

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

Cserhádi Ilona

PhD, az ECOSTAT GTI
osztályvezetője

E-mail: ilona.cserhati@ecostat.hu

Keresztély Tibor,

az ECOSTAT GTI tudományos
munkatársa

E-mail: tibor.keresztely@ecostat.hu

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

TÁRGYSZÓ:

Mikroszimulációs modell.
Makroökönómia.
Háztartásstatisztika.

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

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

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

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

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

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

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

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

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

¹ Ennek eredményeit *Cserháti et al.* [2007] munkája tartalmazza.

² Lásd *Cserháti–Péter–Varga* [2009].

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

1. Külső adatok imputálása

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

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

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

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

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

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

1.1. A személyijövedelemadó-bevallás adatainak imputálása

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

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

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

1.2. A tulajdonosi jövedelmek számítása

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

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

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

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

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

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

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

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

2. Átsúlyozás

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

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

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

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

2.1. Az eredeti súlyrendszer kialakítása

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

Magát a mintát a települések szerint alakítják ki, és ehhez meghatározzák az induló súlyokat (*Éltető–Mihályffy* [2002]). A válaszmegtagadások következtében a felvételben közreműködő háztartásokban élő személyek különféle szempont szerint vett megoszlása eltérhet az alapsokaságtól, ezért az általános statisztikai gyakorlatnak megfelelően az esetleges alul- és felülreprezentáltságot a súlyok további kalibrálásával küszöbölik ki (lásd még *Molnár* [2005]). A kalibrálás alapja a makroszintű adatokhoz való illesztés. A gyakorlatban ehhez a KSH demográfiai adatokat használ. Ilyenek a régió és a településtípus szerinti elhelyezkedésen kívül a nemek és életkorcsoportok szerinti megoszlás, az iskolai végzettség, a gazdasági aktivitás⁷ és a gyerekszám szerinti megoszlás. A kalibrálásnál az ún. Darroch–Ratcliff-féle iteratív skálázás módosított változatát alkalmazzák (*Darroch–Ratcliff* [1972]). Maga az eljárás egy adott valószínűség-eloszlás meghatározására irányul lineáris feltételek mellett, amellyel egyébként biztosítható a súlyok nemnegativitása, és az is, hogy az egy háztartásban élő személyek azonos súlyt kapjanak. Az eljárás a következő matematikai programozási feladat megoldását igényli:

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

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

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

$$l \leq w_j \leq u, \quad j = 1, \dots, n,$$

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

2.2. Az átsúlyozás algoritmus

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

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

Legyen $\mathbf{w}_k = \mathbf{w}$, $\mathbf{u} = \mathbf{1} \in \mathbf{R}^m$, ahol $\mathbf{1}$ az összegzővektor, valamint $h=1$, és képezünk egy \mathbf{u}_1 vektort, amelynek i . eleme az \mathbf{X} mátrix sorösszesen értéke:

$$\mathbf{u}_1 = \left(\sum_{j=1}^n x_{1j} \quad \cdots \quad \sum_{j=1}^n x_{mj} \right)^T \in \mathbf{R}^m,$$

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

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

Képezzük a \mathbf{c} vektort a következő módon:

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

Ha teljesül, hogy

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

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

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

definíció alapján, és jelölje ennek általános ele-

mét s_{ij} .

Legyen

$$\mathbf{u}_2 = \left(\sum_{j=1}^n s_{j1} \quad \cdots \quad \sum_{j=1}^n s_{jm} \right)^T \in \mathbf{R}^m,$$

azaz a vektor i -edik eleme az \mathbf{S} mátrix i . sorösszege.

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

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

3. A számítások eredményei

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

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

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

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

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

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

*

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

Irodalom

- DARROCH, J. N. – RATCLIFF, D. [1972]: Generalized Iterative Scaling for Log-Linear Models. *The Annals of Mathematical Statistics*. 43. évf. 5. sz. 1470–1480. old.
- CSERHÁTI I. ET AL. [2007]: A háztartások jövedelemalakulásának elemzése mikroszimulációs modellel. *A gazdaságelemzés módszerei*. II. sz. ECOSTAT – Központi Statisztikai Hivatal. Budapest.
- CSERHÁTI I. – PÉTER I. – VARGA ZS. [2009]: A lakosság jövedelmi rétegződésének tendenciái 2008–2009-ben. *Fejlesztés és Finanszírozás*. 3. sz. 70–78. old.

- ÉLTETŐ, Ö. – MIHÁLYFFY, L. [2002]: Household Surveys in Hungary. *Statistics in Transition*. 5. évf. 4. sz. 521–540. old.
- ATKINSON, A. B. ET AL. [1999]: *An Introduction to EUROMOD*. EUROMOD Working Paper No. EM0/99. <http://ideas.repec.org/p/ese/emodwp/em0-99.html> (Elérés dátuma: 2010. június 11.)
- MOLNÁR GY. [2005]: Az adatállomány és a rotációs panel. In: *Kapitány Zs.–Molnár Gy.–Virág I.: Háztartások a tudás- és munkapiacra*. KTI Könyvek. MTA Közgazdaságtudományi Intézet. Budapest. 141–147. old.
- ONYF (ORSZÁGOS NYUGDÍJBIZTOSÍTÓ FŐIGAZGATÓSÁG) [2010]: *Nyugdíjban, nyugdíjszerű ellátásban részesülők állománystatisztikai adatai*. http://www.onyf.hu/index.php?module=news&fname=onyf_left_menu_kiadvany&root=ONYF (Elérés dátuma: 2010. június 11.)
- ZAIDI, A. – RAKE, K. [2001]: *Dynamic Microsimulation Models: A Review and Some Lessons for SAGE*. ESRC SAGE Research Group's Discussion Paper No. 2. London School of Economics. London.

Summary

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