KÖZGAZDASÁGI SZEMLE, LXIV. ÉVF., 2017. NOVEMBER (1148–1159. o.)

VINCZE JÁNOS

Információ és tudás

A *big data* egyes hatásai a közgazdaságtanra

Az informatikai forradalom és az ezzel összefüggő *big data*-jelenség a tudományokat és a tudományos kutatást is megváltoztatja. Ez az írás néhány olyan tényezőre mutat rá, amelyek a közgazdaságtant érintik. A dezaggregáltabb és strukturálatlan adatok intenzív használatától összességében az várható, hogy az empirikus közgazdaságtan módszertana megváltozik, ami hatással lesz az elmélet és az empíria kapcsolatára is. Journal of Economic Literature (JEL) kód: A12, B41, C01.

Big data, adatelemzés és közgazdaságtan

Hal Variannek, a mikroökonómia egyik legnépszerűbb tankönyvszerzőjének Mikroökonómia középfokon címmel megjelent könyvéből Magyarországon is sok évfolyam közgazdászhallgatói tanultak már. Talán kevésbé ismert, hogy ő a Google vezető közgazdásza is. A 2014-ben a Journal of Economic Perspectivesben Big Data: New Tricks for Econometrics címmel megjelent cikkében a következőt állítja: „Nagyon gyümölcsöző együttműködés alakult ki informatikusok és statisztikusok között nagyjából az elmúlt évtizedben, arra számítok, hogy a jövőben az informatikusok és az ökonométerek közti kooperáció is nagyon termékeny lesz." (*Varian* [2014] 3. o.)

Az állítás második része egy rejtett kritikai megjegyzést tartalmaz, nevezetesen azt, hogy ez a bizonyos gyümölcsöző együttműködés eddig még nem nagyon érte el a közgazdaságtant. A cikk további részében ennek okait is taglalja a szerző, amellett hogy kifejti véleményét a jövőben várható kooperációról. 2014 óta eltelt néhány év, és a kooperáció, úgy tűnik, felgyorsult. Ebben a cikkben kifejtem azzal kapcsolatos véleményemet, hogy ez milyen irányban fogja befolyásolni a közgazdaságtan jövőjét, hangsúlyozva azt is, hogy mit nem várhatunk tőle.

Nem szeretnék feleslegesen definíciókkal bajlódni, és ezekkel az olvasót untatni. Tudjuk, hogy az informatika (hardver és szoftver) fejlődése következtében képesek vagyunk megőrizni olyan nagy méretű adathalmazokat, amelyeknek bájtokban

Vincze János a Budapesti Corvinus Egyetem egyetemi tanára és az MTA KRTK Közgazdaságtudományi Intézet tudományos tanácsadója.
A kézirat első változata 2017. szeptember 18-án érkezett szerkesztőségünkbe.
DOI: http://dx.doi.org/10.18414/KSZ.2017.11.1148

kifejezett nagysága (petabájt, terabájt stb.) legtöbbünknek elképzelhetetlen. A *big data* használata feltételez tartalomtól független hardvermegoldásokat (szuperszámítógépek, „felhők"), valamint újfajta adatbázisokat, illetve azokat kezelni képes általános szoftvereket (ezekről ad egy felsorolást *Varian* [2014] 4. és 5. o.). Ezek az eszközök elválaszthatatlan részei a *big data*-jelenségnek, de alapvetően informatikai *(computer science)* problémák. A *big data*-jelenségnek természetesen számos más oldala is van (például digitalizáció, sávszélesség-növekedés), amelyek kétségtelenül hatással lesznek a közgazdászok munkavégzésére, de várhatóan a közgazdaságtan szemléletmódját alapvetően nem befolyásolják.[1]

Az itt elmondottak középpontjában az empirikus közgazdaságtan lesz, amely megfigyelésekkel, adatok beszerzésével-létrehozásával és ezek feldolgozásával, illetve értelmezésével foglalkozik. Az adatfeldolgozás ma jelentős részben, ha nem is kizárólag, olyan statisztikai módszerek használatát jelenti, amelyet hagyományosan ökonometriának nevezünk, de nem azonos azzal. Itt elsősorban a *big data*-jelenség ökonometriára való hatásának vizsgálatáról lesz szó: milyen módon változnak majd meg azok a statisztikai-adatfeldolgozó eljárások, amelyek segítségével tartalmat próbálunk az adatokból kicsiholni. Varian cikkében az 5. oldal alján ér ehhez a problémához, és innentől kezdve a maradék 20 oldalon ezzel foglalkozik.

Több népszerű kifejezés is forgalomban van ezen a téren (statisztikai tanulás, gépi tanulás, adatbányászat). Például *Varian* [2014] a „gépi tanulás" kifejezést használja, miközben többek között egy olyan könyvre hivatkozik, amelynek címében a „statisztikai tanulás" és az „adatbányászat" kifejezés szerepel (lásd *Hastie és szerzőtársai* [2013]). Mindezeket induktív statisztikai megközelítésnek fogom nevezni, mivel a gépi tanulás kifejezést sokan a mesterséges intelligenciával asszociálják.[2] Biztonsággal merem állítani, hogy ha olyan kézikönyvet olvasunk, amelynek címében ezen terminusok valamelyike szerepel, akkor nyilvánvaló lesz, hogy más szemlélettel találkozunk, mint amit egy hagyományos statisztika- (matematikai statisztikai, bayesi statisztikai vagy ökonometriai) könyv tükröz. Van persze átfedés a fogalmak, a célok, az alkalmazott módszerek tekintetében, amelyek gyakran ugyanazt a problémát igyekeznek leírni, megoldani. De egy induktív statisztikai kézikönyvben számos fejezet jól megvan bármilyen valószínűségszámítási fogalom nélkül, a nagy számok törvénye még csak az indexben sem igen fordul elő, és lehet, hogy a *t*- és *F*-statisztikák is hiányoznak. Aki tanult hagyományos statisztikát vagy ökonometriát, az tudja, hogy ezek ott alapfogalmak. Ugyanakkor, ha valaki belemélyed ezekbe a könyvekbe, és hajlandó

[1] A *big data* szokásos definícióiban gyakran beszélnek négy (néha három vagy esetleg öt *V*-ről) (lásd ttps://en.wikipedia.org/wiki/Big_data). A következőkben elsősorban a nagyságnak *(Volume)* és esetleg a változatosságnak *(Variety)* lesz jelentősége, a sebesség *(Velocity)* és a megbízhatóság *(Veracity)* nem fog szerepet játszani.

[2] Felvethető, hogy a mesterséges intelligencia, ami részben „gépi tanuláson" alapul, megváltoztatja-e a közgazdaságtan problematikáját. A hagyományos közgazdaságtan *homo oeconomicusa* úgy jellemezhető, mint egy ágens egyszerű célokkal és szofisztikált problémamegoldó képességgel. Az utóbbi évtizedekben a pszichológia eredményeit adaptáló viselkedési közgazdaságtan hatására egyre inkább bonyolult célrendszerű, ám nem annyira szofisztikált egyénként tekintünk a döntéshozókra. Egy mesterségesen intelligens gazdasági ágenst viszont úgy *tervezhetünk,* hogy céljai legyenek egyszerűek, miközben problémamegoldó képessége sok tekintetben felülmúlhatja az emberét (lásd *Parkes–Wellmann* [2015]).

„történeti" utalásokat is elolvasni, akkor azt találja, hogy a leírt eljárások jelentős része korábbi, mint a *big data*-korszak, és ezekben a könyvekben számos olyan példát fog találni, amelyek nem dolgoznak különösen nagy adathalmazzal.

Miről van itt szó tulajdonképpen? Véleményem szerint arról, hogy az induktív statisztikai módszerek népszerűségének növekedése a *big data*-jelenség egyfajta mellékjelensége. A *big data*-jelenség mintegy kínálja magát ezeknek a megközelítéseknek, amelyek diadalaikat elsősorban nagyon nagy adathalmazok elemzésénél érték el. A szűkebb értelemben vett közgazdaságtanban azonban még nem igazán számolhatunk be egyértelmű sikerekről, ami részben talán annak is betudható, hogy a közgazdászok számára rendelkezésre álló adathalmazok nem annyira nagyok, mint mondjuk a biológusok vagy csillagászok számára rendelkezésre állók. De mivel az indukció (azaz tényekből való általánosítás) nemcsak nagyon nagy adathalmazokban vezethet sikerre, megszületett az igény a módszerek használatára.[3]

Felmerül azonban az a kérdés is, hogy mit várhatunk ezektől a módszerektől. Az elvárások kezelése sem lényegtelen, hiszen a túlzott várakozások csalódáshoz – és „a gyerek kiöntése a fürdővízzel együtt" jelenséghez – vezethetnek. A következőkben elsősorban arról lesz szó, hogy az induktív statisztikai módszerek alkalmazása (*big datával* vagy a nélkül) mit jelenthet a jövő közgazdaságtana szempontjából, egyáltalán nem foglalkozom a *big data* tisztán informatikai vonatkozásaival.

Mit hozhat a jövő?

Aggregálás és adattömörítés

Fizikusok, biológusok, kémikusok és csillagászok adathalmazainak mérete gyakran csak petabájtokban mérhető (1 petabájt $= 2^{15}$ bájt). Ekkora adathalmazzal közgazdászok nem szoktak találkozni. Ezért aztán a gazdasági adatok kezeléséhez a modern informatikai lehetőségeket nem kell a „határon" kihasználni, habár ezek lényegesen meggyorsíthatják a feldolgozás idejét. Ugyanakkor a közgazdasági adathalmazok potenciális méretei jóval nagyobbak a ma létezőknél, például az otthon elhelyezett online okosmérőknek köszönhetően. Ezért a korszerű adatfeldolgozó módszerek alkalmazása felkészíthet arra, hogy a jövőbeli igazán nagy adathalmazokat is tudjuk használni.

Az információelmélet egyik kulcsfogalma az adattömörítés (lásd például *MacKay* [2003] 2. fejezet), ami nagyon fontos szempont a modern adatelemzési módszereknél, és ahol a cél az, hogy a tömörített adatokból az eredeti adatok (elég jól) visszanyerhetők legyenek. Ez egészen más nézőpont, mint az, amely a hagyományos statisztikai hivatalok eljárásait irányítja. A hivatalos statisztikák is tömörítenek, sok adatból keveset állítanak elő. Ez a tömörítés (aggregáció) azonban előfeltevések szerint zajlik (ár- és mennyiségi indexek, nemzetiszámla-fogalmak), azt számolják ki, amit lényegesnek tartanak, de ezek az aggregátumok nem alkalmasak arra, hogy akár közelítően

[3] A legfontosabb módszerekről jó áttekintést ad közgazdászok számára *Varian* [2014] vagy *Einav–Levin* [2014].

visszanyerjük belőlük az alapadatokat. Ezt régen indokolhatták a technológia (az adattárolás fizikai és informatikai) korlátai, ma már azonban egyre kevésbé. Lehetőség van tehát arra, hogy statisztikai hivatalok által gyűjtött nyers adatokat szinte feldolgozatlanul is a felhasználók rendelkezésére bocsássanak. Miért lenne ez üdvös?

A hivatalos statisztikákat szinte minden közgazdász használja tudományos meggyőződésétől és közgazdasági filozófiájától függetlenül. De elgondolkoztak-e valaha azok, akik bírálják a „főárami" közgazdaságtant, hogy a hivatalos statisztikák milyen mértékben alapulnak főárami elméleti megfontolásokon? Ha végig tanulmányozzuk például a nemzeti számla készítéséhez használt kézikönyveket, és persze ismerjük a walrasi általános egyensúlyelméletet, akkor bizony azt találjuk, hogy a nemzeti számlákat walrasi elvek alapján igyekeznek létrehozni, és ahol ezeket a számlákat közvetlen módon nem lehet előállítani, ott igyekeznek walrasi kényszerzubbonyba húzni az adatokat.[4]

A gazdasági adatok persze nemcsak statisztikai hivatalokban jönnek létre, hanem vállalatoknál is. A vállalati kettős könyvelés sem más, mint egy aggregációs rendszer, amely a vállalatok életében keletkező mikroadatokból *lényegesne*k gondolt tömörített statisztikákat állít elő. A kérdés az, hogy kinek és miért lényeges? A kettős könyvelés középkori eredetű, a kereskedelmi partnerségek igényei hozták létre (több befektető által finanszírozott kereskedelmi vállalkozás), ahol a „vállalat" mint entitás elkülönül a befektetőktől, akik szeretnék tisztán látni, mennyi jövedelmet termel számukra a cég (lásd *Kam* [1990] 1. fejezet). A könyvvitel további története lényegében ezeknek a középkori elveknek a kiterjesztése, alkalmazása új körülményekre, illetve új szereplők igényeinek a kielégítésére szolgál. Például a modern könyvviteli szabályok valószínűleg mások lennének, ha nem lennének jövedelem- vagy vagyonadók, illetve befektetővédelmi megfontolások. A *big data*-módszerek potenciálisan forradalmasíthatják a könyvelést is, mint ahogyan a vállalatok marketingtevékenységét már forradalmasították is.[5]

Induktív statisztikai szemlélet a közgazdaságtanban

Egy 1980-as évekbeli ökonometria-kézikönyv egyik fejezete (*Griliches* [1986]) előtt a következő mottó állt: „Nagyapám soha nem evett fasírtot; az étteremben azért nem, mert nem tudta, mi van benne, otthon azért nem, mert tudta." Ez azt hiszem, találóan jellemezte számos kutató álláspontját, és részben magyarázhatja is az (akkori?) közgazdaságtan elméleti elfogultságát. Az informatikai forradalom előtt az adatok

[4] Csak egy részlet az *SNA* [2009] kézikönyvből: „Egy piaci rendszerben a különböző javak és szolgáltatások piaci árainak a termelés relatív költségeit és a vásárlóknak való relatív hasznosságát kell tükrözniük…" (296. o. 15. fejezet, B rész, 15.12. paragrafus.) Mi ez, ha nem színtiszta neoklasszikus közgazdaságtan? Másfelől a látszólag elméletmentes mérlegekről és reálaggregátumokról Joel Mokyr a tartós stagnálás *(Secular Stagnation)* hipotézise kapcsán a következő megjegyzést teszi: „A történet része az, hogy a közgazdászok hozzá vannak szokva ahhoz, hogy egy főre jutó GDP-ben és a belőle levezetett tényezőtermelékenységben gondolkodjanak. Ezeket a mértékeket egy acél-ésbúza gazdaságra dolgozták ki, nem pedig olyanra, ahol az információ és az adatok a legdinamikusabb szektor." (*Mokyr* [2014] 88. o.)

[5] Az alapadatok visszanyerésének persze vannak jogi korlátai, amivel itt nem foglalkozom.

gyűjtéséhez és feldolgozásához sok emberre volt szükség, a bonyolultabb aggregátumokat egyénileg nem lehetett létrehozni. Az aggregálás szükségszerű volt, viszont az aggregálás nem lehetséges előfeltevések nélkül. Ma a kutatóknak egyre inkább módjuk van saját maguknak elkészíteni a fasírtot. Kérdés, hogy képesek vagyunk-e rá. Vannak-e megfelelő elméleteink ahhoz, hogy kikeverjük magunknak a hozzávalókat, illetve ha nincsenek, akkor szükség van-e ezekre? A közgazdaságtant még ma is jól jellemzik a közgazdászviccek, arról például, hogy a lámpa alatt keressük a kulcsot, amelyet a sötét kertben veszítettünk el. Viszont az egyre több adat és a gyorsan fejlődő induktív statisztikai módszerek lehetővé teszik, hogy kisebb súlyt adjunk a meglehetősen gyenge lábon álló elméleteinknek, és képesek legyünk az adatokban összefüggéseket, mintákat felfedezni. Ez persze csak lehetőség, ha valakinek van fémdetektora, nem biztos, hogy talál aranyat a homokban.

A hagyományos ökonometriai megközelítésben nagy szerepet játszott az elméletek (hipotézisek) tesztelése. Nem mondható el, hogy ez a program hatalmas sikernek bizonyult, mindmáig nagyon széles tartományban mozognak a paraméterbecslések szinte minden fontos témakörben, és számos alternatív elmélet létezik, amelynek képviselői tudni vélik, hogy valamilyen teszt igazolta őket. Kérdés, hogy rendelkezünk-e valóban tesztelhető elméletekkel, és létezhetnek-e ilyenek egyáltalán? Komplex rendszerekkel foglalkozó természettudományok sokkal nagyobb hangsúlyt fektetnek az adatokban rejlő szabályosságok felismerésére, általánosításokra, mint elméletek tesztelésére. Nem biztos, hogy a közgazdaságtan számára nem volna célszerű egy ilyen irányváltás, és a sok adat és az új módszerek legalábbis megteremtik ennek az irányváltásnak a lehetőségét. A közgazdaságtanban gyakorlatilag soha nincs biztos *fenntartott hipotézisünk* (a klasszikus statisztika kiindulópontja), amelyet kísérletileg igazolva látnánk, és bármennyire is szeretnénk, *a priori* ismereteinket sem igazán tudjuk mindenki számára elfogadható *a priori* eloszlások formájában megfogalmazni, ami a bayesi statisztika kiindulópontja.[6] Mit is jelentenek pontosabban ezek az induktív módszerek, amelyekhez segítségért fordulhatunk?

Hogyan lehet jellemezni az induktív és a hagyományos statisztikai (ökonometriai) szemlélet különbségét? A kiindulópont azonos: vannak adataink valamilyen jelenségről, és a cél az, hogy találjunk egy olyan matematikai összefüggést, amely jól írja le az adatokat. Az induktív módszerek általában rugalmasabbak, az általuk számításba vett modellcsalád (azaz a matematikai összefüggések családja) bővebb, a becslésnél figyelembe vett *a priori* (értsd: a kérdéses adatokat nem használó) feltevések kevésbé korlátozók. Ennél talán fontosabb különbség az, hogy olyan matematikai összefüggést igyekeznek meghatározni, amely általánosítható, vagyis azt tapasztaljuk, hogy az összefüggés akkor is „érvényes", amikor olyan adatokra alkalmazzuk, amelyek nem vettek részt a becslési eljárásban. Tehát a leírás jóságának a legfontosabb kritériuma a „prediktív" képesség, a becslés és a lehetséges modellek közti választás ennek van alárendelve.[7]

[6] Dacára a bayesi statisztika térnyerésének a makroökonometriában, ami sokkal inkább a technikai megvalósíthatóság terén elért fejlődésnek tudható be, mint tartalmi megfontolásoknak vagy felhasználói igényeknek.

[7] Az induktív módszereket gyakran valóban predikcióra (előrejelzésre) használják. A predikció azonban lehet tartalmilag retrodikció is, a hangsúly „a becslésben nem részt vevő" kifejezésen van.

Mondhatnánk persze, hogy a predikció minden hagyományos statisztikai módszertannak is része. A klasszikus ökonometria azonban jóval nagyobb szerepet ad a modellek jóságának eldöntésénél azoknak a mutatóknak vagy teszteknek, amelyek lényegében a modell mintán belüli illeszkedését és bizonyos előfeltevéseinknek való megfelelését mérik. Egy másik lényeges különbség az, hogy az induktív statisztika nem „szégyelli" a visszacsatolásokat, a feltevések korrekcióját a becsült modellek predikciós képessége alapján. A klasszikus statisztikai alapelvekhez – amelyeket persze nem tartottak be a gyakorlatban – tartozott, hogy a lehetséges modellcsalád létrehozása legyen független az adott vizsgálatban használt adatoktól, ne legyen visszacsatolás, előre tervezzenek meg minden lépést.

Több, általánosan használt algoritmus létezik, amelyek matematikai leírását megtalálhatjuk például *Hastie és szerzőtársai* [2013]-ban, de néhányat közülük informálisan *Varian* [2014] is ismertet. Ezek a módszerek tisztán matematikai formájukat tekintve sokkal kevésbé egységesek, mint a hagyományos statisztikai eljárások, és állandóan újak is keletkeznek. Az adatelemzés általános filozófiája különbözteti meg az induktív statisztikát a hagyományos statisztikától, amely önmagukban a matematikai algoritmusokban nem látható.

Egy fontos célja az induktív statisztikának az adatokban rejlő minták felfedezése, a megfigyelések osztályozása. Természetesen közgazdászok is ismerik a klaszteranalízist, de túlzás nélkül állíthatjuk, hogy ez nem nagyon megbecsült módszer az ökonometriában. Konkrétan nem ismerek olyan ökonometria-tankönyvet, amely akár egy fejezetet is szentelne ennek a kérdéskörnek. Ezzel szemben az empirikus természettudományok egyik alapvető funkciója a jelenségek osztályozása. Az orvostudományban például létezik páciensszegmentáció, amely igyekszik a betegeket besorolni abból a célból, hogy mely gyógymód alkalmas egy bizonyos betegség kezelésére, és ehhez ma már induktív statisztikai módszereket is használnak (*Lemon és szerzőtársai* [2003]). Induktív módszerekkel elemeznek például ökológiai idősorokat, minták és rejtett folyamatok felfedezése céljából (lásd *De'ath–Fabricius* [2000]). Vissza lehet utalni a fémdetektorra: garancia nincs arra, hogy találjunk valamit, de ezek a módszerek bizonyították, hogy sok esetben működnek.

Egyébként az osztályozás jelen van a közgazdaságtanban is. Vannak ágazati vagy termékcsoportok szerinti besorolások, vagy gyakran alkotnak országcsoportokat például az egy főre jutó GDP alapján. A klasszifikáció gyakorlatilag minden téren szükséges része a gondolkodásunknak. Ezek az osztályozások azonban szinte sohasem adat vezérelte módon történnek, inkább a hétköznapi józan észt vagy valamilyen ellenőrizhetetlen előfeltevést tükröznek. Adat vezérelte osztályozásra egy példa *Durlauf–Johnson* [1995], amelyben a szerzők növekedési regressziókhoz használtak CART- (klasszifikációs és regressziósfa-) módszerrel „felfedezett" országcsoportokat. Sokat nem veszíthetünk azzal, ha kipróbálunk alternatív (sok adaton alapuló) osztályozási módokat is, feltéve, ha tudjuk, hogy mi a célja az osztályozásnak.[8]

[8] *Wu és szerzőtársai* [2008] egy összefoglalás a legnépszerűbb „adatbányász" algoritmusokról.

Egy példa – gazdasági potenciál és induktív módszerek

Tegyük fel, hogy a gazdaság alul- vagy túlfűtöttségének mérési problémáját akarjuk megoldani. Először is abból kell kiindulni, hogy pontosan mit keresünk. Fogadjuk el, hogy a túl- (alul-) fűtöttség definíciójának azt tekintjük, hogy a reálpiac hozzájárulása az inflációhoz pozitív (negatív). Az ennél pontosabb definíciók mindig valamilyen specifikus modellen alapulnak.

Szokás a túl- vagy alulfűtöttséget a GDP-réssel, azaz a GDP természetes szintjétől való eltéréssel mérni. Az egyik lehetséges eljárás az, hogy vesszük az aggregált GDP idősorát, majd valamilyen idősorelemzési technikával trendet számolunk, és a trendtől való eltérés méri a túl- vagy alulfűtöttséget (lásd például *Darvas–Vadas* [2003]). Ezt sokan túlságosan elmélet nélküli megoldásnak tartják, ezért használnak több-kevesebb modellezést (például termelési függvény becslését), de komplett modellekből is lehet potenciális (vagy természetes) GDP-t számolni (egy áttekintést ad *Dupasquier és szerzőtársai* [1999]). Ezek a módszerek már sokkal több információt használnak, viszont szinte biztosan olyan feltevéseken alapulnak, amelyek legalábbis gyanúsak, de semmiképpen sem tekinthetők bizonyítottnak. Hogyan lehetne sok információt felhasználó, ám relatíve elméletmentes becslést találni?

Vegyünk számos relevánsnak tűnő gazdasági változót, amelyek lehetnek elég dezaggregáltak is, és amelyekről azt gondoljuk, hogy bizonyos eséllyel szerepet játszanak az inflációs folyamatban. A vizsgálat célváltozója legyen valamilyen aggregált inflációs mérték, itt tehát megmaradunk egy hagyományos aggregátumnál. Ezután készítsünk egy „validált" előre jelző modellt valamilyen induktív módszerrel az inflációra úgy, hogy mindezeket a változókat használjuk, majd pedig úgy, hogy azokat, amelyeket a reálpiaci „egyensúly" szempontjából relevánsnak tartunk (például munkanélküliség, betöltetlen állások száma, különböző kapacitáskihasználási mutatók, készletek), kihagyjuk a modellből. Ezt a második modellt is validáljuk. Ezután a két modell előrejelzéseinek a különbségét tekinthetjük a gazdaság fűtöttségét jelző mutatónak egy adott időpontban.

Látszik, hogy ez az eljárás nagyon hasonló azokhoz a hagyományos statisztikai eljárásokhoz, ahol bizonyos változókat kihagyva becslünk meg modelleket, majd a két modellből vett előrejelzést összevetve mérjük meg a kihagyott változók hatását. Természetesen lehetne azon tépelődni, hogy a hatás kauzális vagy sem, de nem érdemes, mivel nem az a kérdésünk, hogy milyen beavatkozás vezet a fűtöttség megváltoztatásához. Itt még csak egy diagnózis felállításáról volt szó. A probléma (ha van) kezelése már más kérdés. Ne felejtsük el azt sem, hogy összefüggések kauzalitását sohasem tudjuk pusztán adatelemzés által eldönteni, mindig szükség van szubsztantív ismeretekre is, például ismernünk kell a mintavétel módját vagy a jelenségek közti időbeli relációkat.[9]

[9] Egy egyszerű példa: ha valaki megfigyeli a benzinárakat és a benzineladásokat, majd ezeket statisztikailag elemzi, akkor naiv módon azt hiheti, hogy a benzineladások növekedése (csökkenése) „okozza" az árak növekedését (csökkenését). Ha viszont *tudjuk,* hogy az árak változását napokkal előbb bejelentik, akkor levonhatjuk azt a következtetést, hogy a várt árak változásai „okozzák" a keresleti változásokat. Semmilyen statisztikai módszer (legyen klasszikus vagy induktív) nem ad *önmagában* alapot ehhez a konklúzióhoz.

Melyek a lehetséges előnyei ennek a procedúrának? Először is, úgy tűnik, számos olyan induktív módszer létezik, amelyekkel jobb előrejelzések adhatók, mint az ökonometriában használt módszerekkel. Ha egy konkrét esetben egy ilyet találunk (és persze miért ne vetnénk össze a modell teljesítményét valamilyen hagyományos modell teljesítményével), akkor ez önmagában már előny. Másodszor, a statisztikai tanulási technikáknak óriási előnye az, hogy megbirkóznak azzal a problémával is, amikor kevesebb megfigyelésünk van, mint változónk. Jó idősoros előrejelzések érdekében makroökonométerek kétségbeesetten igyekeznek minél hosszabb adatsorokat használni. Ez ugyanakkor érdekes módon éppen ellentétes a gyakorlati céllal modellező értékpapírpiaci elemzők azon szokásával, hogy a régebbi adatokat nem használják, tekintettel arra, hogy a folyamatok minden jel szerint az időben instabilak (nem stacionáriusak). Észszerűnek tűnik rövidebb idősorokat, ám sokkal több változót felhasználni az előrejelzésekhez, ami azonban nagyon gyorsan azzal jár, hogy a hagyományos módszerek működésképtelenek lesznek.

Szemléleti változások

Strukturált vagy strukturálatlan adatok • A közgazdászok által hagyományosan használt adatok előzetesen jelentősen strukturálódnak, egyáltalán nem „természetes" megfigyelésként kapjuk meg legtöbbjüket. A *big data*-forradalom egyik lényeges eleme, hogy képes – potenciálisan – strukturálatlan adatokkal is elbánni, amelyeknek egyik fontos fajtája a természetes nyelvi szöveg (lásd például *Einav–Levin* [2014]). Aki már készített kérdőívet, tudja, hogy milyen könnyebbséget jelentene, ha nem kellene arra törekedni, hogy a válaszadók, lehetőleg ugyanazt értve a kérdéseken, könnyen kódolható válaszokat adhassanak. Sőt erre a kódolásra gyakran magukat a válaszadókat kérik fel, amikor „Értékelje ötös skálán …" típusú kérdéseket tesznek fel. Tudjuk, hogy az ötös osztályzat is egészen mást jelent különböző iskolákban, ezért észszerű elkerülni az ilyen sommás szubjektív értékeléseket, de a könnyű kódolhatóság követelménye ezt mégis megköveteli. Ennek a kényszerzubbonynak az oldása jelentősen javíthatná az empirikus adatszerző munkát. Komoly időmegtakarítással járna, ha egy megfelelő programmal az eredetileg sok gigabájtnyi anyagot kutatói fogyasztásra alkalmas formában állíthatnánk elő. Talán nem túlzás azt állítani, hogy a közgazdászokban bizonyos előítélet alakult ki az interjúkkal szemben, a szövegek feldolgozásának képessége oldhatná ezt az ellenérzést.

ABM (ágensalapú modellezés) • A közgazdaságtanban az utóbbi évtizedekben kezd elterjedni az ágensalapú modellezés. Az ABM sok ágens viselkedését szimulálja nagyon sok perióduson keresztül. Így aztán egy ABM matematikai értelemben vett outputja nagyon nagy mennyiségű adat. Léteznek ABM-ek több millió ágenssel is (*Deissenberg és szerzőtársai* [2008]). Ráadásul az ABM-ek gyakran nem ergodikusak, azaz ahhoz, hogy következtetéseket lehessen levonni belőlük, nem elég, ha egy hosszú idősor áll rendelkezésre a statisztikai elemzéshez, ehhez sok hosszú szimuláció eredményeire van szükség. Ez nyilvánvalóan azt jelenti, hogy

a szimulációk adatainak tárolása, ezek elemzése igazi *big data*-probléma. Ezért aztán a *big data*-módszerek könnyebbé tehetik az ABM mint metodológia elterjedését a közgazdaságtanban.

Mit nem várhatunk a *big datától?*

Információ és tudás

A statisztikai tanulási módszereknek van egy nagyon csábító felhasználási területe, ami már most is népszerűvé teszi őket az üzleti életben. Ezt az általános értelemben arbitrázslehetőségek felfedezésének nevezhetjük. Ennek illusztrálására álljon itt az Epagogix esete (lásd *Ayres* [2007]). Az Epagogix egy döntéstámogató rendszer, amelynek olyan kérdéseket tesznek fel, hogy mekkora profitot lehet elérni egy adott forgatókönyvű, rendezőjű, bizonyos színészekkel megtervezett filmmel. De kereshető optimális forgatókönyv, rendező, szereposztás is, ha bizonyos egyéb ismérveket adottnak veszünk. A rendszert nem valamilyen elméletből építették fel, az tisztán múltbeli filmek adataiból tanulta meg, hogy milyen válaszokat adjon az új kérdésekre. Az Epagogixet természetesen üzleti célokra használják, állítólag sokat keresnek vele, így aztán magától értetődően titkos, nem lehet pontosan tudni, hogyan működik. Ez ellenkezik a tudományos kutatás alapelveivel, mivel a tudományos eredményeket akkor fogadjuk el, ha azok reprodukálhatók, hiszen ezzel válnak ellenőrizhetővé és közkinccsé. Sajnos az adatbányászmódszerek alkalmazásának egyik nagy hátulütője, hogy állítólagosan sikeres felhasználásaik egy részére nincs bizonyíték (üzleti titkok), lényegében el kell hinnünk a sikereket olyan embereknek, akiknek érdekük, hogy azokat elhitessék velünk.

Ezt a szempontot azonban most tegyük félre, és fogadjuk el, hogy az Epagogix valóban jól működő rendszer. Az egyik fontos tanulság, amelyet alkotói levontak belőle, az az „általánosítás", hogy fiatal, még nem befutott színészekkel gyakran nagyobb hasznot lehet elérni, mint sztárokkal, akiknek természetesen sztárgázsi is jár. Ezt az eredményt ajánlották, mint ami meglepő, és ami ellene megy a hagyományos hollywoodi sztárgázsi-filozófiának. Tegyük fel, hogy egy stúdió elfogadja és valóban alkalmazza, majd az Epagogix jóslatának megfelelő hatalmas profitra tesz szert. Ezután az összefüggés kitudódik (már ki is tudódott), és más stúdiók is átveszik a stratégiát. Mi történik? Egy idő után megszűnnek a sztárok, és a fontos szerepeket még nem befutott színészek alakítják viszonylag kis fizetésért, akik azonban előbb-utóbb észreveszik, hogy miután látszólag befutottak, nem kapnak szerepet, és foglalkoztatás nélküli sztárok lesznek. Ennek tudata vajon hogyan fogja ösztönözni a fiatal kezdő színészeket, akiknek vélhetően fontos motivációjuk a sztárrá válás és az ezzel járó hatalmas jövedelem, amiért érdemes befektetni kezdőként? Ha az Epagogixot ezek után újra futtatják, lehet, hogy azt találják, hogy a stratégia már nem is működik olyan jól, és valami más „arbitrázslehetőséget" kell felfedezni.

Az arbitrázs szó használatát az is indokolja, hogy számos tanulási algoritmusról állítják azt, hogy képes profitábilis (azaz a piacot verő) kereskedési-befektetési stratégiát

találni, és ezt több ízben bizonyítani is tudták (*Zhang–Zhou* [2004]). Azonban ezek a sikerek mindig is kérészéletűnek bizonyultak, az arbitrázslehetőségek általában bezárulnak (állítólag vannak kivételek). Egyszóval a sikeres predikció könnyen lehet, hogy csak nagyon rövid időre szól, és ezt némi „tradicionális elméleti" okfejtéssel meg is jósolhatjuk. Ettől még az átmeneti arbitrázs is arbitrázs, és lehet, hogy jól meg lehet belőle élni. Viszont első látásra kevés a hozzáadott értéke a közgazdaságtanhoz.

A helyzet nem nagyon különbözik az orvostudomány egyik nagy problémájától. Amikor egy új antibakteriális szert forgalomba hoznak, akkor ezt olyan ellenőrzött kísérletek előzték meg, amelyeknek a statisztikai elemzése plauzibilissé tette a szer hatékonyságát. Mint tudjuk, antibiotikumokra rezisztens baktériumtörzsek ki tudnak alakulni, és ugyanaz a statisztikai vizsgálat lehet, hogy néhány év múlva már nem mutatja ki a készítmény hatékonyságát. A baktériumok evolúciója megszünteti az arbitrázslehetőséget, és a predikció (általánosítás) érvénye átmenetinek bizonyul. Az orvostudomány ismeri a problémát, és dolgozik a megoldáson. Más esetekben viszont nem tudjuk feltétlenül megmagyarázni, hogy valamilyen predikció miért működött egy bizonyos ideig, és miért nem működik azután. Az ideiglenes általánosításoknak is megvan a potenciális haszna, felvetnek egy kérdést, amely megoldandó, de nem statisztikai eszközökkel. Ezek az eszközök lehetnek újfajta adatok utáni kutatások, kísérletek (egyre gyakrabban a közgazdaságtanban is), de szimulációk is, amelyek egyre több tudományban egészítik ki a hagyományos elméleti és empirikus-kísérleti módszereket.

Adatelemzés és adatgenerálás

A *big data,* úgy tűnik, óriási eredményeket ért el a természettudományokban. A hatalmas mennyiségű információ, amelyet az újonnan kialakított módszerekkel nemcsak tárolni, hanem kezelni is tudunk, minden jel szerint nagyszerű eredményeket hozott számos területen. Ilyen például a génkifejeződési *(gene expression)* adatokból való következtetések. Azonban vigyázzunk, az orvostudomány számára nem egyszerűen az adatok mennyisége változott, hanem azok minősége is, és ha lehet, még radikálisabban!

XIII. Lajos gyermekkorában került trónra, és mivel beteges gyerek volt, fontosnak tartották egészségi állapotának állandó monitorozását, ezért egy nagy hírű, tekintélyes orvost (Jean Héroard) rendeltek ki mellé, aki a következő mintegy 20 évben (saját haláláig) gyakorlatilag minden idejét a királlyal töltötte. Ott volt, amikor felkelt, étkezett, aludni tért, rendszeresen vizsgálta, diagnosztizálta. Előírta étrendjét, valamint beavatkozott orvosilag, ha jónak látta. Mindennap részletes feljegyzéseket készített az uralkodó állapotáról, és mivel mellékesen a királlyal kapcsolatos legapróbb történéseket is feljegyezte, több ezer oldalas, hitelesnek tartott feljegyzést hagyott maga után, amely fontos történelmi forrás (*Héroard–de Barthélemy* [1868]). A kérdés az, hogy mi volt az az információ, amely alapján Héroard diagnózisait és gyógymódjait meghatározta. Vizelet- és székletvizsgálat (szag, szín), pulzus, hőmérséklet és így tovább. Héroard rengeteg adatot gyűjtött, képes lehetett volna ennek alapján „előre jelző (diagnosztizáló) modellt" kialakítani, ha az ő korában ilyenek léteztek volna. És az általa feljegyzett adatok alapján utólag mi is megtehetnénk ezt. Nem

valószínű azonban, hogy megérné. Az orvostudomány néhány évszázad alatt sokat fejlődött, egy valamirevaló vizeletvizsgálat ma már sokkal több kémiai információt tartalmaz, mint amit a szag és a szín képesek leírni. Ez azonban még mindig nem az a pont, ahol a *big data* megjelenik, hanem például a génkifejeződési *(gene expression)* adatok, ahol viszont egy megfigyelés több százezer dimenziós vektorként fogható fel (lásd például *Hastie és szerzőtársai* [2013]), és amihez eljutni a biológia és fizika számos nagy felfedezésére volt szükség. Egyszóval, nagyon fontosak azok a fejlemények, amelyek mindenféle informatikai forradalomtól függetlenül olyan „állapottereket" eredményeznek a statisztikusok számára, amilyeneket nemcsak 400, hanem 50 évvel ezelőtt is elképzelhetetlennek tartottunk volna. Vagyis vigyázzunk, ha azt hisszük, hogy pusztán az adatok mennyisége és az új technikák elegendők a sikerhez. A közgazdaságtan automatizálása valószínűleg nem fog bekövetkezni. Fogalmi újításokra is szükség lesz, és az információs forradalom következtében ezeknél kevésbé lesz kötve a kezünk. Például a keresleti döntéseknél valós idejű fiziológiai adatok felhasználása elfelejtetheti velünk a (nem megfigyelhető) hasznossági függvény ósdi fogalmát.

Következtetések

A fenti írás arról szólt, hogy a *big data*-jelenség véleményem szerint milyen katalizáló hatással járhat a közgazdaságtanban. Először is, egy olyan új szemléletre ösztönözhet, amely nagyobb szerepet ad az empíriának az elméletek létrehozásában, és a tesztelésben a predikciós erő fontosságát erősíti. Másodszor, olyan eddig elhanyagolt módszertanokat is népszerűvé tehet, mint az interjúk vagy az ágensalapú szimulációk. Végül igyekeztem óvatosságra is inteni: nagyon hasznos az életben információval rendelkezni, de ez nem feltétlenül tudomány, vagyis átadható, mindenki által felhasználható, általánosítható tudás. A tudomány alapvetően gondolkodás útján fejlődik, amelynek eredményei az újfajta tények, és ezt valószínűleg semmilyen informatika nem fogja kiküszöbölni a belátható jövőben.[10]

Hivatkozások

AYRES, I. [2007]: Super crunchers: Why thinking-by-numbers is the new way to be smart. Bantam Books, New York.

DARVAS ZSOLT–VADAS GÁBOR [2003]: Univariate potential output estimations for Hungary. MNB Working Paper, No. 8. https://www.mnb.hu/letoltes/wp2003-8.pdf.

DE'ATH, G.–FABRICIUS, K. E. [2000]: Classification and regression trees: A powerful yet simple technique for ecological data analysis. Ecology, Vol. 81. No. 11. 3178–3192. o. https://doi.org/10.2307/177409.

DEISSENBERG, C.–VAN DER HOOG, S.–DAWID, H. [2008]: EURACE: A massively parallel agent-based model of the European economy. Applied Mathematics and Computation, Vol. 204. No. 2. 541–552. o. https://doi.org/10.1016/j.amc.2008.05.116.

[10] Habár a szingularitáshívők talán ezt gondolják (*Kurzweil* [2005]).

Dupasquier, C.–Guay, A.–St-Amant, P. [1999]: A survey of alternative methodologies for estimating potential output and the output gap. Journal of Macroeconomics, Vol. 21. No. 3. 577–595. o. https://doi.org/10.1016/s0164-0704(99)00117-2.

Durlauf, S. N.–Johnson, P. A. [1995]: Multiple regimes and cross-country growth behaviour. Journal of Applied Econometrics, Vol. 10. No. 4. 365–384. o. https://doi.org/10.1002/jae.3950100404.

Einav, L.–Levin, J. [2014]: The data revolution and economic analysis. Innovation Policy and the Economy, Vol. 14. No. 1. 1–24. o. https://doi.org/10.1086/674019.

Griliches, Z. [1986]: Economic data issues. Megjelent: *Heckman, J. J.–Leamer, E. E.* (szerk.): Handbook of econometrics. Vol. 3. 1465–1514. o. https://doi.org/10.1016/s1573-4412(86)03005-2.

Hastie, T.–Friedman, J. H.–Tibshirani, R. [2013]: The elements of statistical learning: Data mining, inference, and prediction, Springer, New York.

Héroard, J.–de Barthélemy, E. [1868]: Journal de Jean Héroard sur l'enfance et la jeunesse de Louis XIII (1601–1628) [microforme]. https://archive.org/details/cihm_03764.

Kam, V. [1990]: Accounting theory. Wiley, New York.

Kurzweil, R. [2005]: The singularity is near: When humans transcend biology. Penguin, London.

Lemon, S. C.–Roy, J.–Clark, M. A.–Friedmann, P. D.–Rakowski, W. [2003]: Classification and regression tree analysis in public health: Methodological review and comparison with logistic regression. Annals of Behavioral Medicine, Vol. 26. No. 3. 172–181. o. https://doi.org/10.1207/s15324796abm2603_02.

MacKay, D. J. [2003]: Information theory, inference and learning algorithms. Cambridge University Press, http://www.inference.org.uk/itprnn/book.pdf.

Mokyr, J. [2014]: Secular stagnation? Not in your life. Megjelent: *Teulings, C.–Baldwin, R.* (szerk.): Secular Stagnation: Facts, Causes and Cures. Vox, http://voxeu.org/article/secular-stagnation-not-your-life.

Parkes, D. C .–Wellman, M. P. [2015]: Economic Reasoning and Artificial Intelligence. Science, Vol. 349. No. 6245. 267–272. o. https://doi.org/10.1126/science.aaa8403.

SNA [2009]: System of National Accounts, 2008. EC–IMF–OECD–UN–WB, New York, https://unstats.un.org/unsd/nationalaccount/docs/SNA2008.pdf.

Varian, H. R. [2014]: Big Data: New Tricks for Econometrics. The Journal of Economic Perspectives, Vol. 28. No. 2. 3–28. o. https://doi.org/10.1257/jep.28.2.3.

Wu, X.–Kumar, V.–Quinlan, J. R.–Ghosh, J.–Yang, Q.–Motoda, H.–Zhou, Z. H. [2008]: Top 10 algorithms in data mining. Knowledge and Information Systems, Vol. 4. No. 1. 1–37. o. https://doi.org/10.1007/s10115-007-0114-2.

Zhang, D.–Zhou, L. [2004]: Discovering golden nuggets: Data mining in financial application. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), Vol. 34. No. 4. 513–522. o. https://doi.org/10.1109/tsmcc.2004.829279.