HAVI MUNKAÜGYI ADATOK BECSLÉSÉNEK MÓDSZERTANA A KSH-BAN[1]

HORVÁTH BEÁTA – LOVICS GÁBOR
Központi Statisztikai Hivatal – Magyar Telekom

Bármely ország gazdaságának fontos jellemzője a munkaerőpiaci folyamatainak alakulása. Ezért fontos, hogy a döntéshozók minél gyorsabban, minél pontosabb képet kapjanak az ezt jellemző folyamatokról, amiben a munkaerő-felmérésből származó információknak kiemelten fontos szerepe van. A felvétel klasszikusan negyedéves gyakorisággal ad becslést, de gyorsuló világunkban egyre nagyobb szükség van ennél sűrűbb információkra. Ebben a tanulmányban azt mutatjuk be, hogy a KSH milyen állapottérmodell segítségével készíti el havi becslését a főbb munkaerőpiaci indikátorok tekintetében.

Kulcsszavak: foglalkoztatottság, munkanélküliség, állapottérmodell

1 Bevezetés

A Központi Statisztikai Hivatal (KSH) 1992 óta a nemzetközi ajánlásoknak megfelelően, harmonizált módon teszi közzé a munkaerő-felmérés (MEF) adatait, melyek a foglalkoztatotti és a gazdasági aktivitási információk legfőbb forrása. A MEF mintájának mind a felépítése, mind pedig a nagysága többször változott az idők folyamán. A jelenlegi negyedéves minta az adatgyűjtés szempontjainak megfelelően három, közel azonos nagyságú havi részmintából áll, melyeknek nincs közös részük. A negyedéves mintából közölt adatok megfelelnek a felvétellel szemben támasztott minőségi kritériumoknak. A felhasználói igények azonban, összhangban az Eurostat előírásaival, abba az irányba mutattak, hogy a főbb munkaerőpiaci indikátorok tekintetében szükség van havi frekvenciájú adatközlésre. A nyers havi becslések alakulását az 5. és a 6. ábrák kék vonalai mutatják. Mivel a negyedéves adatfelvétel mögött meghúzódó diszjunkt havi minta meg nem magyarázható volatilitást visz a becslés idősorába, ezért azok közvetlenül nem használhatók statisztikai következtetésekre. A Hivatal eddig ezt háromhavi mozgóátlag publikálásával oldotta meg, ami azonban az adatsorokban jelentkező nagyobb kilengéseket csak késleltetéssel és simítva jeleníti meg.

Tanulmányunk célja, hogy ismertesse, milyen modellezési eszköztárral javítunk a havi munkaügyi adatok becslésén. A szakirodalom áttekintése és előzetes vizsgálatok után arra jutottunk, hogy a probléma legjobb megoldását

[1]Beérkezett 2023. december 8. DOI: https://doi.org/10.15170/SZIGMA.54.1203. E-mail: beata.horvath@ksh.hu, lovixg@gmail.com. A szerzők köszönik Rakovics Mártonnak, hogy segített nekik elindulni az állapottérmodellek megértésének izgalmas utazásában.

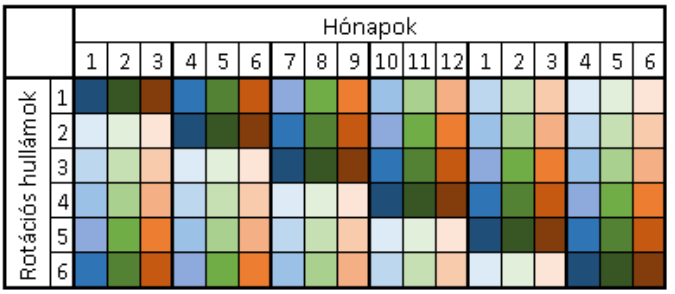
az úgynevezett állapottérmodellek jelentik (Aoki (1990)). Ezek a modellek képesek arra, hogy a rendszeresen mért adatok hibáját korrigálják. A modellkeret kellően rugalmas, ugyanakkor megbízható, stabil megoldásra csak akkor vezet, ha a paramétereinek becslésénél elég sok információ áll rendelkezésre.

A probléma természetesen nem hazai specifikum, számos ország foglalkozott a kérdéssel. Európán belül először a Holland Statisztikai Hivatal, eredményeiket számos alkalommal publikálták (van den Brakel and Krieg (2008), van den Brakel and Krieg (2015), Bollineni-Balabay et al. (2016), Schiavoni et al. (2021)), amit később a Francia Statisztikai Hivatal (Deroyon et al. (2013)) és az Angol Statisztikai Hivatal (ONS (2019)) is felhasznált és alkalmazott. Az USA-ban kisterületi becslés alkalmazására használják a munkaügyi adatok kapcsán (Tiller and Pfeffermann (2006)). Jelenleg az Európai Hivatalos Statisztikai Szolgálat tagjaként Hollandia és Magyarország publikálja a havi munkaügyi adatokat állapottér alapú modellezés segítségével. A többi tagország és az Eurostat az időbeli szétosztási technikák (Cholette (2006)) segítségével oldja meg a problémát. A KSH egy grant projekt keretében (KSH (2018)) szintén körüljárta ezeket a lehetőségeket, azonban az eredmények nem teljesítették a revízióra és volatilitásra vonatkozó kritériumokat[2], így elvetettük ezen eljárások alkalmazását.

Azért, hogy stabil, jól működő, torzítatlan becsléseket kapjunk, a modell építését két irányból is megtámogattuk. Egyrészt a nyers becslés hibájának dinamikájával kapcsolatban előzetes vizsgálatokat végeztünk (Pfeffermann et al. (1998)). Ez jelentősen befolyásolta a modell struktúráját, és így bizonyos számú paramétert előre meg tudtunk határozni. Másrészről igyekeztünk olyan adminisztratív adatforrásokat találni, melyek szintén támogatják a becsléseinket. Az adminisztratív adatokból segéd idősorok lettek, melyeket szintén beépítettünk a modelljeinkbe. A segédsorok beépítése már erősen a magyar sajátosságok figyelembevételével történt. Nagyon hasonló eljárást készítettek az Egyesült Államokban Pfeffermann and Tiller (2006).

A tanulmány szempontjából fontos, hogy megértsük a MEF jelenlegi struktúráját, rotációs panel jellegét. A MEF többlépcsős, rétegzett valószínűségi minta (bővebben KSH (2006)), ahol a folyamatos felvétel során egyszerű rotációt is alkalmaznak. Azaz a kiválasztott háztartásokban negyedévenként, mindig a negyedéven belül azonos sorszámú hónapban, hat egymásután következő alkalommal kerül sor interjúra. Például, ha kiválasztanak egy háztartást 2022. januári adatközlésre, akkor őt még abban az évben áprilisban, júliusban, októberben, valamint 2023. januárban és áprilisban is, azaz összesen hat alkalommal (panel) felkeresi az összeíró. A MEF mintája tehát egy olyan panelminta, amelyben az egymás után következő időszakok öthatoda, majd két időszak múlva négyhatoda, majd háromhatoda stb. – a természetes lemorzsolódástól eltekintve – azonos. Más megközelítésből ez azt jelenti, hogy a mintának mindig az egyhatoda rotálódik ki. Így egy adott időszak (hónap vagy negyedév) becslése hat különböző panelből (hullám) áll össze, lásd *1. ábra,* ahol az azonos színű és árnyalatú egységek az egyes paneleket, míg az azonos színű, de eltérő árnyalatú színek a hullámokat jelölik.

[2]https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32019R2241



1. ábra. A rotációs panel alakulása időben

A KSH 2023. január óta nemenként, havonta publikálja[3] a 15–64, illetve a 15–74 éves népesség gazdasági aktivitására vonatkozóan az alábbi mutatókat: foglalkoztatottak, munkanélküliek, gazdasági aktívak, gazdaságilag nem aktívak, 15–74 éves népesség, aktivitási arány, munkanélküliségi ráta, foglalkoztatási ráta. Az Eurostat havonta közli[4] a munkanélküliségi rátát nemenként a 15–24, illetve a 25–74 éves korcsoportra szezonálisan kiigazítva is. Mivel a népességre vonatkozó nemenkénti korcsoportos adatok ismertek a havi becslés készítésekor, ezért elegendő a munkanélküliek és a foglalkoztatottak számát modelleznünk, mert az összes többi mutató ezekből már kiszámolható.

Az uniós és a hazai publikációk miatt a foglalkoztatottsággal kapcsolatos mutatókat többféle korcsoportos bontásban, és valamennyi korcsoportot nemek szerint bontva kell előállítani. A gyakorlatban ez azt jelenti, hogy becslést készítünk a teljes korcsoportról, a 15–74 évesekről, a magyar nyugdíjszabályokhoz jobban alkalmazkodó 15–64 évesekről, és ugyan külön nem publikáljuk, de így adódik becslésünk a 65–74 évesekre is. Szintén külön becslés készül a fiatalokról, vagyis a 15–24 éves korcsoportról, és így indirekt módon jutunk a 25–74 évesek becsléséhez.

A téma részletes tárgyalása előtt pár szó a jelölésekről. Tanulmányunkban hagyományosan kis latin vagy görög betűkkel jelöljük a számokat. Az oszlopvektorokat vastag latin kis betűkkel, speciálisan $\mathbf{o}$ azt a vektort jelöli, amelyiknek minden komponense 0. A mátrixokat nagy latin betűkkel, a csupa nullából álló mátrixot 0-val jelöljük. A transzponálás műveletére a $\top$ jelölést használjuk jobb felső indexben. A nevezetes számhalmazokat dupla vonalas nagy latin betűkkel jelöljük. Mivel a modellezés során gyakran nagyméretű mátrixokkal dolgozunk, ezért a mátrixok felett és tőle balra jelöljük, hogy az egyes sorok, illetve oszlopok mire vonatkoznak. Példaként nézzük meg a következő kétváltozós dinamikai rendszert:

$$\begin{aligned} a_t &= a_{t-1} + 2b_{t-1} \\ b_t &= 3a_{t-1} + 4b_{t-1} \end{aligned}$$

[3] https://www.ksh.hu/stadat_files/mun/hu/mun0097.html

[4] https://ec.europa.eu/eurostat/databrowser/view/UNE_RT_M/default/table?lang=en&category=labour.employ.lfsi.une

Ebben az esetben a rendszerhez tartozó együtthatómátrixot a következő módon jelöljük:

$$\begin{array}{c} \\ a_t \\ b_t \end{array} \begin{array}{c} \begin{array}{cc} a_{t-1} & b_{t-1} \end{array} \\ \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \end{array}$$

Használjuk még a diag függvényt, amelyik egy n dimenziós vektorból egy $n \times n$-es mátrixot készít úgy, hogy a vektor elemei a mátrix diagonálisába kerülnek, a többi elem pedig nulla. Például

$$\mathrm{diag}(1,2,3) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{pmatrix}.$$

A tanulmány hátralévő része a következőképpen épül fel. A 2. fejezetben röviden ismertetjük az állapottérmodelleket, ami az alapvető keretét adja a becslési eljárásnak. Az ezt követő két fejezetben azt mutatjuk be, milyen előzetes vizsgálatok segítenek stabilabbá és pontosabbá tenni a becslési eljárásainkat. A 3. fejezetben azt taglaljuk, hogy a becslési hiba időbeli alakulásának milyen jellemzőit állapítottuk meg. Az itt kapott eredmények mint fontos fix paraméterek épültek be a modellünkbe. Ezt követően a 4. fejezetben pedig azt mutatjuk meg, hogy milyen segédsorokat építünk be a modellekbe.

Ezután két külön fejezetben tárgyaljuk, hogy milyen modelleket használunk a becslési eljárás során. Az 5. fejezet a foglalkoztatottaknál alkalmazott modellt tartalmazza, míg a 6. fejezet a munkanélküliekre vonatkozót írja le. Legvégül a 7. fejezetben röviden összefoglaljuk, hogy a korábban bemutatott technikák hogyan állnak össze egy komplex becslési eljárássá, és bemutatjuk az eredményeinket. Röviden ismertetjük azt is, hogy mennyire felel meg a modell az előzetes elvárásoknak, illetve azt, hogy milyen továbblépési lehetőségeket látunk a modell fejlesztésére.

A tanulmányban a számítások jelentős része R programnyelvben készült, a becslés alapját képező állapottérmodelleket dlm csomag (Petris (2010), Petris (2009)) használatával számoltuk ki. Ugyanezeket használjuk a KSH hivatalos adatközlései során. Az idősoros előrejelzés Eviews szoftverben készült, míg az idősoros kiigazítás a Demetra szoftver segítségével történt.

2 Az alkalmazott modellkeret

A becslési eljárás során alapvetően az úgynevezett állapottérmodelleket használjuk. Ezeknek a modelleknek a kiindulópontja, hogy vannak időben változó, de közvetlenül meg nem figyelhető jelenségek, amelyeknek az alakulását szeretnénk becsülni. Ezen változókat jelölje $\mathbf{y}_t \in I\!R^n$ vektor. Itt a közvetlenül meg nem figyelhetőség jelentheti azt is, hogy vannak ugyan megfigyeléseink, de ezek hibásak. Vannak viszont olyan folyamatok is, amiket megfigyelünk: jelölje ezek vektorát $\mathbf{z}_t \in I\!R^k$. Feltételezzük továbbá azt is, hogy van valamilyen (időben nem feltétlenül állandó) összefüggés a kétfajta változó között,

ami lineáris és sztochasztikus. Jelölje $F_t \in I\!R^{k\times n}$ azt az időben változó mátrixot, amely ennek az összefüggésnek az együtthatóit tartalmazza. Az állapottérmodell úgynevezett állapotegyenlete a következő alakban írható fel:

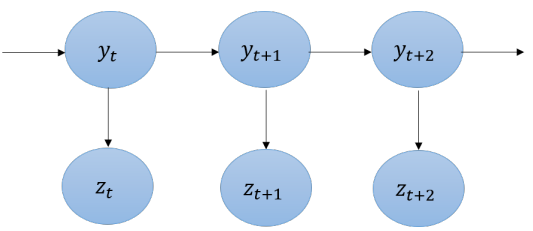
$$\mathbf{z}_t = F_t \mathbf{y}_t + \mathbf{e}_{1,t}\,, \tag{1}$$

ahol $\mathbf{e}_{1,t} \in I\!R^k$, $\mathbf{e}_{1,t} \sim \mathcal{N}(\mathbf{o}, V)$ többváltozós sztochasztikus változó $V \in I\!R^{k\times k}$ kovarianciamátrixszal. Az eljárás szempontjából nem követelmény, hogy F_t és $V \in I\!R^{k\times k}$ mátrixok teljes mértékben ismertek legyenek, de minél kevesebb az ismeretlen tényező, annál nagyobb a modell stabilitása.

Szükséges még, hogy legyenek ismereteink az $\mathbf{y}_t$ időbeli dinamikájáról. Ezt is egy sztochasztikus lineáris összefüggéssel írhatjuk le, ahol az együtthatókat a $G_t \in I\!R^{n\times n}$ mátrix tartalmazza. Ezt az egyenletet a modell átmenetegyenletének fogjuk hívni:

$$\mathbf{y}_{t+1} = G_t \mathbf{y}_t + \mathbf{e}_{2,t}\,, \tag{2}$$

ahol $\mathbf{e}_{2,t} \in I\!R^n$, $\mathbf{e}_{2,t} \sim \mathcal{N}(\mathbf{o}, W)$ egy másik többváltozós sztochasztikus változó $W \in I\!R^{n\times n}$ kovarianciamátrixszal. Ennél az egyenletnél is igaz, hogy sem az együttható-, sem a kovarianciamátrix nem kell, hogy feltétlenül ismert legyen. Az állapottérmodellek sematikus leírását a *2. ábrán* láthatjuk.



2. ábra. Az állapottérmodellek sematikus ábrája

Ez a két egyenletrendszer szükséges ahhoz, hogy legyen egy állapottérmodellünk. Amennyiben ezeket sikerült felírnunk, akkor az úgynevezett Kálmán-szűrők (Kalman (1960); Kalman and Bucy (1961)) segítségével meg tudjuk becsülni az F_t, G_t, V, W mátrixok ismeretlen elemeit, és ennek segítségével szimulálni tudjuk azokat az idősorokat, amikre kíváncsiak vagyunk. Ez lényegében azt jelenti, hogy ez a négy mátrix írja le az állapottérmodellt.

Az állapot- és az átmenetegyenlet is sztochasztikus összefüggéseket tartalmaz ugyan, de speciálisan ez jelenthet determinisztikus összefüggéseket is. Ez technikailag úgy oldható meg, ha valamelyik egyenlethez tartozó hibatagot egy 0 várható értékű és 0 szórású valószínűségi változóként definiáljuk.

Érdemes még megjegyezni, hogy a társadalmi és gazdasági folyamatok időbeli elemzésénél gyakran használt komponensek, mint a trend és a szezonális faktor állapottérmodellekben való reprezentációja jól ismert a szakirodalomban (például van den Brakel and Krieg (2009)). Ezt fel is fogjuk használni a részletes modell leírásakor.

Ahogy azt korábban leírtuk, a becsléseket nemi és korcsoportos bontások szerint is el kell készítenünk. Természetes elvárás, hogy a komplementer részsokaságok összege kiadja a teljes sokaságot, azonban ha külön-külön modellalapú becslést adunk az egyes sokaságokra, akkor ez nem feltétlenül teljesül. Ez igaz azokra a technikákra is, amiket itt alkalmazunk. A probléma viszont többféle módon is orvosolható.

A szakirodalommal összhangban abból indultunk ki, hogy a teljes sokaságra a nagyobb minta miatt pontosabb becslés adható, mint a részsokaságokra. Ezért a teljes sokaságra adott becslésünket elfogadjuk, és a részsokaságokra adott becslést módosítjuk úgy, hogy a kívánt összefüggések teljesüljenek.

A modell eredményeinek vizsgálata során arra jutottunk, hogy máshogy érdemes kezelni azt az esetet, amikor a sokaságot egy nagy és egy kis csoportra bontjuk (például korcsoportos bontás esetén a 15–74 éveseket 15–64 és 65–74 évesekre), vagy ha két, nagyjából egyforma csoportra (például egy tetszőleges korcsoportot férfiakra és nőkre). Előbbi esetben a nagy csoportban keletkező relatív kis hiba is nagy gondokat tud okozni a kis csoportban, míg hasonló probléma nem fordul elő, ha két, közel azonos méretű sokaságra bontjuk a teljeset.

Az adatok vizsgálata közben azt is megállapítottuk, hogy a munkanélkülieknél a 65–74 éves korcsoport egészen egyedinek tekinthető. Ebbe a korcsoportba olyan emberek kerülnek, akik a hazai szabályozás szerint jelenleg ugyan elérték a nyugdíjkorhatárt, mégis aktívan keresnek munkát, munkába is tudnának állni, de nem találnak állást. Ez a teljes populációban is annyira ritka eset, hogy gyakran nem is kerül a mintába egyetlen ilyen adatszolgáltató (háztartás) sem, így ekkor a nyers becslés erre a csoportra vonatkozóan 0. Ugyanakkor az is elmondható, hogy a relatív volatilitása nagyon nagy ennek a csoportnak, úgy, hogy a teljes sokasághoz viszonyítva összeségében nagyon kevesen vannak. Ezért úgy döntöttünk, hogy ezt a sokaságot nem modellezzük, mert ebben a nagyon speciális esetben nem javítható a nyers becslés eredménye modellezéssel.

Mindezek alapján a korcsoportos bontások a következő módon készülnek el. A foglalkoztatottak esetén modellezzük a 15–74, a 65–74 és a 15–24-es korcsoportot, és a többi korosztályt úgy kapjuk, hogy a teljesből kivonjuk a megfelelő, már meghatározott korcsoportokat. A munkanélkülieknél modellezzük a 15–74 és a 15–24-es korcsoportot, valamint a 65–74-es korcsoportnak elfogadjuk a nyers becslését, valamint a foglalkoztatottakhoz hasonlóan képezzük a többi korcsoportot.

A nemi bontások esetén minden modellezett korcsoportban meghatározzuk a nők $v_{n,t}$, a férfiak $v_{f,t}$ és az összes $v_{o,t}$ személy számát, minden $t \in \{1, 2, \ldots, \tau\}$ időpontra. Ezekből képezzük a következő vektorokat:

$$\begin{aligned} \mathbf{v} &= (v_{n,1}, \ldots, v_{n,\tau}, v_{f,1}, \ldots, v_{f,\tau})^\top, \\ \mathbf{v}_o &= (v_{o,1}, \ldots, v_{o,\tau})^\top. \end{aligned}$$

Ezeket a becsléseket, amennyire csak lehet, minimálisan módosítjuk úgy, hogy az összefüggés teljesüljön, vagyis, hogy a nők és a férfiak száma kiadja az

összes megfigyeltet. Ez azt jelenti, hogy keressük azt az

$$\mathbf{u} \quad = \quad (u_{n,1}, \ldots, u_{n,\tau}, u_{f,1}, \ldots, u_{f,\tau})^\top$$

vektort, amelyik optimális megoldását adja a következő (SOP) kvadratikus optimalizálási feladatnak:

$$\left.\begin{array}{c} \min\limits_{\mathbf{u}} (\mathbf{u}-\mathbf{v})^\top (\mathbf{u}-\mathbf{v}) \\ A\mathbf{u} = \mathbf{v}_o \end{array}\right\} \qquad (SOP)$$

A két nemet aggregáló $A \in I\!R^{\tau \times 2\tau}$ mátrix pedig a következő

$$A = \begin{array}{c} \\ v_{o,1} \\ \vdots \\ v_{o,\tau} \end{array} \begin{array}{c} \begin{array}{cccccc} u_{n,1} & \ldots & u_{n,\tau} & u_{f,1} & \ldots & u_{f,\tau} \end{array} \\ \begin{pmatrix} 1 & \ldots & 0 & 1 & \ldots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \ldots & 1 & 0 & \ldots & 1 \end{pmatrix} \end{array}.$$

A feladatra a Lagrange-féle multiplikátor módszert (Klerk et al. (2004)) alkalmazva kapjuk, hogy az optimális megoldás a következő:

$$\mathbf{u}^* = \mathbf{v} + A^\top (AA^\top)^{-1} (\mathbf{v}_o - A\mathbf{v}) \,. \tag{3}$$

Mivel az A mátrix sorai lineárisan függetlenek, ezért az $AA^\top$ mátrix szigorúan pozitív definit, így valóban invertálható. Az eredményül kapott $\mathbf{u}^*$ vektor első felében vannak a nőkre, a második felében a férfiakra vonatkozó korrigált becslések, melyekre már teljesül a kívánt összefüggés.

A foglalkoztatottak esetén összesen kilenc idősort fogunk modellezni: három korcsoportot, és mindegyik esetben a férfiakat, a nőket és az összesent. Minden korcsoportban módosítjuk a nők és a férfiak becslését úgy, hogy kiadják az összesent a (3) egyenlet segítségével. Végül különbségképzéssel előállítjuk az összes többi korcsoportot is, nemi bontásokban is.

A munkanélkülieknél összesen hat idősort modellezünk, mert csak két korcsoportot modellezünk nemenkénti bontásban. Itt is alkalmazzuk az (3) egyenletet, hogy korrigáljuk a nemek szerinti bontást, és a korábban leírtaknak megfelelően a 65–74 évesekre vonatkozó nyers eredményeket nemi bontásonként elfogadjuk. Végezetül a többi korcsoportot ugyanúgy különbségképzéssel állítjuk elő, mint a foglalkoztatottak esetén is.

3 A nyers becslés hibatagjának időbeli dinamikája

A részletes elemzéshez egy nagyon egyszerű összefüggésből indulunk ki, nevezetesen, hogy a valódi adat mindössze egy becslési hibával tér el a becsült értéktől.

$$\hat{v}_t = v_t + r_t \,, \tag{4}$$

ahol

- v_t a becsülni kívánt munkaügyi adat t-edik hónapban;
- $\hat{v}_t$ a nyers becslés a munkaügyi adatra t-edik hónapban;
- r_t a nyers becslés hibája a t-edik hónapban.

Munkánk során feltételeztük, hogy a munkaügyi adatok negyedéves becslései, és így hosszú távon a havi becslések is torzítatlanok, vagyis, hogy $E(r_t) = 0$. Ahogy azonban korábban bemutattuk, a mintavételi technika miatt okkal feltételezhetjük, hogy a hibatag időben nem korrelálatlan, vagyis, hogy

$$\rho_k = \mathrm{corr}(r_t, r_{t-k}) \neq 0$$

minden k-ra. Itt a szakirodalommal összhangban indirekt azt is feltételeztük, hogy a korreláció csak k-tól függ, t-től viszont független.

A fejezet további részében a Pfeffermann et al. (1998) tanulmányban leírtakat követjük, de számításainkat magyar adatokon, a MEF mintavételi eljárásának megfelelően végezzük el. A vizsgálathoz be kell vezessünk néhány új jelölést. A KSH jelenlegi gyakorlatában a becslési folyamat során azt feltételezzük, hogy a végleges becslés az egyes hullámokból készült becslések átlaga. Jelölje $\hat{v}_{it}$ azt a becslést, ami a t-edik hónapban az i-edik hullámból történik, akkor a nyers becslés végső soron

$$\hat{v}_t = \frac{1}{6} \sum_{i=1}^{6} \hat{v}_{it}$$

alakban írható fel. Az egyes hullámokból készült becsléseknek is van hibája, jelölje őket r_{it}, melyek átlaga adja a teljes hibát, hiszen

$$r_t = \hat{v}_t - v_t = \frac{1}{6} \sum_{i=1}^{6} \hat{v}_{it} - v_t = \frac{1}{6} \sum_{i=1}^{6} (\hat{v}_{it} - v_t) = \frac{1}{6} \sum_{i=1}^{6} r_{it} \,.$$

Ahhoz, hogy az elméleti, de meg nem figyelhető hibát jellemezni tudjuk, bevezetjük még a megfigyelhető hiba (q_{it}) fogalmát, ami az egyes hullámokból készült becslés és az összes megfigyelésből készült becslés különbsége

$$q_{it} = \hat{v}_{it} - \hat{v}_t \,.$$

Könnyen látható, hogy a megfigyelhető hiba előáll, mint a korábban bemutatott elméleti hibák különbsége, hiszen:

$$q_{it} = \hat{v}_{it} - \hat{v}_t = \hat{v}_{it} - v_t - (\hat{v}_t - v_t) = r_t - r_{it} \,.$$

Az tehát látszik, hogy ha ismernénk az elméleti hibákat, akkor mindent meg tudnánk mondani a megfigyelhető hibáról. Nekünk azonban bizonyos szintig ennek a fordítottjára van szükségünk, hiszen a megfigyelhető hibából szeretnénk következtetéseket levonni a nem megfigyelhető hibára. Ahogy azt Pfeffermann et al. (1998) megmutatták, bizonyos feltételek mellett ez is lehetséges, legalábbis ρ_k becsülhető a megfigyelhető hibák és azok korrelációi segítségével.

Ahhoz, hogy ezt leírjuk, be kell vezetnünk az előzménypanel fogalmát. Egy adott év júliusában a második panelnak nincs előzménye júniusban és májusban. Ezzel szemben áprilisban az első panel ugyanazokból az emberekből áll, akik a második panelt alkotják júliusban, így ebben az esetben ezt a panelt tekintjük az eredeti előzményének. Újabb három hónappal korábban, januárban már nincs olyan panel, amely ugyanazokból az emberekből áll, mint akikről eddig írtunk. Azonban az ekkori hatodik panelt cseréljük az áprilisi első panelre, így valójában ezt is előzménypanelnek tekintjük. És így tovább visszafelé az időben. Az 1. ábrán, ha egy panel előtt látunk egy azonos színű panelt, akkor az azt jelöli, hogy ebben az esetben ez az eredeti panel előzménye. Általánosságban tehát a következő módon írhatjuk le ezt a fogalmat. Legyen egy adott panel t-edik hónapban az i-edik. Ekkor a $t-k$-ban ennek a panelnek az előzménye az, amelyik a t-ben lesz az i-edik, vagy ha lecseréljük egy panelra, amelyik t-ben az i-edik lesz.

Tegyük fel, hogy $t-k$-ban van előzménye a panelnek, amelyik i-edik t-ben. Jelölje ekkor az ebből a panelből készült becslés hibáját $r_{t-k}^{(it)}$, illetve a hozzá tartozó megfigyelhető hibát $q_{t-k}^{(it)}$. Az elméleti hibák időbeli alakulásával kapcsolatban két további feltevéssel élünk. Az első, hogy ha egy korábbi panel nem előzménye egy későbbinek, akkor azok korrelálatlanok, vagyis

$$\operatorname{cov}(r_{it}, r_{t-k}^{(jt)}) = 0 \qquad i \neq j\,, \quad k = 0, 1, 2, \ldots\,.$$

Érdemes felhívni a figyelmet, hogy a jelenleg alkalmazott mintavételi eljárás tekintetében ez azt jelenti, hogy ha k nem osztható hárommal, akkor a hibák nem korrelálnak. Hiszen ebben az esetben azok a panelek, amelyek szerepelnek a t-edik időszak becslésében, azoknak nem szerepel előzménye a $t-k$-adik időszakban. És ekkor

$$\gamma_k = \operatorname{cov}(r_t, r_{t-k}) = \operatorname{cov}\Big(\frac{1}{6}\sum_{i=1}^{6} r_{it}, \frac{1}{6}\sum_{j=1}^{6} r_{j(t-k)}\Big) =$$

$$= \frac{1}{36}\sum_{i,j} \operatorname{cov}(r_{it}, r_{j(t-k)}) = 0\,.$$

A másik feltevés, hogy ha a hibáknak a saját előzményével való kovarianciáját vizsgáljuk, az nem függ a megfigyelés idejétől, csak attól, hogy mennyi idő telt el a két megfigyelés között, amelyekhez a hibák tartoznak. Formálisan ezt a következő módon fejezzük ki

$$\operatorname{cov}(r_{it}, r_{t-k}^{(it)}) = \gamma_k^i \qquad k = 0, 3, 6, \ldots$$

azokban az esetekben, amikor a kovariancia nem 0.

Ahhoz, hogy becslést tudjunk készíteni ρ_k-ra, szükségünk lesz a megfigyelhető hibák kovarianciáira, vagyis legyen

$$c_k^i = \operatorname{cov}(q_{it}, q_{t-k}^{(it)})\,.$$

Ahhoz, hogy ezt a kovarianciát becsülni tudjuk, határozzuk meg az átlagos megfigyelt hibákat a következő módon

$$\begin{aligned} \bar{q}_i &= \sum_{t=1}^{n} \frac{q_{it}}{n} \\ \bar{q}_{ik} &= \sum_{t=k+1}^{n+k} \frac{q_{t-k}^{(it)}}{n}\,, \end{aligned}$$

ahol n azon hónapok számát mutatja, amelyeknek adatait felhasználjuk a számításokhoz. Ekkor a keresett kovariancia a következő módon becsülhető

$$\hat{c}_k^i = \sum_{t=k+1}^{n} \frac{(q_{it} - \bar{q}_i)(q_{t-k}^{(it)} - \bar{q}_{ik})}{n} \qquad (k = 0, 3, 6, \ldots)\,.$$

Pfeffermann et al. (1998) bizonyították, hogy a korábban bemutatott feltételek mellett a hibák időbeli korrelációja az alábbi képlet segítségével számolható:

$$\hat{\rho}_k = \frac{\sum_{i=1}^{5} \hat{c}_k^i}{\sum_{i=1}^{5} \hat{c}_0^i} \qquad (k = 0, 3, 6, \ldots)\,. \tag{5}$$

Ebben a képletben felhasználtuk, hogy ha egy panel megjelenik, akkor az még 5 alkalommal fog szerepelni a becslésünkben. A képletünket csak azokra a késleltetésekre alkalmazzuk, amikor feltevéseink szerint a korreláció nem 0.

A korrelációk segítenek, hogy a hibák időbeli alakulását egy autoregresszív folyamatként (Hamilton (1994); Brockwell and Davis (2002)) írjuk fel. Ehhez először is arra van szükség, hogy az autokorreláció mellett az r_t hibatagok időbeli kovarianciáját ($\hat{\gamma}_k$) is megbecsüljük, amihez felhasználtuk Pfeffermannék levezetéséből, hogy

$$\hat{\gamma}_k = \frac{1}{30} \sum_{i=1}^{5} \hat{c}_i^k \qquad (k = 0, 3, 6, \ldots)\,.$$

A kovarianciákat felhasználva pedig becsülhetővé válnak a parciális autokorrelációk ($\hat{\alpha}_k$) is. Tudjuk, hogy az $\hat{\alpha}_0 = 1$. A további $\hat{\alpha}_k$-k meghatározásához definiáljuk azt a M_k $k \times k$-ás mátrixot, amelynek az i-edik sorának j-edik eleme $\hat{\gamma}_{i-j}$ (ahol $\hat{\gamma}_k = \hat{\gamma}_{-k}$ teljesül). Legyen továbbá $\mathbf{p}_k = (\gamma_1, \gamma_2, \ldots \gamma_k)^\top$ és

$$\tilde{\mathbf{p}}_k = M_k^{-1} \mathbf{p}_k\,.$$

Az $\hat{\alpha}_k$ ebben az esetben a $\tilde{\mathbf{p}}_k$ vektor utolsó koordinátájával egyenlő.

Mindezeket felhasználva megvizsgáltuk, hogy időben visszamenőlegesen milyen hosszan tekinthetőek szignifikánsnak az autokorelációk és a parciális autokorelációk.

Hasonlóan ahhoz, amit a szakirodalomban találtunk más országok esetén, az eredmények ebből a szempontból az egyes célváltozókon nem voltak azonosak. A foglalkoztatottak esetén kettő, míg a munkanélküliek esetén három

negyedéves késleltetést veszünk figyelembe. Formálisan ez azt jelenti, hogy a foglalkoztatottak esetén a becslési hiba dinamikáját az alábbi alakban írhatjuk le

$$r_t = \phi_1 r_{t-3} + \phi_2 r_{t-6} + \epsilon_{r,t}\,. \tag{6}$$

A munkanélküliek esetében ugyanez a következő formában írható fel:

$$r_t = \phi_1 r_{t-3} + \phi_2 r_{t-6} + \phi_3 r_{t-9} + \epsilon_{r,t}\,. \tag{7}$$

Mindkét egyenletre igaz, hogy $\epsilon_{r,t} \sim \mathcal{N}(0, \sigma_r^2)$.

Ezek után alkalmazhattuk az úgynevezett Yule-Walker egyenleteket, melyek a foglalkoztatottak és a munkanélküliek esetén rendre a következő formában írhatók fel:

$$\begin{aligned} \rho_k &= \phi_1 \rho_{k-3} + \phi_2 \rho_{k-6} \qquad (k = 3, 6)\,, \\ \rho_k &= \phi_1 \rho_{k-3} + \phi_2 \rho_{k-6} + \phi_3 \rho_{k-9} \qquad (k = 3, 6, 9)\,. \end{aligned}$$

Kihasználva, hogy az (5) egyenletből ismert ρ_k becslése, valamint hogy tudjuk, hogy $\rho_0 = 1$ és $\rho_k = \rho_{-k}$, becsülhetők lesznek a ϕ_i együtthatók. Az *1. táblázatban* foglaltuk össze, hogy számításaink alapján hogyan alakultak az együtthatók a különböző sokaságok esetén.

Sokaság	Foglalkoztatottak		Munkanélküliek		
	ϕ_1	ϕ_2	ϕ_1	ϕ_2	ϕ_3
15–74 Összesen	0.5863056	0.02131146	0.49311877	0.128417098	-0.005960687
15–74 Férfi	0.6369951	-0.02169521	0.50367000	0.16012430	-0.09601269
15–74 Nő	0.5488991	0.05460737	0.51508370	0.02655802	0.06359577
15–24 Összesen	0.6576190	-0.04391339	0.44643662	0.06880053	0.03713867
15–24 Férfi	0.6420195	0.02402952	0.46096777	0.04061063	0.05390537
15–24 Nő	0.6929328	-0.07264049	0.37626151	0.08615574	0.05445237
65–74 Összesen	0.5526888	0.0820494			
65–74 Férfi	0.5697033	0.02318538			
65–74 Nő	0.5874505	0.0678526			

1. táblázat. Az autoregresszív folyamat együtthatói

4 Segédsorok előállítása

4.1 A foglalkoztatottak segédsorainak becslése

A foglalkoztatottságra vonatkozóan a járulékbevallásból képzett foglalkoztatási mutató bizonyult alkalmas indikátornak. Azonban a járulékbevallás adatai nem állnak időben rendelkezésre, így közvetlenül nem építhetők be a modellbe, az idősorukat előre kell jelezni. Az előrejelzéshez ARIMAX típusú klasszikus idősormodellezési technikát alkalmaztunk, melyhez felhasználtuk a változásbejelentőből származó alkalmazotti jogviszonyra (közfoglalkoztatottak nélkül) vonatkozó nyitó- és záróegyenlegeket. A vizsgálatok során arra jutottunk, hogy a járulékbevallásból képzett foglalkoztatotti mutató jól korrelál az adott hónap nyitó- és az eggyel korábbi hónap záróadatából képzett

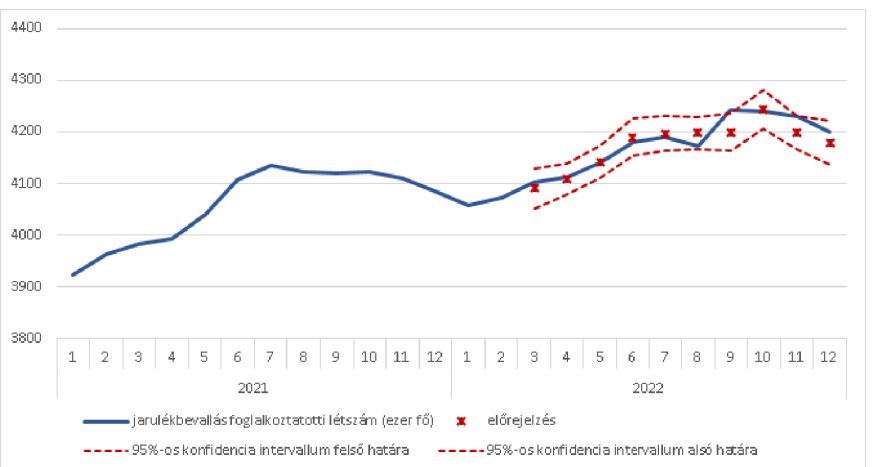
egyenleggel. Így ez az egyenleg ($\hat{x}_{\text{egy}}$) képezte a foglalkoztatottak összesen becslésére vonatkozó magyarázó változót.

A függő változót valamennyi esetben differenciáltuk, a szezonalitást szezonális dummy-k segítségével, a kiugró értékeket pedig additív outlierekkel kezeltük. A különböző bontásokra illesztett modelleket a *2. táblázatban* foglaljuk össze.

Függő változó	Magyarázó változó	AR tag rendje	MA tag rendje	Szezonális dummy	Outlier
d(fogl 15–74)	$\hat{x}_{\text{egy}}$	AR(6)	MA(2)	SEAS(01)	
				SEAS(06)	
d(fogl ffi 15–24)	d(fogl ffi 15–24(-12))	AR(1)	MA(2)		
d(fogl ffi 25–64)	d(fogl 15–74)			SEAS(06)	2021M03
				SEAS(07)	2021M11
				SEAS(09)	2021M12
d(fogl ffi 65–74)	d(fogl 15–74)			SEAS(01)	
				SEAS(07)	
				SEAS(09)	
d(fogl no 15–24)	d(fogl no 15–24(-12))	AR(1)	MA(2)		
d(fogl no 25–64)	d(fogl 15–74)	AR(3)		SEAS(06)	2021M03
				SEAS(07)	
				SEAS(09)	
d(fogl no 65–74)	d(fogl 15–74)			SEAS(01)	
				SEAS(07)	
				SEAS(09)	

2. táblázat. A foglalkoztatottak segédsorának ARIMAX modellje

A táblázatból látható, hogy a 15–74 éves foglalkoztatottakra vonatkozó becslés a többi mutató becslésében magyarázó változóként jelenik meg, így az első lépés az ő előrejelzésük. Ahhoz, hogy teljesüljenek az aggregálási feltételek, a (3) egyenlet segítségével módosítjuk a modellből származó eredményeket. A modellezést és az előrejelzést Eviews szoftverrel hajtjuk végre.



3. ábra. A 15–74 éves foglalkoztatottakra vonatkozó előrejelzése és konfidenciaintervalluma

4.2 A munkanélküliek segédsorainak kiigazítása

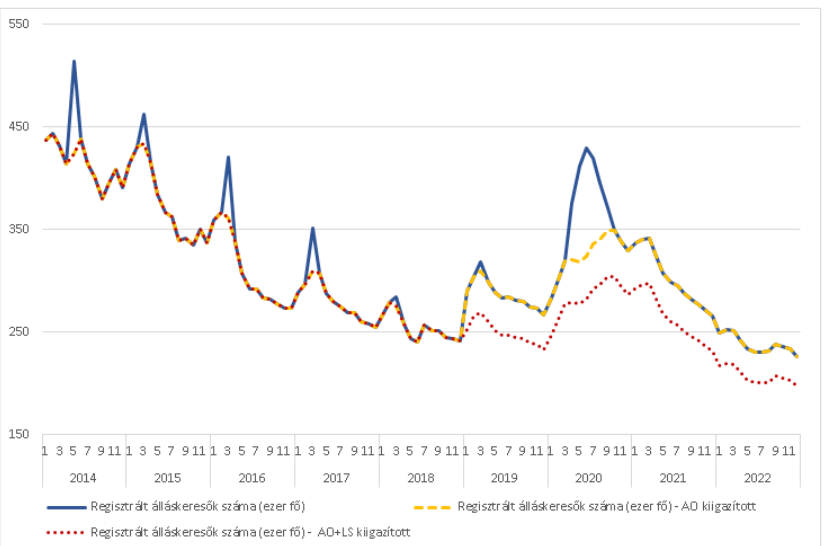
A munkanélküliek becslésének indikátoraként a Foglalkoztatási Hivataltól kapott regisztrált álláskeresők számát használjuk fel nemenkénti és korcsoportos bontásban. A rendelkezésünkre álló idősor megfelelően korrelál a munkanélküli mutatókkal, azonban közvetlen alkalmazáshoz túl zajosak, számos olyan kiugró értéket tartalmaznak, melynek hátterében nem társadalmi, hanem jogszabályi okok húzódnak. Ahhoz, hogy a modellben megfelelően alkalmazni tudjuk az idősort, az alábbi típusú és időpontú outlierek tekintetében kiigazítást hajtottunk végre. Ahol az AO az additív outliereket, azaz az egyszeri kiugró értékeket, míg az LS (level shift) a tartósan fennálló idősortöréseket ragadja meg.

Időszak	Outlier	Mnélkül 15–64	Mnélkül ffi	Mnélkül nő	Mnélkül 15–24	Mnélkül ffi 15–24	Mnélkül nő 15–24
2014M05	AO	×	×	×	×	×	×
2015M03	AO	×	×	×	×	×	×
2016M03	AO	×	×	×	×	×	×
2017M03	AO	×	×	×	×	×	×
2018M03	AO	×	×	×	×	×	×
2019M01	LS	×	×	×			
2019M03	AO	×	×	×	×	×	×

3. táblázat. Outlierek a munkanélküliek segédsorában

Mindemellett kezelnünk kellett a COVID-19 hatását is, mivel a regisztráltak eltérő dinamikát mutattak a nyers havi MEF becsléshez képest. Ebben az esetben a 2020 áprilisától szeptemberéig tartó időszakot igazítottuk ki.

Mivel a munkanélküliek segédsorai – a módosított időszakokat kivéve – teljesítik az aggregálási feltételeket, további módosításra nincs szükség.



4. ábra. A 15–74 éves munkanélküliek kiigazított idősora

5 A foglalkoztatottak dinamikáját leíró modell

Azért, hogy az eredményeinket az állapottérmodell-keretbe helyezhessük, meg kell vizsgálni, hogyan írhatjuk fel a (4) egyenletben szereplő változók dinamikáját. Feltételezésünk szerint az egyes részsokaságok dinamikája ugyanolyan szerkezetű, csak a paraméterekben térnek el egymástól. Ezért most általánosan írjuk le a modellt, amit a modellezett 9 részsokaság esetére alkalmazunk.

A hibatag dinamikáját a 3. fejezetben írtuk le. A foglalkoztatottak esetén ezt a (6) egyenlet tartalmazza. Mivel a negyedéves becsléseinket torzítatlannak gondoljuk, ezért arra számítunk, hogy a becslésünk közel kell, hogy essen a háromhavi mozgóátlaghoz. Ezt a becslésbe úgy építjük be, hogy a keresett hibatag szórását a nyers becslés háromhavi mozgóátlagától vett különbségének szórásával egyenlőnek feltételezzük. Ez azt is jelenti, hogy ezt a szórást előre meg tudjuk becsülni, és a becsléseket a *4. táblázatban* foglaltuk össze.

Sokaság	σ_r^2
15–74 Összesen	864229847.69
15–74 Férfi	396269402.86
15–74 Nő	364447229.06
15–24 Összesen	93474579.70
15–24 Férfi	37606644.32
15–24 Nő	45283803.46
65–74 Összesen	22339111.36
65–74 Férfi	11573527.85
65–74 Nő	7173276.57

4. táblázat. A hibatag szórása, foglalkoztatottak

A becsülni kívánt foglalkoztatotti létszám vizsgálatakor arra jutottunk, hogy az idősor jól együtt mozog a járulékbevallásból származó foglalkoztatottsági mutató idősorával (ξ_t). Az előzetes vizsgálat azt is megmutatta, hogy a két sor dinamikája közötti eltérésnek van egy szezonális komponense, ezért az idősorra a következő összefüggést írtuk fel:

$$\upsilon_t = \alpha_t + \beta_t \xi_t + s_t + \epsilon_{\upsilon,t}\,, \tag{8}$$

ahol

$$\begin{aligned} \alpha_t &= \alpha_{t-1} + \epsilon_{\alpha,t}, & \epsilon_{\alpha,t} &\sim \mathcal{N}(0, \sigma_\alpha^2) \\ \beta_t &= \beta_{t-1} + \epsilon_{\beta,t}, & \epsilon_{\beta,t} &\sim \mathcal{N}(0, \sigma_\beta^2) \\ & & \epsilon_{\upsilon,t} &\sim \mathcal{N}(0, \sigma_Y^2)\,. \end{aligned} \tag{9}$$

Az s_t változó pedig az idősorok dinamikája közötti szezonalitást hivatott megragadni. Az állapottérmodelleket feldolgozó szakirodalomban ismert (például van den Brakel and Krieg (2009)), hogyan kell szezonális komponenst a modellbe építeni. Ennek megfelelően ezt a komponenst újabb 11 ($s_{i,t}, s_{i,t}^*, s_{6,t}\,(i = 1, 2, 3, 4, 5)$) független, periodikusan ismétlődő sztochasztikus komponensre kell bontanunk. Formálisan ezt a következő egyenletekkel

írhatjuk le:

$$s_t = \sum_{i=1}^{6} s_{i,t} \tag{10}$$

és

$$\begin{aligned} s_{i,t} &= \cos\Big(\frac{i\pi}{6}\Big) s_{i,t-1} + \sin\Big(\frac{i\pi}{6}\Big) s^*_{i,t-1} + \epsilon_{s_i,t} \qquad (i = 1,2,3,4,5,6) \\ s^*_{i,t} &= \cos\Big(\frac{i\pi}{6}\Big) s^*_{i,t-1} - \sin\Big(\frac{i\pi}{6}\Big) s_{i,t-1} + \epsilon_{s^*_i,t} \qquad (i = 1,2,3,4,5)\,. \end{aligned} \tag{11}$$

Most már minden készen áll ahhoz, hogy felírjuk mindezt állapottérmodell-szerkezetben. Ehhez az kell, hogy összerakjuk az állapotegyenletet (1), amihez a (4) egyenletbe kell behelyettesítenünk a (8) és (10) egyenleteket. Így a következőt kapjuk:

$$\hat{v}_t = \alpha_t + \beta_t \xi_t + \sum_{i=1}^{6} s_{i,t} + r_t + \epsilon_{v,t}\,.$$

Ez azt jelenti, hogy az állapotegyenlet speciálisan egyetlen egyenletből áll.

Az egyenletben közvetlenül megjelenő ismeretlen idősorok a következők α_t, β_t, $s_{1,t}$, …, $s_{6,t}$, r_t. Nem szabad elfelejteni azonban, hogy a (6), (9) és (11) egyenletek adják a modell átmenetegyenleteit, és ebben szerepelnek még változók $(s^*_{1,t} \ldots s^*_{5,t}, r_{t-1}, \ldots, r_{t-5})$. Ez azt jelenti, hogy az egész rendszerünkben technikailag 19 idősort becsülünk. Az állapotegyenlet együtthatómátrixa ennek megfelelően a következő alakban írható fel:

$$F_t = \begin{array}{c} \begin{array}{cccccccccccc} \alpha_t & \beta_t & s_{1,t} & s^*_{1,t} & \ldots & s_{5,t} & s^*_{5,t} & s_{6,t} & r_t & r_{t-1} & \ldots & r_{t-5} \end{array} \\ \begin{pmatrix} 1 & \xi_t & 1 & 0 & \ldots & 1 & 0 & 1 & 1 & 0 & \ldots & 0 \end{pmatrix}. \end{array}$$

Az állapotegyenlethez tartozó V kovarianciamátrix pedig 1×1-es, $V = \sigma_v^2$, ami az egyenlet ismeretlen paramétere.

Az átmenetegyenlet mátrixa egy 19×19-es mátrix, ami nem változik időben, s amit érdemes kisebb blokkokra bontani a következő módon:

$$G_t = \begin{pmatrix} G_R & 0 & 0 \\ 0 & G_s & 0 \\ 0 & 0 & G_e \end{pmatrix}. \tag{12}$$

Az első blokk egy 2×2-es mátrix, ami lényegében a (9) egyenlet együtthatóit tartalmazza

$$G_R = \begin{array}{c} \begin{array}{cc} \alpha_{t-1} & \beta_{t-1} \end{array} \\ \begin{array}{c} \alpha_t \\ \beta_t \end{array} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \end{array} \tag{13}$$

A második blokk G_s mátrixa fogja tartalmazni a (11) egyenleteket. Ez egy 11×11-es mátrix, amit érdemes további blokkokra bontani, a következő módon:

$$G_s = \begin{pmatrix} G_{s,1} & 0 & \ldots & 0 \\ \vdots & \ddots & & \vdots \\ 0 & \ldots & G_{s,5} & 0 \\ 0 & \ldots & 0 & G_{s,6} \end{pmatrix}. \tag{14}$$

Az egyes blokkokból az első öt 2×2-es:

$$G_{s,i} = \begin{array}{c} \\ s_{i,t} \\ s^*_{i,t} \end{array} \begin{array}{c} \begin{array}{cc} s_{i,t-1} & s^*_{i,t-1} \end{array} \\ \begin{pmatrix} \cos\left(\frac{i\pi}{6}\right) & \sin\left(\frac{i\pi}{6}\right) \\ -\sin\left(\frac{i\pi}{6}\right) & \cos\left(\frac{i\pi}{6}\right) \end{pmatrix} \end{array} \qquad (i = 1, 2, 3, 4, 5)\,, \tag{15}$$

a hatodik pedig 1×1-es

$$G_{s,6} = -1\,. \tag{16}$$

Végezetül a G_e blokk egy 6×6-os mátrix, ami a hibatagot mint egy AR(6) folyamatot írja le. Itt lényegében egy egyváltozós, 6 késleltetéses folyamatot írtunk le, hatváltozós 1 késleltetéses folyamattal. Formálisan a mátrix ekkor a következő alakú:

$$G_e = \begin{array}{c} \\ r_t \\ r_{t-1} \\ r_{t-2} \\ r_{t-3} \\ r_{t-4} \\ r_{t-5} \end{array} \begin{array}{c} \begin{array}{cccccc} r_{t-1} & r_{t-2} & r_{t-3} & r_{t-4} & r_{t-5} & r_{t-6} \end{array} \\ \begin{pmatrix} 0 & 0 & \phi_1 & 0 & 0 & \phi_2 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix} \end{array}. \tag{17}$$

Az egyenletekhez tartozó kovarianciamátrix egy 19×19-es diagonális mátrix. A diagonálisban az átmenetegyenletek szórásai vannak, amiket függetlennek feltételezünk egymástól, ezért a főátlón kívül nem szerepel 0-tól különböző tag sehol. Az átmenetegyenletek közül a hibatag első egyenletének szórása a korábban leírtaknak megfelelően ismert, σ_e. A hibatag többi egyenlete technikai, determinisztikus, így a hozzájuk tartozó valószínűségi változó szórása 0. A többihez ismeretlen szórású valószínűségi változó tartozik, így a kovarianciamátrix a következő:

$$W = \begin{array}{c} \\ \operatorname{diag}(\\ \\ \end{array} \begin{array}{c} \begin{array}{ccccccccccc} \alpha_t & \beta_t & s_{1,t} & s^*_{1,t} & \dots & s_{5,t} & s^*_{5,t} & s_{6,t} & r_t & r_{t-1} & \dots \quad r_{t-5} \end{array} \\ \begin{array}{ccccccccccc} \sigma^2_\alpha & \sigma^2_\beta & \sigma^2_{s_1} & \sigma^2_{s^*_1} & \dots & \sigma^2_{s_5} & \sigma^2_{s^*_5} & \sigma^2_{s_6} & \sigma^2_r & 0 & \dots \quad 0 \end{array} \end{array})\,.$$

6 A munkanélküliek dinamikáját leíró modell

A munkanélküliek esetében a hibatag dinamikáját leíró egyenlet némileg különbözik a foglalkoztatottakétól, és ebben az esetben a (7) egyenletet alkalmazzuk. Hasonlóan a foglalkoztatottakhoz a negyedéves becsléseinket itt is torzítatlannak gondoljuk. Ezért ebben az esetben is arra számítunk, hogy a becslésünk közel kell essen a háromhavi mozgóátlaghoz. Ezt a becslésbe úgy építjük be, hogy a keresett hiba szórását a nyers becslés háromhavi mozgóátlagától való eltérésének szórásával egyenlőnek feltételezzük. Ez azt is jelenti, hogy ezt a szórást előre meg tudjuk becsülni, és a becsléseket az *5. táblázatban* foglaltuk össze.

Sokaság	σ_r^2
15–74 Összesen	245802606.50
15–74 Férfi	132260342.38
15–74 Nő	86746803.97
15–24 Összesen	30530741.74
15–24 Férfi	13690429.76
15–24 Nő	13865747.76

5. táblázat. A hibatag szórása, munkanélküliek

Ebben az esetben segédsornak a regisztrált munkanélküliek idősorának kigazított változatát (ξ_t) használjuk. Azonban ez az idősor csak 2019. év végéig volt teljesen alkalmas a havi becslések megfelelő előállítására. A COVID-19 járvány után azt tapasztaltuk, hogy a becslés nem éri el a kívánt stabilitást. Vizsgálataink során arra a következtetésre jutottunk, hogy a segédsor és az eredeti idősor kapcsolatában törés következik be 2020 májusában. Azért, hogy ezt a törést kezeljük, vezessük be a következő segédváltozót:

$$d_t = \begin{cases} 0, & \text{ha } t \leq 2020.05 \\ 1, & \text{ha } t > 2020.05\,. \end{cases} \tag{18}$$

Ezt a segédsort felhasználva az egyenletünk a következő formában írható fel:

$$v_t = \alpha_{1,t} + \alpha_{2,t} d_t + \beta_{1,t}\xi_t + \beta_{2,t} d_t \xi_t + s_t + \epsilon_{v,t}\,, \tag{19}$$

ahol a foglalkoztatottakhoz hasonlóan

$$\begin{aligned} \alpha_{1,t} &= \alpha_{1,t-1} + \epsilon_{\alpha,1,t}, & \epsilon_{\alpha,1,t} &\sim \mathcal{N}(0, \sigma^2_{\alpha,1}) \\ \alpha_{2,t} &= \alpha_{2,t-1} + \epsilon_{\alpha,2,t}, & \epsilon_{\alpha,2,t} &\sim \mathcal{N}(0, \sigma^2_{\alpha,2}) \\ \beta_{1,t} &= \beta_{1,t-1} + \epsilon_{\beta,1,t}, & \epsilon_{\beta,1,t} &\sim \mathcal{N}(0, \sigma^2_{\beta,1}) \\ \beta_{2,t} &= \beta_{2,t-1} + \epsilon_{\beta,2,t}, & \epsilon_{\beta,2,t} &\sim \mathcal{N}(0, \sigma^2_{\beta,2}) \\ & & \epsilon_{v,t} &\sim \mathcal{N}(0, \sigma^2_Y)\,. \end{aligned} \tag{20}$$

A szezonalitást megragadó változók ugyanúgy épülnek fel, mint a foglalkoztatottak esetében, vagyis, ahogy azt a (10) és a (11) egyenletekben leírtuk. Ebben az esetben is egy darab állapotegyenletünk lesz, amit úgy kapunk, hogy a (4) egyenletbe behelyettesítjük a (10) és (19) egyenletet, és azt kapjuk, hogy

$$\hat{v}_t = \alpha_{1,t} + \alpha_{2,t} d_t + \beta_{1,t}\xi_t + \beta_{2,t} d_t \xi_t + \sum_{i=1}^{6} s_{i,t} + r_t + \epsilon_{v,t}\,.$$

A (11), (7) és (20) egyenletek adják a rendszer átmenetegyenleteit. Ez azt jelenti, hogy ebben az esetben a feladatnak 24 ismeretlen sorozata van ($\alpha_{1,t}$, $\alpha_{2,t}$, $\beta_{1,t}$, $\beta_{2,t}$, $s_{1,t}$, $s^*_{1,t}$, …, $s_{5,t}$, $s^*_{5,t}$, $s_{6,t}$, r_t, …, r_{t-8}). Az állapotegyenlet együtthatóit tartalmazó együtthatómátrix a következő

$$\begin{array}{c} \\ F_t = \end{array} \begin{array}{cccccccccccccc} \alpha_{1,t} & \alpha_{2,t} & \beta_{1,t} & \beta_{2,t} & s_{1,t} & s^*_{1,t} & \dots & s_{5,t} & s^*_{5,t} & s_{6,t} & r_t & r_{t-1} & \dots & r_{t-8} \\ \multicolumn{14}{c}{\Big(} \\ 1 & d_t & \xi_t & d_t\xi_t & 1 & 0 & \dots & 1 & 0 & 1 & 1 & 0 & \dots & 0 \\ \multicolumn{14}{c}{\Big)\,.} \end{array}$$

Az állapotegyenlethez tartozó V kovarianciamátrix pedig 1×1-es, $V = \sigma_v^2$, ami az egyenlet ismeretlen paramétere.

Az átmenetegyenlet együtthatóit tartalmazó mátrix ezúttal 24×24-es, és struktúrája megegyezik azzal, amit láttunk a foglalkoztatottaknál, lásd (12) egyenlet. Ezúttal az első blokk egy 4×4-es mátrix:

$$G_R = \begin{array}{c} \\ \alpha_{1,t} \\ \alpha_{2,t} \\ \beta_{1,t} \\ \beta_{2,t} \end{array} \begin{array}{c} \begin{array}{cccc} \alpha_{1,t-1} & \alpha_{2,t-1} & \beta_{1,t-1} & \beta_{2,t-1} \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \end{array}. \tag{21}$$

A szezonalitást megragadó változókhoz tartozó blokkban nincs változás a foglalkoztatottakhoz képest, ezeket ugyanúgy kapjuk, mint ahogy azt a (14), (15) és (16) egyenletekben meghatároztuk. Az utolsó blokk a hibatag dinamikáját ragadja meg. A mátrix struktúrája itt is megegyezik a foglalkoztatottak esetében leírtakkal, csak nagyobb, ahogy azt a fejezet elején említettük.

$$A = \begin{array}{c} \\ r_t \\ r_{t-1} \\ r_{t-2} \\ r_{t-3} \\ r_{t-4} \\ r_{t-5} \\ r_{t-6} \\ r_{t-7} \\ r_{t-8} \end{array} \begin{array}{c} \begin{array}{ccccccccc} r_{t-1} & r_{t-2} & r_{t-3} & r_{t-4} & r_{t-5} & r_{t-6} & r_{t-7} & r_{t-8} & r_{t-9} \end{array} \\ \begin{pmatrix} 0 & 0 & \phi_1 & 0 & 0 & \phi_2 & 0 & 0 & \phi_3 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix} \end{array}.$$

Az átmenetegyenletekhez tartozó véletlen tag kovarianciamátrixa is 24×24-es, és ugyanúgy diagonális, mint a foglalkoztatottak esetén.

$$W = \operatorname{diag}\begin{array}{c} \begin{array}{cccccccccccc} \alpha_{1,t} & \alpha_{2,t} & \beta_{1,t} & \beta_{2,t} & s_{1,t} & s^*_{1,t} & \dots & s_{6,t} & r_t & r_{t-1} & \dots & r_{t-8} \end{array} \\ (\begin{array}{cccccccccccc} \sigma^2_{\alpha,1} & \sigma^2_{\alpha,2} & \sigma^2_{\beta,1} & \sigma^2_{\beta,2} & \sigma^2_{s_1} & \sigma^2_{s^*_1} & \dots & \sigma^2_{s_6} & \sigma^2_r & 0 & \dots & 0 \end{array}) \end{array}$$

A mátrixban szereplő első 11 szórást becsüli meg a modell, míg a σ_r^2 már ismert a korábban leírt módon.

7 A becslési folyamat, tapasztalatok és továbblépési lehetőségek

Ebben a fejezetben összegezzük, hogy a fent leírtak hogyan állnak össze egy egységes adatelőállítási folyamattá. Kiindulásképpen rendelkezésre állnak a nyers becslések a foglalkoztatottak és a munkanélküliek számáról egy adott hónapban, valamennyi korosztályra nemi bontásban.

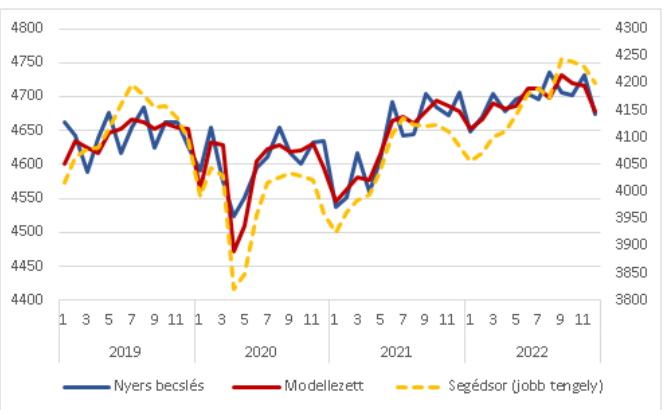
Elérhetőek továbbá a regisztrált munkanélküliek adatai ugyanerre a hónapra, illetve a járulékbevallásból származó foglalkoztatottak adatai egy hónappal korábbi időpontra. Első lépésként a modellekben használt segédsorokat

állítjuk elő. Ehhez a munkanélküliek esetén ellenőrizzük, hogy van-e olyan ugrás a regisztrált munkanélküliek idősorában, amit kezelni kell, és ha igen, ezt megtesszük. A foglalkoztatottak esetén pedig a járulékbevallásban szereplő foglalkoztatottakat jelezzük előre a változásbejelentő adatsorok segítségével. Ennek menetét írtuk le a 4. fejezetben.

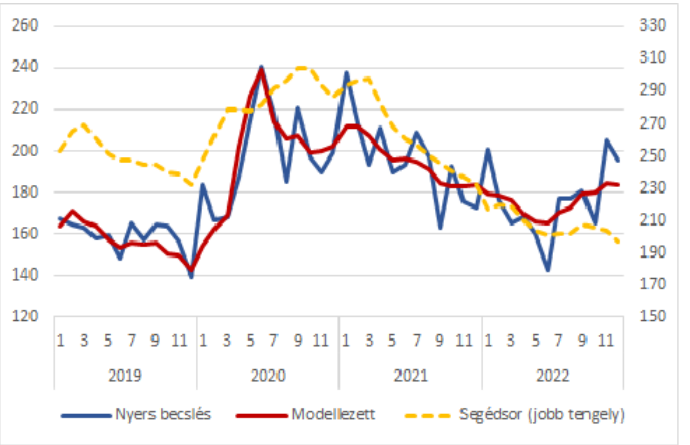
Amikor mind az alapadatok, mind a segédsorok rendelkezésre állnak, le tudjuk futtatni a foglalkoztatottakra az 5. fejezetben ismertetett modellt, az 1. fejezetben ismertetett sokaságokra. Hasonlóan lefuttatható a 6. fejezetben bemutatott modell a munkanélküliekre, minden 1. fejezetben leírt sokaságra. Az eljárás során becsült idősorokból valójában kizárólag az r_t-t használjuk fel, amit a (4) egyenletben definiáltunk. Ezt átrendezve kapjuk, hogy $v_t = \hat{v}_t - r_t$, és ez fogja adni a modellezett eredményeket. A 15–74 korosztályra vonatkozó eredményeket az *5. és 6. ábrákon* mutatjuk be. A bemutatott állapottérmodelleket bayesi eljárással becsüljük, melynek részletei megtalálhatóak a Petris (2009) hivatkozott irodalomban.

Ezen a ponton a modellezett korcsoportok esetén nem lesz igaz, hogy a férfiak és a nők száma kiadja a teljes sokaságra adott becslésünket. Ezért az összes esetben használjuk a (3) egyenletünket, hogy a nemek szerinti bontás konzisztens legyen. Végezetül indirekten előállítjuk a még el nem készült korcsoportokra a becslést.

Az így kapott eredményekből aztán meghatározható az aktívak száma, illetve ki lehet számolni a becsülni kívánt rátákat is. A kapott eredmények különböző ellenőrzési szempontok mentén szakmai validálás után kerülnek publikálásra.



5. ábra. A 15–74 korosztályban foglalkoztatottakra vonatkozó idősorok (ezer fő)



6. ábra. A 15–74 korosztályban munkanélküliekre vonatkozó idősorok (ezer fő)

A kapott eredmények megfelelnek a szakmai elvárásoknak. A múltbéli teszteléseken túl, az eljárással a KSH közel egy éve élesben is előállítja a havi munkaügyi adatokat. Azokat az Eurostat számára is megküldi, és az adatelőállítási folyamat minden esetben határidőre elkészült.

Az Eurostat mérhető objektív minőségi elvárásokat[5] is megfogalmazott a modellezett eredményekkel, egészen pontosan az ezekből számítható munkanélküliségi rátával kapcsolatban. Ezek alapvetően azt mérik, hogy mennyire sikerült megszüntetni a volatilitást az eredményekből. Ennek érdekében két olyan kritériumot fogalmaztak meg, amiknek az adatsornak meg kell felelnie. Mindkét kritérium a munkanélküliségi ráta egymást követő hónapjai közötti százalékos változást vizsgálja a tárgyidőszakot megelőző 36 hónapra vonatkozóan.

Az első elvárás az, hogy ezen adatsor, valamint annak egy hónappal való eltoltjának korrelációja -0,3 és 0,7 közé kell, hogy essen. A másik elvárás az, hogy legfeljebb az esetek 5 százalékában fordulhat elő az, hogy az idősor kétszer közvetlenül egymás után előjelet vált, és a havi munkanélküliségi ráta jelentős mértékben, 0.2 százalékponttal változik.

Mind a két elvárást 2020. januári tárgyidőszakig visszamenőleg ellenőriztük. Az első esetben elmondható, hogy a korreláció minden esetben a kívánt értékek között volt, egészen pontosan 0,12 és 0,56 között mozgott. A második kritériumról pedig azt, hogy egyik tárgyidőszakban sem történt olyan, hogy akár egyszer is előfordult volna, hogy az idősor kétszer közvetlenül egymás után előjelet vált, és a változás mértéke jelentős. Összességében ez alapján elmondható, hogy az Eurostat elvárásainak maximálisan eleget teszünk.

A pozitív eredmények mellett azért látunk lehetőséget a bemutatott eredmények jövőbeli fejlesztésére is. A modellezés során abból indultunk ki, hogy a közös elemet nem tartalmazó részminták hibái egymástól függetlenek. Ezt alapvetően azért tehetjük meg, mert ezt a mintavételezés biztosítja a

[5]https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32019R2241

számunkra. Ez valójában teljes mértékben csak a tervezett mintára igaz. A nem válaszolásból eredően azonban a megvalósult mintára már nem teljesen igaz ez a feltevés. A korábban hivatkozott holland tanulmányok ezen problémakör egy speciális részére, az úgynevezett panelkopásból eredő torzítás kiküszöbölésére mutatnak egy lehetséges megoldást. Ennek beépítése elképzelhető a hazai becslésbe is.

Egy másik feltételezés, amivel éltünk, hogy a modellezés hibái normális eloszlásúak. Ezen feltételezésünket Q-Q plot segítségével ellenőriztük. Azt kaptuk, hogy a hibák nulla várható értékűek, és az eloszlásuk szimmetrikusak ugyan, de a normális eloszláshoz képest vastagabb farkúak. Ez azt jelenti, hogy túl sok kiugró érték van ezen eloszlásokban ahhoz, hogy normálisnak tekinthessük őket. Ezen kiugró hibaértékeket a jövőben érdemes lesz alaposabban is vizsgálni és javítani.

Irodalom

1. Masanao Aoki (1990). State Space Modeling of Time Series. Springer.
2. Oksana Bollineni-Balabay, Jan van den Brakel, and Franz Palm (2016). State space time series modelling of the Dutch labour force survey: Model selection and mse estimation, – extended version. Discussion Paper Statistics Netherland, (13).
3. Peter J. Brockwell and Richard A. Davis (2002). Introduction to Time Series and Forcasting. Springer-Verlag.
4. Estela Bee Dagum Pierre A. Cholette (2006). Benchmarking, Temporal Distribution, and Reconciliation Methods for Time Series. New York: Springer Science+Business Media, LLC.
5. T. Deroyon, A. Montaut, and P.-A. Pionnier (2013). A monthly estimation method of ilo unemployment: a state-space framework. Documents de Travail de l'Insee – INSEE Working Papers, (G2013/01-F1301).
6. James D. Hamilton (1994). *Time Series Analysis.* Princton University Press.
7. Rudolph Emil Kalman (1960). A new approach to linear filtering and prediction problems. Transactions of the ASME – *Journal of Basic Engineering,* 82(Series D), 35–45.
8. Rudolph Emil Kalman and Richard S Bucy (1961). New results in linear filtering and prediction theory. *Journal of Basic Engineering,* 83(1), 95–108.
9. Etienne De Klerk, Cornelis Roos és Terlaky Tamás (2004). *Nem lineáris optimalizálás.* Aula.
10. KSH (2006). A munkaerő-felmérés módszertana 2006. KSH, Módszertani füzetek, 46.
11. KSH (2018). Quality improvement of the monthly unemployment rate. (07131. 2017.003-2017.593).
12. ONS (2019). Experimental model-based single-month estimates for the labour force survey: methods explained.
13. Giovanni Petris (2010). An R package for dynamic linear models. *Journal of Statistical Software,* 36(12), 1–16.
14. Patrizia Campagnoli, Sonia Petrone, and Giovanni Petris (2009). Dynamic Linear Models with R.

15. Danny Pfeffermann and Richard Tiller (2006). Small-area estimation with state: Space models subject to benchmark constraints. *Journal of the American Statistical Association,* 101(476).
16. Danny Pfeffermann, Moshe Feder, and David Signorelli (1998). Estimation of autocorrelations of survey errors with application to trend estimation in small areas. *Journal of Business Economic Statistics,* 16(3), 339–348.
17. Caterina Schiavoni, Siem Jan Koopman, Franz Palm, Stephan Smeekes, and Jan van den Brakel (2021). Time-varying state correlations in state space models and their estimation via indirect inference. Discussion Paper Statistics Netherland.
18. Richard B. Tiller and Danny Pfeffermann (2006). State-space modeling with correlated measurements with application to small area estimation under benchmark constraints. *Journal of the American Statistical Association,* 101 (476).
19. Jan van den Brakel and Sabine Krieg (2008). Estimation of the monthly unemployment rate through structural time series modelling in a rotating panel design. Discussion paper Statistics Netherland, (08003).
20. Jan van den Brakel and Sabine Krieg (2009). Estimation of the monthly unemployment rate through structural time series modelling in a rotating panel design. *Survey Methodology,* 35(2), 177–190.
21. Jan A. van den Brakel and Sabine Krieg (2015). Dealing with small sample size, rotation group bias and discontinuities in a rotating panel design. *Survey Methodology,* 41(2), 267–296.

METHODOLOGY OF MONTHLY LABOR MARKET DATA ESTIMATION AT THE HUNGARIAN CENTRAL STATISTICAL OFFICE

One important characteristic of any country's economy is the evolution of its labor market processes. Therefore, it is important for decision-makers to obtain picture of these processes as fast and accurate as possible. The labor force survey has a particularly important role where information can be derived from. Traditionally, estimates are made on a quarterly basis, but in our rapidly changing world, there is an increasing need for more frequent information. In this study, we discuss how the Hungarian Central Statistical Office (HCSO) prepares its monthly estimates using a state space model for the key labor market indicators.