A KISTERÜLETI MUNKAÜGYI STATISZTIKA MÓDSZERTANA ÉS ENNEK ALKALMAZÁSA (II.)*

A tanulmány előző két fejezetében (*Statisztikai Szemle*. 2000. évi 7. sz. 497–507. old.) rámutattunk arra, hogy az országos, a regionális és kisterületi szintű munkanélküliségi ráták becslései fontos mutatói az országos avagy a helyi (regionális) gazdasági helyzetnek. Ezért fontos feladat a munkaügyi információs rendszer továbbfejlesztése a különböző szintű munkanélküliségi ráták számítása. E feladat különböző becslésekkel oldható meg.

Az egyesült államokbeli LAUS adatbecslő-rendszer tapasztalatai alapján egy olyan kisterületi munkanélküliségi statisztikai rendszer (KMSR) kidolgozására tettünk javaslatot, amely képes a magyarországi mintegy 180 munkaügyi körzetre, a 19 megyére és a fővárosra érvényes havi munkanélküliségi ráták becslésének elkészítésére, a becslések statisztikai megbízhatóságának megadására, és végül megfelelő összesítések on-line rendszerű elkészítésére.

Ebben a fejezetben bemutatjuk, az előzőben megfogalmazott javaslat alapján, a tesztelési periódusban megvalósíthatónak talált és megvalósított módszereket tartalmazó elemző rendszert.

A JAVASLAT ALAPJÁN KIALAKÍTOTT KMSR FUNKCIONÁLIS BEMUTATÁSA

A rendszer fő feladatait az alábbiakban összegezhetjük. A nemzetközi előírásoknak – azaz a Nemzetközi Munkaügyi Szervezet (International Labour Organisation – ILO) definíciójának – megfelelő munkanélküliségi és foglalkoztatottsági becslések és előrejelzések készítése ötféle területi bontásban:

1. Magyarország megyéi (Budapesttel kiegészítve).
2. A KSH által meghatározott régiók.
3. A Munkaügyi Központok kirendeltségeihez tartozó kisebb területek.
4. A KSH által meghatározott, az előző pontbelihez hasonló körzetek.
5. A 10 000 főnél nagyobb lélekszámú települések.

A becslések negyedévente – hónapokra és negyedévekre nézve –, az előrejelzések havonta készülnek.

* A tanulmány szerzői: *Banai Miklós*, *Koleszár Kázmér*, *Lázár György*, *Lukács Béla*, *Prisznyák Miklós*, *Varga István*.

A rendszer bemenő adatai. *1*. Az Országos Munkaügyi Kutató és Módszertani Központ (OMMK) által a regisztrált munkanélküliek számáról szolgáltatott adatok, melyek havonta érkeznek. *2*. A Központi Statisztikai Hivatal által végzett, az ILO meghatározásának megfelelő gazdasági aktivitást felmérő adatok. Ezek negyedévente érkeznek, havi bontásban.

A számítási módszer lényege. *1*. Olyan becslőfüggvények alkalmazása, melyek kis területekre nézve is torzítatlan, elfogadhatóan kis standard hibájú eredményeket adnak. Ezekkel a megyékre és Budapestre vonatkozó adatokat kapjuk. *2*. A megyei eredmények kisebb körzetekre vonatkozó leosztása a lakosságiigény-módszer szerint. A régiókra vonatkozó adatokat a megyei eredmények összeadásával kapjuk. *3*. A hibaszámítás egy újravételezési eljárás, a jackknife-módszer alapján történik. *4*. Idősorelemző módszer, konkrétan a Kalman-szűrős eljárás alkalmazása a becslőfüggvények eredményidősoraira. Ennek célja a mintavételi hiba leválasztása, így megbízhatóbb, kisebb ingadozású és hibájú becslésekhez jutunk. Amint az elemezni kívánt idősor magyarázó változója előállíthatóvá válik, előrejelzéseket készítünk a negyedévek közbenső hónapjaira. *5*. Év végén a munkanélküliségre és a foglalkoztatottságra vonatkozó különböző gyakoriságú idősorok összeigazítása a Denton-féle, illetve az additív Cholette–Dagum kiigazítási (benchmark) módszerrel.

A kimenő adatok. Az ILO meghatározásának megfelelő munkanélküliségi és foglalkoztatottsági becslések, előrejelzések és ezek hibái a kisterületi becslőfüggvények szerint, valamint ezek Kalman-szűrős eljárással módosított idősorai; mind a régiókra, mind a megyékre, a KSH és az OMMK irodai körzeteire és a fentebb említett településekre vonatkozóan. (Részletesen lásd a következő alfejezetben.)

Jelentések. A kimenő adatokból különböző jelentések készülnek. Ezek a következők: *1*. A foglalkoztatottság és a munkanélküliség becsléseinek Kalman-szűrős eljárással javított változata régiókra, megyékre, a kétféle (OMMK, illetve KSH) bontás szerinti irodai körzetekre és a nagyobb településekre vonatkozólag. Minden új negyedév elején készül az előző negyedév hónapjairól külön-külön. *2*. Negyedéves összesített adatok, melyeknél csak a becslőfüggvény és a hibák számítandók (Kalman-szűrést itt nem végzünk), szintén minden, az előző pontban felsorolt területi bontásban. *3*. A negyedévek közbenső hónapjaiban Kalman-szűrős előrejelzés készül havonta minden területi bontásban. *4*. Az *1*. pontban említett becslések és a *2*. pontban szereplő előrejelzések összehasonlítása negyedévente.

Megyei becslések és a becslési eljárás megválasztása

Az egyesült államokbeli LAUS rendszer adaptációja során az első és legfontosabb kérdés az volt, hogy a szakirodalomban (*Small Area Statistics*; 1987) javasolt különféle becslőfüggvények közül melyik vagy melyek használandó(k) Magyarország esetében.

A KSH által használt direkt becslőfüggvény minden olyan esetben hatékonyan használható, amikor az adott kisterületre megfelelő számú megfigyelés jut. Ha az adott kisterületre a megadott mintába kerülési valószínűségek helyesek, a megfigyelések megbízhatók, s elhanyagolhatók az eredményt befolyásoló egyéb nem mintavételi hibák, a direkt becslés torzítatlan becslést ad, vagyis a becslés várható értéke, eltérésének várható értéke a sokasági értéktől nulla. A torzítatlanság (vagy kis torzítottság) kívánatos tulajdonsága valamely becslésnek, de ugyanakkor nem ad információt a mintaeloszlás másik fontos

tulajdonságáról, a szórásról. Amikor választani kell különböző becslőfüggvények között, azt célszerű választanunk, amelynek mintaeloszlása az adott mérendő mennyiségre nézve az (ismeretlen) valódi érték körüli szűk tartományban helyezkedik el. Ennek alapján a „kicsi" átlagos négyzetes hiba (mean square error) kritériumát használhatnánk a becslőfüggvények közötti választásra, mivel ha $MSE(\hat{\theta}) = V(\hat{\theta}) + [B(\hat{\theta})]^2$ kicsi, akkor nagy valószínűséggel várható, hogy a becslési eredmény közel lesz a keresett valódi (sokasági) értékhez. Itt $V(\hat{\theta})$ a mért érték varianciája, míg $[B(\hat{\theta})]^2$ a becslési eredmények várható értékének a valódi eredménytől való távolságát, a becslés torzítottságának négyzetét jelöli.

Csupán az *MSE* nagyságának vizsgálata azonban nem elég, mert az *MSE* kis értéke mellett azt is biztosítani kell, hogy a torzítottság kicsi legyen a standard hibához viszonyítva. A probléma az, hogy az *MSE* összetevőit a priori nem ismerjük, hiszen a populációs értéket a teljes körű felvétel adja meg.

Definiáljuk a torzítottsági hányadost a

$$BR(\hat{\theta}) = \frac{B(\hat{\theta})}{\sqrt{V(\hat{\theta})}}$$

módon. Amennyiben $B(\hat{\theta}) = 0$, illetve $BR(\hat{\theta}) = 0$, úgy a becslési értékhez rendelt konfidencia-intervallum megbízhatósági szintje megegyezik azzal a valószínűséggel, amellyel a konfidencia-intervallum a becsült értéket tartalmazza. (Másképp fogalmazva, 95 százalék annak valószínűsége, hogy a becslőfüggvény a várható értéktől legfeljebb a variancia négyzetgyökének 1,96-szorosával tér el.) Minél nagyobb $BR(\hat{\theta})$ értéke, annál kisebb lesz a fenti tartalmazási valószínűség. Gauss-eloszlásokat feltételezve, $|BR(\hat{\theta})| = 1$ esetén a tartalmazási valószínűség $1-\alpha = 95$ százalék konfidenciával is magas marad (0,83). ($BR(\hat{\theta})$=2 esetén ez a valószínűség már csak 0,5 lesz.) Ez annyit jelent, hogy minden olyan becslőfüggvényt használhatónak tekintünk, amelynek várható értéke a (torzítatlannak feltételezett) direkt becslőfüggvény várható értékétől legfeljebb a direkt becslőfüggvény standard hibájának várható értékével tér el. (Ez utóbbi mennyiséget az időátlagra vett empirikus variancia négyzetgyökével közelítjük.) Vagyis egy becslőfüggvény alkalmasabb becslést ad a direkt becslésnél, ha torzítása nem nagyobb, mint varianciájának négyzetgyöke és varianciája kisebb a direkt becslésnél.

Részletesen vizsgáltunk ún. szintetikus (ezen belül hányadosbecslés és regressziós becslés típusú) becslőfüggvényeket, vagyis olyan becslési eljárásokat, amelyekben a becslőfüggvény valamely paraméterét egy nagyobb területre vonatkozó megfigyelési adatok alapján becsüljük, adott esetben a megyére vonatkozó becsléseket országos szintű adatok felhasználásával állítjuk elő. Ezen becslőfüggvények akkor adnak torzítatlan becslést, ha az országos adatsorból becsült paramétereik nem térnek el szignifikánsan a megyei adatokból becsült paraméterértékektől. Ezen feltétel teljesülését különféle statisztikai próbákkal ellenőriztük.

Keresztkorrelációs vizsgálatokkal megbizonyosodtunk arról is, hogy statisztikailag a vizsgált becslőfüggvények időben nagyon hasonlóan viselkednek. (*Kirk M. Wolter*; 1985, *P. A. Cholette*; 1992)

Általánosan érvényesül, hogy a szintetikus becslőfüggvények varianciája kisebb, mint a „megfelelő" nem szintetikus becslőfüggvényeké. A szintetikus becslőfüggvények közül a hányados típusú becslőfüggvények általában kisebb varianciájúak, mint a regressziós függvények. Kis varianciához viszont nagy torzítás tartozhat. Ezért vizsgálatainkat a szintetikus regressziós becslőfüggvényekre fókuszáltuk, ugyanis megmutatható (*Särndal–Svenson–Wretman*; 1992), hogy „nagy" minták esetén ezek közelítőleg torzítatlanok. Ugyanakkor vizsgáltunk szintetikus hányadosbecsléseket is.

Valamennyi (szintetikus vagy nem szintetikus, illetve hányados típusú vagy regressziós) becslőfüggvény rétegzett változataival is kísérleteztünk. A rétegzéseket nem és életkor, illetve nem és iskolai végzettség szerinti bontásban, többféle rétegzéssel is vizsgáltuk. Kiderült (*Kirk M. Wolter*; 1985), hogy a vizsgált minta esetén a rétegeken belüli mintaelemszám-csökkenésből származó fluktuáció gyakorlatilag lerontja a rétegzésből származó varianciacsökkenést, ezért *Occam* borotva elve alapján kizártuk a rétegzett becslőfüggvényeket.

A fentiek értelmében a becslőfüggvényekre vonatkozó statisztikai vizsgálatoknak két területre kellett kiterjedniük. Először is vizsgálnunk kellett a varianciák (pontosabban mivel a torzítottság mértéke nem megadható, így az *MSE*-k) nagyságát és azok relatív nagyságát. Ebből megállapítható, hogy az egyes megyékben az egyes becslőfüggvények mennyire képesek a direkt becslés hibáját csökkenteni. Másodszor pedig vizsgálnunk kellett a becslési eredmények torzítottságát. (Ez utóbbit természetesen csak közelítő mérésekkel tudjuk megállapítani.)

Azt, hogy a szintetikus és nem szintetikus, hasonló alakú becslőfüggvények országos és megyei paraméterei szignifikánsan eltérnek-e egymástól, az alábbi statisztikai próbákkal vizsgáltuk:

a) *t*-próbával teszteltük a $\hat{\beta}$ és $\hat{\beta}_a$ regressziós paraméterek eltérésének szignifikáns voltát. H_0 hipotézisnek a H_0:$\hat{\beta} - \hat{\beta}_a = 0$ -t tekintettük. A tesztelést a rétegzett becslőfüggvények rétegzett $\hat{\beta}_g$ és $\hat{\beta}_{g,a}$ értékeire is elvégeztük.

b) χ^2 -próbával vizsgáltuk a MEF-adatfelvétel regisztrált munkanélküliek és az ILO-munkanélküliek együttes eloszlásának illeszkedését országos és megyei szinten. H_0 hipotézisnek az országos és megyei eloszlások egyezését tekintettük.

c) Mann–Whitney-féle rangösszeg-próbával vizsgáltuk, hogy az egyes megyékre értelmezett $\hat{\beta}_a$ értékek idősorai tendenciózusan eltérnek-e valamilyen irányban az országos $\hat{\beta}$ idősorától, vagy körülötte véletlenszerűen ingadoznak. H_0 hipotézisünk szerint az eltérés iránya véletlenszerű.

A statisztikai próbák eredménye az volt, hogy a megyék többségénél használhatók az országos adatokból származó paraméterek a megyei adatokból származó paraméterek helyett (az 1994–1996-os időszakban), azaz a szintetikus modellek joggal alkalmazhatók a nem szintetikus modellek helyett. Néhány megye esetében a statisztikai próbák kevésbé voltak meggyőzők az adott referenciaidőszakra, de a statisztikai próbák alapján a paraméterek eltérése nem bizonyítható.

A statisztikai teszteken alapuló szignifikancia-vizsgálatok mellett, a torzítottsági hányados vizsgálatára elemeztük azt is, hogy az egyes becslőfüggvények értékei időátlagban mennyire térnek el a direkt becslés torzítatlannak tekinthető becslési értékeinek idő

átlagától. Az átlagos eltérést viszonyítottuk a direkt becslés hibájához. Hasonló módon vizsgáltuk a szintetikus regressziós becslőfüggvényhez viszonyított torzítottságot is.

A relatív hibákat, valamint a relatív hibák arányát megvizsgálva arra a következtetésre jutottunk, hogy minden becslőfüggvénnyel csökkenthető a variancia a direkt becsléshez képest. Ugyanakkor nem szintetikus regressziós függvényekkel egyes megyék esetén, hasonló varianciacsökkenés mellett, a vizsgált referenciaidőszakra a szintetikus regressziós változatnál kisebb torzítás tapasztalható. Mégis, mivel a szintetikus eljárás is statisztikailag hasonló eredményt adott, az egyszerűségre törekedve, minden megyére a szintetikus regressziós becslőfüggvény használatát javasoltuk.

A vizsgálatok eredményeként azt találtuk, hogy a szintetikus regressziós becslőfüggvény szolgáltatja statisztikailag a legjobb becslést, mivel mindenkor elfogadható torzítás mellett csökkenti a becslés hibáját, ezért a becslőfüggvények közül Magyarországon ezt célszerű használni. Ezután ezt a becslést bemeneti adatsorként használva az idősorelemző eljárás becslést ad a sokasági értékre, illetve annak hibájára is.

Becslőfüggvények

A programrendszer a becslőfüggvényeket a megyékre vonatkozólag, az egyes körzetekre (OMMK, KSH irodai körzetek és a települések) lebontva, a megyei becslőfüggvényekből számítja. A régiók adatait a megyei adatok összegezésével számoljuk, a területadditivitást kihasználva. A kisebb területi egységekre való lebontás a lakossági igény módszerrel történik. A kisterületi becslőfüggvények módszere a KSH felméréseinek direkt becslését külső, egyidejű varianciamentes adatok (az OMMK adatai) bevonásával korrigálja a variancia csökkentése céljából.

Több mint 26 féle becslőfüggvényt teszteltünk a rendszer kialakításakor. Ezek leírása, a különböző szempontú (például kor, nem, iskolai végzettség szerinti) változataikkal együtt megtalálható a régebbi dokumentációkban (*Kisterületi…*; 1993). Most csak a vizsgálatok alapján kiválasztottakat, vagyis a legegyszerűbb ún. direkt becslést (amelyet a KSH is használ) és a legjobbnak (azaz a legkisebb torzításúnak és varianciájúnak) talált, jelenleg a rendszerben ténylegesen használt korrigált szintetikus regressziós becslést ismertetjük.

A direkt becslés a Központi Statisztikai Hivatalban hivatalosan használt módszer, amely megfelel az ILO-definíciónak és torzítatlan becslést nyújt. Viszont kisterületeken kevésbé alkalmazható, mivel a viszonylag kicsi megfigyelési szám miatt nagy varianciájú. Jelenleg az év végi kiigazítási (benchmark) eljárás során az egyéves összesített adatok kiszámításakor használjuk. Ehhez történik a negyedéves, a régebben (*Kisterületi…*; 1993) 13. sorszámmal jelölt korrigált szintetikus regressziós becslőfüggvénnyel kiszámolt adatok hozzáigazítása.

A direkt becslés alakja:

$$\hat{Y}_{e,a} = \sum_i y_i w_i , \qquad \left(y_i = \begin{cases} 1, \text{ha GAKT} = 2 \\ 0 \quad \text{egyébként} \end{cases} \right)$$

ahol:

$\hat{Y}_{e,a}$ – a munkanélküliek becsült száma egy adott időpontban (*e* a dátumot jelölő index) és egy adott megyében (*a* a megyéket jelölő index, melynek lehetséges értékei: 1, 2, …, 20),

GAKT – a gazdasági aktivitás státusa, GAKT =2 a munkanélküliek,

w_i – WKORR (továbbvezetett népességre korrigált súly).

A továbbiakban a szokásoknak megfelelően, a kalapos mennyiségek becsült adatokat, a kalap nélküliek statisztikai hibával nem rendelkező tényadatokat jelölnek.

A korrigált szintetikus regressziós becslés

A vizsgált becslőfüggvények közül az ismertetett torzításvizsgálat segítségével választottuk ki (*Kisterületi*…; 1996/a) azt a független adatforrást (a regisztrált munkanélküliek létszámadatait) hasznosító – és szintén torzítatlan – kisterületi becslő függvényt, amelynek segítségével a becslés (standard) hibája csökkenthető. Alakja:

$$\hat{Y}_a = \hat{Y}_{e,a} + \hat{B}_e(X_a - \hat{X}_{e,a}),$$

ahol:

$\hat{Y}_{e,a}$ – a direkt becslésből származó adat,

X_a – a regisztrált munkanélküliek a megyében (OMMK adat),

$\hat{X}_{e,a}$ – a regisztrált munkanélküliek számának direkt becslése,

$\hat{B}_e$ – a két utóbbi mennyiség közötti lineáris regressziós együttható országos adatokból számolva:

$$\hat{B}_e = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})w_i}{\sum_i (x_i - \bar{x})^2 w_i}, \text{ és } x_i = \begin{cases} 1, \text{ ha NYILV} = 1 \\ 0 \quad \text{egyébként} \end{cases}.$$

NYILV=1 esetén a MEF-felmérésbeni megkérdezett nyilvántartott munkanélküli.

A megyei becslőfüggvények hibája

A becslőfüggvények (standard) hibájának (empirikus) becslésére – az irodalomban javasolt – újravételezési eljárást, az ún. jackknife-módszert használjuk. Az eljárás – amely a mintavételi eljárás sajátosságai miatt (*Idősorelemzés*…; 1996/b) csak megyei szinten alkalmazható – röviden a következő. A KSH munkaerő-felmérése során használt mintavételi alapegységek (Elsődleges Mintavételi Egységek, PSU) H halmazából képezzük azok H_i részhalmazait oly módon, hogy egyesével kihagyunk egy-egy PSU-t (az így kapott halmazok a jackknife-részminták). Egy (területadditív) statisztikai mennyiséget a megyére a népesség szerinti lineáris interpolációval kaphatunk meg a H_i jackknife-részmintán felvett $\tilde{x}_i$ értékből kiindulva:

$$x_i = \tilde{x}_i + \frac{x_M}{N_M}(N_M - N_i),$$

ahol:

$\tilde{x}_i$ – a mennyiség *i*-edik részmintán kiszámolt értéke,

N_M – az adott megye lakossága,

N_i – az adott részminta lakossága,

x_M – a mennyiségnek az egész megyén felvett értéke.

Ezt az interpolációt csak a KSH-ból kapott statisztikai mennyiségekre végezzük, az OMMK adataira nem.

Az így kapott x_i -ekkel mint bemenő adatokkal kiszámoljuk a becslőfüggvény teljes megyére vonatkozó értékét, ezt $f_i^{(k)}$ -val jelöljük. Legyen a k -adik típusú becslőfüggvénynek az adott megyén felvett tényleges értéke $f^{(k)}$. Kiszámítjuk az

$$f_{J,i}^{(k)} = nf^{(k)} - (n-1)f_i^{(k)}$$

mennyiségeket, melyek a *k*-adik becslőfüggvény H_i jackknife-részminta szerinti becslései. Ezekből a „pszeudoértékekből" képezett variancia négyzetgyöke (σ) adja a standard hibát. A varianciát a következő formulával számoljuk, ahol a $\langle\rangle$ a H_i -kre való átlagolást jelöli:

$$(\sigma^{(k)})^2 = \frac{1}{n-1}\langle f_{J,i}^{(k)} f_{J,i}^{(k)}\rangle - \langle f_{J,i}^{(k)}\rangle\langle f_{J,i}^{(k)}\rangle .$$

Megjegyezzük, hogy a direkt becslés és a korrigált szintetikus becslés teljesítőképességét a következő módon mértük össze. Kiindultunk a BLS által használt elfogadhatósági kritériumból, miszerint egy becslés akkor publikálható hivatalosan, ha 6 százalékos munkanélküliségi ráta esetén a becslés relatív hibája legfeljebb 10 százalék. Kiértékeltük a direkt és a korrigált szintetikus regressziós becslést 1995 minden negyedévére, minden megyére és a fővárosra. Azt találtuk, hogy a direkt becslés az esetek 45 százalékban, míg a korrigált szintetikus regressziós becslés az esetek 74 százalékban teljesítette az elfogadhatósági kritériumot. A javaslatban szereplő centrumot is igyekeztünk megvalósítani. Bár az az elméletileg várható tulajdonsággal rendelkezett, viszont az alkalmazása során időbeli instabilitások léptek fel, azaz ugyanazon becslőfüggvény-kombinációból álló centrum az egymás után következő időpontokban nem volt mindig kiértékelhető. Változó összetétel esetén pedig az egymást követő becslések simasága nem volt garantálható. Így a centrum gyakorlati alkalmazásától sajnos el kellett tekintenünk.

Kisterületi leosztás – Lakosságiigény-módszer

Az egyes irodai körzetekre és a településekre vonatkozó becslőfüggvény-értékeket a megyei értékekből a program a lakosságiigény-módszer (*Kisterületi*…; 1993) alkalmazásával számítja ki. Eszerint a foglalkoztatottak száma a kisterületre, adott időpontban:

$$F_{it} = \frac{F_{in}}{\sum_{i=1}^{h}\left[F_{in}\frac{N_{it}}{N_{in}}\right]} \times \frac{N_{it}}{N_{in}} \times F_t ,$$

ahol:

h – a megyén vagy fővároson belüli kisterületek teljes száma,
F – a foglalkoztatottak száma,
N – a népesség 15–74 éves része,
n – a legutolsó népszámlálás indexe,
i – a kisterületek indexe az adott nagyobb területen belül,
t – az időpont megjelölése, amire a becslés vonatkozik.

A munkanélküliek számát a körzetre vagy települési szintre egyszerűen a megyei becslésnek a regisztrált munkanélküliek arányában történő leosztása adja:

$$M_i = \frac{R_i'}{R'} M_F ,$$

ahol:

i – az adott nagyobb terület (megye, főváros) kisterületeinek indexe,
M_i – az i-edik kisterületbeli munkanélküliek becslése,
M_F – a nagyobb terület munkanélküliségének független becslése (a megyei szintű becsléseket adó modellből érkezik),
R_i' – az i-edik kisterületen regisztrált munkanélküliek száma,
R' – a nagyobb területen regisztrált munkanélküliek száma.

A területi lebontás településekre való alkalmazhatósága

A megyéken belül a munkaügyi irodákhoz tartozó néhány tízezernyi gazdaságilag aktív személyt felölelő irodai körzetekre az eddigi adatszolgáltatás tanúsága szerint is nehézség nélkül alkalmazható az előzőkben ismertetett lakosságiigény-módszer. Igény van azonban a nemzetközi szabványoknak megfelelő munkaerő-piaci becslésekre települések esetén is. A mintaméret csökkenésével azonban – még a becslés torzítottságának kérdésétől eltekintve is – természetszerűleg (körülbelül a mintaméret négyzetgyökével arányosan) nő a statisztikai ingadozások szerepe, ezért meg kell vizsgálni, hogy az adatszolgáltatás milyen kis méretű településekre terjeszthető ki. Az összetett statisztikai eljárásoknak alávetett, megyei szintű, torzítatlan becslések felhasználásával (mint amilyen a regisztrált munkanélküliek száma, a népszámlálási vagy továbbvezetett népességadatok, illetve a gazdaságilag aktív lakosság száma) bonthatók le a megyéken belüli földrajzi egységekre vonatkozó becslésekre. Az a priori értéket becslő direkt eljárások híján a lebontás torzításáról csak indirekten, bizonyos feltevésekre építve mondhatunk bármit.

Ha a munkanélküliek számát vizsgáljuk, akkor az

$$r = \frac{M}{G}$$

hányadossal (ahol r a ráta, M az ILO-definíció szerinti munkanélküliek száma, G pedig a gazdaságilag aktívak száma) definiált munkanélküliségi ráta relatív hibája jó közelítéssel megegyezik a munkanélküliség relatív hibájával. Ez azért igaz, mivel a foglalkoztatottak és a munkanélküliek mintavételi hibája ellentétesen fluktuál, más szóval a gazdaságilag aktívak száma gyakorlatilag mentes a mintavételi hibától.

Következésképpen, a szélsőségesen nagy munkanélküliségi rátáktól eltekintve, a ráta szempontjából ugyanazon kritériumot szabhatjuk meg, mint a munkanélküliek száma esetében.

A továbbiakban tehát a munkanélküliek számának relatív hibájával foglalkozunk.

Az *M* ILO-definíció szerinti munkanélküliség-becslés relatív hibáját három tényező határozza meg:

1. az idősoros (szezonális),
2. a tisztán statisztikus (véletlen, vagyis mintavételi) és
3. a területi inhomogenitásokból eredő hiba.

Tegyük fel, hogy az idősor stacionárius, így az első összetevőt elhanyagoljuk. Mivel a megyén belüli inhomogenitásokra a leosztás módszere nem adhat információt, hiszen az a leosztó faktorok szerinti homogén eloszlást feltételez, ezért erre a tényezőre legfeljebb a megyék közötti inhomogenitások segítségével készíthetünk nagyságrendi becslést. Ezt a

$$\frac{\delta\overline{M}}{\overline{M}} \approx \frac{\delta(a\overline{R})}{(a\overline{R})} \approx \frac{\delta a}{a} = \frac{\delta(\frac{\overline{M}}{\overline{R}})}{\frac{\overline{M}}{\overline{R}}}$$

relatív hiba jellemzi, ahol $\overline{M}$ és $\overline{R}$ az ILO-, illetve a regisztrált munkanélküliek száma (a szezonalitás kiküszöbölése miatt) valamely megyére, egy évre átlagolva, *a* pedig e két mennyiség várható értékeinek hányadosa, ami rendszerint 1-hez közel álló érték. A relatív hibát a megyékre számított varianciából kapjuk. Például 1996-ra ez a relatív hiba 24 százalék (illetve egy másik, alább ismertetendő módszerrel 21 százalék) volt, és feltételeztük, hogy a megyén belüli területi inhomogenitás hatása ennél kisebb lesz.

A kétféle munkanélküliség kapcsolatára egy másik módszerrel is következtethetünk. Az

$$\frac{\left[M^t - aR^t\right]^2}{N^t} \approx const \equiv c$$

összefüggésnek a fluktuációk négyzetgyökös törvénye miatt közelítőleg teljesülnie kell. (A mennyiségek itt egy hónapra vonatkoznak, M^t a településre a lakossági igénymódszer szerint (azaz a regisztrált munkanélküliek arányában) számolt ILO-munkanélküliség, N^t a továbbvezetett népesség az adott évben és a *t* index a települések szerinti bontásra utal. Ekkor a *c* mennyiséget a legkisebb négyzetek módszere szerinti paraméterillesztéssel határozhatjuk meg:

$$\sum_{t}^{\text{megye}} \left\{\left(M^t - aR^t\right)^2 - cN^t\right\}^2 = min\,.$$

Innen *c* és *a* megyénként numerikusan (vagy az adódó egyenletrendszer megoldásával) meghatározható.

Vizsgáljuk meg, hogy a statisztikus fluktuáció milyen korlátot szab. A kiindulási összefüggés alapján az ILO-munkanélküliség relatív hiba négyzetére adódó

$$\left(\frac{\delta M^t}{M^t}\right)^2 \cong c\frac{N^t}{(R^t)^2}$$

összefüggés korlátozhatná a mintaméretet (vagyis független kritériumot adhatna a legki-

sebb település népességére). Az illesztések numerikus elvégzéséből (1997. január hónapra) és legfeljebb 10 százalékos relatív hibát megengedve a munkanélküliségre az adódik, hogy minden településre legfeljebb is csak 100 lakos lehetne ez a korlát. Ezt a kritériumot, azonban másképp is megszabhatjuk:

$$t_M^2 = \left(\frac{M^t - aR^t}{M^t} \right)^2$$

A relatív hibanégyzetet az egyes megyékre a regisztrált munkanélküliek függvényében ábrázolva látható, hogy a $t_M^2 \leq 0{,}01$ kritériumot minden megyére a legfeljebb 10 regisztrált munkanélkülivel rendelkező, azaz a (fentieknek megfelelő) körülbelül 100 lakosú települések is teljesítik.

A két független módszerrel kapott *a*-ra az eltérés minden megye esetében legfeljebb csak mintegy 10 százalék, amiből arra lehet következtetni, hogy ezen értékek a valódi mennyiséget jól közelíthetik.

Hátra van még a megyén belüli területi inhomogenitások hatásának vizsgálata. Ha valamely település munkanélküliségi eloszlásfüggvénye azonos a teljes megyéével, akkor a fenti korlát érvényes. Mintavétel nélkül viszont nem ismerhetjük az adott településre jellemző eloszlásfüggvényt, így ennek paramétereit is becsülnünk kell.

Ezért tételezzük fel, hogy a településeken is mintavételt végzünk $n < N$ személy bevonásával.[1] A munkanélküliek számát kizárólag sztochasztikusan változó mennyiségnek tekintve (*r* munkanélküliségi ráta esetén) a ténylegesen talált munkanélküliek száma az $M=nr$ várható érték körül $nr(1-r)$ szórásnégyzettel fluktuál, a binomiális eloszlás tulajdonságainak megfelelően. *M* hibája ennek megfelelően

$$(\delta M)^2 = (1-r)nr = (1-r)M \approx R^m.$$

Itt R^m a mintában levő regisztrált munkanélküliek száma, és a munkanélküliségi ráta kicsi 1-hez képest. (Itt kihasználtuk, hogy $a(1-r) \approx 1$.) Mivel a mintavétel reprezentatív, a mintaelemek relatív szórása várható értékben megegyezik a sokasági relatív szórással. Innen kapjuk, hogy a munkanélküliek számának relatív hibájára a

$$\frac{\delta M}{M} \cong \frac{\sqrt{R}}{R} = \frac{1}{\sqrt{R}}$$

közelítő kritérium alkalmazható. Vagyis 10 százalék relatív hibát megengedve, csak azon településekre alkalmazható a leosztásos módszer, ahol $R \geq 100$ teljesül. Ez a település összlakosságára nézve azt jelenti, hogy általában 1200-1600 (de esetenként 2-3000 személy) lakosú településekre szolgáltatható adat az ILO-munkanélküliségi hányadról. A statisztikai fluktuációk és a területi inhomogenitás szabta kritériumok közül természetesen az erősebbet kell használni.

[1] Az itt következő elemzés is település szintű értékekre épül, de az egyszerűség kedvéért a *t* felső indexet a továbbiakban elhagyjuk.

Megjegyezzük, hogy a munkanélküliek, illetve foglalkoztatottak abszolút számát az említett bizonytalansági tényezők miatt, valamint azért, mert a leosztási módszer a regisztrált munkanélküliek számára támaszkodik, csak egy nagyságrenddel nagyobb lélekszámú, azaz a körülbelül tízezres lakosságú nagyközségekre és kisvárosokra látszik célszerűnek közölni.

Idősoros elemzés

Az eddig leírt statisztikai eljárások közös jellemzője, hogy egyidejű adatokon dolgoznak, szemben az itt következő idősoros módszerekkel. Az időben egymást követő, rendelkezésünkre álló kérdőíves felmérésekben azonban olyan többletinformáció rejlik, amely lehetővé teszi az adatfelvétel mintavételi hibájának hatékonyabb kiszűrését, s a sokasági értékek pontosabb becslését.

A többletinformáció nyerése azon a feltevésen alapul, hogy a sokasági értékeket és a mintavételi hibát jól meghatározott és elkülöníthető folyamatok generálják, amelyek statisztikai tulajdonságai időben lassan változnak. Így az egymás utáni adatfelvételek egyfajta mintanövekedéssel egyenértékűek, tehát a becslések megbízhatóságát növelik.

Az általunk használt idősoros elemzés (*Harvey*; 1991) – szemben az egyébként széleskörűen használt ARIMA-módszerekkel – az idősor ún. strukturált modellezésén alapul. Ez azt jelenti, hogy a mért idősort több különböző statisztikai tulajdonságú folyamat összegeként modellezzük, míg az ARIMA-modellek egyetlen, meghatározott autoregresszív szerkezetű folyamatként modellezik a méréseket.

A munkaügyi adatok szűrésére egy jel+zaj modellt alkalmaztunk, amelyet az Egyesült Államokban már sikeresen használnak hasonló célokra. Ebben a zajt a mintavételi hiba jelenti, melyet ARMA-folyamatként írunk le. A jelet (sokasági érték) regressziós összetevőre, hosszú távú trendre és szezonális komponensre bontjuk. A regressziós összetevő a munkaügyi adatokat a regisztrációban szereplő adatokkal próbálja kapcsolatba hozni. A trend és a szezonális összetevők a regisztrált és a valós adatok közötti különbség hosszú távú és szezonális jellegét ragadják meg.

A modellparaméterek becslése, illetve független módszerekkel történő meghatározása után a modell felhasználható a meglévő idősor szűrésére, illetve előrejelzésére is.

Strukturált idősorelemzés

A strukturált idősorelemzés a megfigyelt (mért) idősort egy ismeretlen állapotvektor függvényeként modellezi:

$$y_t = \mathbf{Z}_t\boldsymbol{\alpha}_t + \varepsilon_t\,, \qquad /1/$$

ahol y_t a megfigyelt idősor értéke a t időpontban, $\boldsymbol{\alpha}_t$ az állapotvektor ($n\times1$), $\mathbf{Z}_t$ a modellre jellemző együttható vektor ($1\times n$), míg ε_t egy a modellre jellemző 0 átlagú és h varianciájú fehérzaj $N(0,h)$.

Ezen ismeretlen állapotvektor időben az alábbi Markov-folyamat szerint változik:

$$\boldsymbol{\alpha}_t = \mathbf{T}\boldsymbol{\alpha}_{t-1} + \mathbf{R}\xi_t\,, \qquad /2/$$

ahol $\mathbf{T}$ a modellre jellemző együtthatómátrix (nxn), ξ_t az idősor véletlenszerűségét jel-

lemző sztochasztikus folyamatok vektora ($1\times m$), ezek korrelálatlanok és normáleloszlást követnek a modellre jellemző **Q** ($m\times m$) kovarianciamátrixszal, **R** pedig szintén a modellre jellemző együtthatómátrix.

Az /1/ és /2/ egyenletek határozzák meg az adott folyamat ún. állapottér alakját. Az ebben a modellben szereplő **Z**, **T**, **R**, és **Q** mátrixokat rendszermátrixoknak nevezzük, ezek egyértelműen definiálják az állapottérmodellt.

A rendszermátrixok meghatározása egyrészt az idősorról rendelkezésre álló külső információkból, illetve magából az idősorból történhet; ez utóbbi esetben becsült paraméterekről beszélünk.

A rendszermátrixok és az állapotvektor kezdeti eloszlásának ismeretében a Kalman-szűrő algoritmussal állíthatjuk elő az állapotvektor későbbi időpontokra vonatkozó becsléseit. Ez két lépésben történik. Először az állapotvektor $t-1$ időpontbeli értékéből megbecsüljük a t időpontbeli értéket a /2/ átmeneti egyenlet alapján:

$$\mathbf{a}_{t|t-1} = \mathbf{T}\mathbf{a}_{t-1}\,, \qquad /3/$$

ahol $\mathbf{a}_{t|t-1}$ az állapotvektor becslése a $t-1$ időpontig bezárólag rendelkezésre álló megfigyelt értékek alapján.

Hasonlóképpen számítjuk ki az állapotvektor becslésének kovarianciamátrixát is:

$$\mathbf{P}_{t|t-1} = \mathbf{T}\mathbf{P}_{t-1}\mathbf{T}' + \mathbf{R}\mathbf{Q}\mathbf{R}\,. \qquad /4/$$

A következő lépésben ezt a becslést igazítjuk ki a t időpontban megfigyelt értéket (y_t) felhasználva:

$$\mathbf{a}_t = \mathbf{a}_{t|t-1} - \mathbf{P}_{t|t-1}\mathbf{Z}_t' F_t^{-1}(y_t - \mathbf{Z}_t\mathbf{a}_{t|t-1}), \qquad /5/$$

$$\mathbf{P}_t = \mathbf{P}_{t|t-1}\mathbf{Z}_t' F_t^{-1}\mathbf{Z}_t\mathbf{P}_{t|t-1}\,, \qquad /6/$$

ahol:

$$F_t = \mathbf{Z}_t\mathbf{P}_{t|t-1}\mathbf{Z}_t' + h\,. \qquad /7/$$

A Kalman-szűrő algoritmusának egy kiterjesztése, a Kalman-simítás segítségével pedig az állapotvektor olyan becslését állíthatjuk elő, amely a teljes megfigyelt idősor információtartalmán alapul. Ez azt jelenti, hogy a korábbi időpontokhoz tartozó becslésekhez a később megfigyelt értékeket is felhasználjuk.

A fenti algoritmus természetesen csak akkor használható eredményesen, ha a rendszermátrixok és a kezdeti eloszlás megfelelő becslései rendelkezésre állnak. A kezdeti értékek becslésére magát a Kalman-szűrőt is használhatjuk. Ilyenkor ún. diffúz kezdeti feltételekkel indítjuk el az algoritmust:

$$\mathbf{a}_0 = 0, \qquad /8/$$

$$\mathbf{P}_0 = k\mathbf{I},\ k>>1, \qquad /9/$$

amely az állapotvektor dimenziószámának megfelelő lépésszám után „rátalál” a megfelelő becslésre.

A rendszermátrixok becslése úgy történik, hogy a likelihood függvény ún. intervenciós alakját maximalizáljuk valamilyen numerikus optimalizáló módszerrel. A mi esetünkben az ún. EM-algoritmust alkalmaztuk (*Harvey*; 1991), amely sikerrel kombinálható más általánosabb módszerekkel, például a BFGS kvázi-Newton-módszerrel.

Jel + Zaj modell

Az általunk feldolgozott munkaügyi adatok (foglalkoztatottak és munkanélküliek száma) megfigyelt, azaz a munkaerő-felvételből származó idősoraira a következő állapottérmodellt találtuk alkalmazhatónak:

$$y_t = \beta_t X_t + T_t + S_t + E_t + \varepsilon_t . \qquad /10/$$

ahol X_t magyarázó változó, jelen esetben a regisztrált munkanélküliek száma, β_t random változó regressziós együttható, T_t lokális lineáris trend-összetevő, S_t 12 havi szezonális összetevő, E_t a mintavételi hibát leíró ARMA-idősor, ε_t pedig a megfigyelési egyenlet fehérzaj (vagy másképpen irreguláris) összetevője.

Ebből a modellből a munkaügyi adatok sokasági értékeit szeretnénk kiszűrni, tehát a „jel" a mi esetünkben az E_t összetevőn kívüli összes összetevő összegét jelenti:

$$\theta_t = \beta_t X_t + T_t + S_t + \varepsilon_t = y_t - E_t . \qquad /11/$$

Nézzük meg, mi a szerepe a modell egyes összetevőinek.

Regresszor-összetevő. A KMSR-rendszer idősoros módszereinek alapgondolata, hogy a kisszámú mintán végzett munkaerő-felvétel adatai a nagyszámú adaton alapuló munkaügyi regiszter adataival hasonló módon változik az időben. Ezért a regiszteradatokkal való összehasonlítás lehetőséget nyújt a mintavételi hiba által okozott ingadozások kiszűrésére. Ezért a modell leghangsúlyosabb eleme a regisztrált munkanélküliek számát magyarázó változóként használó regresszor-összetevő. A regressziós együttható időbeli változását is megengedjük, σ_b^2 varianciával. σ_b^2 a rendszermátrixokban megjelenő, maximum likelihood módszerrel becsült érték.

A tapasztalat azt mutatta, hogy az előzetes várakozásnak megfelelően, a munkanélküliség és a foglalkoztatottság becsült értékének 60–70 százalékát a regresszor-összetevő adja.

Trendösszetevő. A regisztrált munkanélküliek és a munkaerő-felvétel által mutatott „tényleges" munkanélküliség természetesen nem mindig arányos egymással. Az arányosságon túl meglévő hosszú távú eltéréseket egy lineáris trend-összetevővel becsüljük. Ezt sok esetben nehéz volt szétválasztani a regressziós együttható lassú változásától, ebből adódóan egyes megyéknél konvergenciaproblémák adódtak. Ha azonban ezen összetevő teljes elhagyásával próbálkoztunk, az eredmények lényegesen romlottak.

Mind a trendnívó, mind a ráta lassú véletlen változását megengedtük, σ_t^2 illetve σ_r^2 varianciával.

Szezonális összetevő. A regisztrált és tényleges munkanélküliek számának eltérésében jellegzetes szezonális mintázatot vártunk. Ennek egyrészt a nyári ideiglenes (nem bejelentett) munkavállalók nagyobb száma, valamint a tanév végén a munkaerőpiacon megjelenő új munkaerő lehet az oka. Valóban, minden esetben sikerült egy kismértékű, de

jellegzetes szezonális ingadozást kimutatnunk. A szezonális mintázat lassú, σ_s^2 varianciájú változását megengedtük.

Mintavételi hiba (ARMA) összetevő. A mintavételi hibát egy ARMA-folyamatot követő egységnyi varianciájú (e_t) és egy változó varianciát leíró tényezőre (γ_t)bontjuk:

$$E_t = \gamma_t e_t . \qquad /12/$$

A γ_t idősort ismertnek tételezzük fel, azt a területi becslőfüggvények varianciájával tesszük egyenlővé. Ez ebben az esetben megfelelő közelítés, hiszen az idősorok varianciájának fő forrása a mintavételi hiba. Az e_t idősor ARMA paramétereit szintén egy független módszerrel becsüljük meg, magát az idősort viszont már a Kalman-szűrő algoritmus állítja elő.

Irreguláris összetevő. Az /1/ megfigyelési egyenletben szereplő ε_t fehérzajtag esetünkben a jelhez tartozik, hiszen mi csupán a mintavételi hibát szeretnénk leválasztani, a sokasági értékben megmaradó megmagyarázatlan ingadozást nem. Ez az összetevő annál kisebb, minél jobb a modell, azaz az idősor minél nagyobb részét képes megmagyarázni. Ha azonban túl nagyok (>1-2%) vagy kirívóan nem normáleloszlásúak ezek a maradékok, akkor a modellünk nagy valószínűséggel nem helytálló.

Modellválasztás, diagnosztika

Az EM-algoritmus alkalmas arra, hogy adott struktúrájú modell esetén kiszámítsa a paraméterek optimális értékeit, nem mond azonban semmit arról, hogy vajon a leírni kívánt folyamatot mennyire jól reprezentálja a modell. Különböző modellekkel végzett idősorbecslések összehasonlítására szolgálnak a modellszelekciós kritériumok.

Ilyen kritériumnak használható az Akaike Information Criterium (AIC), vagy a Prediction Error Variance (PEV). Minél kisebb ezen mérőszámok értéke, annál jobb az illeszkedés a modell és a valós adatok között.

A diagnosztikai eljárások azt mutatják meg, hogy mennyire sikerült a modellel leírni a rendszer szisztematikus viselkedését. Ehhez a becslés maradékait vizsgáljuk meg, amelyeknek tökéletes modell esetén független (korrelálatlan) normáleloszlást kell követniük.

A maradékok autokorrelációját teszteli a Ljung–Box-féle Q* statisztika (*Harvey*; 1991), normalitását a Bera–Jarque-teszt, mely az eloszlás ferdeségét (skewness) és lapultságát (kurtosis) kombinálja egy 2 szabadságfokú χ^2 eloszlást követő statisztikává. A heteroszkedaszticitást az idősor első és utolsó harmadának varianciáját összehasonlító *F*-teszttel vizsgáltuk.

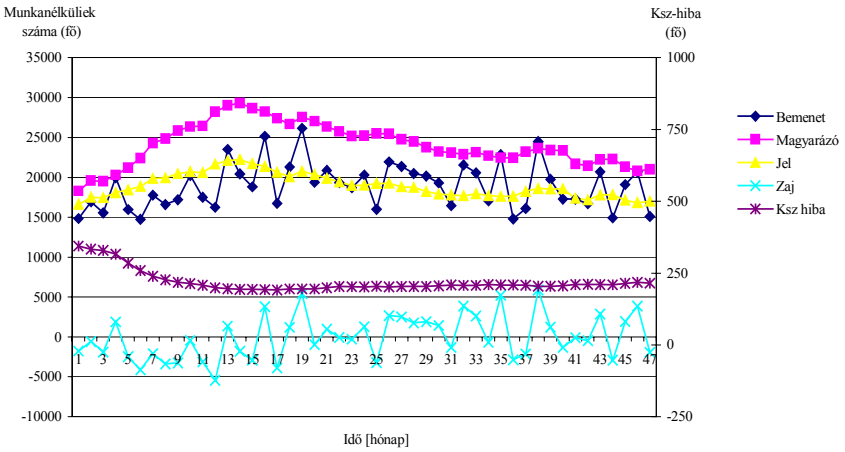
A modell előrejelző képességét ún. poszt-minta teszt segítségével vizsgáljuk. Ennek lényege az, hogy az idősor egy adott pontjától előrejelzést készítünk, s ennek a valós értékektől való eltéréseit (előrejelzési hiba varianciája) hasonlítjuk össze az idősor első szakaszán mért egylépéses előrejelzésihiba-varianciával. Minél jobb a modell előrejelző képessége, annál kisebb a két variancia hányadosa.

A Kalman-szűrős becslési eljárás illusztrációjaként mutatjuk be a következő, Zala megyére vonatkozó példát (lásd az ábrát).

A „Bemenet” nevű idősor a becslési eljárás során előállított ún. egyidejű becslésből származik, tehát az adott területi egység (Zala megye) adatait az ország más területein

mért értékekkel korrigálja, de csak az adott időpontra vonatkozó adatokat használja. A Kalman-szűrő ezt a bemenő idősort bontja fel „jel" és „zaj" komponensekre, az előbbiekben leírt strukturált idősormodell alapján. A „zaj" összetevő az adatfelvétel mintavételi hibáját reprezentálja. A modell magyarázó változóként az OMMK által regisztrált munkanélküliek adatait használja. Ezek láthatóan jóval simább idősort adnak, tehát jó okkal feltételezzük, hogy a bemenő idősorunk változékonysága jórészt a mintavételi hiba eredménye. A strukturált idősormodell teszi lehetővé, hogy a regisztrált munkanélküliek és a tényleges munkanélküliek közötti eltérést ne csupán egy állandó együtthatós regresszióval írjuk le, hanem figyelembe vehessük ezen eltérés szezonális ingadozását és lassan változó trendjét is.

A Kalman-szűrő eljárás jellemző idősorai



A Kalman-szűrő által előállított becslés standard hibáját a „Ksz-hiba" feliratú idősor mutatja, ez esetünkben (megye szintű felbontás) jellemzően körülbelül 1 százalék.

Idősoros kiigazítás

Az idősoros kiigazításon (benchmark) két különböző forrásból származó idősor utólagos összehangolását értjük. Adott tehát ugyanannak a változónak két idősora, melyek a mintavétel gyakoriságában különböznek. (Például a munkanélküliségre vonatkozó havi gyakoriságú, illetve évente végrehajtott felmérésekből származó adatok.) Lehetséges, hogy ez a két különböző forrásból származó adatsor nincs összhangban egymással, vagyis, például a havi idősorból számolt éves adat nem egyezik meg kielégítő pontossággal a direkt éves adattal.

Az idősoros kiigazítás olyan eljárás, amely optimális módon megteremti az összhangot a két idősor adatai között. Az optimális mód azt jelenti, hogy úgy érjük el a kellő konzisztenciát, hogy közben a lehető legkevésbé („alakot megőrizve") változtatjuk meg az adatokat.

A legszélesebb körben használatos módszer a *Denton-féle benchmarking* (*Cholete*; 1992, *Kisterületi*…; 1993) eljárás, amely matematikailag a korlátozott kvadratikus minimalizálás keretébe tartozik.

Reprezentálja a nagyobb gyakoriságú, kiigazítandó idősort a $\mathbf{z} = \left[z_1, z_2, ..., z_{p \cdot m}\right]$ vektor, a másik, nagyobb megbízhatóságú idősort pedig $\mathbf{y} = \left[y_1, y_2, ..., y_m\right]$.[2]

Keressünk **z** helyett olyan új

$$\mathbf{x} = \left[x_1, x_2, ..., x_{p \cdot m}\right]$$

vektort, amely

a) minimalizálja az eredeti **z** idősortól való eltérést egy célfüggvény segítségével (a Denton-módszer esetében ez az első differenciákból képzett négyzetösszeg),

b) valamint teljesíti azt a feltételt, hogy mindegyik évre az új idősor éven belüli értékeinek összege az arra az évre vonatkozó, másik forrásból származó éves összértékkel egyenlő.

Tehát a minimumot a

$$\sum_{i=1}^{p} x_{i,m} = y_m, \quad m = 1,2 \ldots M \qquad /13/$$

mellékfeltétellel keressük, ahol *p* az éven belüli periódusok száma és *m* éven belül vizsgálhatók felül az adatok.

A Denton-módszer a benchmark-értéket hiba nélkülinek tekinti, így a kiigazított adatok megbízhatóságáról sem szolgáltat információt.

A következőkben ismertetendő additív Cholette–Dagum-modellel hiba is becsülhető. Ebben a nagyobb gyakoriságú (évközi) idősort olyan összegnek fogjuk fel, melynek egyik tagja a keresett kiigazított idősor, a többi tag pedig konstans eltérést és sztochasztikusan viselkedő hibát ír le. Az éves idősort, tehát amihez hozzáigazítjuk a másikat, szintén összeggel modellezzük, melynek egyik tagja a megkövetelt kiigazítási kényszerfeltételt jelenti, másik tagja sztochasztikus hiba:

$$s_t = a + \theta_t + e_t, \quad E(e_t) = 0, \quad E(e_t e_{t-k}) = \sigma_{e_t} \sigma_{e_{t-k}} \rho_k, \quad (t = 1,...,T),$$

$$y_m = \frac{(\sum_{t \in m} \theta_t)}{p_p} + w_m, \quad E(w_m) = 0, \quad E({w_m}^2) = \sigma^2_{w_m}, \quad (m = 1,...,M). \qquad /14/$$

A /14/ modellben s_t a vizsgált évközi idősort jelenti, amely az igazi, de ismeretlen θ_t évközi érték, az ismeretlen konstans a eltérés és az autokorrelációs e_t hiba összege. Ugyanitt y_m az éves idősor, a p pedig a kétféle idősor kapcsolatát írja le, végül w_m az éves idősorhoz tartozó hiba.

A hibatagot bizonyos körülmények között mintavételi hibaként értelmezhetjük, mely heteroszkedasztikus lehet, azaz varianciája változhat az időben. A ρ_k autokorreláció sta

[2] Megjegyezzük, hogy ebben az egyszerűsített leírásban *p*-t (az egy éven belüli megfigyelések számát) konstansnak tekintettük. A valóságban előfordulhatnak olyan esetek is, amikor ez a *p* évente változik (például az egyik évben havi, a másik évben negyedéves megfigyelésekkel dolgozunk). Ugyancsak előfordulnak nem teljes évet felölelő megfigyelések is, amikor a nagyobb gyakoriságú idősor elemeinek száma nem *p·m.* A módszer megértése szempontjából azonban ez az egyszerűsített eset elegendőnek látszik.

cionárius és invertálható ARMA-modellnek felel meg, amit a felhasználó lát el paraméterértékekkel. Vagyis e_t olyan folyamatot követ, melyre

$$e_t = \sigma_t \varepsilon_t \ , \qquad /15/$$

és ahol ε_t egy kiválasztott stacionárius ARMA-modell szerint alakul:

$$\varepsilon_t = (\eta(B) / \phi(B)) \nu_t . \qquad /16/$$

A /16/ modellben $\eta(B)$ és $\phi(B)$ a mozgó átlag és az autoregresszív polinom ν_t pedig az ARMA-folyamat által generált zaj.

Az additív modell megközelítőleg visszaadja a Denton-féle eljárást, ha

– a konstans eltérés paramétere = 0,
– a ritkább idősor, amihez igazítunk, kötött, zéró varianciájú,
– az évközi idősor varianciája konstans,
– a hibataghoz választott ARMA-modell véletlen bolyongást ír le.

E módszer előnye, hogy figyelembe veszi annak az idősornak a varianciáját, amihez a kiigazítást végezzük, ezáltal a kiigazított idősor variancia-idősorát is megkaphatjuk. Megjegyezzük, hogy a KMSR képes még az ún. multiplikatív Cholette–Dagum-módszer szerint is a kiigazítást elvégezni.

IRODALOM

WOLTER, K. M. (1985): *Introduction to Variance Estimation*. Springer-Verlag, New York.
CHOLETTE, P. A. (1992): *Users Manual of Programme Bench*. Statistics Canada, Ottawa.
Kisterületi munkanélküliségi statisztikai rendszer kialakításának vizsgálata I–II. (1993) (Megvalósíthatósági tanulmány.) MultiRáció Szolgáltató Szövetkezet. Budapest.
Kisterületi munkanélküliségi statisztikai rendszer kifejlesztése. (1996a) (Világbanki zárótanulmány). MultiRáció Kft. Budapest.
Idősorelemzés alkalmazhatósága munkaügyi adatokra kisterületi szinten. (Jelentés), (1996b) MultiRáció Kft. Budapest.
HARVEY, A. C. (1991): *Forecasting, structural time series models and the Kalman filter*. Cambridge Univ. Press, Cambridge.
BANAI MIKLÓS ÉS TÁRSAI (2000): A kistérségi munkanélküliségi statisztikai rendszer és alkalmazása. *Területi Statisztika*, 40. évf. 2. sz. 108–125. old.
SINGH, M. B.– GAMBINO, Y. – MANTEL, H. (1992): *Issues and Options in the Provision of Small Area Data*. Statistics Canada, Working Papers K1AOT6.
SÄRNDAL, C-E. – SWENSSON, B. – WRETMAN, Y. (1992): *Model Assisted Survey Sampling*. Springer-Verlag, New York.
PLATEK, R. ÉS TSAI. (Szerk.) (1987): *Small Area statistics*. John Wiley and Sons, New York.
HIDIROGLOU, M. A. – MORRY, M. – DOGUM, E. B. – RAO, Y. N. K. – SÄRNDAL, C-E. (1985): *Evaluation of Small Area Estimators Using Administrative Records*. Statistics Canada, Working Papers TSRA-85-024E.

SUMMARY

This two-part study reviews the so-called small area unemployment statistical system developed for the calculation of small area labour force data. It also reviews the process of the development of this system. The estimating system combines the small area data of the labour force survey with those data of the registered unemployment statistics to produce the monthly or quarterly labour force small area estimation. The goal of the combination is to increase the reliability of data estimated in this way by decreasing the statistical error coming from the small sample size. The estimating system reaches this goal by applying small area estimators for cross sectional data and time series methods for historical data.