GRÁBICS ÁGNES – LOVÁSZY LÁSZLÓ GÁBOR

Az adatvezérelt állam és a világ megkettőződése – avagy a Big Data perspektívája a népszámlálás szolgálatában

Absztrakt

A statisztikai adatgyűjtések eredményeként előálló adatállományok és a különböző adminisztratív nyilvántartások, regiszterek előre meghatározott céllal, strukturáltan kerülnek előállításra, jól behatárolható és minőségű információk alapján. Ezzel szemben a Big Data meghatározás alá eső hatalmas adattömeg létrejöttét semmilyen cél nem határozza meg előre, sőt, az adatok elsődleges forrása akár ismeretlen is lehet a felhasználó számára, így rendszerezettségéről sem beszélhetünk. Ez az a két fontos jellemző, aminek hiánya miatt alapvetően különbözik a statisztikai és adminisztratív adatállományoktól. Mindemellett folyamatosan és egyre gyorsabb ütemben keletkeznek új adatok, exponenciálisan bővül az a halmaz, amelynek immár célirányos kiaknázása hatalmas lehetőség a hivatalos statisztika számára. A közadatok mellett a nagy tech- és más cégek adathalmazai, a magánadatok széles köre bevonható ebbe a felhasználási folyamatba. A Big Data célzott felhasználásával kapcsolatban még nagyon sok megoldandó kérdés és probléma van, amely feladatot generál mind az adatbányászok, mind a statisztikusok számára. Tanulmányunkban felvillantunk néhány kérdést ezzel kapcsolatban, kiindulva a népszámlálások történetéből és a lakossági adatgyűjtések bizonytalansági tényezőiből, megemlítve a hazai statisztikusok korábbi, a témához kapcsolódóan igen előremutató módszertani munkáit is. Összességében úgy látjuk, hogy mielőtt a Big Data beállhat a népszámlálások szolgálatába, rengeteg módszertani és technikai fejlesztésnek kell még megtörténnie, amelyeket először a reprezentatív adatgyűjtésekben kell és érdemes meghonosítani. Az írás célja rámutatni arra, hogy a statisztikusi szemlélet és a Big Data „bányász" homlokegyenest eltérő munkamódszere még távol áll egymástól, ugyanakkor hosszabb távon létre fog jönni a világ „megkettőződése", hiszen immáron a valóság egyre nagyobb részének lesz digitális megfelelője, ami már mindkét szakmát arra fogja kényszeríteni, hogy ne egymás ellenében, hanem egymást értve, mi több: egyesülve jöjjön létre egy egészen új szakma, amit – Isaac Asimov sci-fi-író életművének is tisztelegve – nevezhetünk akár kiberstatisztikusnak is.

Kulcsszavak: népszámlálás, adatbázis, hibaforrás, regiszterek, Big Data, módszertani fordulat, adatbányászat, sci-fi, kiberstatisztikus

Abstract

The datasets made from statistical data collections and the various administrative registers are produced from a predefined purpose in a structured way, based on well-defined and quality information.

In contrast, the huge amount of data under the definition of Big Data isn't created for a specific purpose, in fact the primary source of the data may even be unknown to the user, so it isn't systematic. These two important characteristics make it fundamentally different from statistical and administrative data sets. Nevertheless, new data are constantly being generated at an ever faster pace, and the data set is expanding exponentially, and the potential for targeted exploitation of this data set is now overwhelming for official statistics.

In addition to public data, a wide range of data sets from big tech and other companies as well as private data can be included in this exploitation process. There are still many questions and problems to be solved added to the targeted use of Big Data, that generates a challenge for both data miners and statisticians. In our study we outlined some of the issues regarding this, starting from the history of censuses and the uncertainty factors in population data collections and also mentioning previous methodological work by Hungarian statisticians that had been very forward-looking on the subject. In conclusion, we believe that before Big Data can be used in censuses, a lot of methodological and technical developments has to be made, which should be implemented in representative data collections first.

The aim of this paper is to point out that both the statisticians and the "Big Data miners" diametrically different working methods, approaches are still far apart, but in the longer term there will be a "duplication" of the world, as more and more of reality will have a digital counterpart, which will force both professions to work not against each other but understand each other, and even more: a whole new profession which – in homage to the work of science fiction writer Isaac Asimov – could be called cyberstatistics can appear.

Key words: census, database, source of error, registers, Big Data, methodological turn, data mining, science fiction, cyberstatistics

Bevezetés

A 21. század elején a modern információtechnológiák új generációja már nemcsak a tudományt, az űr- és hadiipart, a médiát, a szórakoztatóipart, hanem a kormányzati infrastruktúrákat és kormányzóképességet is alapvetően befolyásolja. Az Egyesült Államok és azon belül a Szilícium-völgyben működő óriásvállalatok, így különösen a nagy techcégek – GAFA (Google, Amazon, Facebook, Apple) – élen járnak nemcsak a globális IT ipar, hanem az innovációk és azok társadalmasítása terén is, amire már a kormányzati politikának is reagálnia kell.

Julia Lane, a New York Egyetem közgazdászprofesszora tavaly nyáron az MIT kiadójának gondozásában megjelent, az adatok demokratizálásáról szóló művében (*Democratizing Our Data: A Manifesto*)[1] kifejti, hogy az Egyesült Államok népszámlálási adatfelvételi módszerei és intézményei elavultak, széttagoltak és hibás adatokat szolgáltatnak, ráadásul valójában nem a legfontosabb társadalmi folyamatokat mérik, ugyanis az USA jelenleg képtelen a saját népességének változásait, a gazdasági és társadalmi folyamatokat naprakészen nyomon követni. Lane szerint az USA nem rendelkezik pontos adatokkal – többek közt – a saját GDP-jét, vagy akár az országos munkanélküliségi statisztikákat illetően, nem beszélve a tagállamok költségvetéseinek nyomon követéséről. Ez Lane szerint már a demokratikus, átlátható és szakszerű kormányzati döntéshozatalt veszélyezteti. A professzornő azt javasolja, hogy szükség lenne egy tízéves nemzeti adatfelvételi kutatási programra, valamint egy szövetségi szintű adatgyűjtésért és kezelésért felelős intézmény felállítására. A magyar Központi Statisztikai Hivatalhoz hasonló intézmény felállításának időszerűségét nem lehet elvitatni, Lane azonban egy érdekes elemmel toldotta meg javaslatát: szerinte az amerikai kongresszusnak létre kellene hoznia egy, a közösségi adatokért felelős nemzeti intézményt is.

Ami az Egyesült Államokban még csupán igényként merült föl, az Magyarországon már megvalósulni látszik. A kormány a Digitális Jólét Programhoz és a Nemzeti Digitalizációs Stratégia (NDS) 2021–2030 alapelveihez kapcsolódóan jogszabálytervezetet dolgozott ki a nemzeti adatvagyon mint gazdasági jószág minél szélesebb körű felhasználására, szorosan kapcsolódva az Európai Unió vonatkozó jogszabályaihoz és irányelveihez, melyek a személyes adatok védelméről, a közszféra információinak további felhasználásáról szólnak.[2] Felelős nemzeti intézményként létrehozta a Nemzeti Adatvagyon Ügynökséget,[3] melynek legfőbb feladata a nem-

1 Lane 2020.

2 Az Európai Parlament és a Tanács (EU) 2016/679 rendelete a természetes személyeknek a személyes adatok kezelése tekintetében történő védelméről és az ilyen adatok szabad áramlásáról, valamint a 95/46/EK rendelet hatályon kívül helyezéséről; az Európai Parlament és a Tanács 2003/98/EK irányelve (2003. november 17.) a közszféra információinak további felhasználásáról; az Európai Parlament és a Tanács (EU) 2019/1024 irányelve (2019. június 20.) a nyílt hozzáférésű adatokról és a közszféra információinak további felhasználásáról; az Európai Parlament és a Tanács (EU) 2018/1807 rendelete a nem személyes adatok Európai Unióban való szabad áramlásának keretéről.

3 A kormány 440/2020. (IX. 23.) rendelete a Digitális Jólét Program végrehajtásával összefüg-

zeti közadatportál üzemeltetése, a nemzeti közadatkataszter létrehozása és gondozása. A magyar kormány ezzel a lépéssel az adatvagyon-gazdálkodás terén a világ élvonalába kerülhet, megteremtve annak lehetőségét, hogy elinduljon Magyarország adatgazdasága.

Egy hónappal Lane könyvének megjelenése után, 2020 októberében a Marion Fourcade – Jeffrey Gordon szerzőpáros, a Berkeley, valamint a Yale kutatói a *Journal of Law and Political Economy*ben megjelent tanulmányukban (*Learning Like a State: Statecraft in the Digital Age*)[4] az „adatvezérelt állam" modelljét ismertetve arra mutattak rá, hogy a közeljövőben teljesen másképp fog működni az az állam, amely az adatbázisokra építi kormányzását. A tudósok szerint az adatvezérelt állam a mai kormányzatoknál kevésbé lesz politikailag elszámoltatható, illetve a stratégiai tervezés is teljesen átalakul, hiszen gyakorlatilag nem lesznek hosszú távú akciótervei, hanem lényegében állandó válságkezelésre fog fókuszálni. A stratégiai célhoz vezető előre eltervezett lépések kerülnek folyamatosan átértékelésre és változtatásra. Az adatvezérelt államban a magánszereplők és nem ritkán külföldi gazdasági szervezetek által birtokolt adatokért fokozódó verseny folyik majd, hiszen az adatokra fog épülni egy adott kormányzat sikere. Ez pedig azt jelenti a kutatók szerint, hogy az államnak – és így a hivatalos statisztikai szervezeteknek is – meg kell tanulnia a polgárok szemével látni a problémákat, és számukra is értelmezhetővé tenni az adatokat.

A társadalom számára a legegyszerűbben elfogadható adatok a népességre vonatkozó számok. Ezekhez nem kell különösebb magyarázat. Hányan vagyunk, milyen korúak, hol élünk? És ezek azok az adatok, amelyek már régóta rendelkezésünkre állnak.

gő egyes feladatokról, valamint a Kormányzati Informatikai Fejlesztési Ügynökségről szóló 268/2010. (XII. 3.) korm.-rendelet módosításáról szóló 127/2017. (VI. 8.) korm.-rendelet módosításáról.

4 Fourcade–Gordon 2020.

Röviden a népszámlálásokról

Az feltehetően sokak előtt ismert, hogy a népszámlálások több ezer éves múltra tekintenek vissza. Az első ilyen jellegű adatgyűjtések még nem feleltek meg teljesen a ma ismert összeírási követelményeknek, a teljeskörűségnek. Többnyire adózási, katonai célokat szolgáltak, így például a nőket, gyermekeket nem írták össze, ugyanakkor sokszor kiterjedtek a vagyontárgyak – ennek megfelelően a rabszolgák – számbavételére is.[5]

A Római Birodalom bukása után ezek a cenzusszerű adatgyűjtések több évszázadig szüneteltek. Az első, mai értelemben is népszámlálásnak tekinthető összeírásokra a 14–16. században került sor Európában, de ezek sem fedtek le teljes országokat, csupán néhány városállamot, fejedelemséget. A mai követelményeknek már többé-kevésbé megfelelő, teljes körűnek tekinthető adatgyűjtéshez a statisztika mint tudományág fejlődésére is szükség volt. A Központi Statisztikai Hivatal honlapján található információk szerint a világon az első, statisztikai céllal szervezett népszámlálást az észak-amerikai Québecben hajtották végre 1665-ben, Európában pedig 1749-ben Finnországban. Ezek már megfeleltek az egyidejűség, az azonos időpontra vonatkozás és a teljeskörűség feltételeinek.

Magyarországon az 1784–1787-ben végrehajtott népszámlálást[6] tekintjük az első, statisztikai céllal végrehajtott népesség-összeírásnak, ami kiterjedt a demográfiai, társadalmi jellemzőkre. Mindenkit össze kellett írni, aki az eszmei időpontban – az az időpont, amelyre az adatok vonatkoznak – életben volt, mégpedig azon az összeírási helyen, ahol az eszmei időpontban megfelelt az összeírás körébe tartozó személy kritériumának. Az adatgyűjtés tényleges egyidejűségét ekkor még nem sikerült biztosítani. Az adatok feldolgozásánál a katonai szempontok még fontos szerepet játszottak.

Az adatgyűjtés szervezete, végrehajtása lényegében megegyezett az ezt követő népszámlálásokéval. A feldolgozási módszerek, lehetőségek rengeteget fejlődtek, de magának a végrehajtásnak a módja alapjaiban nem változott az időközben eltelt több mint kétszáz évben. Olyannyira nem, hogy eleink még a „reklámra" is gondoltak, amire az első magyarországi

5 KSH 2006, 7–9.

6 KSH 1960, 5–7.

népszámlálásról szóló kötet szerint szükség is volt, mivel a nemesség ellenállással viseltetett az összeírással szemben.

> „Magától értetődik, hogy a honoratiorok, nemesek és mágnások közül senkinek sem kell húzódoznia attól, hogy családiával összeírják, és kastélyait megszámozzák, minthogy magát a Császári Palotát is megszámozták, és politikai tekintetben mindenki fogyasztónak számít."[7]

A népszámlálások sorában fordulópontot jelentett az 1930-as cenzus, hiszen ekkor használtak először statisztikai gépeket, ahogy az szerepel a megjelent kiadványban.[8] Ezek Powers-féle gépek voltak, automatikus billentyűs lyukasztógépek és számlálószerkezetes rendezőgépek. A leírás szerint az adatok közlése részletesebb volt, mint a nyugati államok által közölt adatok. A háború alatt végrehajtott 1941-es népszámlálás terveiben a korábbiaknál részletesebb adatgyűjtés szerepelt, de ezt a háborús körülmények felülírták.

Napjainkban a modern technika eszközei segítik az adatgyűjtést, a módszertanok nemzetközi összehangolása – ami lehetővé teszi az országok közötti összehasonlítást, és aminek igénye már a 19. század második felében felmerült – megtörtént, ezek egyike az ENSZ által 1998-ban megjelentetett „Népszámlálási elvek és ajánlások" (*Principles and Recommendations for Population and Housing Censuses,* United Nations, New York, 1998).

2011 óta kérdezőbiztosok közreműködése nélkül is lehetőség van (interneten keresztül) a válaszadásra, a részvételre törvény kötelezi az állampolgárokat. Az előkészítés, a kérdőívek kialakítása, a terepmunka lebonyolítása alapjaiban nem változott. Mindezek mellett az sem változott, hogy az adatok forrása továbbra is maga a kérdőív, amit a cenzusok során – a modern technika nyújtotta lehetőségek felhasználásával – még ma is többnyire kérdezőbiztosok töltenek ki, akik nagyban hozzájárulnak az összegyűjtött adatok pontosságához vagy pontatlanságához, kisebb-nagyobb mértékben befolyásolva az eredményeket. Persze nem szándékosan, elegendő egy általuk rosszul megfogalmazott mondat vagy egy olyan

7 KSH 1960, 439–440.

8 KSH 1932, 3–4.

válasz, amit nem tudnak értelmezni, az előírásoknak megfelelő részletezettséggel leírni. Egy teljes körű felvételnél elvileg nem jön szóba mintavételi hiba, mégis számtalan tényező alakítja a kapott adatok minőségét, még akkor is, ha tudjuk, hogy a hibák negatív hatásai a feldolgozás során mérsékelhetők.

A statisztikai adatok megbízhatóságáról

A nem mintavételi hiba nehezen mérhető, számszerűsítésére korábbi tapasztalatok, analógiák, illetve szakértői becslések állnak csak rendelkezésre.[9] Írásunk szempontjából releváns hibaforrás a kérdező, az adatgyűjtés módja és a kérdőív, továbbá – a törvényi kötelezettség ellenére – a nemválaszolás és maga a válaszadó. A hivatkozott értekezés szerint a kérdőív olyan hibákat hordozhat magában, mint a kérdés megfogalmazása, a kérdések hossza, magának a kérdőívnek a hossza, a kérdések sorrendje, a válaszkategóriák, nyitott és korlátozott válaszlehetőségek megadása. A felmérések lebonyolítása eltérő technikákat követel meg különböző adatgyűjtési módszerek esetén. A postai-, az elektronikus, a személyes adatgyűjtések, illetve a naplózáson alapuló felmérések mind-mind sajátos hibaforrásokat rejtenek magukban. A kérdezőbiztos által okozott válaszadási hibák a válaszadó kiválasztásából, a kérdezés módjából, a rögzítésből és a csalási hibákból erednek. Emellett a kérdezési hiba azokat az eseteket jelenti, amikor a kérdésfeltevés során követnek el hibákat, vagy amikor nem kérdeznek rá tovább valamire, holott több információra lenne szükség. Például a kérdezőbiztos a kérdéseket nem a kérdőív szóhasználatával teszi fel. A válaszadásból akkor származik hiba, ha a válaszadók pontatlan válaszokat adnak, vagy válaszaikat rosszul rögzítik, illetve azokat félreértelmezik. A válaszadás hibáját „elkövethetik" a kutatók, a kérdezőbiztosok vagy a válaszadók. A válaszadó által elkövetett hibák a képtelenségből és a válasz megtagadásából eredhetnek. A képtelenségből származó hibákat az jelenti, hogy a válaszadó nem tud pontos válaszokat adni. Ismeretlen a téma, fáradt, unatkozik, rosszul emlékszik, nem

9 Szilágyi 2011, 33.

érti a kérdést, vagy más miatt ad pontatlan feleletet. A válaszmegtagadási hibák abból erednek, hogy a válaszadó nem hajlandó pontos információt adni. A válaszadó szándékosan „félreválaszolhatja" a kérdéseket, mert társadalmilag elfogadható válaszokat akar adni, vagy el akarja kerülni, hogy megütközzenek a válaszán, zavarba jöjjön, vagy egyszerűen a kérdezőbiztosnak akar imponálni.

Az ilyen hibák aránya sokszor meglehetősen magas lehet, de előfordulási arányuk pontos meghatározása – éppen a statisztikai célú adatgyűjtésből fakadóan – gyakorlatilag lehetetlen. A statisztikai célú adatgyűjtés során a kérdezőbiztos nem kérhet írásos bizonyítékokat, az ő tevékenységéből fakadó hibákat pedig utólag, a feldolgozás során csupán részlegesen lehet kiküszöbölni. Éppen ezért fontos ezeknek a hibáknak a lehető legnagyobb mértékű megelőzése, részben a kérdőív kialakítása, részben a kérdezőbiztosok felkészítése során, továbbá az adatszolgáltatók minél sokrétűbb tájékoztatásával, érdekeltté tételével, ahogy erre már az első hazai népszámlálásnál is gondot fordítottak.

A nem mintavételi hibákon belül többnyire az összeírók által elkövetett hibák jelentik a legnagyobb hibaforrást, ugyanakkor ezeket a legnehezebb tetten érni, a szakirodalom sem tudja pontosan meghatározni ennek nagyságát. Ehhez társul továbbá, hogy egyre nehezebb – különösen a cenzusokhoz szükséges nagy tömegben – megbízható kérdezőbiztosokat találni, és ha képzésük mégoly alapos is, a felvétel iránti feltétlen elkötelezettségük, a minél jobb eredmény érdekében végzett lelkiismeretes és pontos munkájuk – tisztelet a kivételnek – sokszor várat magára. Napjaink technikai lehetőségeivel néhány sorozathiba kiküszöbölhető, de mindenre nincs megoldás.

A hitelességi problémáról és a Big Data jellemzőiről

A statisztikák korántsem írják le olyan megbízhatóan a szociológiai folyamatokat, mint azt hisszük, mindazonáltal minden korlátjuk ellenére fontos szerepük van a társadalom megismerésében. Az abban való hit, hogy a társadalmi változások, az emberek cselekedetei a statisztikai számítások segítségével könnyen megjósolhatók, veszélyes következményekkel

járhat. Ennek elkerülésére tesz javaslatot Hannah Fry egy 2019-ben megjelent tanulmányában.[10] Ebben leírja, hogy valószínűleg a Big Data korai megjelenése volt a francia igazságügyi minisztérium által a bűnügyi nyilvántartások 1825-ben elrendelt nemzeti gyűjteményének létrehozása. Ennek adatait elemezte Quetelet belga csillagász, és megállapította, hogy a bűncselekmények száma és összetétele évről évre ugyanannyi volt, minden ehhez kapcsolódó intézkedéstől függetlenül. Ennek nyomán kidolgozta elméletét arról, hogy az emberi élet számszerűsíthető és megjósolható.

Gondoljunk bele, ez minden felmerülő kétség ellenére ma is elfogadható állítás, és nem is kell másra hivatkoznunk, mint a nagy számok törvényére. És ez az, amit szem előtt kell tartanunk, hiszen a sokaságra igaz állításokat nem vonatkoztathatjuk az egyénekre. Átlagos egyén nem létezik, az átlag az egyének különbözőségének összegzése nyomán alakul ki. Fry szerint erre a Big Data sem ad jó választ, ígérete ellenére az egyén élete továbbra is kiszámíthatatlan marad. A statisztikai adatok bizonytalanságát a Big Data korszaka csak tovább súlyosbította. Minél több adat áll rendelkezésre, minél több a kereszthivatkozás, minél több összefüggést keresünk, annál könnyebb hamis következtetésekre jutni. Ezért válik egyre fontosabbá a tudomány ösztönzőinek megváltoztatása, egyre fontosabbá kell hogy váljanak a mások munkájának megismétlésére épülő tanulmányok. Fry azzal zárja írását, hogy a bizonytalanság világában a statisztika soha nem fogja eloszlatni a kétségeket, de mindenképpen jó alap a kezdéshez.

Ez a kezdet lehetne egy módszertani alapjaiban megváltoztatott statisztikai munka, ami a Big Data és a mesterséges intelligencia korában, a fentebb már említett nagy techcégek (GAFA) adatvagyona mellett előnyt tudna kovácsolni a módszertani fordulatból, kiküszöbölve azokat a hibákat, amelyek éppen a statisztika deduktív módszeréből, a frissen, célirányosan gyűjtött adatokból fakadnak. A Big Data nyújtotta lehetőségek mellett nem kell előre megtervezni az adatgyűjtést, nem kell adatokat gyűjteni, mivel a felhasználni kívánt adatok már rendelkezésre állnak. Ugyanakkor ennek ára van, hiszen megoldandó feldolgozási és elemzési problémák merülnek fel.

10 Fry 2019.

Statisztikai szakmai körökben már több évtizeddel ezelőtt felmerült a kérdezőbiztosokkal végrehajtott cenzusok, a hagyományosnak tekinthető terepmunka „leváltása", az adminisztratív adatforrások mind szélesebb körű bevonása. Ennek megvalósítására talán a legjobb példa a holland statisztikai hivatal, ahol már négy évtizede digitálisan szervezik a népszámlálásokat, az adminisztratív adatforrások lehető legszélesebb körét felhasználva. A cenzus végrehajtásához számlálóbiztosok helyett az önkormányzati statisztikákat és a társadalomstatisztikai adatbázisok rendszeréből származó adatokat használják fel, további népszámlálási adatgyűjtés nélkül, összevonva a rendelkezésre álló forrásokat, megspórolva ezzel sok millió eurót.[11] E módszer nem keverendő össze a Big Data használatával, az abból fakadó lehetőségek teljes kiaknázásával, de a regiszteralapú népszámlálás mindenképpen előrelépés annak irányába, összekapcsolva a rendelkezésre álló forrásokat egy virtuális népszámláláshoz.

Már ennek a módszernek a hazai megvalósítása is komoly, összehangolt előkészítést igényel, amit a gyakorlati megvalósítás érdekében már a legutóbbi népszámlálás befejezésekor el kellett volna kezdeni. Ez is egyfajta kiaknázása lehetne a Big Data nyújtotta lehetőségeknek, de ehhez el kell érni, hogy az adminisztratív adatforrások megbízhatósága magas szintű legyen. Ennek megvalósítását szolgálja a már említett nemzeti adatvagyonnal összefüggő kormányzati intézkedések sora.

A hivatalos statisztika – úgy tűnik – továbbra sem használja ki teljes mértékben az adatforradalom nyújtotta lehetőségeket,[12] kívülről nézve az erre irányuló szándék sem érhető tetten, bár Németh Zsolt szociológus elemzése megemlíti az Eurostat lépéseit e téren.[13]

Annak ellenére, hogy jelenleg még nincs teljeskörűen kiaknázva a lehetőségek tárháza, néhány évtizeddel ezelőtt a Big Data előfutárának is tekinthető ún. kisterületi becslések terén a hazai statisztika igencsak úttörő szerepet töltött be, és fontos tanulmányok születtek a témában, többek közt Marton Ádám[14] és Mihályffy László[15] tollából.

11 van der Sangen 2021.
12 Németh 2021, 18–19.
13 Németh 2021, 21.
14 Marton 2006.
15 Mihályffy 2018.

A KSH munkatársai egy, a *Statisztikai Szemlé*ben megjelent írásban javaslatot tettek a Big Data hazai alkalmazásának néhány lehetőségére.[16] Mindeközben a KSH maga is hozzájárul ahhoz, hogy egyre több adathoz férjünk hozzá, saját – a hagyományos módszertan szerint összegyűjtött és feldolgozott – adatvagyonát egyre szélesebb körben és egyre inkább felhasználóbarát módon osztja meg a mind nagyobb számú érdeklődővel.

Ugyanakkor egyre több kétség merül fel a hivatalos statisztikai adatok valóságtartalmával kapcsolatban, amely valóságra és tárgyilagosságra törekszik a hivatal immár több mint százötven éve. Egyre többen – köztük például Matolcsy György jegybankelnök is – gondolják úgy, hogy meg kellene reformálni a statisztikát,[17] hogy felgyorsult világunkban, a ránk zúduló információdömpingben fel tudja venni a versenyt a statisztikai szervezet annak érdekében, hogy gyorsan és ténylegesen a valóságot tükröző adatokat állítson elő, lehetőleg mind a társadalomra, mind a gazdaságra vonatkozóan. Ne kerüljön ellentmondásba a tapasztalás és a közölt adat, ugyanis napjainkban gyakorta láthatjuk, hogy a közösségi médiában egyre többen legyintenek a megjelenő inflációs adatokra, hiszen a saját érzetük, a pénztárcájuk mást mutat.[18] És itt sem arról van szó, amit nem szakmabeliek közül sokan és sokszor mondanak, hogy „hazudik a statisztika". Csupán annak vagyunk tanúi, hogy a közösségi média, az okos infokommunikációs eszközök világában az emberekre zúduló adatdömpingben a laikusok nem tudnak különbséget tenni a manipulatív adatok és a hivatalos statisztika között. Erre vonatkozóan közölt érdekes cikket 2020 őszén a nature.com.[19]

Ezzel párhuzamosan egyre nagyobb a megbízható adatok iránti igény, de az alapot jelentő adatok megismerése mind nagyobb akadályokba ütközik, különösen ott, ahol ezek megadása önkéntességen alapul. Az egyének egyre inkább elzárkóznak az adatszolgáltatástól, miközben a világhálón élik életük nagy részét. A hivatalos statisztikában az aktuális kormányzatot látják, a válaszmegtagadásban sokan egyfajta állampolgári engedetlenséget testesítenek meg.

16 Giczi–Szőke 2017.
17 Matolcsy et al. 2019.
18 Sebők 2020.
19 Cinelli et al. 2020.

A hivatalos statisztikának a hitelesség és a felhasználói bizalom megőrzése érdekében mielőbb választ kell adnia a jelen kihívásaira, természetesen az általa képviselt magas minőség és átlátható módszertan mellett. Erről a válaszadásról írt Walter J. Radermacher, amikor 2018-ban arról értekezett,[20] hogy a Big Data által kínált lehetőségek az adatok sokkal gyorsabb és gyakoribb terjesztése; a felhasználók speciális kéréseire sokkal relevánsabb válaszok, mivel a hagyományos statisztikai adat-előállítási folyamatok hiányosságai eltűnnek; a meglévő intézkedések finomítása, új mutatók kidolgozása és új kutatási lehetőségek megnyitása; az adatszolgáltatókra nehezedő teher jelentős csökkenése és a nemválaszolás arányának csökkenése. Végül, de nem utolsósorban a Big Data-hozzáférés jelentősen csökkentheti a statisztikák előállítási költségeit, az erőforrások és a kiadások drasztikus szűkítése idején. A Big Data jelenség azonban bizonyos számú kihívást is felvet. Ezek az adatok nem a szokásos gyakorlatnak megfelelően tervezett statisztikai előállítási folyamat eredményei. Nem felelnek meg a módszertanoknak, osztályozásoknak és definícióknak, ezért nehéz őket összehangolni, és integrálni a meglévő statisztikai struktúrákhoz. A statisztikusoknak továbbra is olyan módszerekbe és algoritmusokba kell befektetniük, amelyek javítják a felhasználók igényeihez igazodó statisztikai szolgáltatások adatminőségét.

Nyilvánvaló – ahogy az W. Radermacher írásából is kitűnik –, hogy a Big Data statisztikai célú alkalmazása előtt még mindig számos nehézség tornyosul. Ilyen az is, amiről Diego Kuonen statisztikus írt,[21] miszerint a legtöbb adatbányász és statisztikus továbbra is gúnyosan kritizálja egymást. Ez mindkét tudományágnak káros. Sajnos az antistatisztikus hozzáállás azt is megakadályozza, hogy az adatbányászat kiaknázza tényleges lehetőségeit – ugyanis az adatbányászat szintén tanulhat a statisztikákból. Az adatbányászat és a statisztika a közeljövőben óhatatlanul közeledni fog egymás felé, mert az adatbányászat statisztikai gondolkodás nélkül nem válik tudássá, a statisztikák pedig nem tudják sikeresen felhasználni a robusztus és összetett adathalmazokat adatbányászati ismeretek nélkül.

20 Radermacher 2018.

21 Kuonen 2004.

A jelenlegi kérdőíves adatgyűjtések felvázolt problémáira megoldást jelenthet a Big Data „beengedése" a hivatalos statisztikába, de ez paradigmaváltással kell hogy járjon a statisztikai módszertan területén, méghozzá a nem is oly távoli jövőben. Ugyanakkor látnunk kell, hogy a jelenlegi lehetőségek egyelőre nem alkalmasak a népszámlálással szemben támasztott követelmények, adatigények teljes körű kiváltására.

Következtetések és a jövő – avagy mit adott nekünk Asimov a statisztikában?

A Fourcade–Gordon alapján bemutatott adatvezérelt állam azt is jelenti, hogy a célokhoz vezető társadalompolitikai eszközök is átértékelődnek, hiszen az államnak folyamatosan mozgó célpontra kell lőnie, miközben tartania kell(ene) magát a stratégiai célokhoz is. Ez alapján az adatokhoz való jobb hozzáférés miatt megnő a nyomás a permanens politikai kommunikációra is, valamint – ahogy korában láttuk – az állami szerveknek meg kell tanulniuk a polgárok szemével látni a problémákat, és számukra is értelmezhetővé tenni az adatokat.

Kevesen tudják, de – Ian Stewart *Social physics*[22] című írása szerint – az egyik legnagyobb sci-fi-íróként számon tartott Isaac Asimov fantáziájának köszönhetően válik valóra a mesterséges intelligencia és a Big Data-elemzés tökéletesedésével az ún. „pszichohistória" tudománya. A Warwicki Egyetem professzor emeritusa szerint Asimov 1951-ben írt *Alapítvány* című mesterművének sokkal több a tudományos alapja, mint hinnénk, hiszen a fő karakter, Hari Seldon figuráját – a korábban már említett – 18. századi belga származású matematikus Adolphe Queteletről mintázta, aki matematikai módszerekkel elemezte az emberi viselkedést. Az időközben regényfolyammá bővülő *Alapítvány*-sorozatban megjelenő pszichohistória lényege, hogy a matematika és a pszichológia házasításával megjósolhatók a jövő eseményei, hasonlóan a szociofizikai kutatásokhoz, melyek kiindulópontja a népsűrűségre vonatkozó számadat. Stewart professzor szerint ez, majd a francia matematikus Pierre-Simon de Lap-

22 Stewart 2019.

lace munkássága is hozzájárult ahhoz, hogy az emberi viselkedési minták társadalmi szinten történő értelmezését is jobban megértsük. (De Laplace volt az első, aki összevetette a népszámlálást megelőző évben történt születések számát, valamint ennek az összegnek a teljes népességhez viszonyított arányát.)

Miért fontos a tudomány múltjára és a sci-fi-irodalom érdekes fantáziáira is figyelni? Asimov zsenialitásán túlmenően ma már látjuk, hogy a modern technológia lehetővé teszi a különböző rendszerek, eszközök, folyamatok permanens és virtuális modellezését. Ez gyakorlatilag azt jelenti, hogy a bővülő adatbázisokkal nemcsak egy-egy komplex folyamat válik egyre jobban nyomon követhetővé, érthetőbbé, hanem azok mintázatai is láthatóbbá válnak. A pekingi Beihang Egyetem Automatizálási és Villamosmérnöki Karának professzora és PhD-hallgatója, Fei Tao és Qinglin Qi közös cikkükben már 2019-ben úgy gondolták, hogy a világ mintegy meg fog kettőződni. A *Nature*-ben 2019 szeptemberében publikált írásukban (*Make more digital twins*)[23] a „digitális iker" lényegében létre fogja hozni egy fizikai eszköz, rendszer vagy termék virtualizált, informatikai mását, amely lehetővé teszi azok különböző célból történő szimulációját. Ez azt is fogja jelenteni, hogy a társadalomtudományok számára is rendelkezésre fognak állni olyan eszközök, amelyek eddig csak a természettudományok számára álltak rendelkezésre: laboratóriumi körülmények közötti kutatások lehetősége. Ehhez viszont az szükséges – ahogy a szerzők kifejtik –, hogy elengedhetetlenné váljon a legkülönbözőbb tudományágak és szakemberek átfogó, szoros együttműködése, méghozzá egy önálló virtuális és fizikai tér kialakításával, amelyben a szakértők képesek interdiszciplinárisan kommunikálni egymással, valamint megoszthatják tudásukat és fejlesztéseiket.

Talán ez is lehet az egyik – ha nem éppen a legfontosabb – feladata egy nemzeti statisztikai hivatalnak azzal, hogy új típusú tudással és képességekkel felvértezett szakemberekre, azaz „kiberstatisztikusok"-ra lehet szükség a jövőben.

23 Tao–Qi 2019.

Összefoglaló

Bármennyire is hihetetlennek tűnik, az adatvezérelt állam koncepciója a 20. század eleji sci-fi-irodalomban leírtak felé tart, s akár Asimov munkáiból is kisarjadhatott. Az adatvezérelt intelligens megoldások kikényszerítik a statisztikai tudomány, a statisztikai módszertan elkerülhetetlen megújulását. A modern információtechnológiák, a hatalmas adatbázisok már nem csupán a tudományt, hanem a kormányzóképességet is befolyásolják. Az információs versenyben a hivatalos statisztika sem maradhat le, saját, költséges adatgyűjtései mellett és/vagy helyett fel kell tudnia használni a nemzeti adatvagyon adminisztratív adatbázisait, és ki kell aknáznia a Big Data nyújtotta lehetőségeket is.

Mindezek segítségével egyre gyorsabban, az aktuális viszonyokat jól tükröző adatokat kell előállítania, egyesítve a statisztikai gondolkodás és az adatbányászati ismeretek teljes tárházát, kiküszöbölve azokat a hibákat, amelyek a statisztika eddigi deduktív módszeréből, a frissen, célirányosan gyűjtött adatokból fakadnak. Végső soron a különböző tudományágak szoros együttműködése nyomán egy önálló virtuális tér kialakítása – azaz a valóság megkettőződése – lehet a nemzeti statisztikai hivatal feladata, amelynek megvalósítása egy új foglalkozás és szakembergárda, a „kiberstatisztika" és „kiberstatisztikusok" megjelenése nélkül aligha képzelhető el.

Irodalom

II. József 1784: Leirat gróf Esterházy Ferenc magyar főkancellárnak. In *Az első magyarországi népszámlálás* (1784–1787) 1960: Budapest, Központi Statisztikai Hivatal, Függelék 439. https://library.hungaricana.hu/hu/view/NEDA_1784_elso_magyar/?pg=0&layout=s

Az 1930. évi népszámlálás 1932: Budapest, Központi Statisztikai Hivatal. https://library.hungaricana.hu/hu/view/NEDA_1930_01/?pg=2&layout=s

Az első magyarországi népszámlálás (1784–1787) 1960: Budapest, Központi Statisztikai Hivatal. https://library.hungaricana.hu/hu/view/NEDA_1784_elso_magyar/?pg=0&layout=s

Cinelli, Matteo – Quattrociocchi, Walter – Galeazzi, Alessandro – Valensise, Carlo Michele – Brugnoli, Emanuele – Schmidt, Ana Lucia – Zola, Paola – Zollo, Fabiana – Scala, Antonio 2020: *The COVID-19 social media infodemic.* https://www.nature.com/articles/s41598-020-73510-5#Sec6

Digitális Jólét Program 2020: Megalakult a Nemzeti Adatvagyon Ügynökség. https://digitalisjoletprogram.hu/hu/hirek/megalakult-a-nemzeti-adatvagyon-ugynokseg

Fourcade, Marion – Gordon, Jeffrey 2020: *Learning Like a State: Statecraft in the Digital Age.* https://escholarship.org/uc/item/3k16c24g

Fry, Hannah 2019: *What Statistics Can and Can't Tell Us About Ourselves.* https://www.newyorker.com/magazine/2019/09/09/what-statistics-can-and-cant-tell-us-about-ourselves?fbclid

Giczi, Johanna – Szőke, Katalin 2017: Hivatalos statisztika és a Big Data. *Statisztikai Szemle.* http://www.ksh.hu/statszemle_archive/2017/2017_05/2017_05_461.pdf

Kuonen, Diego 2004: *Data Mining and Statistics: What is the Connection?* http://www.statoo.info/en/publications/articleDM4TDAN.pdf

Lane, Julia 2020: *Democratizing Our Data: A Manifesto.* Cambridge, MIT Press. https://mitpress.mit.edu/books/democratizing-our-data

Marton Ádám: *Országos reprezentatív felvételek – (kis)területi becslések*. http://www.ksh.hu/docs/hun/xftp/terstat/2006/06/ts2006_06_02.pdf

Matolcsy, György – Nagy, Márton – Palotai, Dániel – Virág, Barnabás 2019: *Az infláció mibenléte – ideje a mérőrendszereinket újragondolni.* https://www.mnb.hu/letoltes/matolcsy-gyorgy-nagy-marton-palotai-daniel-virag-barnabas-az-inflacio-mibenlete.pdf

Mihályffy László: *Kisterületi becslések: rövid áttekintés a korszerű módszerekről*. http://www.ksh.hu/statszemle_archive/all/2018/2018_01/2018_01_091.pdf

Németh, Zsolt 2020: A hivatalos statisztika válsága az adatforradalomban. *Replika,* 18–19, 21. https://www.replika.hu/system/files/archivum/replika_117-118-08_nemeth.pdf

Radermacher, Walter 2018: *Official statistics in the era of big data opportunities and threats.* https://link.springer.com/article/10.1007/s41060-018-0124-z

Sebők, Miklós 2020: *Statisztikai adatok és elemzés a Big Data korában.* https://www.mstnet.hu/cikkek/_doku/06_Sebok_Miklos_200122.pdf

Statisztikai Módszertani Füzetek, 48 2006: Budapest, Központi Statisztikai Hivatal. http://www.ksh.hu/docs/hun/xftp/idoszaki/pdf/nepcsoport.pdf

Stewart, Ian 2019: *Social physics.* https://aeon.co/essays/our-behaviour-in-bulk-is-more-predictable-than-we-like-to-imagine?fbclid=IwAR0jrHDSMrI56AmbpX_PKHvNvm4SrXxag-NKgbSQVNq9P0_DOh-XxQA95XeU

Szilágyi, Roland 2011: *Mintavételen alapuló becslések hibáinak kezelése különös tekintettel a nemválaszolás okozta problémákra.* Miskolc, Miskolci Egyetem 33. http://midra.uni-miskolc.hu/document/12105/4124.pdf

Tao, Fei – Qi, Qinglin 2019: *Make more digital twins.* https://www.nature.com/articles/d41586-019-02849-1

van der Sangen, Miriam: *CBS in the starting blocks for the 2021 Census.* https://www.cbs.nl/en-gb/corporate/2021/20/cbs-in-the-starting-blocks-for-the-2021-census