, valamint az a körvonalazódni látszódó tény, hogy a válság a GDP összetevőiben nem egy időszakra vonatkozóan jelenik meg, és a 2008 IV. negyedévében bekövetkező erőteljes csökkenés 2009 I. negyedévében tovább folytatódott. Azonban két, időben közvetlen egymás után jelentkező kiugró érték módszertanilag nehezen értelmezhető.

Az alapkérdés, azaz a gazdasági válság kiugró értékekkel vagy azok nélküli kezelésének dilemmája, 2009 II. negyedévében, egy újabb adat rendelkezésre állásával ismét megvitatásra került. Ugyanis világossá vált, hogy nemzetközi viszonylatban egyre több ország mesterségesen alkalmazza a gazdasági válság szinteltolódással való kezelését, ami rontotta az adatok nemzetközi összehasonlíthatóságát.

A különböző elemzések, vizsgálatok, valamint a szezonális kiigazítást végző hazai intézményekkel folytatott szakmai konzultációk[10] azt támasztották alá, hogy a gazdasági válság kezelésére nem alkalmazhatók kiugró értékek a magyarországi GDP idősoraiban.

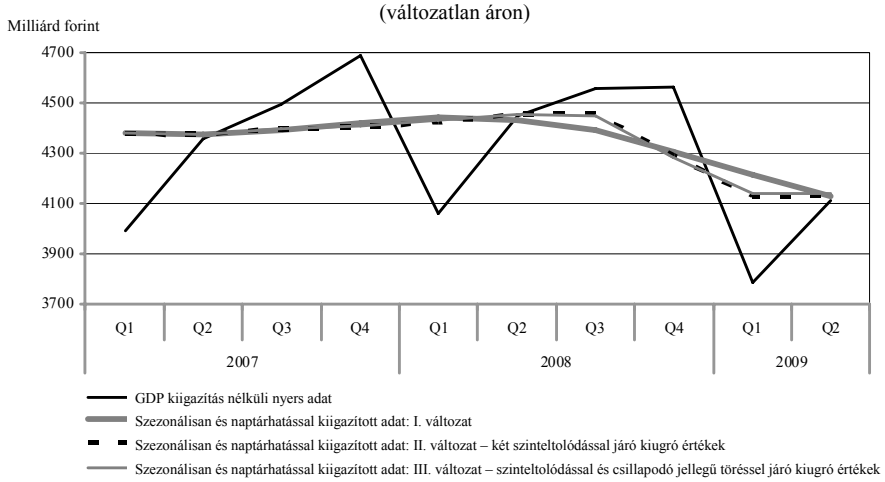
A 2. ábrán látható a GDP idősorának 2009 II. negyedévéig a három különböző eljárás alapján történt igazítása, melyek az idősor végén igen eltérő eredményeket ad-

[9] Az Airline modell (SARIMA(011)(011)) kedvező tulajdonsága, hogy viszonylag jól közelít más modelleket, és kevés paraméterrel jól illeszthető a legtöbb idősorra.

[10] A KSH Statisztikai kutatási és módszertani főosztálya, a Nemzeti számlák, illetve a Szektor számlák főosztállyal közösen a szezonális kiigazítás gyakorlati tapasztalataival foglalkozó szakértői konzultációkat hívott össze 2009 szeptemberére az Ecostat, valamint 2009 decemberére az MNB, a Pénzügyminisztérium (PM), a Nemzeti Fejlesztési és Gazdasági Minisztérium (NFGM), az Ecostat munkatársai részvételével, melyek központi témája a gazdasági válság idősorokban megjelenő hatásának kezelése volt.

tak. A vizsgálatok során olyan eredmények is születtek, melyek a GDP összesen soránál „trendfordulót” mutattak ki. Ugyanakkor matematikai-statisztikai szempontból ismert tény, hogy a trend lefutása az idősor végén nem egyértelmű, hiszen az a későbbi adatok ismeretében és a becslési eljárás következtében megváltozhat.

2. ábra. Negyedéves GDP, 2007. I negyedév–2009. II. negyedév
(változatlan áron)



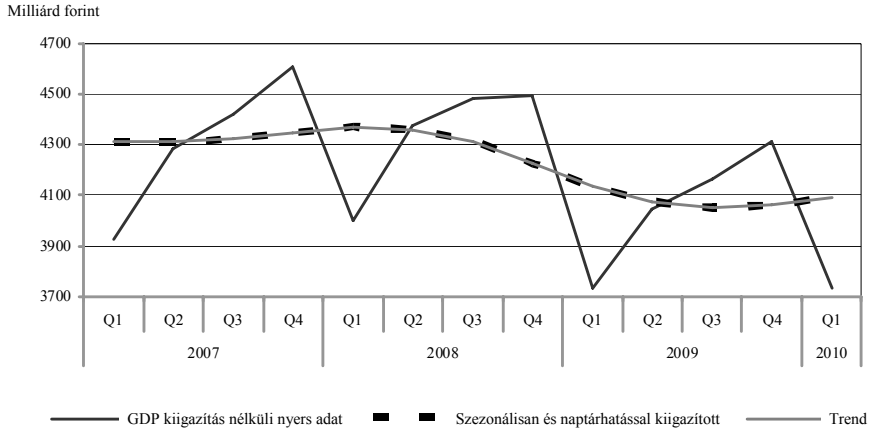
Forrás: KSH.

Az első változat a 2009. I. negyedéves rögzített paraméterekkel történő futtatás eredménye, amely megfelel a KSH évközi kiigazítási gyakorlatának. Ez megegyezik az automatikus tesztelés eredményével, azaz nem jelenik meg kiugró érték az idősorban. A második vizsgálatnál olyan modellel történt a kiigazítás, mely 2008 IV., illetve 2009 I. negyedévében egyaránt szinteltolódást mutatott ki. A harmadik eljárás annyiban különbözik az előzőtől, hogy 2009 I. negyedévében csillapodó jellegű törés került beállításra a szinteltolódás helyett, míg 2008 IV. negyedévére maradt a szinteltolódás. A további kiugró értékre vonatkozó vizsgálatok során, mint például abban az esetben, amikor egyetlen kiugró érték jelenne meg az említett két időszak valamelyikén, a program nem nyújtott matematikai-statisztikai szempontból elfogadható eredményeket.

Mint a 2. ábrából is látható a II. illetve III. változat eredményeként az előző negyedévhez viszonyított szezonálisan kiigazított adatokban stagnálás, illetve enyhe növekedés mutatkozott. Ugyanakkor a kapcsolódó idősorok nem támasztották alá a negyedéves indexek növekedésének hihetőségét, mivel mind a kiskereskedelmi mind más szolgáltatási adatok továbbra is erőteljesen csökkenő tendenciát mutattak. Ennek következtében a szezonális kiigazítás során kiugró értékek nem kerültek alkalmazásra a GDP idősorában.

A gazdasági válság kirobbanása és annak gazdasági életben való hazai megjelenése óta a tárgyalt alapkérdés valamennyi negyedévben felülvizsgálatra kerül. Mindezek eredményeképp a KSH a gazdasági válság kezelésére jelenleg sem alkalmaz kiugró értékeket a GDP idősorában. (Lásd a 3. ábrát.) A válság ilyen típusú kezelésének ugyanakkor természetes következménye, hogy a trend fokozatosan alkalmazkodik az idősor változásaihoz, melynek hatása a többi komponensben is tükröződhet.

3. ábra. Negyedéves GDP, 2007. I negyedév–2010. I. negyedév
(változatlan áron)



Amennyiben a jelenlegi tendencia tartós marad az idősorban, a továbbiakban nem várható jelentős mértékű revízió a szezonálisan kiigazított adatokban. A felhasználóknak azonban továbbra is figyelembe kell venniük, hogy a szezonálisan kiigazított adatok revíziójának oka csak egyrészről rejlik magában a folyamatban, hiszen nagyban befolyásolhatja azokat az alapadatok revíziója is. Az éves és negyedéves nemzetiszámla-adatok felülvizsgálatára évente egyszer kerül sor, a KSH részletes tájékoztatást ad a felhasználóknak a revíziók okairól és hatásairól.[11]

3. A gazdasági válság hatása a Magyar Nemzeti Bank idősoraira

Az MNB saját statisztikai publikációit tekintve alapvetően két szakterületet, a fizetési mérleget és a monetáris statisztikát érinti a szezonálisan kiigazított adatok közzététele, és a két szakterület együttműködik e munka során. Ennek alapja a ko-

[11] http://portal.ksh.hu/pls/ksh/ksh_web.meta.objektum?p_lang=HU&p_menu_id=110&p_ot_id=100&p_obj_id=QPT&p_session_id=89972292

rábban említett Eurostat-ajánlások figyelembe vétele: modellrögzítés, direkt kiigazítás, ünnepnap hatások figyelembe vétele, kiugró értékek vizsgálata, minőségi kritériumok teljesítése, előre meghirdetett revíziós politika. A szezonálisan kiigazított adatok mellett az MNB a szezonális kiigazítás főbb alapelveit is közzéteszi a honlapján, a sajtóközleményeiben. A fizetési mérleg és a monetáris statisztika szezonálisan kiigazítási gyakorlatai alapesetben ugyanakkor egy, a gyakorlati következményeket tekintve nem jelentős pontban eltérnek egymástól: míg a monetáris statisztikai idősorok esetében – a KSH gyakorlatával egyezően – a modellrögzítés mellett paraméterrögzítésre is sor kerül, addig a fizetési mérleg idősorok esetében csak a modell kerül rögzítésre, a paramétereket a szakértők minden időszakban újrabecslik. Ez utóbbi eljárás szintén megfelel az Eurostat ajánlásainak.

A gazdasági válság az MNB által kiigazított idősorokban is életre hívta azoknak a problémáknak többségét, amelyekről az eddigiekben szó esett. A revíziók, a kiugró értékek kezelése az MNB szakértői számára is fontos feladatként jelentkeztek, mert a kiigazított idősorok jelentős részénél a publikációt nem lehetett elhagyni (mint majd láthatjuk, volt, ahol igen, és az idősor alakulása ezt indokolta is). Ez magával hozta az alkalmazott módszerek eredményeinek a különböző időszakok közötti összehasonlítását, a szerzett tapasztalatok áttekintését és azok lehetséges dokumentációját.

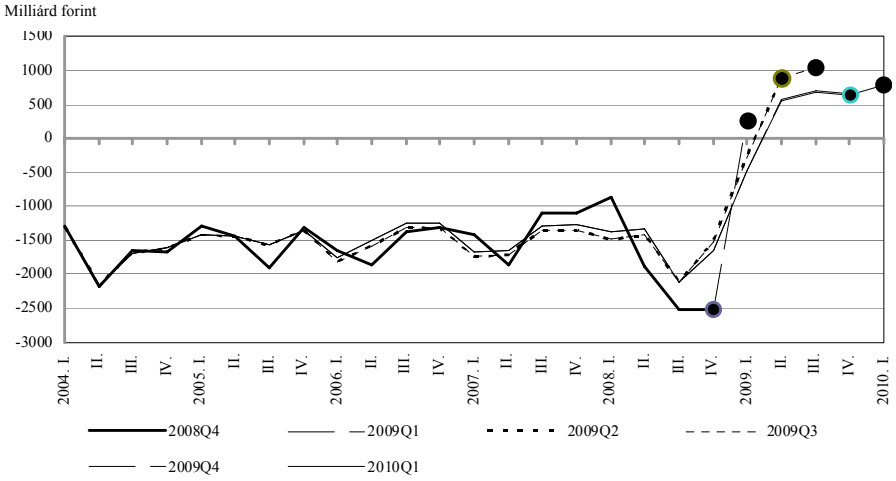
3. 1. Szezonális kiigazítás a fizetési mérleg idősorokban

A szezonális kiigazítás eredményeinek az időbeli változásain túl a fizetési mérleg idősorok összeállításának természetes velejárója az alapadatok revíziója is, az alkalmazott adatgyűjtési és feldolgozási rendszerből fakadóan. Az alapadatok ezen revíziója pedig önmagában is módosíthatja a szezonális kiigazítás eredményeit. A 4. ábrán az alapadatok revízióját tekinthetjük át a külső finanszírozási igény[12] adatain keresztül, amelyek az elemzők érdeklődésére talán leginkább számot tartanak. Az ábra alapján látható, hogy az alapadatok komolyabb revíziójára a 2009. II. negyedév publikációjakor került sor, a 2009. IV. negyedév publikálásakor végrehajtott alapadatrevízió, bár hosszú időszakot érintett visszamenőlegesen, leginkább csak az előző két negyedévi adatot érintette. Ugyanakkor a kiigazítatlan adatok alapján az idősor alakulása 2009 I. negyedévében éles fordulatot vett a gazdasági válság eredményeképpen: a külső finanszírozási igény jelentősen visszaesett, sőt ekkortól Magyarország

[12] A külső finanszírozási igény röviden azt jelzi, hogy a nemzetgazdaság mennyiben kényszerül külföldről forrásokat bevonni, illetve mennyire képes a külföldet forrásokkal ellátni. Technikailag ez lehet „felülről", illetve „alulról" számított, az előbbi a folyó fizetési mérleg és a tőkemérleg együttes egyenlegét, az utóbbi a pénzügyi mérleg (devizatartalékok változását is figyelembe vevő) egyenlegét jelenti. A két számítás elméletileg azonos eredményt ad, ám a fizetési mérleg statisztikai hibája a gyakorlatban eltérést okoz közöttük. A továbbiakban a „felülről" számított külső finanszírozási igény alakulását mutatjuk be, amint az a fizetési mérleg sajtóközleményekben is szerepel.

vált a külföld finanszírozójává, és a válság hatása megjelent a fizetési mérlegben. Ez természetesen nem hagyta érintetlenül a szezonálisan kiigazított adatokat sem, ahogy az az 5. ábrán látható.

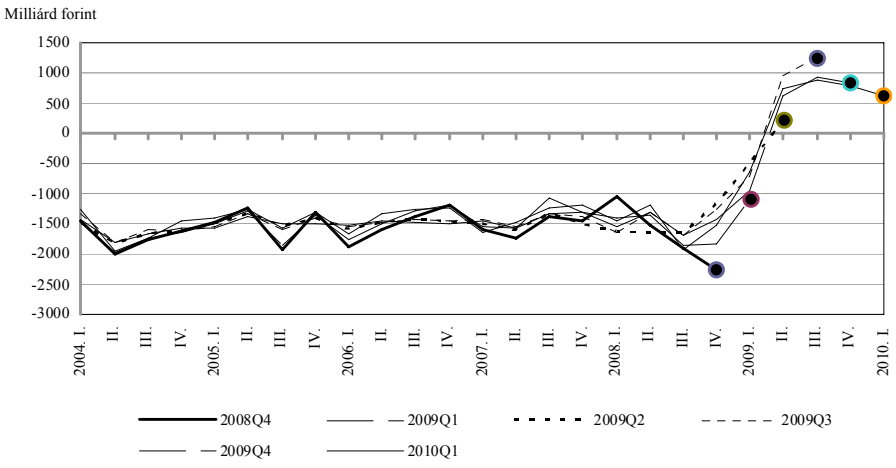
4. ábra A külső finanszírozási igény alakulása (eredeti adatok)



Megjegyzés. Itt és a következő ábráknál a fekete pontok azt jelzik, hogy az egyes időszakok utolsó értékei hol helyezkedtek el.

Forrás: MNB.

5. ábra A külső finanszírozási igény alakulása (szezonálisan kiigazított adatok)



Forrás: MNB.

A 2009. I. negyedévi hirtelen irányváltás abban az időszakban még nem változtatta meg a szezonális kiigazító program által becsült modellbeállítást: az általa korábban adott Airline modell továbbra is használhatónak bizonyult. Noha az általános gazdasági információk elhúzódó recesszióról szóltak, és a külső finanszírozási igény a szezonálisan kiigazított adatok alapján is jelentősen csökkent az előző negyedévhez viszonyítva, a szezonális kiigazítás még nem jelzett kiugró értéket. Ez természetesen nehezítette az interpretációt, bár az idősor alakulása összhangban állt más gazdasági változókkal.

A későbbi szezonális kiigazítások – utólag – érdemben már nem befolyásolták a külső finanszírozásról „irányváltozásáról" kialakult képet, azonban 2009 III. negyedévében modellváltásra került sor, mivel az idősorban a kiugró érték jelenléte és típusa (szinteltolódás) ekkor már szemmel is látható volt, és az addig használt Airline modell ezt továbbra sem tartalmazta. Az új modellbeállítás eredményeképpen a külső finanszírozási igény alakulását, úgy tűnik, immár az autoregresszív folyamatok írják le statisztikailag elfogadhatóbban. Ez a modell ugyanis érzékeli – összhangban az egyéb gazdasági változókkal –, hogy tartós változás következett be a külső finanszírozás nagyságrendjében, amit kiugró értékként (szinteltolódásként) kell kezelni, és 2009 II. negyedévére ezt a szinteltolódást azonosította is.

Összességében a külső finanszírozási igény szezonális kiigazítása során a 2009. III. negyedéves adatok publikálásakor az MNB a kiugró értékek alkalmazása mellett döntött, mert az idősor alakulása és a változások mértéke ezt indokolttá tette. Ez egyben modellváltással is együtt járt, azonban ez nem okozott lényeges revíziót a szezonálisan kiigazított idősorban.

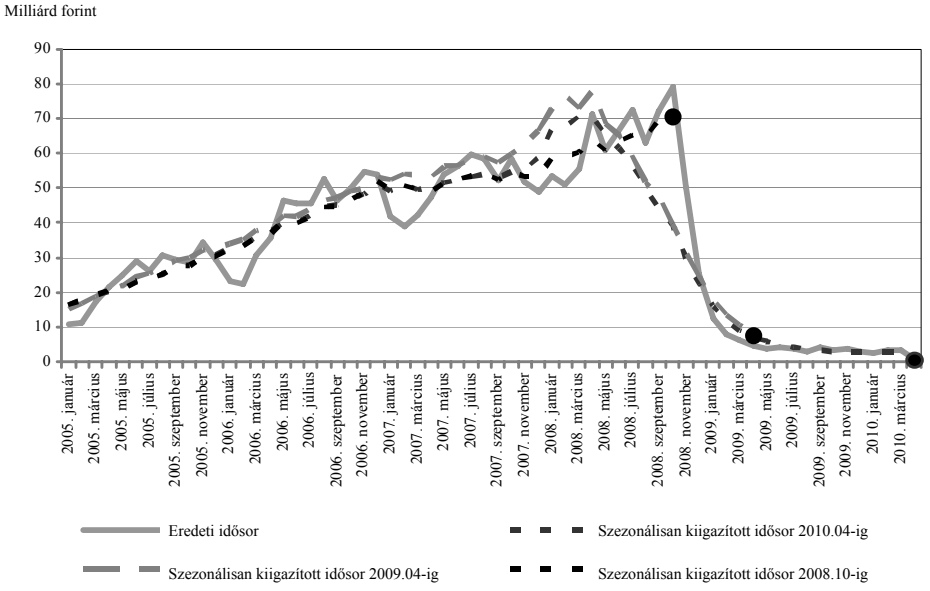
3.2. Szezonális kiigazítás a monetáris statisztikai idősorokban

A válság hatására erőteljes változások történtek a pénzügyi idősorokban is, és ennek következtében a sokszor szinteltolódással járó jelenségek problémát okoztak a szezonális kiigazítás során. Egyes esetekben sikerült pontosan kiszűrni a szezonalitás hatását az idősorokból, azonban előfordult, hogy a pénzügyi válság hatásainak következtében a szokásos kiigazítási módszerek nem bizonyultak használhatónak. Előbbire példa a háztartások hiteltranzakcióit, utóbbira pedig a háztartások svájci frank alapú lakáscélú hitelek új szerződéseinek értékét leíró idősor.

A háztartások svájci frank alapú lakáscélú hiteleinek új szerződéses értékének szezonális kiigazításakor az MNB szakértői azon korábban már említett, az Eurostat ajánlásainak megfelelő és az MNB-ben is alkalmazott gyakorlatot követték, hogy minden év elején új modellt illesztenek az addig megfigyelt adatokra, és ezzel a modellel hajtják végre az adott idősor szezonális kiigazítását egész évben. Az 2009. év elején rögzített modell alapján a 2008 szeptemberében kezdődő és 2009 márciusáig-áprilisáig tartó erőteljes csökkenés következtében azonban 2008 szeptemberében, ok-

tóberében és novemberében kritikus mértékű szezonhatás adódott,[13] így annak publikálhatósága kérdésessé vált. Miután világossá vált, hogy az említett modellel nem végezhető el az idősor elfogadható igazítása, a paraméterek megváltoztatására, valamint automatikus modellillesztésre is sor került, azonban így sem sikerült kielégítő eredményt elérni (lásd a 6. ábrát), ellentétben a háztartások hiteltranzakcióit leíró idősorral, ahol a megfelelő modell illesztésével már nem jelentett gondot a válság hatására módosult idősori értékek kezelése (lásd a 7. ábrát). Elfogadható szezonális kiigazításra csak azután kerülhetett volna sor, hogy az idősor egy erőteljes csökkenést követően 8–10 hónapig nagyságrendileg azonos szinten ingadozik. Azonban a pénzügyi válság miatt a svájci frank alapú lakáscélú hitelek új szerződéses értéke olyan alacsony szintre csökkent, hogy bár az idősor szezonális kiigazítása technikailag a 2010-ben újonnan illesztett modellel ma már megoldható lenne, de annak jelentősége gazdasági értelemben beszűkült, így az eredmények publikálásától eltekintettünk.

*6. ábra. Háztartások svájci frank alapú lakáscélú hitelei új szerződéses értékének alakulása**
(szezonálisan kiigazított adatok)

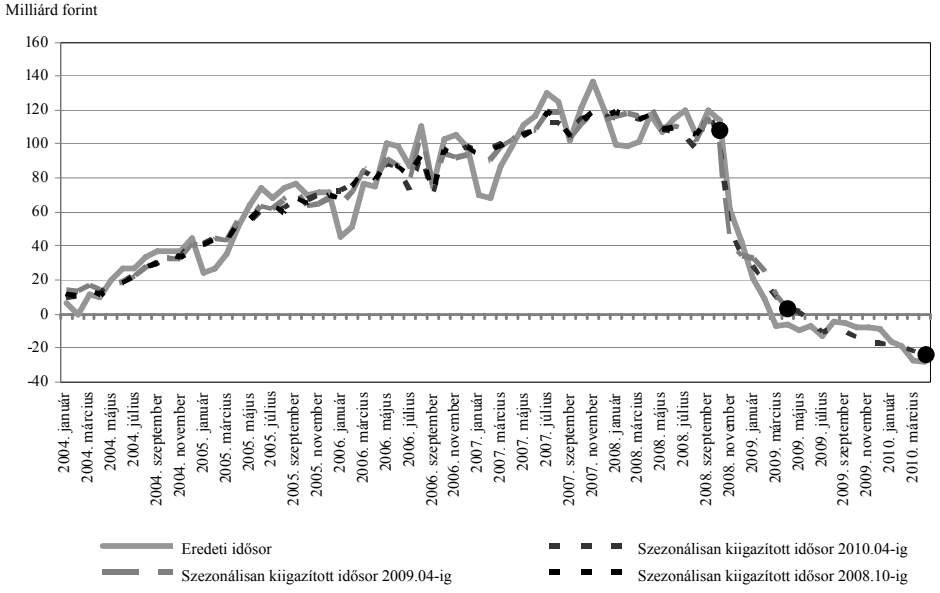


* Az ábra csupán illusztráció, az itt 2009-ben, illetve 2010-ben végződő idősorok nem kerültek publikálásra.
Megjegyzés. Valamennyi igazítás az adott év elején illesztett modellel történt.
Forrás: MNB.

[13] A 2009. év elején illesztett modellben a három őszi hónap szezonális faktora a 2005–2007-es évek 100–110 milliárd forintos értékeiről 2008-ban 160–200 milliárd forintra növekedett, majd a 2010-es illesztés során 2009-ben ismét visszatért a korábbi szintjére.

A háztartások hiteltranzakciói esetében a teljes idősor (2001. június–2010. április) rövidítésével – az első 31 hónap elhagyásával – sikerült elérni a megfelelő minőségű szezonális kiigazítást. A 2008-ban használt idősoros modell mindaddig megbízható eredményekkel szolgált, amíg a pénzügyi válság első jelei miatt nem változott meg az idősor. Ezen negatív hatások kiküszöbölésének érdekében 2008 decembere után a kiigazítás első időpontjának 2004. januárt tettük meg. Ekkor az alkalmazott TRAMO/SEATS-módszer által produkált eredmények már megfeleltek a tesztelés során felállított mindenkori kritériumoknak, és a 2009. valamint a 2010. év elején illesztett új modellek nem eredményeztek jelentős mértékű visszamenőleges revíziót a szezonálisan kiigazított idősorban.

7. ábra. Háztartások hiteltranzakcióinak alakulása
(szezonálisan kiigazított adatok)



Megjegyzés. Valamennyi igazítás az adott év elején illesztett modellel történt.
Forrás: MNB.

A teljes időszak hármas tagozódását igazolta a szezonális faktorok utólagos vizsgálata. Annak ellenére, hogy a 2001-ig visszamenőleges idősor additív modellel történő – kiigazítása negatív diagnosztikát eredményezett, a szezonalitás mintája változásának közelítő becslésére megfelelőnek bizonyult.[14] A tíz év egyes hónapjaihoz

[14] E modell szintén nem került publikálásra, lefuttatása és vizsgálata elemzési célokat szolgált.

tartozó szezonalitás abszolút és relatív értékeinek minden hónapra kiszámított normalizált szórása alapján egyértelműen kitűnik a három időszak eltérő jellege.[15] (Lásd a táblázatot.) Ugyanis ezen mutató abszolút értéke a 2004-től 2007-ig tartó időszakban inkább a relatív, míg a 2008-tól 2010-ig tartó időszakban inkább az abszolút faktorok esetén volt alacsonyabb, így mutatva a szezonalitás aktuális jellegzetességét (azaz, hogy az additivitás inkább jellemző volt a 2008–2010, mint a 2004–2007 közötti időszakban); az első három évre (2001–2003) számított értékek alapján nem szűrhető le ilyen egyértelmű megállapítás.

A tapasztalati szezonális faktorok normalizált szórása

Hónap	Abszolút faktor esetén			Relatív faktor esetén		
	2001–2003	2004–2007	2008–2010	2001–2003	2004–2007	2008–2010
Január	0,6567	–0,5688	–1,2104	–3,8493	–0,1792	–0,6363
Február	–0,3391	–0,3298	–0,9222	–0,1042	–0,7231	–1,3202
Március	0,8174	–0,4193	–0,3301	0,9246	–0,3397	–2,2795
Április	–0,4605	–0,8130	–0,8024	–0,0313	–1,5061	–1,9279
Május	–3,8713	0,6080	–1,1266	–3,8606	0,3633	1,4150
Június	–5,1809	0,3493	2,7556	1,6909	0,4352	1,0805
Július	1,3777	0,6103	1,8992	–1,8419	0,1095	0,2291
Augusztus	1,4584	0,2162	0,3674	–1,8380	0,4758	–1,9546
Szeptember	0,8829	13,3631	5,7722	–2,8688	2,7008	–1,2860
Október	3,0236	0,5658	0,0305	2,0207	0,7873	–1,8444
November	–2,0619	0,9354	0,3426	18,4632	0,5864	–6,8574
December	–1,5702	0,8455	0,0347	–2,0407	0,8166	–3,9449
Átlag	–0,4389	1,2802	0,5675	0,5554	0,2939	–1,6106

Forrás: MNB.

4. Záró megjegyzések

Általánosságban megállapítható, hogy amíg a gazdaság belső átrendeződése folyamatban van, azaz nem alakul ki egy új, stabil állapot, addig a szezonális kiigazítási eljárás nem feltétlenül tud megfelelő pontosságú képet adni a gazdasági folyamatok alakulásáról, mivel a modellezési eljárások a megfigyelésekre, a múltbeli adatok-

[15] Abszolút faktoron a megfigyelt idősori érték és a szezonálisan igazított érték különbségét, míg relatív faktor esetén azok hányadosát; normalizált szóráson pedig az adott hónaphoz tartozó szezonális faktorok szórásának azoknak átlaguk egységére vetített értékét értjük.

ra és azok összefüggéseire támaszkodnak. A szezonális kiigazításra vonatkozó nemzetközi ajánlások célja többek között az, hogy az adatelőállítási folyamat önkényességét csökkentse, és azok transzparenciáját növelje. Mindezek ellenére a szezonális kiigazítás magában hordozza azt a konfliktust, mely szerint az adatelőállítónak mérlegelni kell, hogy a rögzített vagy az újrabecsült közgazdasági és matematikai-statisztikai szempontoknak esetleg jobban megfelelő modellt válassza, amennyiben ezek nem azonos következtetésekre vezetnek.

Ezen megfontolások a különböző statisztikák esetében a gazdasági válság kezelését illetően eltérő eredményeket hoztak:

– az *ipar és számos havi gyakoriságú idősor* esetén a szezonális kiigazítás során szinteltolódás típusú kiugró érték került bevonására, az évi egyszeri modell és paraméterrögzítés mellett,

– a *GDP* szezonális kiigazításakor a KSH revíziós politikájának megfelelően történt a modellek és a paraméterek újrabecslése, matematikai-statisztikai és közgazdasági szempontok alapján nem került sor kiugró értékek bevonására,

– a *külső finanszírozási igény* kiigazításakor, az idősorok alapos közgazdasági és matematikai-statisztikai elemzését követően, az évközi modellváltás és egy kiugró érték (szinteltolódás) bevonása volt indokolt,

– a *monetáris statisztika* idősorok esetében évközi modellváltásra a vizsgált idősoroknál nem került sor, vagy azért, mert már az sem igazán segített volna a kiigazítatlan idősor extrém alakulása miatt, vagy azért, mert más eszközökkel – például a kiigazítandó idősor kezdőpontjának a módosításával – sikerült biztosítani az elfogadható szezonális kiigazítást.

A nemzetközi ajánlások és követett gyakorlatok mindezt lehetővé teszik, azonban nagyon fontos egyrészt a már említett transzparencia, másrészt az idősorok közötti konzisztencia annak érdekében, hogy a szezonális kiigazítás elérje célját, azaz tiszta, hiteles és pontos információt tudjon szolgáltatni a felhasználók és a döntéshozók számára.

Irodalom

BAUER P. – FÖLDESI E. [2004]: A szezonális kiigazítás harmonizációja a Központi Statisztikai Hivatalban. *Statisztikai Szemle*. 82. évf. 8. sz. 691–704. old.
http://www.ksh.hu/statszemle_archive/2004/2004_08/2004_08_691.pdf

EUROSTAT – KSH (KÖZPONTI STATISZTIKAI HIVATAL) [2007]: *Seasonal Adjustment Methods and Practices*.
http://epp.eurostat.ec.europa.eu/portal/page/portal/ver-1/quality/documents/SEASONAL_ADJUSTMENT_METHODS_PRACTICES.pdf (Elérés dátuma: 2010. július 6.)

EUROSTAT [2009]: *ESS Guidelines on Seasonal Adjustment.*
http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-RA-09-006/EN/KS-RA-09-006-EN.PDF (Elérés dátuma: 2010. július 6.)

FERENCZI B. – JAKAB M. Z. [2002]: *Kézikönyv a magyar gazdasági adatok használatához*. http://www.mnb.hu/engine.aspx?page=mnbhu_egyebkiadvanyok_hu (Elérés dátuma: 2010. július 6.)

KSH (KÖZPONTI STATISZTIKAI HIVATAL) [2005]: *Szezonális kiigazítás.* Statisztikai módszertani füzetek, 43. http://portal.ksh.hu/pls/ksh/docs/hun/xftp/idoszaki/pdf/szezonkiig.pdf (Elérés dátuma: 2010. július 6.)

KSH (KÖZPONTI STATISZTIKAI HIVATAL) [2010]: *Minőségi irányelvek a Központi Statisztikai Hivatal statisztikai munkafolyamatainak egyes szakaszaira*.
http://portal.ksh.hu/pls/ksh/docs/bemutatkozas/hun/minosegi_iranyelvek.pdf

Koroknai Péter,
az MNB vezető közgazdasági elemzője
E-mail: koroknaip@mnb.hu

Pellényi Gábor,
az MNB junior elemzője
E-mail: pellenyig@mnb.hu

Szezonális kiigazítás a gazdasági válságban – felhasználói szemmel

A makrogazdasági elemzők napi munkájához nélkülözhetetlenek a megbízható szezonálisan igazított idősorok. Az elemzők feladata a legfrissebb gazdasági folyamatok azonosítása és a jövőbeli trendek előrejelzése. Az első célnak leginkább a szezonálisan igazított adatokból számított rövid bázisú indexek (például a bruttó hazai termék előző negyedévhez viszonyított volumenváltozása) felelnek meg. Az elemzői, előrejelzői munka során használt statisztikai, ökonometriai modellek pedig jóval egyszerűbbek lehetnek, ha eleve szezonálisan igazított adatokat tartalmaznak, és nem egyenként modellezik a változók szezonalitását.

A szezonális igazítás nem pusztán statisztikai kérdés. Például a kiugró értékek (outlierek) szűrése során implicite közgazdasági feltevéseket teszünk azzal kapcsolatban, hogy a rendhagyó megfigyeléseket miként kell értelmezni. A szezonális igazítás paramétereinek megválasztásakor amellett is letesszük voksunkat, hogy egy-egy változó szezonalitását mennyire tekintjük időben stabilnak. E feltevések jelentésével az elemzőknek akkor is tisztában kell lenniük, ha a szezonális igazítást nem maguk, hanem a statisztikai adatszolgáltatók végzik. Az elmúlt években kibontakozó nemzetközi gazdasági válság idején számos erőteljes hatás érte a gazdasági idősorokat, melyek részben előzmény nélküliek voltak. Így ezek szezonális igazítása során új problémák merülhettek fel, illetve a régi, lezáratlan kérdések jelentősége felértékelődhetett. Mindezek eredményeként megnőtt a szezonális igazítás egyébként is meglévő végponti bizonytalansága, azaz a legfrissebb folyamatok a szokásosnál nagyobb bizonytalansággal mérhetők fel.

A bizonytalanság erősödése elemzői szempontból főleg azért fontos kérdés, mert az adatrevíziók nyomán megváltozhatnak az idősorokból levont következtetések. Így az aktuális adatok segítségével meghozott üzleti és gazdaságpolitikai döntések később tévesnek bizonyulhatnak. *Orphanides* [2001] bemutatja, hogy a valós idejű adatokra alapozott monetáris politika érdemben eltérhet a revideált adatokra alapozott optimális monetáris politikától. Az üzleti és gazdaságpolitikai döntéshozók optimálistól eltérő döntései pedig jelentős reálgazdasági költségekkel is járhatnak. Tehát a statisztikai adatszolgáltatás, a szezonális igazítás és az elemzői munka minőségének számottevő jóléti hatásai lehetnek.

Cikkünkben néhány makrogazdasági idősor példáján keresztül mutatjuk be, hogy a gazdasági válság milyen nehézségek elé állította az elemzőket. A problémák felvázolásán túl a hazai és nemzetközi tapasztalatok alapján bemutatjuk, milyen technikákkal lehet mérsékelni a szezonális igazítás végponti bizonytalanságát.

1. Reálgazdasági idősorok – a fordulópontok azonosításának nehézségei

A bruttó hazai termék (GDP) talán a legfontosabb reálgazdasági mutató, mely a gazdaság egészének folyamatait igyekszik megragadni. Jelentőségét növeli, hogy több fontos viszonyszám (például hiánymutatók, adósságráták) nevezője. Így a makrogazdasági helyzet értékelésében kiemelt jelentőséggel bír a GDP alakulása. Elemzők számára központi kérdés az üzleti ciklusok fordulópontjainak minél kisebb késéssel való azonosítása. Egy gyakran hivatkozott hüvelykujjszabály szerint a recesszió kezdetét (végét) az jelzi, ha a szezonálisan igazított GDP szintje két negyedéven keresztül csökken (növekszik).

A gyakorlatban nem ilyen egyszerű a fordulópontok felismerése, mivel a szezonálisan igazított GDP negyedéves lefutása minden új adatpont beérkeztével jelentősen változik. Ez arra vezethető vissza, hogy a negyedéves frekvenciájú idősorok rövidek, így a szezonális igazítás paraméterbecsléseinek statisztikai bizonytalansága jelentős lehet. Másrészt a GDP szezonálisan igazítatlan idősorait rendszeresen revideálják, mely a változó szintjét és dinamikáját is érintheti.[1]

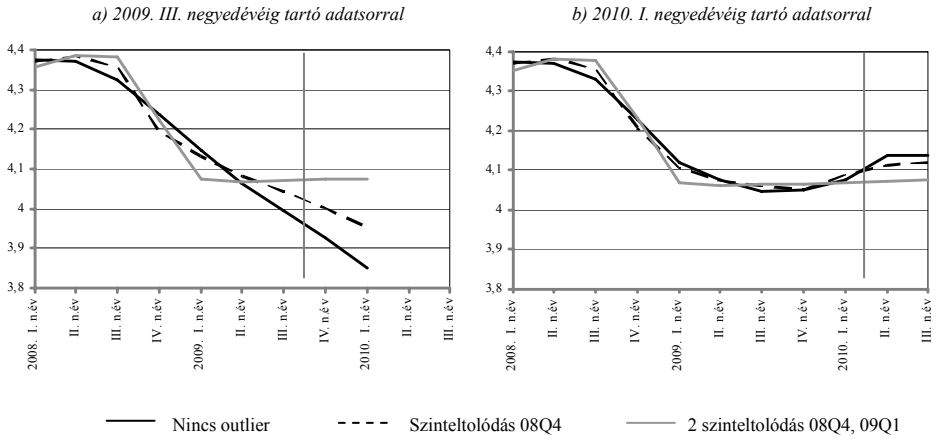
A szezonális igazítás bizonytalansága a 2008-ban kirobbant nemzetközi válság idején is megfigyelhető volt. E bizonytalanság jelentős részben abból fakadt, hogyan ítéljük meg a válság természetét. Szólnak érvek amellett, hogy a válság a GDP idősorában szinteltolódást okozhatott: pénzügyi válságok nyomán jellemzően permanensen csökken a potenciális kibocsátás szintje, majd általában visszatérhet a válság

[1] Nemzetközi tapasztalatok szerint a szezonálisan igazított GDP idősoraiban végrehajtott revíziók fő oka az igazítatlan adatok változása, de szignifikáns a szezonális igazítás paramétereinek változásából fakadó revízió is (lásd például *Fixler–Grimm* [2002], *Mainwaring–Skipper* [2007]; *Mehrhoff* [2008]).

előtti növekedési ütem (*Abiad et al*. [2009]). Ám akadnak ellenérvek is: a válság idején valószínűsíthető volt, hogy Magyarországon elhúzódik a recesszió és csak lassú kilábalásra számíthattunk.[2] E logikából viszont nem egyszeri szinteltolódás, hanem tendenciaszerű visszaesés következik.

A visszaesés időszakában a szezonális kiigazító program automata módban két szinteltolódást azonosított (2008. IV. és 2009. I. negyedév, lásd az 1. ábra *a*) panelt). E beállításokkal a GDP szintje 2009 közepén már stagnálást mutatott volna. Szinteltolódás nélkül (vagy egyetlen szinteltolódással) a visszaesés folytatódására utalt a szezonális igazító eljárás. Ám 2010 I. negyedévéből visszatekintve már azt láthatjuk, hogy a két szinteltolódásos modell a válság után tartósan stagnáló GDP-t sugallna (lásd az 1. ábra *b*) panelt). Más beállításokkal ennél valószerűbb forgatókönyv – fokozatos, lassú felívelés – képe rajzolódik ki. Így egy adott negyedévben optimálisnak tűnő paraméterezés visszatekintve már nem feltétlenül tűnik közgazdaságilag értelmesnek. Ám a gyakori paraméterváltás okozta revíziók csak tovább fokozzák a szezonális igazítás körüli bizonytalanságot.

1. ábra. A szezonálisan igazított GDP szintje a válság idején, az outlierekre vonatkozó különböző feltevésekkel
(ezer milliárd forint, 2000. évi áron)



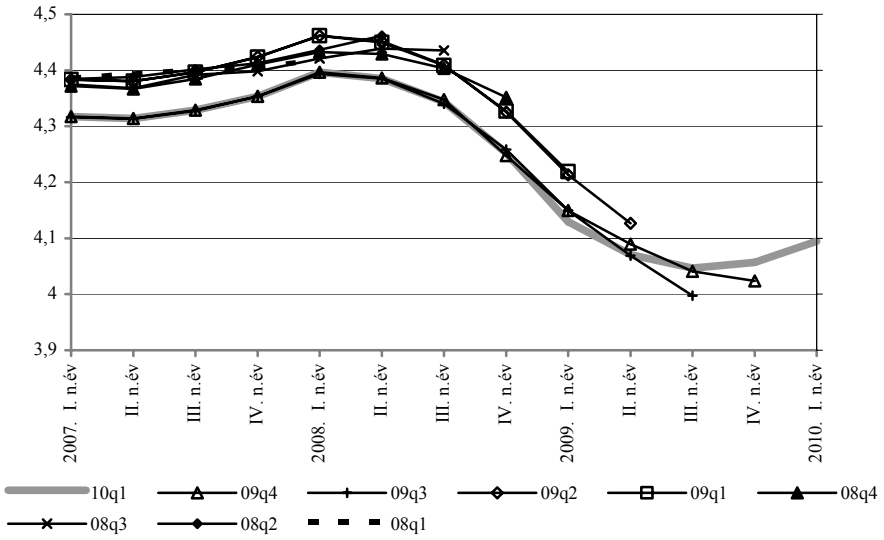
Megjegyzés. A függőleges vonalig a tényadatok, utána az ARIMA-előrejelzés látható.

Mikor azonosították a KSH által publikált, szezonálisan igazított idősorok a konjunktúra fordulópontjait? A cikk írása idején rendelkezésünkre álló információk szerint a hazai konjunktúra 2008 I. negyedévében érte el csúcsát. A mélypont hat ne-

[2] Ennek fő oka az volt, hogy a nagy adósságállományokból fakadó sérülékenység miatt Magyarországon a gazdaságpolitika prociklikus módon volt kénytelen reagálni a válságra, valamint a bankrendszer prociklikus viselkedése is erőteljesebb lehetett, mint más országokban.

gyedévnyi visszaesés után 2009. III. negyedéve lehetett, ezután indulhatott meg a következő növekedési fázis. (Lásd a 2. ábrát.)[3] A recesszió kezdete csak a 2008. IV. negyedévi (előzetes) adatközléssel vált ismertté, 2009. február 16-án. Ez „valós időben", a konjunktúra tetőpontját jelentő negyedév közepéhez viszonyítva éppen egy évnyi késést jelent. Ám pusztán a GDP adatközlésekre támaszkodva is 6 hónap a késés, hiszen már a 2008. II. negyedévi adatközlésből érzékelhető lett volna a visszaesés a megelőző negyedévhez képest. A mélypont jelenleg ismert időpontjára először a 2010. I. negyedévi előzetes adatból következtethettünk, amely 2010. május 12-én jelent meg: ezúttal 9 hónap a „valós idejű" késés, és 3 hónap a szezonálisan igazított GDP-adat végponti bizonytalansága miatti késés.[4]

2. ábra. A szezonálisan igazított és kiegyensúlyozott GDP alakulása az egyes adatközlések alkalmával
(ezer milliárd forint, 2000. évi áron)



Más megközelítésben, a recesszió kezdetén (2008. II. negyedévében) a szezonálisan igazított GDP idősora még töretlen növekedési trendet jelzett. Hasonlóan, a 2009. IV. negyedévben publikált adat a GDP folytatódó (bár mérséklődő) visszaesését jelezte. A példa tanulsága, hogy a szezonálisan igazított GDP nagy késéssel azonosíthatja a konjunktúra fordulópontjait, és valós időben félrevezethető információt

[3] A jövőben beérkező adatok hatására természetesen ismét változhat a szezonálisan igazított idősor lefutása, és ennek hatására a fordulópontok is módosulhatnak.

[4] Az előzetes GDP-adat körülbelül másfél hónappal a tárgynegyedév vége után kerül publikálásra. Így a bemutatott késések elsősorban nem az adatközlés késéseivel, hanem a szezonális igazítás bizonytalanságaival magyarázhatók.

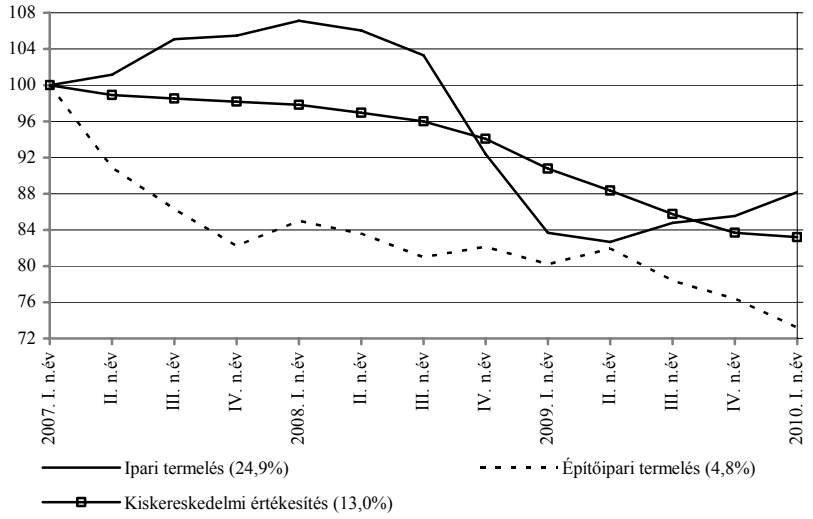
adhat a gazdaság trendfolyamatairól. Ez pedig hibás üzleti vagy gazdaságpolitikai döntésekhez is vezethet, melynek válságos időszakban súlyos következményei is lehetnek. Másfelől e késések összefügghetnek azzal is, hogy különböző megfontolások (közgazdasági érvek, modell stabilitás) alapján választott paraméterek mellett a szezonális igazítás fokozatosan „tanulja" az idősor trendjét.

Hogyan kezelhető a szezonálisan igazított GDP-idősor körüli bizonytalanság? Egyértelmű recepttel nem rendelkezünk, de több módszerrel pontosíthatjuk a gazdasági folyamatok irányáról alkotott képünket. Egyrészt a GDP alakulása összevethető más, dezaggregáltabb makrogazdasági mutatók dinamikájával. A GDP számos mutató aggregátumaként áll elő. A termelési, felhasználási vagy jövedelmi oldali résztételekről számos információval rendelkezünk, negyedéves vagy akár havi frekvencián. A résztételekben megfigyelt folyamatoknak végül az aggregált GDP alakulásában is tükröződniük kell. Így a résztételek szezonális igazítása során nyert információk (például azonosított kiugró értékek, szezonalitás változásai) felhasználhatók a GDP szezonális igazításakor is.

A gyakorlatban a termelési oldali folyamatokról rendelkezünk a legtöbb és legfrissebb információval. Már néhány szektor havi kibocsátási statisztikáiból is levonhatók kvalitatív következtetések a GDP alakulásáról. (Lásd a 3. ábrát.)

*3. ábra. Az ipari termelés, az építőipari termelés és a kiskereskedelmi értékesítések alakulása**

(szezonálisan igazított szintek, 2007. I. negyedév = 100 százalék)



* MNB szezonális igazítás.

Megjegyzés. A jelmagyarázat zárójeleiben az ágazatok 2009. évi alapáras hozzáadott értékből vett részesedései szerepelnek.

Az ipari termelés havi idősora 2008 decemberében szinteltolódást tartalmaz, mely világszerte megfigyelhető jelenség volt, és a nemzetközi kereskedelem hirtelen leállásával, a készletek erőteljes leépítésével magyarázható. 2009 második felében azonban már élénkült az ipari konjunktúra, melyet a vállalati készletek újbóli felépítése mellett a nemzetközi keresletélénkítő lépések (például roncsautó-programok) is támogattak. Ezzel szemben a belső kereslettől függő építőipar és kiskereskedelem elhúzódó visszaesést mutatott 2007 óta, bár a kiskereskedelmi értékesítések visszaesésének üteme 2010 elején már mérséklődni látszott. Mivel ismert a szektorok megtermelt hozzáadott értékből való részesedése, ezért az ágazati folyamatokat megfelelően súlyozva lehet következtetni a GDP alakulására is.

Ugyanakkor a GDP dinamikájával kapcsolatban információt hordozhatnak különböző bizalmi indexek, makrogazdasági változók is. Az efféle megfigyelések segíthetnek ellenőrizni, hogy mennyire „hihető" eredményt ad a GDP szezonális igazítása. Nagyszámú indikátor információtartalmát statisztikai eszközök, többek között bridge modellek (*Baffigi et al*. [2004]) vagy faktormodellek (*Forni et al*. [2005], *Stock–Watson* [2002]) útján lehet egyszerre hasznosítani.

Végül a statisztikai és közgazdasági előrejelző irodalom régóta foglalkozik a várható adatrevíziók mértékének előrejelzésével, az ismert előzetes adatközlés függvényében (*Koop et al.* [2008]). E módszert jellemzően az előzetes és végleges adatközlések közötti eltérés becslésére használják (nem a szezonális igazításból adódó visszamenőleges revíziók mérésére), és sikeres alkalmazásához kellően hosszú idősorokra van szükség. Ám egyszerű technikákkal hasonló eredmény érhető el. Ha külső információval (például megbízhatónak tartott előrejelzéssel) rendelkezünk a változó jövőbeli értékeiről, akkor a tényadatok szezonális igazítása összevethető a jövőbeli értékekkel meghosszabbított idősor igazításával. Hasonlóan, értékelhető a szezonális igazító program által adott előrejelzés közgazdasági értelme is. Például a GDP-re adott ARIMA-előrejelzéstől recesszió idején elvárható, hogy ne tartós visszaesést mutasson, hanem előbb-utóbb jelenjen meg benne a fordulópont (igaz, fentebb láttuk, hogy ez gyakran csak utólag lesz azonosítható).

A hazai és a nemzetközi tapasztalatokból az a következtetés adódik, hogy a gazdasági folyamatok értelmezésekor – így szezonális idősorok igazítása során is – célszerű minél több idősorban rejlő információt felhasználni. Többek között az amerikai National Bureau of Economic Research is számos idősorra támaszkodva határozza meg a konjunktúra fordulópontjait. Hasonló céllal fejlesztették ki többek között az eurózóna gazdasági teljesítményét nyomon követő (New) Eurocoin indikátort is (*Altissimo et al*. [2007]).[5]

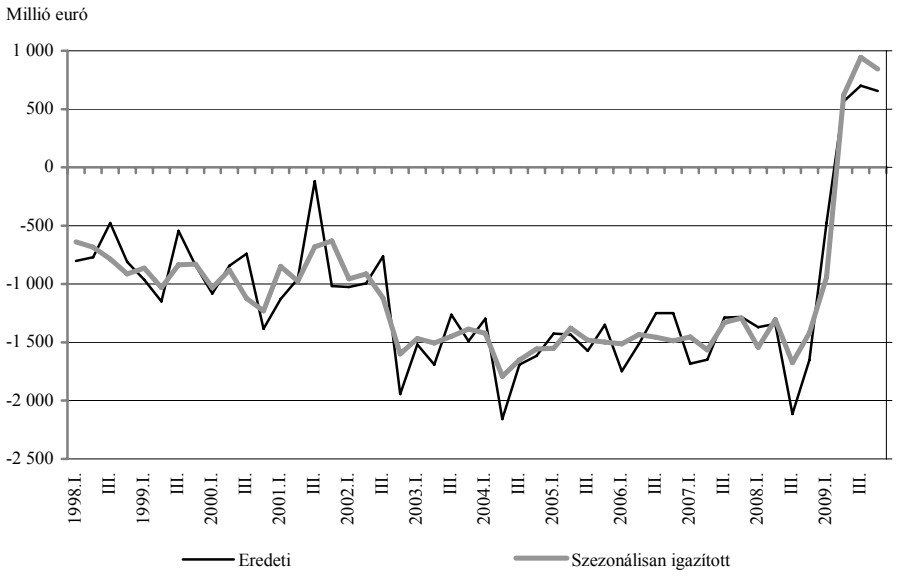
[5] E megközelítésben a GDP egy nem megfigyelt változó (a konjunktúra) zajos indikátorának tekinthető. Így a fő cél nem feltétlenül a GDP rövid távú változásainak minél pontosabb előrejelzése, hanem a mélyebb konjunktúrafolyamatok azonosítása, melyeket a GDP csak tökéletlenül jelez.

2. Külső finanszírozási igény – megnövekedett végponti bizonytalanság

Válság esetén a külső egyensúly korrekciójának mértéke és időtartama a válság jellegétől, illetve az arra adott gazdaságpolitikai válaszoktól függően lehet tartós vagy átmeneti. A feltörekvő országokban a külső finanszírozási igény korrekciója elsősorban a tőkebeáramlás megtorpanásával áll összefüggésben.[6] A külső finanszírozás visszaesése az esetek többségében az árfolyam jelentős leértékelődésével és a gazdasági növekedés visszaesésével jár együtt, aminek következményeként a folyó fizetési mérleg hiánya válság esetén jelentős mértékben csökken. Ezzel összefüggésben ugyanakkor a külső egyensúly korrekciójának jellege is nagyban függ attól, hogy a tőkeáramlás mennyi idő alatt éri el korábban jellemző szintjét, a gyengébb árfolyam mennyi ideig marad fenn, a gazdasági visszaesés mekkora mértékű és mennyire tartós.

Magyarországon a jelenlegi válság hatására a külső finanszírozási igény nagy valószínűséggel tartósan és jelentősen alacsonyabb lehet a korábban jellemzőnél. A válság 2007-es első jeleinek idején, illetve a válság 2008-as eszkalálódásakor még nem volt egyértelmű, hogy a pénzpiaci krízis milyen súlyosan érinti a magyar gazdaságot.

4. ábra. A külső finanszírozási képesség alakulása



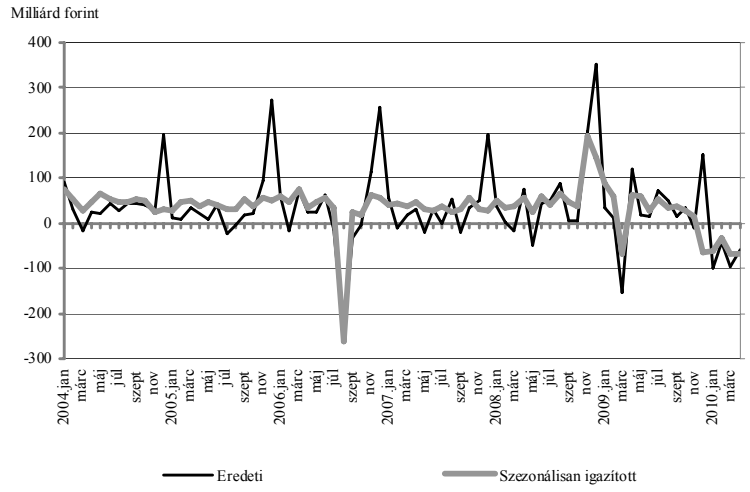
[6] A folyó fizetési mérlegben bekövetkezett korrekciók tapasztalatairól lásd: *Tóth* [2005], *Edwards* [2001].

A válság kitörését követően ugyanakkor egyértelművé vált, hogy az árfolyam előreláthatóan tartósan gyenge marad, és nem várható rövid időn belül fordulat a gazdasági növekedésben. Így az is valószínűsíthető volt, hogy a külső finanszírozási igényben tapasztalt visszaesés 2009 elején szinteltolódásként (level shift) fog megjelenni a szezonálisan igazított idősorban (lásd a 4. ábrát), ami rávilágít az egyes idősorok statisztikai és elemzői felhasználásában rejlő különbségekre. Míg elemzői szempontból szinteltolódás volt várható, a statisztika igazításában – a kialakult módszertan szerint rögzített modell eredményeként, illetve a végpontnál jelentkező nagyfokú bizonytalanság miatt – csupán a III. negyedéves adat publikációjakor, 2009 decemberében jelent ez meg. Kérdéses ugyanakkor, hogy a külső egyensúlyi folyamatokban bizonyosan bekövetkező fordulatra mikor kerül sor és milyen mértékű lesz, ugyanis ez az idősor jelenlegi szezonálisan igazított értékeinek akár jelentősebb változását is implikálhatja.

3. Lakossági betétek – a szokásostól eltérő lefutás összes esete

A válság kitörése és a helyzet normalizálódása is jelentős mértékben befolyásolta az egyébként viszonylag stabilan alakuló lakossági betételhelyezést. A válság 2008. szeptemberi kitöréséig – a kamatadó bevezetésének időszakától eltekintve – a lakosság banki betéteinek szezonálisan igazított tranzakciója egyenletesen alakult. Ezt követően a háztartási szektor több alkalommal is nagymértékben átrendezte portfólióját, ami jól nyomon követhető a banki betétek idősorában is. (Lásd az 5. ábrát.)

5. ábra. A háztartási szektor nettó banki betételhelyezése



Fontosnak tartjuk ugyanakkor hangsúlyozni, hogy ami a lakossági betétek esetében kiugró értékként jelent meg, az a lakosság pénzügyi követeléseiben, illetve a banki források alakulásában közel sem jelentett ilyen látványos változást. Egyrészt a banki betétek szokatlan változása – a portfolióátrendeződés eredményeként – a lakosság más pénzügyi követeléseinek alakulását is jelentős mértékben befolyásolta. Másrészt a betétkivonás hatására a bankokból kiáramló források is közvetve visszaáramlottak a hitelintézetekbe, hiszen a befektetési alapokban elhelyezett megtakarítások egy része is bankbetétként kerül elhelyezésre.

Lecsengő kiugró érték

2008 szeptemberében a piaci árak esése és a bizonytalanság növekedése mellett nagymértékben emelkedtek a betéti kamatok, és hangsúlyosabbá vált a piaci szereplők kockázatkerülő magatartása. Ebben a környezetben a lakosság igyekezett megszabadulni befektetési jegyeitől, és az így felszabadult megtakarításokat bankbetétbe helyezte el. Ez a folyamat 2009 elejéig tartott, aminek eredményeként a szezonális igazításban lecsengő kiugró érték jelent meg.

Egyszeri kiugró érték

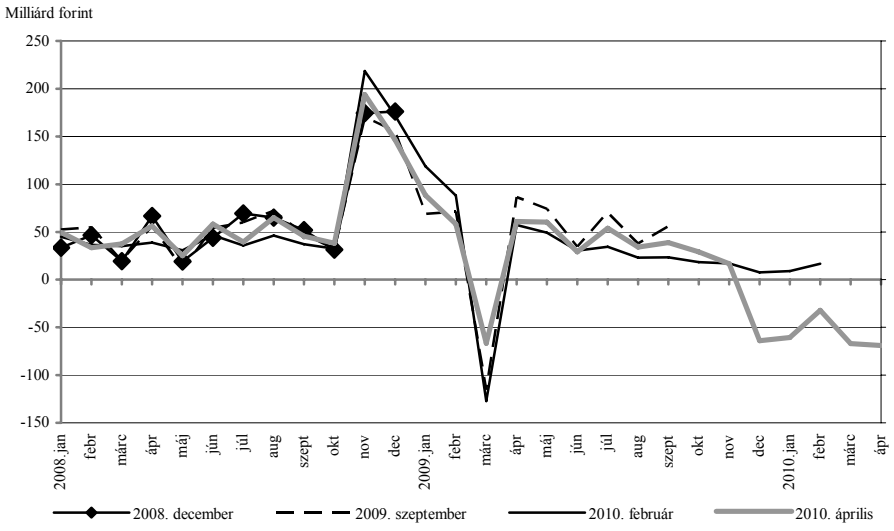
2009 márciusában az euróárfolyam és a magyar csődkockázat megugrását követően a lakosság készpénzbe – ezen belül is elsősorban valuta-készpénzbe – menekítette banki betéteinek jelentős részét. Az esetleges bankcsőd valószínűsége hamar visszaesett, így áprilisban már korrekció következett be, aminek következményeként ebben a hónapban szokatlanul nagy betételhelyezés volt megfigyelhető.

Szinteltolódás

2009 végétől a gazdasági kilátások javulásával párhuzamosan mérséklődött a kockázatkerülő magatartás, és korábban nem tapasztalt szintre csökkent a jegybanki alapkamat. A betéti kamatok csökkenése és az értékpapírok árainak növekedése azzal a következménnyel járt, hogy az elmúlt hónapokban a válság kitörésekor tapasztalttal ellentétes irányú portfolióátrendeződés történt: a lakosság folyamatosan építette le banki betéteinek állományát, amit aztán döntő részben befektetési jegyek vásárlására fordított. A szokatlan mértékű és tartósságú betétkivonás hátterében ugyanakkor a portfolióátrendeződés mellett két másik jelenség is meghúzódik: egyrészt a nehéz gazdasági helyzetben a lakosság összes pénzügyi követelése is a korábbinál kisebb mértékben emelkedik, másrészt pedig a kamatok csökkenése miatt a betéteken keletkező kamatbevétel is jóval kisebb a 2009 első felében jellemzőnél.

A lakossági betétek szezonális igazítása során az új adatpontok többször megváltoztatták a kiugró értékek jellegét. A 2008 végi jelentős betételhelyezések alapján a szezonális igazító program még szinteltolódást azonosított az idősorban, ezért a novemberi és decemberi szezonálisan igazított érték is igen magas lett. (Lásd a 6. ábrát.) Elemzői szempontból ugyanakkor az tűnt valószínűnek, hogy a betételhelyezés nem stabilizálódik hosszabb távon is ilyen magas szinten, hanem a válsághangulat mérséklődésével, illetve a nemzetközi hitelkeret biztosítását követően a megtakarítások szerkezete normalizálódik, és a betétek lassuló növekedése lecsengő kiugrást fog eredményezni. A 2009 eleji hónapok adataival kiegészített idősor szezonális igazítása során kapott eredmények később alátámasztották ezt a feltevést. Az elmúlt hónapokban bekövetkezett betétkivonást pedig a Demetra hosszú ideig még csak folyamatos csökkenésként azonosította, és csupán áprilisban váltott át az igazítás egy decemberben bekövetkezett szinteltolódásra, ami természetesen a korábbi hónapok igazított adatainak értékét is befolyásolta.

6. ábra. A háztartási szektor nettó banki betételhelyezésének szezonális igazítása különböző időpontokban

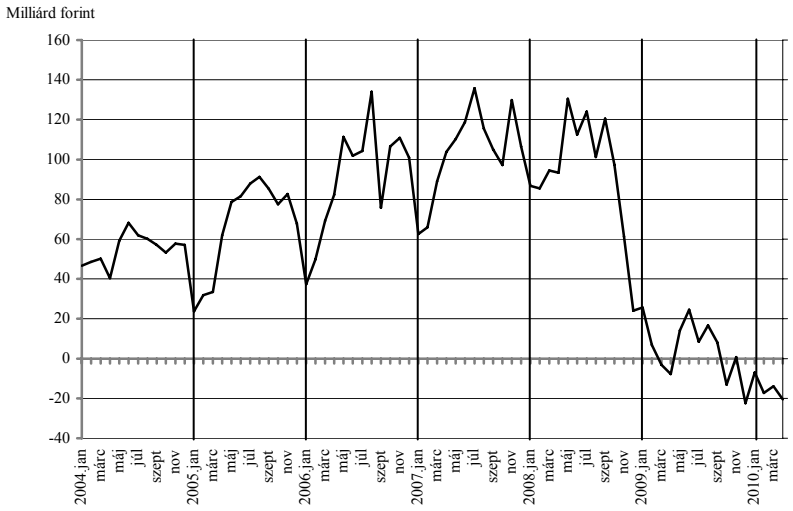


4. Lakossági hitelek – megváltozó szezonalitás

A szezonalitás megváltozása talán a lakosság banki hitelfelvételének idősorán a legszembetűnőbb, amiben az idősort determináló tényezők alakulása játszhatja a döntő szerepet. A háztartási szektor banki hitelfelvételének éves lefutása jellemzően úgy alakult, hogy januárban – a karácsonyi vásárlást követően – érte el mélypontját, majd fo-

lyamatos emelkedéssel a nyáron tetőzött – vélhetően az ingatlan-beruházásokkal összefüggésben –, majd az év végére újra csökkenésnek indult. (Lásd a 7. ábrát.) A válság egyrészt jelentősen visszavetette a lakosság hitelfelvételét, másrészt azonban – a korábban megszokott szezonalitással ellentétben – 2009 és 2010 első hónapjaiban további csökkenés volt megfigyelhető, ami az idősort jellemző szezonalitás megváltozására utal. A jelenség hátterében az állhat, hogy míg korábban, a likviditásbőség időszakában alapvetően a lakosság hitelkereslete határozhatta meg a hitelfelvétel nagyságát – hiszen a bankok külföldi hitelfelvételből azt meg tudták finanszírozni –, a válság kitörését követően a források beszűkülése azzal a következménnyel járt, hogy a lakosság hitelfelvétele nagyobb mértékben függött a bankok hitelezési hajlandóságától. A hitelkínálat által meghatározott hitelfelvétel ugyanakkor azt a kérdést is felveti, hogy érdemes-e egyáltalán igazítani a lakosság hitelfelvételének idősorát, hiszen azt ezek szerint már teljesen más szezonalitás jellemzi.

7. ábra. A háztartási szektor nettó banki hitelfelvétele



5. Összefoglaló

A makrogazdasági idősorok szezonális igazításában megjelenő szokásos bizonytalanság a válságban jelentősen megnövekedett. Ez különösen a végponti becslések estében állította nehézségek elé az elemző felhasználót, hiszen az új adatok fényében revideálódó eredmények megnehezítik, illetve késleltetik a fordulópontok – a recesszió kezdetének illetve a kilábalás elindulásának – azonosítását.

A tanulmányban arra igyekeztünk felhívni a figyelmet, hogy a probléma kizárólag statisztikai szempontok alapján nem kezelhető. Egy-egy idősor szezonális igazí-

tásának eredményét érdemes egy konzisztens keretben, a szorosan kapcsolódó makrogazdasági idősorokból és indikátorokból leszűrhető információk figyelembevételével értékelni.

Irodalom

ABIAD, A. ET AL. [2009]: *What's the Damage? Medium-term Output Dynamics after Banking Crises*. International Monetary Fund Working Paper WP/09/245. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1512251 (Elérés dátuma: 2010. július 7.)

ALTISSIMO, F. ET AL. [2007]: *New Eurocoin: Tracking Economic Growth in Real Time*. Banca d'Italia. Temi di discussione. 631. sz. http://www.bancaditalia.it/pubblicazioni/econo/temidi/td07/td631_07/td631/en_tema_631.pdf (Elérés dátuma: 2010. július 7.)

BAFFIGI, A. – GOLINELLI, R. – PARIGI, G. [2004]: Bridge Models to Forecast the Euro Area GDP. *International Journal of Forecasting*. 20. köt. 3. sz. 447–460. old.

EDWARDS, S. [2001]: *Thirty Years of Current Account Imbalances, Current Account Reversals and Sudden Stops*. NBER Working Paper 10276. http://www.nber.org/papers/w10276 (Elérés dátuma: 2010. július 19.)

FIXLER, D. J. – GRIMM, B. T. [2002]: Reliability of GDP and Related NIPA Estimates. *Survey of Current Business*. 82. sz. 9–27. old. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.7.8182&rep=rep1&type=pdf (Elérés dátuma: 2010. július 10.)

FORNI, M. ET AL. [2005]: The Generalized Dynamic Factor Model: One-Sided Estimation and Forecasting. *Journal of the American Statistical Association*. 100. sz. 830–840. old. http://www.jstor.org/stable/2646650?seq=1 (Elérés dátuma: 2010. július 7.)

KOOP, G. M. – GARRATT, A. – VAHEY, S. [2008]: Forecasting Substantial Data Revisions in the Presence of Model Uncertainty. *Economic Journal*. 118. köt. 530. sz. 1128–1144. old. http://www3.interscience.wiley.com/journal/119879360/abstract?CRETRY=1&SRETRY=0 (Elérés dátuma: 2010. július 10.)

MAINWARING, H. – SKIPPER, H. [2007]: GDP(O) Revisions Analysis System: Overview and Indicative Results. *Economic and Labour Market Review*. 1. évf. 10. sz. 36–42. old. http://212.58.231.21/elmr/10_07/downloads/ELMR_Oct07.pdf#page=36 (Elérés dátuma: 2010. július 7.)

MEHRHOFF, J. [2008]: Sources of Revisions of Seasonally Adjusted Real Time Data. http://www.oecd.org/dataoecd/47/9/40671433.pdf (Elérés dátuma: 2010. július 7.)

ORPHANIDES, A. [2001]: Monetary Policy Rules Based on Real-Time Data. *American Economic Review*. 91. köt. 4. sz. 964–985. old. http://www.jstor.org/stable/2677821 (Elérés dátuma: 2010. július 7.)

STOCK, J. H. – WATSON, M. W. [2002]: Forecasting Using Principal Components from a Large Number of Predictors. *Journal of the American Statistical Association.* 97. köt. 460. sz. 147–162. old. http://www.jstor.org/stable/3085839 (Elérés dátuma: 2010. július 7.)

TÓTH M. B. [2005]: Jelentős külső egyensúlytalanságok következményei – nemzetközi tapasztalatok. *MNB Háttértanulmányok* 5. MNB. Budapest.

A statisztikai módszertan jelenlegi helyzete az Eurostatnál

Új feladatainak és növekvő jelentőségének megfelelően a statisztikai módszertan új szervezeti kereteket kapott az Európai Unióban.

A szakstatisztikai szervezeti egységek mellett a statisztikai hivatalok felépítésében jellemzően egy központi módszertani egység is megtalálható, mely az általános módszertani kérdésekkel, fejlesztésekkel, speciális szaktudást igénylő területekkel (például mintavétel, becslés, szezonális kiigazítás) foglalkozik. Az EU statisztikai hivatalában is hagyományosan jelen voltak a különböző szakterületeket felölelő egységek, de az önálló igazgatóság, mely központilag foglalkozik a minden szakterületet érintő, átfogó módszertani kérdésekkel, csak 2009-ben jött létre.

Walter Radermacher, az Eurostat 2008-ban kinevezett vezetője megfogalmazta az európai statisztikai rendszer jövőképét, és ennek szellemében az Eurostat szervezeti felépítését is. A jövőkép szerint az európai statisztikai rendszer az ún. „kályhacső modellről" – amikor minden szakstatisztika szabályozása és az adat-előállítás a többi szakterülettől szinte függetlenül történik – az integrált rendszer felé mozdul el. Az új típusú integrált rendszerben a különböző céllal gyűjtött adatok egységes elvek alapján, közös adatbázisokba szerveződnek, mikroszinten kombinálhatók, így az új és növekvő igények gyorsabban, rugalmasabban lesznek kielégíthetők. Ehhez azonban a statisztikai fogalmak, folyamatok, eljárások nagyfokú standardizálására van szükség.

Az új szervezetben, létszámban is megerősödve, a „B" Minőség, módszertan- és információrendszerek igazgatóság lesz a változások mozgatórugója. Ehhez az igazgatósághoz tartozik a statisztika minősége, a metarendszer és az osztályozások, a módszertan és az informatika. Az igazgatóság munkáját a tagországok képviselőiből álló, rendszeresen ülésező csoportok támogatják. Két vezetői csoport: a módszertani és az informatikai vezetőké, továbbá két támogatói csoport (sponsorship group): a minőség és a tervezett standardizálás tartozik ide. Egyes aktuális témákban konkrét célok elérésére, határozott időre munkacsoportokat (task force-okat) állítanak fel.

1. DIME (Directors of Methodology)

A szervezeti változással egyidejűleg jött létre az EU-tagországok statisztikai hivatalainak módszertani vezetőit összefogó fórum, a DIME (Directors of Methodology – Módszertanért felelős igazgatók).

A DIME megbízatása elsősorban a statisztikai módszertan stratégiai kérdéseire terjed ki. A globális, kompetitív, technológiavezérelt információs világban a statisztika elé állított feladatokra megfelelő válaszok szükségesek. Az Európai Statisztikai Rendszeren (European Statistical System – ESS) belül a DIME a módszerek és az eszközök közös tárházának segítségével a folyamatok és a módszerek fejlesztését és standardizálását; a képzés- és tudásmenedzsmentet tartja szemmel. Ezeknek a feladatoknak a megoldásához statisztikai módszertannal összefüggő programokat, együttműködéseket (task force, ESSnet javaslatok

stb.) kezdeményez és szervez. Az ESSnet-programokat, melyek elsősorban a tagországok statisztikai hivatalainak közös munkájaként készülnek el, az Eurostat finanszírozza, vezetésüket pedig egy-egy tagország vállalja el.

Mint már korábban említettük, a DIME megbízatásához tartozik a standardizálással kapcsolatos feladatok kialakítása, a standardokkal kapcsolatos döntéshozatal, a standardok dokumentálása és karbantartása, továbbá alkalmazásainak támogatása és nyomon követése. Ilyen standardok többek között a szezonális kiigazítás irányelvei, a felfedés elleni védelem keretrendszere, a pontossági követelmények megfogalmazása, a törzsváltozók meghatározása a statisztikai és adminisztratív adatgyűjtésekhez, a mintavételi hiba számítása, valamint a metaadat-standardok definiálása adatokra és folyamatokra. A DIME feladatai közé tartozik még a kijelölt folyamatok feltárása és dokumentálása, valamint a jó gyakorlatok kiemelése.

Az EU kutatási célú hetedik keretprogramja (FP7) a kutatás, az oktatás és az innováció alkotta „tudásháromszög" együttesét öleli fel. Ezek együttes kezelését tűzte ki célul a „B" Igazgatóság és a DIME is. Ehhez kapcsolódó program a CROS (Cooperation between researches and official statisticians – A hivatalos statisztikusok és kutatók együttműködése), melynek munkatervében szerepel egy együttműködési platform kialakítása az egyetemek és a hivatalos statisztika között a kutatási lehetőségekről és a közös fellépésről. Tervezik továbbá a diákok és a statisztikusok cseréjét gyakorlatra és oktatásra, valamint egy hivatalos statisztikai laboratórium szervezését az Isprában (Olaszország) székelő Közös Kutatási Központ (Joint Research Centre) keretében. Egy másik kezdeményezés „A hivatalos statisztika európai mestere" szak megszervezésének előkészítésére, a statisztikai hivatalok és az érdeklődő felsőoktatási intézmények képviselőnek 2010 júniusában, Southamptonban (Egyesült Királyság) szervezett munkaértekezlet. Bár több országban folyik már statisztikusképzés, a jelenlegi javaslatban az európai megközelítés újnak mondható.

Tenderkiírásokra kerül sor még 2010-ben a hivatalos statisztika kutatási igényeinek felmérésére, azonosítására, a CROS-platform létrehozására, az ESSnet projektek eredményeinek jobb hasznosulására, továbbá az ESSnet-projektek honlapjának létrehozására. Még ebben az évben elindul a BLUE-ETS FP7-es kutatási projekt a hivatalos üzleti statisztika témájában, a MEETS (Modernisation of European enterprise and trade statistics – Európai vállalkozás- és kereskedelemstatisztikák modernizációja) eredményeinek kiegészítésére és alkalmazására.

Az Európai Statisztikai Rendszer (ESS) változásai kemény módszertani feladatokat adnak, ilyen például a különböző forrásból származó adatok növekvő arányú használata, a nemmintavételi hiba arányának növekedése, mérése, kezelése, a koherencia, a mintavételi keretek stb. terén. Az új adatgyűjtési módszerek, adatellenőrzési stratégiák, a robusztus módszerek, a modellezés, az adatelemzés, a minőségi összetevők közötti cserearányok, az adattárházak és mikroadatok (Stiglitz-jelentés az eloszlásról és egyenlőtlenségről), valamint a földrajzi (térinformatika) kódok használatával kapcsolatos módszertani kérdések szintén a kiemelt feladatok közé tartoznak.

A DIME két szakterületi (a munkaerő-felmérés és Információs és kommunikációs eszközök) projekt általános módszertani ajánlásainak kidolgozására hozta létre a „Pontosság" (lakossági felvételek) munkacsoportot, melynek keretében az EU háztartási felvételeire általános pontossági ajánlás is készül.

A DIME Háztartási költségvetési felvétel megújítására szerveződött munkacsoporthoz kapcsolódó szerepe még nem pontosan tisztázott. Ez a felvétel a fogyasztói árindex és a GDP-számítás mellett, a Stiglitz-jelentés

(http://www.ksh.hu/statszemle_archive/2010/2010_03/2010_03_305.pdf) szempontjából is alapvető fontosságú. Ugyanakkor az is nyilvánvaló, hogy az adatgyűjtés jelentős adatszolgáltatói teherrel jár. Két kiemelt célterületet terveznek, az egyiket a válaszadási arány javítására, a másikat az újrasúlyozás, kalibrálás módszereinek felülvizsgálatára.

A DIME kezdeményezésében, olaszországi vezetéssel indult a „Kisterületi becslés" és az „Adatintegráció" projekt. A „Kisterületi becslés" projekt célja, hogy segítsék a területi és társadalmi-demográfiai ismérvek szerint részletezett statisztikai információk előállítását és az eddigi tudományos eredmények átültetését a hivatalos statisztikai gyakorlatba. Az „Adatintegráció" projekt a 2007–2008-ban futott ESSnet ISAD-projekt (Integration of surveys and administrative data – Adatfelvételek és igazgatási adatok integrálása) folytatása, melynek eredményeit 2008 novemberében mutatták be az SPC (Statistical Programme Committee – Statisztikai Programbizottság) előtt. Az eredmények az elemi szintű azonos vagy különböző egységekre vonatkozó rekordok különböző technikákkal történő összekapcsolására (rekord linkage), statisztikai összekapcsolásra (statistical matching) és a mikrointegrált adatok feldolgozására vonatkoztak. A jelenlegi projektben is ezeken a területeken gyűjtik össze az ismereteket, speciális fejlesztéseket hajtanak végre, hogy a hivatalos statisztikában való használatot segítsék.

Még 2010-ben indul az „Adatelemzés" ESSnet-projekt azzal a céllal, hogy a statisztikai hivatalok az adatelemzési és szemléltetési technikákat szélesebb körben és egységesebben használják. Az indítást előkészítő műhelykonferencia 2010. május 27–28-án volt Bécsben. Erről további információk a http://www.statistik.tuiwen.ac.at/edavis/ honlapon találhatók.

Előkészítés alatt van a standardizálással foglalkozó ESSnet-projektet előkészítő műhelykonferencia is, melybe a Központi Statisztikai Hivatalt (KSH) is bevonták. A lehetséges feladatok között szerepel a meglevő standardok áttekintése, értékelése, a standardizálási folyamat felülvizsgálata, fejlesztési javaslatok, új infrastruktúra kialakítása.

Az ESSnet-projektek növekvő száma szükségessé teszi összehangolásukat. Ebben a DIME-nek is szerepet kell kapnia, hogy tanácsokat adjon az ESSC-nek (European Statistical System Commitee – Európai Statisztikai Rendszer Bizottsága) a prioritások kijelölésében, valamint a futó és záruló projektek értékelésénél a módszertani szempontok érvényesítésében.

A DIME évenkénti egyszeri plenáris ülésének előkészítést az Eurostat szakértői és az ún. DIME steering group végzi, az utóbbiban az Egyesült Királyság, Olaszország, Spanyolország, Hollandia, Szlovénia és Magyarország statisztikai hivatalának képviselői vesznek részt.

A DIME életre hívása és működése arra hívja fel a figyelmet, hogy a minőségi statisztika, különös tekintettel a hivatalos statisztikára, nem nélkülözheti módszertani tudásunk folyamatos fejlesztését és az új eredmények, fejlesztések beépítését a statisztikai adatelőállítási folyamatba. Mindazok, akik mélyebben érdeklődnek a téma iránt, bővebb információkat kaphatnak a témáról az Eurostat honlapján valamint az ismertetések íróitól.

2. Új módszertani projektek az Eurostatnál

A következőkben ismertetjük, az Eurostatnál folyó, már részben említett legfontosabb módszertani projekteket.

Pontosság (Accuracy).[1] Az Eurostat jelenleg szakstatisztikai területenként határozza

[1] A témával kapcsolatban további információ: *Fraller Gergely* (gergely.fraller@ksh.hu).

meg a különböző becslések pontosságára előírt követelményeit. Azonban az Európai Statisztikai Rendszer integrált rendszerré alakításával ismételten felmerült az igény a pontossági követelmények újradefiniálására, a szórásbecslés témakörének harmonizálására. Ennek kapcsán jelenleg két szakstatisztikai területen folyik fejlesztés.

A munkaerő-statisztikán belül a munkaerő-felmérés területén tervezik a pontossági előírások és követelmények felülvizsgálatát, valamint a teljesítésének ellenőrzésére vonatkozó eljárás kidolgozását. Megvizsgálják a különböző módszereket a szórásbecslések közös alapra helyezésének lehetőségéről, valamint a longitudinális becslések pontosságának méréséről és a pontossági követelmények megfogalmazásáról. Az információstatisztika területén harmonizált szórásbecslő eljárás bevezetését tervezik.

Felismervén, hogy az említett igények a legtöbb háztartási felvételt érintik, a DIME Pontossággal foglalkozó munkacsoport felállítását kezdeményezte azzal a céllal, hogy általános ajánlásokat fogalmazzon meg az EU háztartási felvételeire kirótt pontossági követelményekről és becslésekről. Az általános ajánlások speciális alkalmazását a munkaerő- és az információstatisztika területén egy-egy külön munkacsoporton belül valósítják meg.

A munkacsoport céljai a következők:

– javaslat készítése az EU-előírások pontossági követelményeire, figyelembe véve a felvételek sajátosságait. A követelmények megfogalmazásának egyértelműnek, egységesen értelmezhetőnek kell lennie,

– javaslat a követelmények és előírások teljesítésmérésének folyamatára,

– a szórásbecslő eljárások áttekintése egy egységesebb, harmonizált szórásszámítás kialakítása érdekében,

– javaslat a mintavételi hibák elérhetőségének növelésére az ESS-en belül (beleértve a követelményeket és metaadatokat).

A DIME Pontosság munkacsoportja már elkészítette előzetes jelentését, melyben ajánlásokat tesz többek között a pontosság mérésére az indikátor típusától függően és a pontossági követelmények megfogalmazására. A mintavételi terv és a statisztika jellege szerint különböző szórásbecslő eljárásokat javasolnak. A két említett (munkaerő-, információstatisztika) szakstatisztikai munkacsoport munkáját követően, a DIME Pontosság munkacsoportja a visszacsatolások nyomán készíti el végleges jelentését. A munkacsoportban hat tagországi statisztikus, két tudományos szakértő, két-két Eurostatot képviselő szakterületi és módszertanos szakértő vesz részt.

Kisterületi becslés (Small area estimation – SAE).[2] A felhasználók egyre inkább igénylik a minél részletesebb bontású, megbízható adatokat. A statisztikai hivatalokban működő mintavételen alapuló eljárások a minta mérete, illetve gyakorisága miatt bizonyos részletezettségi szint felett önmagukban nem alkalmasak megbízható becslésekre. A modellezésen alapuló kisterületi becslési eljárásokkal azonban ezek az igények is kielégíthetők.

Ennek érdekében jött létre az ESSnet keretén belül 2009-ben a kisterületi becslések projektje, melynek fő eredménye olyan ajánlások és informatikai eszközök kialakítása, amelyek alkalmasak kisterületi becslések készítésére. A projekt további célja egyrészt a kisterületi módszerek újraértékelése, a tapasztalat és tudás összegyűjtése, másrészt a meglévő tudományos elméletek és azok gyakorlatba történő átültetésének feltérképezése a megfelelő tu-

[2] A témával kapcsolatban további információ: *Horváth Beáta* (beata.horvath@ksh.hu).

dásátadás és kommunikáció kialakítása a nemzeti statisztikai hivatalok között. Az ESSnet felmérte a társadalmi felvételeknél alkalmazott kisterületi becslési eljárásokat, illetve az igényeket, elvárásokat. A projekt keretein belül olyan minőségértékelési rendszert kívánnak kifejleszteni, amely alkalmas lesz a különböző módszerek, eljárások összehasonlítására, figyelembe véve különböző szempontokat, mint például a modellválasztást, a torzítást és az átlagos négyzetes hibát. A helyzetfelmérésből adódó eljárásokból kialakításra kerülnek majd az ún. „legjobb gyakorlatok".

A jövőbeni feladatok közé tartozik a már említett minőségértékelés, a speciálisan kifejlesztett, ún. nyílt forráskódú szoftverek kialakítása és az ajánlások kidolgozása. Az ismeretek átadására munkahelyi képzéseket, tanfolyamokat és konferenciákat szerveznek. A kétéves projektbe az Olasz Statisztikai Hivatal vezetése mellett Franciaország, Németország, Hollandia, Norvégia, Lengyelország, Spanyolország valamint az Egyesült Királyság delegált tagokat.

Adat-összekapcsolás (Data Integration).[3] Napjainkban, amikor a statisztikai adatgyűjtések tervezése során egyre fontosabb szempontá válik a költséghatékonyság és a felhasználói terhek csökkentése (miközben természetesen senki nem kíván lemondani az eddig megszokott minőségi követelmények teljesítéséről), mind többen ismerik fel és próbálják kihasználni a különböző adminisztratív céllal gyűjtött adatok felhasználásában, illetve a statisztikai adatok újrahasznosításában rejlő lehetőségeket. Ezek kiaknázására sokszor érdemes (a jogi keretek adta lehetőségeken belül) a több adatbázis összekapcsolása útján létrehozott rekordokat elemezni. Ez az igény az adatösszekapcsolással összefüggő számos módszertani kutatást és alkalmazást hívott életre. Ezzel a kérdéssel kapcsolatban említést érdemlő projekt az Olasz Statisztikai Hivatal által 2007 és 2008 között koordinált ISAD, amelynek legfontosabb célja a különböző forrásokból származó adatok integrációjával kapcsolatos alkalmazások közös módszertani alapjainak megteremtése volt. Ennek folytatásaként – szintén az Olasz Statisztikai Hivatal vezetésével – 2009-ben a DIME jóváhagyta egy új projekt indítását, amely az elemi szintű adatok összekapcsolása és a mikrointegrációs adatok feldolgozása területén koordinálja az ESS-országok együttműködését. Az előbbi témával kapcsolatban két fontos módszertani terület, a record linkage és a statistical matching tanulmányozását és fejlesztését nevesítik a projekt tervezői. A record linkage során legalább két adathalmazt kívánunk összekapcsolni olyan közös azonosító segítségével, amely mindegyikben megtalálható egyéb adatok alapján hozható létre pontosan vagy valószínűségi megfeleltetéssel. Az adathalmazok ugyanazon egyedek adatait tartalmazzák. A statistical matching viszont eleve különböző egyedekről szóló adatok összekapcsolását tűzi ki célul, kihasználva, hogy vannak minden táblában azonos tartalmú változók. A projekt célja kettős, egyrészt szeretne felállítani egy közös tudástárat, amely tartalmazza a record linkage és a statistical matching területén elért legújabb eredményeket, áttekinti a mikrointegrált adatfeldolgozás lépéseit és módszereit, megadja a kulcsfontosságú fogalmakat és azok tartalmát, összegyűjti az alkalmazásokat, másrészt pedig módszertani eszközöket, szoftvereszközöket fejleszt, esettanulmányokat és dokumentációkat tesz közzé, valamint oktatásokat tart a téma iránt érdeklődő hivatalokban. Ez utóbbiak (ún. továbbképzések) a record linkage és a statistical matching területeit ölelik fel. A munka folyik, az említett tréningre már le is zárult a jelentkezés, a módszertani kérdések-

[3] A témával kapcsolatban további információ: *Kővári Zsolt* (zsolt.kovari@ksh.hu).

ben érintett szakemberek a DIME honlapján keresztül kísérhetik figyelemmel az eseményeket, és kommentálhatják azokat. Az érdeklődők az Eurostat honlapján találhatnak az ESSnet-ről és az aktuális projektekről részletesebb ismertetést. (Lásd: http://epp.eurostat.ec.europa.eu/portal/page/portal/essnet/essnet_projects/running_ESSnet_projects)

Felismerve a hivatalos statisztikai adatok iránt megnövekedett felhasználói igényeket a DIME, összehangolt módszertani fejlesztésekkel, standardok kidolgozásával, a legjobb gyakorlatok bemutatásával igyekszik segíteni a hivatalos statisztikai adatok minőségi követelményeinek betartását, sőt megszilárdítását. Ezért is tartottuk fontosnak az új módszertani intézmény főbb tevékenységi irányainak ismertetését.

Szép Katalin
kandidátus, a KSH főosztályvezetője
E-mail: Katalin.Szép@ksh.hu

Fraller Gergely,
a KSH szakmai tanácsadója
E-mail: Gergely.Fraller@ksh.hu)

Horváth Beáta,
a KSH tanácsosa
E-mail: Beata.Horvath@ksh.hu)

Kővári Zsolt,
a KSH főtanácsosa
E-mail: Zsolt.Kovari@ksh.hu)

Az MTA Statisztikai Bizottságának 2010. április 27-i ülése

Az MTA Statisztikai Bizottsága 2010. április 27-i ülésének témáját *Rappai Gábor* „A statisztikai modellezés filozófiája" című cikke adta.[1] A szerző-előadó írásában a statisztika néhány alapvető kérdését érinti: az oktatást, a statisztika tudományos besorolását, a sztochasztikus modell alapgondolatát, célját, gyakorlati alkalmazási kérdéseit, valamint az eredmények értelmezését.

A bizottsági ülés célja a felvetett kérdések bizottsági keretek közötti megvitatása volt, hogy világossá váljon, mely pontokon van egyetértés a szakmai keretekben, és milyen, a továbblépést, cselekvést segítő tapasztalatok, gondolatok fogalmazhatók meg.

Rappai Gábor, „Merre tart a statisztikai modellezés?" című bevezetőjében a cikkben felvetett kérdések közül a modellezés néhány alapkérdését emelte ki. A modellezés általános célja, hogy a valóságban megfigyelt változékonyságból a modellel magyarázott rész minél nagyobb legyen, de hogy ez pontosan hibaminimalizálást (maradéktag), a találatmaximalizálást vagy mást jelent, az már az adott helyzetben dől el. A dilemma, hogy a modell eredménye minél többször pontosan egyezzen meg a valósággal, vagy nem a pontos egyezés a cél, hanem az eredmény általában közel legyen a valósághoz, és még az is kérdés, hogyan mérjük ezt a közelséget. Megint más cél a váratlan események előrejelzése. A statisztikai modellezés másik alapkérdése a megfelelő adatbázis megtalálása, illetve összeállítása, mely alapvetően behatárolja az alkalmazható modellek kialakítását, a model-

[1] *Rappai G*. [2010]: A statisztikai modellezés filozófiája. *Statisztikai Szemle*. 88. évf. 2. sz. 121–140. old. http://www.ksh.hu/statszemle_archive/2010/2010_02/2010_02_121.pdf

lek relevanciáját. Az adatbázis minőségét a modellezőnek ismernie kell, mivel munkája során ezt figyelembe kell vennie. A jellemzők leírása az adat-előállító statisztikus feladata, erről a Bizottság 2009. decemberi ülésén volt szó a statisztika minősége kapcsán.

Bevezetőjében kitért a modellek érvényességére, annak szükségességére, hogy tisztázzuk a modell mely tartományon értelmezhető, milyen hipotéziseken alapul, például a ceteris paribus elv mire, milyen mértékben érvényes. Hasonlóképpen nehézséget okoz, de nélkülözhetetlen az eredmények értelmezéséhez a becslés pontossága (hibahatár) és megbízhatósága (valószínűségi szint) összefüggésének megértése.

A problémák ismeretében felmerül a kérdés, mit tehetünk. A válasz első része kézenfekvő: megfelelő színvonalú oktatást kell biztosítani. A garanciát az jelenti, ha a megfelelő képzést a statisztika tárgy keretében kapják a hallgatók, és csak másodlagosan más, statisztikai módszereket alkalmazó tárgyak anyagaként. Példaként a marketingkutatás, illetve a befektetési döntések tárgyakba foglalt terjedelmes statisztikai fejezeteket említette. További garanciát jelentene a statisztikusok képzése, akik megfelelő tudással rendelkeznének. Jelenleg nincs önálló statisztikusképzésünk, doktori iskolánk, kutatóintézetünk. Felmerül a statisztikatudomány helye, elismertsége is. Ez a kérdés legutóbb 1999-ben *Hunyadi László* és *Rappai Gábor* közös cikkében is megfogalmazódott.[2]

Hunyadi László korreferátumában ehhez a ponthoz kapcsolódóan kiemelte, hogy a statisztika tudományági besorolása nem pusztán elméleti kérdés. Azzal, hogy az Akadémia a statisztikát a IX. osztályához sorolja, a Statisztikai Bizottságból kimaradnak az orvosok, matematikusok, fizikusok, általában a természettudósok, holott a statisztikatudomány alkalmazásában egyaránt érintettek. A statisztikusok örök feladata, a profi és a jószándékú (laikus) statisztikusok közötti híd megteremtése, a felhasználó segítése türelmes magyarázattal. De a statisztika jelentőségének növekedése, felhasználói körének bővülése szükségessé teszi a statisztikai ismeretek mind magas szintű, mind népszerű, közérthető formában való terjesztését. Ezt célozta a KSH kiadásában indított „Statisztikai módszerek a társadalmi és gazdasági elemzésekben" című könyvsorozat, amely az első öt kötet megjelenése után leállt.

Bár nem jártunk sikerrel a Mindentudás Egyetemén való megjelenéssel, hasonlóképpen szükség lenne a statisztikai ismeretek népszerű stílusban való terjesztésére könyv formájában, ami a legegyszerűbb példákon keresztül mutatja be a statisztikát vagy az interneten, például a Wikipédiához hasonló digitális enciklopédia formájában. A XXI. században az internet nagyon jó eszköz arra, hogy népszerűsítse, hozzáférhetővé tegye a statisztikát.

A vitában szinte minden, statisztikával is foglalkozó tudományos műhely képviselője részt vett. A legújabb kutatások szerint az új generáció hálózatban, közös munkában gondolkodik és legfeljebb 20 percig képes egy dologra koncentrálni. Rappai Gábor erre alapozva hangsúlyozta, hogy a statisztikában is a felhasználók felé kell fordulni, be kell vonni őket, például közösségben készíteni az enciklopédiát.

Herman Sándor, az MST elnökeként fáradozott a Mindentudás Egyetemén való megjelenéssel, bár elvileg befogadták a javaslatot, de a sorozat megszűnéséig mégsem került megvalósításra. *Pukli Péter*, a KSH korábbi elnöke hangsúlyozta, hogy az ismeretterjesztésnek nemcsak a sztochasztikus modellekre, hanem általában a statisztikai ismeretek széles körére is ki kell terjednie. A Nemzetközi Statisztikai Intézet (ISI) legutóbbi, 2009. évi Dél-Afrikában tartott konferenciáján a statisztikai jártasság

[2] *Hunyadi L. – Rappai G.* [1999]: Gondolatok a statisztikáról. *Statisztikai Szemle*. 77. évf. 1. sz. 5–15. old. http://www.ksh.hu/statszemle_archive/1999/1999_01/1999_01_005.pdf

(statistical literacy) széleskörű fejlesztését tartotta az előttünk álló egyik legjelentősebb kihívásnak. Ez azt jelenti, hogy el kell érni, hogy a lakosság széles rétegei legyenek képesek a statisztikai információk olvasására, helyes értelmezésére. Ebben a feladatban az oktatási intézményeknek, a hivatalos statisztikai szolgálatnak és a médiának is együtt kell működnie.

Szilágyi György – aki a *Statisztikai Szemlében* megjelent cikkében foglalkozott hasonló kérdésekkel[3] – azt hangsúlyozta, hogy a modellezés mindig egy elméletnek megfelelően jellemzi a valóságot. Világosan meg kell tudni különböztetni a valóságot és a statisztikai modellel leírt, jellemzett megfigyelt képet. A statisztikus felelőssége, hogy hangsúlyozza a mérés célját, az alkalmazott absztrakciót. A tudományok kategóriákba sorolása terén végig kell gondolni, hogy a statisztika módszertudományi besorolása milyen következményekkel járna.

Művészet vagy tudomány – art or science, vetette fel a kérdést *Marton Ádám*, és mindjárt a fogalom bizonytalanságára hívta fel a figyelmet, hisz nem egységes, mikor mit értünk statisztika alatt. Az árindex példáját hozta fel. Az árindex esetében a valós inflációt nem tudjuk mérni, csak az indexformuláknak megfelelően a statisztikai modellel jellemzett inflációt. Ugyanakkor a lakosság által érzékelt infláció megint más.

Belyó Pál, a KSH elnöke emlékeztetett arra, hogy az MTA Statisztikai Bizottsága napirendjén az elmúlt 20 évben többször megjelent az a kérdés, hogy tudomány-e a statisztika és mi lenne a helyes besorolása az MTA kategóriái szerint. Egyetértett a statisztikai jártasság fejlesztésének fontosságával, de a statisztikai adatokhoz való hozzáférést ugyanolyan fontosnak ítélte. A KSH-nak mindkét területen nagy a felelőssége. Ennek tudatában ajánlotta fel a KSH honlapját, mint médiát a statisztikai ismeretek népszerűsítésére, terjesztésére. Példával illusztrálta, hogy a KSH vezetői mindig készek az iskolák megkeresésének eleget tenni, és tanórákon bevezetni a diákokat a statisztika rejtelmeibe, az adatok elérési technikáiba.

Sándorné Kriszt Éva a Budapesti Gazdasági Főiskola rektora, *Ferenczi Zoltán* és *Kovács Péter* oktatók a felsőoktatási intézményeik oktatási gyakorlatáról, a statisztika önálló témaként történő elismerési lehetőségeiről számoltak be.

A Bizottság tagjai egyetértettek azzal, hogy a statisztika tudományjellegének elismertetése nagyobb lendületet kaphatna, ha a tudományterület több képviselője szerezné meg az MTA doktora kitüntető címet, így erőfeszítéseinkben ennek kell prioritást adni. Így az egyetemi tanszékeken is biztosítani lehetne a minősített vezetőket, és ez a doktori iskolák indításának is előfeltétele.

Mindez azt is jelenti, hogy a magas szintű statisztikusképzéshez (statisztikai PhD) vezető út elején tartunk. Rövidebb távon fenn kell tartani, illetve fejleszteni kell a statisztika oktatását a nappali képzésben, a sztochasztikus modellek oktatásában fel kell hívni a figyelmet a valóság és a statisztikai modell kapcsolatára, a statisztikai modellek mögötti hipotézisekre, a modell eredményeinek helyes értelmezésére, érvényességének pontos bemutatására. A felsőfokú oktatás jó szakkönyvei érdekében a statisztikai sorozatot célszerű lenne folytatni. A statisztikai ismeretterjesztés közérthető, népszerűsítő eszközök alkalmazását igényli. Ehhez jó egy a Wikipédiához hasonló fórum, illetve az internet. Sajnos konkrét munka felajánlása, javaslat nem született, így ezekre a kérdésekre, további előkészítés után kell visszatérni.

[3] *Szilágyi Gy.* [2000]: Érteni a számok nyelvén… *Statisztikai Szemle.* 78. évf. 1. sz. 5–12. old. http://www.ksh.hu/statszemle_archive/2000/2000_01/2000_01_005.pdf

Szép Katalin
kandidátus, a KSH főosztályvezetője
E-mail: Katalin.Szép@ksh.hu

Hírek, események

Lemondás. *Dr. Belyó Pálnak*, a Központi Statisztikai Hivatal elnökének megbízatása – lemondása miatt – 2010. június 10-i hatállyal megszűnt.

Kinevezés. *Dr. Orbán Viktor*, a Magyar Köztársaság miniszterelnöke (45/2010 (VI. 11.) ME határozata szerint) – 2010. június 11-ei hatállyal, hat évi időtartamra – *dr. Vukovich Gabriellát* a Központi Statisztikai Hivatal elnökévé nevezte ki.

Megbízás visszavonása – megbízás. *Dr. Vukovich Gabriella*, a KSH elnöke 2010. június 14-ei hatállyal visszavonta *Hársfai Ferencné* gazdálkodási főosztályvezetői megbízását, és 2010. június 15-étől *dr. Soós Lőrincet* bízta meg a Hivatal Gazdálkodási főosztályának vezetésével. A volt főosztályvezető 2010. június 15-étől főosztályvezető-helyettesi teendőket lát el.

Lemondás. *Dr. Balogh Miklós* 2010. július 1-jei hatállyal lemondott elnökhelyettesi posztjáról és korengedményes nyugdíjazását kérte.

Kinevezés. *Dr. Orbán Viktor*, a Magyar Köztársaság miniszterelnöke (62/2010. (VII. 22. ME határozata szerint), a közigazgatási és igazságügyi miniszter javaslatára – 2010. július 12-ei hatállyal, hat évi időtartamra – *dr. Németh Zsoltot* a Központi Statisztikai Hivatal elnökhelyettesévé nevezte ki.

Jutalom. Közszolgálati jogviszonyban töltött ideje alapján jubileumi jutalomban részesült 2010. június hónapban 25 éves szolgálatért: *Domonkosné Papp Ágnes* (KSH Veszprémi Igazgatóság); *Kocsis István* (KSH Pécsi Igazgatóság); *Ferencz Józsefné* (Informatikai főosztály); *Nogula Erzsébet* (Népességstatisztikai főosztály); *Káli Gabriella* (Statisztikai kutatási és módszertani főosztály); 30 éves szolgálatért: *Wirth Ferencné* (KSH Pécsi Igazgatóság); *Tóth Éva* (Árstatisztikai főosztály); *Pechinger Antalné* (Népességstatisztikai főosztály); *Torba Erzsébet* (Nemzeti számlák főosztály); 35 éves szolgálatért: *Gróf Csongorné* (Gazdálkodási főosztály); *Kiss Györgyné* (KSH Debreceni Igazgatóság); 40 éves szolgálatért: *Sármány Erzsébet* (Informatikai főosztály); *Sulykosné Papp Edit* (Vállalkozásstatisztikai főosztály); *Meyndt Györgyné* (Informatikai főosztály).

Az ENSZ Statisztikai Bizottsága a 2010. február 23. és 26. között tartott 41. ülésén 2010. október 20-át Statisztikai Világnapnak minősítette, amiről a Közgyűlés 2010. június 3-án hozott határozatot (A/RES/64/267. sz.). A megemlékezés témájaként a statisztika eredményeit, alkalmazását, szakszerűségét és hitelességét jelölték meg. A Nemzetközi Statisztikai Intézet (ISI) ezért körlevélben hívta fel a nemzeti statisztikai hivatalok és társaságok, a szakmai tudományos élet szervezetei, valamint a statisztikaoktatók figyelmét a méltó ünneplésre.

A 2010. évi hazai rendezvénysorozat programtervezetét (http://portal.ksh.hu/pls/ksh/docs/szolgaltatasok/hun/sajto/statisztikai_vilagnap_programja.pdf) a Központi Statisztikai Hivatal, a Magyar Tudományos Akadémia Statisztikai Bizottsága és a Magyar Statisztikai Társaság közösen dolgozta ki. Olvasóink a KSH honlapján – sok statisztikai témájú érdekesség mellett – a 2010. július 14-én indított Statisztikai szellemi TOTÓ-t is megtalálhatják, mely az október 20-i ünnepi ülésig hétről-hétre új 13+1 kérdést tartalmaz.

A Statisztikai Világnapról további információk olvashatók az ENSZ Statisztikai Bizottságának megemlékezésről szóló önálló honlapján (http://unstats.un.org/unsd/wsd/Default.aspx), valamint az ISI honlapján (http://isi-web.org/news/world-statistics-day).

Az Eurostat új szervezeti felépítése 2010. július 1-jével hatályos, mely a http://epp.eurostat.ec.europa.eu/portal/pls/portal/!PORTAL.wwpob_page.show?_docname=2260731.PDF honlapon érhető el.

A Központi Statisztikai Hivatal Vezetői Kollégiuma 2010. június 16-án tartott kibővített ülést a Hivatal Keleti Károly-termében. Először *dr. Belyó Pál*, a KSH volt elnöke elköszönt a Hivatal munkatársaitól, sok sikert és jó egészséget kívánt az új elnöknek és a dolgozóknak. Majd *dr. Vukovich Gabriella*, a KSH elnöke ismertette programját az összegyűltekkel. Elsődleges céljaként a szakmai munka hatékonyságának és minőségének javítását jelölte meg, hangsúlyozva, hogy érdemi, értékrendbeli és stílusbeli változásokat is tervez ennek érdekében. Az utóbbiak fő irányaként a szakmai munka előtérbe helyezését, az adatok minőségének központba állítását, a Hivatal munkatársaira és vezetőire háruló adminisztratív terhek csökkentését, az adatszolgáltatók, a felhasználók, valamint a szakmai és civil közélet felé nyitást tűzte ki. Az elnök többek között elmondta, hogy szeretné, ha a KSH „számgyár" helyett intelligens munkahellyé válna, és ismét elemző hivatalként működne a felhasználók igényeinek nagyobb fokú kielégítése, az adatok jobb hasznosulása, valamint annak érdekében, hogy az adatgyűjtések és -feldolgozások hiányosságai, illetve a minőségi problémák az elemzések révén is feltárhatók legyenek. Az ülést kötetlen beszélgetés zárta, melynek keretében a jelenlevők kérdéseket, észrevételeket tettek az elhangzottakkal kapcsolatban.

A szolgáltatási kibocsátási árindexek publikálásáról tartottak fórumot a KSH Keleti Károly-termében 2010. június 30-án. A fórum résztvevői *Süveges Éva*, a KSH főosztályvezetőjének bevezetője után *Hamvainé dr. Holocsy Ildikó* osztályvezető előadását hallhatták.

A Tadzsik Statisztikai Hivatal Mezőgazdasági statisztikai főosztályának vezetője, *Mutalibjon Abdulloev* két munkatársa kíséretében tett látogatást a KSH-ban 2010. május 25. és 27. között. A tanulmányút során, melyet egy európai uniós projekt keretében szerveztek, a küldöttség megismerkedett a magyar statisztikai rendszerrel, és áttekintést kapott a magyar mezőgazdasági statisztika jellemzőiről, a mezőgazdasági összeírások során használatos adatgyűjtési, -feldolgozási és -elemzési módszerekről, valamint különböző publikálási és módszertani kérdésekről.

Az Eurostat delegációja tett látogatást 2010. július 6-án és 7-én a KSH-ban. Az EU jogszabályi előírása szerint rendszeresen, kétévente kerül sor konzultációs megbeszélésre, amelyen a szakértők áttekintik a túlzott hiány eljárás keretében Magyarország által összeállított jelentést. A küldöttséggel a KSH, az MNB és a Nemzetgazdasági Minisztérium szakértői folytattak tárgyalást.

A Magyar Statisztikai Társaság Statisztikatörténeti Szakosztályának szakmai üléssel egybekötött tisztújító közgyűlésére 2010. június 9-én került sor a KSH Keleti Károly-termében. A közgyűlés megnyitását követően *Lencsés Ákos*, a KSH Könyvtár osztályvezetője tartott előadást. Majd *dr. Faragó Tamás*, a Szakosztály elnöke beszámolt a Statisztikatörténeti Szakosztály elmúlt három éves munkájáról, és a résztvevők megválasztották az új elnökséget. Az elnöki tisztséget továbbra is *dr. Faragó Ta-*

más tölti be; az új titkár *dr. Lakatos Miklós*, a KSH szakmai főtanácsadója, a vezetőség tagjai pedig *Benoist György* osztályvezető, *Grábics Ágnes* főosztályvezető-helyettes, *dr. Nemes Erzsébet*, a KSH Könyvtár főigazgatója és *Szalainé Homola Andrea*, a Miskolci Igazgatóság igazgatója lettek.

A Népesedési Világnap alkalmából *dr. Vukovich Gabriella*, a KSH elnöke és *dr. Spéder Zsolt*, a KSH Népességtudományi Kutatóintézet igazgatója tartott sajtóbeszélgetést 2010. július 9-én a Hivatal Fényes Elek-termében. A rendezvény programja a következő volt: Az adatok és adatgyűjtések fontossága (dr. Vukovich Gabriella); Egy nemzetközi összehasonlító kutatási program néhány tanulsága: szándékok és gyakorlat – a termékenységi döntések európai összehasonlításban (dr. Spéder Zsolt); KorFa online: a vándorlási veszteség Magyarországon az elmúlt évtizedben (*Gödri Irén*, az NKI tudományos munkatársa).

Az ECOSTAT Gazdaság- és Társadalomkutató Intézet és a Magyar Közgazdasági Társaság 2010. június 2-án rendezett kerekasztal-rendezvényt, amelyen *dr. Gertler János*, a Magyar Tudományos Akadémia külső tagja, az egyesült államokbeli George Mason University (Fairfax) egyetemi tanára „A termelés külföldre telepítésének hatása az amerikai gazdaságra – makrogazdasági elemzés" címmel tartott előadást az Intézet tanácstermében.

A Nemzetközi Statisztikai Intézet (International Statistical Institute – ISI) fontosabb konferenciaajánlatai

(A teljes ajánlatlista megtalálható a http//:isi.cbs.nl/calendar honlapon.)

Kampala, *Uganda*. 2010. október 12–15.

Ötödik Nemzetközi Mezőgazdasági Statisztikai Konferencia. (*Fifth International Conference on Agricultural Statistics*.)

Honlap: www.icas-v.org

Aarhus C, *Dánia*. 2010. október 13–14.

Ole E. Barndorff-Nielsen 75. születésnapja tiszteletére rendezett műhelykonferencia. (*Workshop in Honour of Ole E. Barndorff-Nielsen's 75th Birthday.*)

Honlap: http://www.thiele.au.dk/events/conferences/2010/oebn75

Spa, *Belgium*. 2010. október 13–15.

A Belga Statisztikai Társaság 18. konferenciája. (*18th Annual Conference of the Belgian Statistical Society*.)

Információ: *F. Thomas Bruss* elnök

E-mail: tbruss@ulb.ac.be

Washington, D.C., *Egyesült Államok*. 2010. október 19–20.

Az „Előrejelző Analítika Világa" konferencia. („*Predictive Analytics World" Conference*.)

Telefon:+1-(717)-798-3495

E-mail: registration@predictiveanalyticsworld.com

Honlap: www.predictiveanalyticsworld.com/register.php

Vińa del Mar, *Chile*. 2010. október 19–22.

Statisztikai Társaságok IX. Latin-amerikai Kongresszusa. (*IX Latin American Congress of Statistical Societies*.)

Információ: IES-PUCV 56-32-2274051, DEUV-UV 56-32-2508320

E-mail: info@clatse.org

Honlap: www.clatse.org

San Francisco, *Kalifornia*, *Egyesült Államok*. 2010. október 20–22.

2010. évi Nemzetközi Modellezési, Szimulációs és Irányítási Konferencia. (*International Conference on Modeling, Simulation and Control 2010*.)

Information: IAENG Secretariat (Mérnökök Nemzetközi Szövetségének Titkársága)
E-mail: wcecs@iaeng.org
Honlap: http://www.iaeng.org/WCECS2010/ICMSC2010.html

San Francisco, *Kalifornia*, *Egyesült Államok*. 2010. október 20–22.

„Előrejelző üzleti, marketing és internetes analitika" elnevezésű képzési program. („*Predictive Analytics for Business, Marketing and Web*" *training program*.)

Information: Prediction Impact, Inc.
Telefon: +1-(415)-683-1146
E-mail: training@predictionimpact.com
Honlap: www.predictionimpact.com/predictive-analytics-training.html

Santiago, *Chile*. 2010. október 20–22.

A Hivatalos Statisztika Nemzetközi Társasága (International Association for Official Statistics – IAOS) hivatalos statisztikáról és környezetről szóló konferenciája „Megközelítések, kérdések, kihívások és kapcsolatok" címmel. (*IAOS Conference on Official Statistics and the Environment*: *Approaches, Issues*, *Challenges and Linkages*)

E-mail: iaos2010.surs@gov.si
Honlap: http://www.ine.cl/iaos2010/eng/index.html

Folyóiratszemle

Daalmans, J. – de Waal, T.:

A másodlagos cellaelnyomás átfogóbb megközelítése

(A General Formulation of the Secondary Cell Suppression Problem.)

A tanulmány elérhető:
http://www.cbs.nl/NR/rdonlyres/993654A5-C5BC-4469-ACC1-6760D5F67AE7/0/201009x10pub.pdf

Statisztikai adatok nyilvánosságra hozatala sok esetben táblázatos formában történik. Az egyes cellák értéke attól függően, hogy gyakorisági vagy értékösszeg tábláról van szó, lehet az adott cellához hozzájárulók száma vagy a hozzájárulások értékeinek összege.

Mielőtt egy (értékösszeg vagy gyakorisági) tábla nyilvánosságra kerül, mindig gondoskodni kell annak védelméről, azaz arról, hogy a tábla alapján ne lehessen olyan információhoz jutni, amit a táblához hozzájárulók nem szeretnének felfedni. (Az ilyen adatok ún. érzékeny, másképpen fogalmazva védendő cellákban vannak.) A táblázatos adatok védelmére több lehetőség kínálkozik, melyek közül a tárgyalt tanulmány az értékösszeg-táblákra vonatkozó cellaelnyomást vizsgálja. Egy tábla akkor biztonságos, ha nincs benne érzékeny cella. Amennyiben tartalmaz legalább egy védendő cellát, akkor már nem nevezhető biztonságosnak. Adatvédelmi feladat, hogy csak biztonságos táblák kerülhessenek nyilvánosságra.

Azokat a cellákat, amelyekről megállapítható, hogy érzékenyek, „el kell nyomni”, azaz az értékeiket ki kell törölni és egy elnyomásra utaló szimbólummal kell helyettesíteni. Ezt nevezzük elsődleges elnyomásnak. A másodlagos cellaelnyomásra az elsődleges után van szükség, mert egy ún. „támadó” az elsődlegesen elnyomott cella oszlopára/sorára összegzett adatokból kiszámíthatja az elsődlegesen elnyomott cella értékét azáltal, hogy ismeri az adott oszlop/sor többi cellájának értékét. Ilyenkor tehát az adott oszlopban/sorban el kell nyomni más cellát, cellákat is.

A szerzők példán keresztül mutatják be a cellaelnyomást. Vegyük a következő táblát, amely az elsődleges elnyomás utáni állapotot mutatja.

	C_1	C_2	C_3	Összeg
R_1	x_{11}	1	3	104
R_2	x_{21}	2	1	103
R_3	70	3	2	75
Összeg	*270*	*6*	*6*	*282*

Legyenek az elsődlegesen elnyomott cellák $R_1 \times C_1$ és $R_2 \times C_1$, ezek értékei x_{11} és x_{21}. Másodlagos elnyomás nélkül a támadó a sorokból könnyen következtethet az értékekre: $x_{11}=104-1-3=100$, illetve $x_{21}=103-2-1=100$.

A másodlagos elnyomásra mutat egy lehetőséget a következő tábla.

	C_1	C_2	C_3	Összeg
R_1	x_{11}	1	x_{13}	104
R_2	x_{21}	2	x_{23}	103
R_3	70	3	2	75
Összeg	*270*	*6*	*6*	*282*

Megjegyzés. A Folyóiratszemlét a KSH Könyvtár (*Lencsés Ákos*) állítja össze.

Ebből a támadó x_{11}, illetve x_{21} pontos értékére nem következtethet, mert már nem ismeri az x_{13} és x_{23} értékeket sem.

Másodlagos elnyomás esetén lehetőség van az elnyomásra kerülő cellák megválasztására. Figyelembe vehetjük például, hogy az elnyomott cellák értékeinek összege a legkisebb legyen, vagy a lehető legkevesebb cellát nyomjuk el, esetleg minimalizálhatjuk az elnyomott cellákhoz hozzájárulók számát.

Felvetődik a kérdés, hogy mikor nevezünk egy cellát érzékenynek. A cella érzékenységét többféle módon lehet mérni, választhatjuk például az ún. (p,q)-szabályt ($0 \le p < q$). Ez a következőket jelenti. A tábla nyilvánosságra kerülése előtt q százalékos hibával lehet becsülni az egyes hozzájárulók cellából való részesedéseit. Egy cella érzékeny, ha a nyilvánosságra kerülés után valaki (akár a cella valamelyik hozzájárulója is) p százalékon belüli pontossággal tudja becsülni valamelyik (másik) hozzájáruló részesedését a cellából. Minél nagyobb p-t választunk, annál nagyobb lesz egy védett cella hozzájárulójának biztonsága. Ha egy cella kis p érték mellett is érzékeny, az azt jelenti, hogy van olyan támadó, aki legalább egy hozzájárulást nagy pontossággal tud becsülni.

Legyen az i-edik elnyomott cella értéke x_i, ehhez a cellához a hozzájárulások legyenek x_i^1, x_i^2,…, x_i^R. (R a táblához hozzájárulók száma. Ha az r-edik hozzájáruló az i-edik cellához nem járul hozzá, akkor $x_i^r = 0$.) Jelölje ezek közül az r-edik legnagyobbat $x_i^{[r]}$, azaz $x_i^{[1]} \ge x_i^{[2]} \ge \ldots \ge x_i^{[R]}$.

Így tehát $x_i = \sum_{r=1}^{R} x_i^r = \sum_{r=1}^{R} x_i^{[r]}$.

Tegyük fel, hogy az s-edik hozzájáruló szeretné becsülni a t-edik hozzájáruló részesedését az i-edik cellában, azaz az s-edik hozzájáruló támadja a t-edik hozzájárulást. Ilyenkor úgy jár el, hogy alsó becslést ad (a (p,q)-szabály feltevése alapján) az x_i^r hozzájárulásokra ($r \ne s,t$), majd ezeket és a saját hozzájárulását, x_i^s -t levonja a cella értékéből, x_i -ből. Például nemnegatív hozzájárulások esetén az alsó becslés x_i^r -re $x_i^r - (q/100)\sum_{r \ne s,t} x_i^r$. Így a felső becslés x_i^t -re:

$$U_s\left(x_i^t\right) = x_i^t + (q/100)\sum_{r \ne s,t} x_i^r \ .$$

A cella a (p,q)-szabály szerint akkor érzékeny, ha $U_s\left(x_i^t\right) \le x_i^t + (p/100)x_i^t$ valamilyen s-re és t-re. Másképpen, a cella akkor védett, ha $x_i^t + (q/100)\sum_{r \ne s,t} x_i^r > x_i^t + (p/100)x_i^t$ minden s-re és t-re.

Mivel

$$(q/100)\sum_{r \ne s,t} x_i^r \ge (q/100)\sum_{r \ne 1,2} x_i^{[r]} \quad \text{és}$$

$$(p/100)x_i^{[1]} \ge (p/100)x_i^t$$

bármilyen s, t választás esetén, ezért a védettség meglétéhez elegendő azt ellenőrizni, hogy igaz-e a

$$(q/100)\sum_{r \ne 1,2} x_i^{[r]} > (p/100)x_i^{[1]}$$

egyenlőtlenség, vagyis azt, hogy a legnagyobb hozzájáruló megfelelően védett-e a második legnagyobb hozzájáruló támadásával szemben.

Az érzékenységet tehát egy ún. érzékenységi függvénnyel lehet mérni:

$$S_{p,q}\left(x_i\right) = px_i^{[1]} - q\sum_{r=3}^{R} x_i^{[r]} \ .$$

Az i-edik cella védendő, ha $S_{p,q}\left(x_i\right) \ge 0$.

A tábla védettségéhez a gyakorlatban minden i-re ellenőrzik a $S_{p,q}\left(x_i\right) < 0$ egyenlőtlenséget.

A másodlagos cellaelnyomás után, nemnegatív hozzájárulások esetén egy újabb probléma adódik. Ez az ún. elnyomási intervallumok szélességének kérdése. Az érzékeny cella elnyomási intervalluma a legtágabb olyan intervallum, amelybe a cella értéke eshet. Ez az egy oszlopba/sorba eső elnyomott cellák értékeinek összege, az adott oszlop/sor teljes összege, valamint a nemnegativitási feltétel alapján határozható meg.

Ha a korábbi másodlagos elnyomás tábláját tekintjük, akkor az

$$x_{11}+x_{13}=103,$$

$$x_{21}+x_{23}=101,$$

$$x_{11}+x_{21}=200,$$

$$x_{13}+x_{23}=4$$

egyenlőségekből az $x_{11}, x_{13}, x_{21}, x_{23} \geq 0$ feltételek mellett az adódik, hogy $99 \leq x_{11} \leq 103$, illetve $97 \leq x_{21} \leq 101$, vagyis a megfelelő elnyomási intervallumok $[99,103]$, illetve $[97,101]$. Ezek a becslések a támadó szemszögéből nagyon jónak mondhatók.

A támadó tehát akkor szerezhet érdemi információt, ha az elnyomási intervallum szűk. Ezért egy tábla közlésekor figyelni kell arra is, hogy az érzékeny cellák elnyomási intervallumai kellően szélesek legyenek. A biztonság érdekében például megkövetelhető, hogy az elnyomási intervallum felső végpontja legalább akkora legyen, mint amekkora érték esetén a cella védve lenne például a (p,q)-szabály szerint.

A szerzők példákat hoznak azokra az esetekre, amikor egy tábla védett a (p,q)-szabály szerint, de nem védett, ha az elnyomási intervallumot tekintjük, illetve amikor védett az elnyomási intervallumok szélessége szerint, de nem védett a (p,q)-szabály alapján. Ezért megalkotnak egy új kritériumot.

Ennek a fő gondolata az, hogy ha az egy oszlopban/sorban elnyomott cellák értékeit összegezve kialakítunk egy képzeletbeli cellát, akkor az hasonló tulajdonságokkal bírhat, mint az egyes táblabeli cellák. Ehhez a képzeletbeli cellához a hozzájárulások az elnyomott cellákhoz való hozzájárulások lesznek.

Eddig a táblában az egyes cellákat mindig külön-külön védtük. A szerzők megfelelőbbnek gondolják, hogy az így kialakított képzeletbeli cellákat védjék a táblabeli cellák védelmére alkalmazott szabály szerint. A táblát tehát akkor tekintik védettnek, ha minden ilyen képzeletbeli cella védett. A tanulmány ezt a (p,q)-szabály kiterjesztésének veszi.

Tudjuk, hogy egy támadás akkor a legveszélyesebb, ha egy cellában a második legnagyobb hozzájáruló támadja a legnagyobb hozzájárulót. Észrevehetjük, hogy a képzeletbeli cellában a legnagyobb hozzájárulás és a második legnagyobb hozzájárulás származhat különböző elnyomott cellákból. Így előfordulhat, hogy a legnagyobb hozzájárulásra egy másik cellából történő támadás a legveszélyesebb.

Az új kritérium fontos tulajdonsága, hogy ha a képzeletbeli cellához való hozzájárulások megfelelően védettek, akkor a táblában az egyes elnyomott cellákhoz való hozzájárulások is azok. Ez azt jelenti, hogy az egyes elnyomott cellákra már nem szükséges ellenőrizni a védettséget, az automatikusan teljesülni fog, ha a képzeletbeli cella védett.

Teljesül az ún. szubadditivitás is, mely szerint, ha a táblából csupa nem érzékeny elnyomott cellára történik (egy oszlopban/sorban) egy összegzés, akkor nem érzékeny képzeletbeli cellát kapunk. Ennek a tételnek a bizonyításában szerepet kap a korábban látott érzékenységi függvény is.

Összességében elmondható, hogy a szerzők (mivel szerintük a korábbi megoldások

nem minden esetben kielégítők) a tanulmányban egy új utat keresnek a táblák biztonságosabbá tételére. Arra ösztönzik az olvasót is, hogy próbáljon kidolgozni olyan, eddig nem alkalmazott módszereket a táblák védelmére, amelyek a cellaelnyomást használják.

Antal László,
a KSH gyakornoka
E-mail: Laszlo.Antal@ksh.hu

Greulich, M.:

Egy nemzetközi standard osztályozás nemzeti adaptációjának egyes kérdései

(When is a Category in the International Reference Classification Significant for a National Activity Classification – Some Practical Suggestions) – *Classification Newsletter*. 2007. évi 20. sz. 2–3. old.

A tanulmány letölthető: http://unstats.un.org/unsd/class/intercop/newsletter/newsletter_20e.pdf

Matthias Greulich – a német statisztikai hivatal munkatársa – cikkében az osztályozások nemzeti változatának kidolgozásakor a részletezettség megfelelő szintjének kialakításához kínál szempontrendszert, amelyet a gyakorlatból vett példákkal szemléltet.

A globális világgazdaságban az egyes országok és régiók gazdasági és politikai döntéseihez szükséges statisztikai adatoknak összehasonlíthatónak kell lenniük. Többek között emiatt elengedhetetlen, hogy a nemzeti osztályozás közvetlenül összehasonlítható, konzisztens legyen a nemzetközi standard osztályozásokkal. Ez nem jelenti szükségképpen azt, hogy a nemzeti tevékenységi osztályozásnak pontosan követnie kell a kapcsolódó nemzetközi standard osztályozás struktúráját. Az is elegendő, ha annak elemeit egyértelműen hozzá lehet rendelni a nemzetközi standard osztályozás megfelelő elemeihez. Természetesen az is elvárás, hogy a nemzetközi standard osztályozás módszertani elveit nemzeti szinten is figyelembe vegyék.

A nemzetközi standard osztályozás struktúrája alapján, az egyes országok olyan nómenklatúraelemeket (kategóriákat) kívánnak létrehozni, amelyek megfelelnek nemzeti igényeiknek. Ez jelentheti a nemzetközi standard osztályozás kategóriáinak összevonását is, ha egy bizonyos tevékenység az adott országban vagy gazdasági térségben nem létezik, vagy csekély jelentőségű. Erre példaként szolgálhat az ISIC Rev.4 „tengeri és tengerpart menti vízi szállítás" három számjegyű kategóriája, amelyet egy tenger nélküli ország számára felesleges lenne alábontani. Más esetben előfordulhat, hogy a nemzetközi standard osztályozást nemzeti szinten további tételekkel kell bővíteni. A kérdés: mi alapján hozhatunk döntéseket a nemzeti alábontást illetően.

Egy 1991-ben, az Egyesült Államokban, Williamsburgban tartott konferencia[1] foglalkozott ezzel a kérdéssel. A nemzetközi standard osztályozás összevonandó vagy alábontandó kategóriáinak meghatározására szolgáló egyik eljárás a vonatkozó kategória relatív súlyának, fontosságának mérése a következők szerint: az adott tevékenységet végző statisztikai egységek száma; a tevékenységből származó hozzáadott érték vagy forgalom; az adott tevékenységet végző foglalkoztatottak száma stb. Ebben az esetben a szóban forgó kategóriaelem adott változója értékének arányát az osztályozás következő (magasabb) hierarchiaszintjének átlagához képest minden egyes változóra kiszámolják. Ez lehet például egy adott ISIC szakágazat (4 számjegy) statisztikai egységeinek száma viszonyítva az ágazaton belüli (3 számjegy) ösz-

[1] U.S. Department of Commerce, Bureau of Economic Analysis, Economic Classification Policy Committee. Issues Paper No. 4, „Criteria for Determining Industries" Washington D.C. October 1993.

szes szakágazat statisztikai egységeinek átlagához. Ha például ez az arány 0,5 és 1,5 közé esik, akkor ennek a kategóriának van létjogosultsága. Ha az arány ennél kisebb, akkor a nemzetközi standard osztályozás ezen szakágazatát össze lehet vonni egy másik szakágazattal. Ha az arány magasabb értéket mutat, akkor további alábontás indokolt. Az eljárás továbbfejlesztése lehet a súlyozott arányok számítása a gazdasági változók kombinációihoz.

Egy új osztályozás kidolgozásakor ezek a számítások szükségképpen becsült adatokon vagy nem statisztikai adatforrásból származó adatokon alapulnak. Minden egyes esetben előre meg kell határozni a küszöbértékeket (a korábbi példában ez 0,5 és 1,5) és a változók súlyait. Ezek kétségtelenül hátrányai az említett módszernek. Mindemellett tény, hogy a meglévő részletezettség szintje hatással van a küszöbérték átlépésének a valószínűségére. Egy osztályozás olyan területén, amely már eleve nagy részletezettségű könnyebb átlépni a megadott küszöbértéket (ebből következően létrehozni egy szakágazatot), mint az osztályozás kevésbé részletezett területein. Ráadásul ez a formális megközelítés figyelmen kívül hagyja annak vizsgálatát is, hogy az érintett iparág csökkenő vagy növekvő tendenciájú.

Egy másik lehetséges megközelítés az összevonandó vagy részletezendő kategóriák meghatározására a létrehozandó osztályozási tételhez tartozó részsokaság heterogenitásának vizsgálata az ISIC Rev.3 bevezetőjében foglaltak szerint.[2] A módszer nyilvánvaló hátránya, hogy a tevékenységeknek egyrészt nincsen kellően pontosan körülírt és átfedésmentes definíciója, másrészt nincsen elegendő használható adat a számításhoz. Ezek a problémák újra reflektorfénybe kerültek a Nemzetközi Gazdasági és Társadalmi Osztályozások Szakértői Csoport Technikai Alcsoportjának 2007 áprilisában tartott ülésén, ahol megvitatták ezt a koncepciót.

Németországban gyakorlatiasabb megközelítést alkalmaztak az ISIC Rev.4-re, illetve az európai tevékenység osztályozásra – a NACE Rev.2-re – épülő nemzeti osztályozás kidolgozásakor. A megközelítés lényegi elemei a következők:

– Nemzeti alábontást abban az esetben lehet létrehozni, ha ennek szükségét az adatok felhasználói jelezték, és ha adatgyűjtés céljára szolgál.

– Az adatvédelmi szempontokat előre figyelembe kell venni az új alcsoport és a fennmaradó többi alábontás esetén egyaránt.

– Figyelembe kell venni a statisztikai hivatalt és az adatszolgáltatókat érintő, az osztályozás kibővüléséből eredő megnövekedett terheket.

A statisztikai rendszeren belüli adatfelhasználókat (például nemzeti számlák) és külső intézményeket (például minisztériumokat, szakmai szervezeteket és kutató intézeteket) felkérték, hogy véleményezzék a nemzetközi standard osztályozást és jelezzék igényeiket a nemzeti alábontást illetően. Minden javaslatot világos indoklással kellett alátámasztani. Továbbá kötelezően meg kellett becsülni az új javasolt alcsoportba tartozó főtevékenységet végző statisztikai egységek számát és forgalmát. Nem határoztak meg szigorú küszöbértéket az új alcsoportokba tartozó statisztikai egységeket és a forgalmat illetően annak érdekében, hogy figyelembe lehessen venni a növekvő jelentőségű iparágakat (például napenergia felhasználása), illetve a megfelelően indokolt felhasználói igényeket. Azonban a nemzeti és regionális szintű adatvédelmi problémák elke-

[2] Lásd a 154–159. bekezdéseket. A bevezető elérhető a következő linken: http://ec.europa.eu/eurostat/ramon/nomenclatures/index.cfm?TargetUrl=LST_CLS_DLD&StrNom=ISIC_3&StrLanguageCode=EN&StrLayoutCode=HIERARCHIC

rülése érdekében nem minden javaslatot lehetett megvalósítani.

Amennyiben szükséges volt, az adott javaslat indoklását írásos konzultáció és kétoldalú találkozók keretén belül vitatták meg. A nemzeti (német) osztályozás végső változatát az adatfelhasználók és statisztikusok képviselőiből álló konzultációs bizottság hagyta jóvá. Ez az eljárás valóban nagy kihívást jelentett, és olykor heves viták kísérték. Ennek eredményeként az ISIC Rev.4., illetve a NACE Rev.2 általánosan elfogadott nemzeti verziója készült el, amely a statisztikák alapjául szolgál majd az elkövetkező években.

Sápi András,
a KSH tanácsosa
E-mail: Andras.Sapi@ksh.hu

Kiadók ajánlata

FICHET, B. ET AL. (EDS.) [2010]: *Classification and Multivariate Analysis for Complex Data Structures*. (Komplex adatszerkezetek osztályozása és többváltozós elemzése.) Springer. New York.

A adat-előállítási és -gyűjtési lehetőségek miatt sürgető igény jelentkezett új technikák, illetve eszközök iránt a statisztikai információk elemzése, osztályozása és összesítése, a trendek bemutatása és jellemzése, valamint a szabálytalanságok automatikus kategorizálása érdekében. Az összetett struktúrát képező többdimenziós adatok elemzési módszereinek legújabb vívmányairól szóló kötet tanulmányai a Frankofón Osztályozási Társaság, illetve az Olasz Statisztikai Társaság Osztályozási és Adatelemzési Csoportjának első közös ülésén elhangzott előadások közül kerültek ki. A szerkesztők külön figyelmet szenteltek a mind elméleti, mind gyakorlati szempontból új, klaszterezéssel, osztályozással, idősorelemzéssel, többdimenziós adatok elemzésével, nagy adatállományokon alapuló ismeretszerzéssel és térbeli statisztikával foglalkozó módszertani dolgozatoknak.

JIANG, J. [2010]: *Large Sample Techniques for Statistics*. (Nagymintás statisztikai eljárások.) Springer. New York.

A kötet széleskörű áttekintést ad a nagymintás statisztikai eljárásokról. Ami azonban ennél is fontosabb: inkább a gondolkodás képességére összpontosít, mint a használandó képletek kiválasztására; részletes bizonyítások helyett motivációt és betekintést nyújt; nagyon egyszerű technikákkal indít; illetve érdekes módon ötvözi az elméletet és az alkalmazásokat. Az első öt fejezet néhány olyan egyszerű technikát tekint át, mint az alapvető epszilon-delta érvelés, a Taylor-féle kifejtés, a különböző konvergenciatípusok és egyenlőtlenségek; a következő öt pedig a határeloszlás-tételeket a megfigyelési adatok sajátos helyzeteiben tárgyalja. Az első tíz fejezet mindegyikének legalább egy bekezdése közöl esettanulmányt. A szerző az utolsó öt fejezetet a különleges alkalmazási területeknek szentelte. Az esettanulmányokról szóló részek és az alkalmazásokat bemutató fejezetek részletesen szemléltetik a nagymintás elmélet alapján kidolgozott módszerek alkalmazásának miként jét különböző, „nem tankönyvi" helyzetekben.

A kötetet számos feladat is kiegészíti, amelyek megoldásával az olvasóknak bőven van lehetőségük a tanultak gyakorlására. A mátrixokkal és a matematikai statisztikával kapcsolatos háttér-információkat nyújtó függelékekkel együtt a könyv jobbára önálló, füg-

getlen egységet képez. A nagyközönség számára íródott, az alapképzés magasabb évfolyamainak diákjaitól kezdve a PhD-fokozattal rendelkező kutatókig. Megértéséhez alapfeltétel a matematikai statisztikai alap- és az analíziskurzusok elvégzése.

CHEN, M.-H. ET AL. (EDS.) [2010]: *Frontiers of Statistical Decision Making and Bayesian Analysis*. (A statisztikai döntéshozatal és a bayesi elemzés határai.) Springer. New York.

A bayesi elemzéssel és a statisztikai döntéselmélettel kapcsolatos kutatás bővülése és változatossá válása gyors folyamat, ami miatt a kutatók egyre nehezebben tudnak lépést tartani az összes mai kutatási területtel. Ezért a könyv napjaink ilyen jellegű kihívásait és lehetőségeit tekinti át. Noha kimerítően nem tud foglalkozni minden jelenlegi kutatási területtel, a legtöbb példaértékű tárgyalását tartalmazza. A témák között szerepel az objektív Bayes-féle következtetés, a zsugorodás (ún. „shrinkage") becslése, egyéb döntésalapú becslések, modellválasztások és -tesztelések, a nemparaméteres Bayes-féle módszerek, a bayesi és a gyakoriságon alapuló következtetés kapcsolata, az adatbányászat és a gépi tanulás, a kategorikus, az osztályközös és a területi-időbeli adatelemzési, valamint a posteriori szimuláció módszerei. A kötet számos fontos alkalmazási területtel foglalkozik: számítógépes modellekkel, a klinikai vizsgálatok bayesi tervezésével, járványtannal, filogenetikával, bioinformatikával, éghajlat-modellezéssel, valamint politikatudományi, pénzügyi és marketingalkalmazásokkal. Egy jelenleg folyó, bayesi elemzéssel kapcsolatos kutatást áttekintve megteremti az elmélet és az alkalmazások közötti egyensúlyt. Ez utóbbiak egyértelmű elhatárolásának hiánya a bayesi statisztika kutatásának alkalmazott és erősen interdiszciplináris jellegét fejezi ki. A kötet célja a bayesi statisztikával foglalkozó kutatók, köztük azon nem statisztikusok ismereteinek felfrissítése, akik a bayesi következtetést más tudományterületek érdemi kutatásaiban alkalmazzák. Rajtuk kívül a mai kutatási területeket megismerni szándékozó végzős hallgatók, illetve ösztöndíjas statisztika- és biostatisztika-kutatók számára is hasznos lesz.

JAWORSKI, P. ET AL. (EDS.) [2010]: *Copula Theory and Its Applications*. (A kopulák elmélete és alkalmazásai.) Springer. New York.

A kopulák olyan matematikai objektumok, amelyek teljes mértékben megragadják a véletlen változók közötti függőségi viszonyt, és így nagy rugalmasságot nyújtanak a többváltozós sztochasztikus modellek kialakításában. Az 1950-es évek elején történt bevezetésük óta meglehetősen népszerűvé váltak az alkalmazott matematika számos területén, például a pénzügyi életben, a biztosítási matematikában és a megbízhatóság-elméletben. Napjainkban többek között a piaci és a hitelmodellek, a kockázataggregálás, a portfolióválasztás jól bevett eszközeinek számítanak. A könyv két fő részre tagolódik: az I. rész („Áttekintés") 11 fejezetet tartalmaz, melyek a kopulamodellek lényeges szempontjairól nyújtanak naprakész beszámolót. A II. rész („Tanulmányok") pedig a varsói műhelykonferencián bemutatott dolgozatok közül válogatott hat beszámoló bővített változatát gyűjti egybe.

VAN MONTFORD, K. – OUD, J. H. L. – SATORRA, A. (EDS.) [2010]: *Longitudinal Research with Latent Variables*. (Longitudinális kutatás látens változókkal.) Springer. New York.

A kötet, miközben kiemelt hangsúlyt helyez az egyes módszerek alkalmazási módjára, a longitudinális és a látens változós kutatásokat köti össze (például elmagyarázza, hogy milyen módon kell lefolytatni a longi-

tudinális kutatásokat látens változókkal megfogalmazott célok mellett). Kilenc fejezetet tartalmaz, mivel a látens változós longitudinális kutatás napjainkban eltérő előzményekre épülő különböző megközelítéseket, illetve különféle kutatási kérdéstípusokat és számítógépes programokat használ az elemzéshez. Bizonyos (rövid történettel és fő kiadványokkal támogatott) háttérinformációkból kiindulva, minden fejezet leírja az adott módszerrel megválaszolható kutatási kérdéstípusokat, statisztikai és matematikai magyarázatot nyújt az adatelemzési modellekhez, értelmezi az alkalmazott programok inputját és outputját, valamint egy vagy több, jellemző adatkészletre támaszkodó mintapéldát ad, melyek segítségével az olvasók maguk is végrehajthatják az egyes programokat.

Társfolyóiratok

Economie et Statistique

A FRANCIA GAZDASÁGI ÉS PÉNZÜGYMINISZTÉRIUM, VALAMINT A STATISZTIKAI ÉS GAZDASÁGKUTATÓ INTÉZET FOLYÓIRATA

2009. ÉVI 427–428. SZÁM

Barlet, M. – Blanchet, D. – Crusson, L.: Globalizáció és munkaerő-áramlás.

Lorenceau, A.: Az adómentesség hatása a vállalkozásindításra és a foglalkoztatottságra a vidéki Franciaországban.

Carbonnier, C.: Adócsökkentés és adójóváírás hazai munkaerő alkalmazása esetén.

Fack, G. – Landais, C.: Hatékonyak az adományok után járó adókedvezmények?

JASA
Journal of the American Statistical Association

AZ AMERIKAI STATISZTIKAI TÁRSASÁG FOLYÓIRATA

2008. ÉVI 489. SZÁM

Morton, S. C.: Statisztika – tényektől a politikáig.

ver Hoef, J. M. – Peterson, E. E.: Mozgó átlag alkalmazása folyamhálózatok térbeli statisztikai modellezésében.

McLean Sloughter, J. – Gneiting, T. – Raftery, A. E.: Probabilisztikus szélerősségelőrejelzés együttes és Bayes-féle modellátlagolás segítségével.

Choudhury, K. R. et al.: A hullámmagasság és a dőlésmezők szabályos rekonstrukciója víztükörképből.

Rac, M. J. – Sedransk, A. J.: Bayesi és gyakoriságelvű módszerek az ellátás minőségével kapcsolatos információk terjesztésében orvosi eredmények kockázatkorrigált értékelésével.

McCormick, T. H. – Salganik, M. J. – Zheng, T.: Hány embert ismerünk? A személyes kapcsolati háló hatékony becslése.

Peress, M. – Spirling, A.: A kritika mérése. A filmkritikák rejtett dimenzióinak kimutatása valószínűségi tesztelmélet segítségével.

Glynn, A. N. – Richardson, T. S. – Handcock, M. S.: Döntés vitatott választási eredmények esetén – a szavazási eredményekkel kapcsolatos adatok korlátozott ereje a szavazást követően.

Hering, A. S. – Genton, M. G.: A tér- és időalapú szélelőrejelzés bevezetése.

Chen, S. X. – Tang, C. Y. – Mule Jr., V. T.: A kettős rendszerek pontosságával kapcsolatos

utólagos helyi rétegzés és az Egyesült Államok népszámlálásának lefedettségi elemzése.

Nandram, B. – Choi, J. W.: Testtömegindex-adatok bayesi elemzése kis vizsgálati kör alapján, jelentős nemválaszolás mellett.

Balabdaoui, F. et al.: Egyszeres és többszörös transzmembránáram szemléltetése az *Escherichia coli* SecYEG-pórus segítségével.

Cerioli, A.: Többváltozós szélsőértékek kimutatása.

Tan, Z.: Tényezőváltozós marginális és beágyazott szerkezeti modellek.

Koyama, S. et al.: Az állapot-tér modellek közelítő módszerei.

Leng, C. – Zhang, W. – Pan, J.: Félparaméteres átlagos kovarianciájú regresszióelemzés longitudinális adatok esetén.

Qu, A. – Lindsay, B. G. – Lu, L.: Hatékony aggregált torzítatlan becslő függvények korreláló adatokhoz, véletlen adathiány mellett.

Ishwaran, H. et al.: Magas dimenziójú változók kiválasztása túlélési adatokhoz.

Shao, X.: A függő „wild-bootstrap" módszer.

Hooten, M. B. – Wikle, C. K.: Ágensalapú statisztikai modellek diszkrét tér- és időbeli rendszerekhez.

Kato, S. – Jones, M. C.: Köreloszlás-családok Möbius-transzformációval kapcsolatos hivatkozásokkal és alkalmazásokkal.

Padoan, S. A. – Ribatet, M. – Sisson, S. A.: Likelihood-alapú következtetések max-stabilis eljárásokhoz.

Xia, Y. – Zhang, D. – Xu, J.: Dimenziócsökkentés és félparaméteres becslés túlélési modelleknél.

Christensen, R. – Sun, S. K.: Alternatív illeszkedési vizsgálatok lineáris modellek esetén.

Mao, M. – Wang, J.: Félparaméteres hatékony becslés általánosított kumulatív logit-modellek egy osztályához.

Zhang, Y. – Li, R. – Tsai, C.: Regularizációs paraméter kiválasztása általánosított információs kritériumokon keresztül.

Braun, M. – McAulifee, J.: Variációs következtetés nagyméretű diszkrét választási modellek esetén.

Chen, B. – Yi, G. Y. – Cook, R. J.: Súlyozott általánosított becslő függvények longitudinális válaszokhoz véletlen adathiány mellett.

Choi, N. H. – Li, W. – Zhu, J.: Változóválasztás és a Lasso-módszer kiterjesztése.

Cao, J. – Ramsay, J. O.: Lineáris kevert hatások modellezése paraméterek egymásba ágyazásával.

Wasserman, L. – Zhou, S.: A differenciális adatvédelem statisztikai kerete.

Zhou, L. et al.: Redukált rangú kevert hatású modellek térbeli szempontból összekapcsolt hierarchikus függvényadatokhoz.

Qiao, X. et al.: Súlyozott DWD és aszimptotikus jellemzői.

Savchuk, O. Y. – Hart, J. D. – Sheather, S. J.: Indirekt keresztérvényesség sűrűségbecslés esetén.

Wu, Y. – Zhang, H. H. – Liu, Y.: Robusztus modellfüggetlen többosztályos valószínűségi becslés.

POPULATION

A FRANCIA DEMOGRÁFIAI INTÉZET FOLYÓIRATA

2009. ÉVI 4. SZÁM

Gourdon, V. – Rollet, C.: Halvaszületés a XIX. század Párizsában – egy statisztikai kategória társadalmi, jogi és orvosi megközelítésben.

Larmarange, J. et al.: Homoszexualitás és biszexualitás Szenegálban – egy másik valóság.

Dommaraju, P.: Változások az indiai nők iskolázottságában és házasságkötési életkorában.

Donzeau, N. – Pan Ké Shon, J.: A francia lakosság költözködési szokásai 1973 és 2006 között.

Royer, J.: Az ismétlődő és a visszatérő migráció becslése franciaországi alpopulációk körében.

Slovenská štatistika a demografia

ŠTATISTICKÝ ÚRAD SLOVENSKEJ REPUBLIKY

A SZLOVÁK STATISZTIKAI HIVATAL FOLYÓIRATA

2009. ÉVI 4. SZÁM

Kabát, L.: Gazdasági növekedés és szocio-ökonómiai fejlődés a Stiglitz–Sen-jelentés alapján – a statisztika új feladatai.

Katerinková, M.: Bevándorlás Szlovákiába – a hazai és a nemzetközi módszertan összevetése.

Petrášová, A.: Szociális védelem – családok és gyermekek 2007-ben.

Pflüger, A.: A 2010. évi mezőgazdasági gazdaságszerkezeti összeírás.

Gerhardtová, A.: A 2009. évi európai lakossági egészségfelmérés.

Статистика Statistics

A BOLGÁR STATISZTIKAI HIVATAL FOLYÓIRATA

2008. ÉVI 1. SZÁM

Tzonev, V. – Seykova, I.: A statisztika, mint a tömegjelenségekkel foglalkozó adatfelvételek tervezését, szervezését, végrehajtását segítő tudományág.

Vesselinov, R.: A bolgár gazdasági ciklus időrendje.

Seykova, I.: Eltéréselemzésen alapuló innovációs ötletek a statisztikában.

Rusev, B.: A házassági táblák szerkesztésének módszertana a népszámlálások alatt és után.

Gencheva-Dimova, Y.: A gazdaságilag aktív lakossággal (munkaerővel) foglalkozó projektek kidolgozásának alapmódszerei és modelljei.

2008. ÉVI 2. SZÁM

Jekova, S.: A szegénység alakulása Bulgáriában.

Sedlarski, T.: Tranzakciós költségek és a gazdasági növekedés.

Dimitrova, D.: Statisztikák a szociális védőhálóról.

Iakimova, E.: A doktori fokozattal rendelkezők életpályája.

Jordanova, E.: Európai egészségfelmérés.

Barbolova, J.: Az EU-tagállamok vásárlóerő-paritása.

2008. ÉVI 3. SZÁM

Hristov, E.: A faktorváltozók meghatározásának elégséges feltételei mennyiségi értékek esetén.

Lilova, C. – Sugareva, M.: A nettó reprodukciós ráta felbontása két fő összetevőre: termékenység és halálozás egyes európai országokban 1990 és 2005 között.

Dobromirova, M.: Statisztikai kutatási témák az információs társadalomról.

Ilkova, A.: Az európai statisztikai információk felhasználóinak támogatása.

2008. ÉVI 4. SZÁM

Kotzeva, M.: 129 éves a bolgár statisztika.

Tzonev, V.: Elemi indexszámelmélet, mint a nemzeti számlák rendszerének biztos alapja.

Hristov, E.: Átlagszintű faktorvariációk meghatározásának elégséges feltétele.

Petkov, P.: Bulgária aggregált termelési függvénye.

Marinova, D. – Genadiev, H.: Módszertani változások az IMF „Pénzügyi mérleg és nem-

zetközi befektetési pozíció" című kézikönyvének hatodik kiadásában.

AZ OROSZ ÁLLAMI STATISZTIKAI BIZOTTSÁG FOLYÓIRATA

2010. ÉVI 4. SZÁM

Ponomarenko, A. N.: A nemzeti statisztikai rendszer modernizációjának lehetséges irányai.

Eliseeva, I. I. – Kapralova, E. B. – Schirina, A. N.: A nanotechnológia és a nanotermelés statisztikai elszámolása.

Belyaevskii, I. K.: Demográfiai marketing – tudomány és gyakorlat.

Ilyshev, A. M. – Shubat, O. M..: A ciklikus populációdinamika gazdasági és statisztikai kutatása.

Rodionova, L. A.: Népességpolitika és a nők termékenysége Oroszországban – ökonometriai elemzés.

Baranov, S. V.: Oroszország északi és déli területei közötti gazdasági fejlettségbeli különbségek statisztikai becslése.

Gromyko, G. L. – Spiridonova, E. M.: Egy régió lakosságának társadalmi differenciálása.

Bokov, V. A.: A banki szektor empirikus elemzéséhez használt adatok tárolásának kérdései.

Sokolova, E. S.: A pénzügyi adatok minőségbiztosításának statisztikai módszerei.

Smirnov, V. S.: Az orosz gazdaság erőforrásigénye a statisztikák tükrében.

Zhandarov, A. M. – Shiller, F. F.: A gazdaság kormányzati szabályozása – a csak egy iparággal rendelkező orosz városok problémái.

Ganiev, A. M. – Gataullin, R. S.: A Baskírföldi Statisztikai Hivatal megalapításának 175. évfordulója.