Területi Statisztika

Közzététel: 2019. február 12.

A tanulmány címe:

Tér és társadalom Big Data szemüvegen keresztül

Szerző:

Ságvári Bence, Magyar Tudományos Akadémia Társadalomtudományi Kutatóközpont, E-mail: sagvari.bence@tk.mta.hu

https://doi.org/10.15196/TS590102

Tér és társadalom Big Data szemüvegen keresztül*

The Big Data perspective of space and society

Ságvári Bence
Magyar Tudományos Akadémia
Társadalomtudományi
Kutatóközpont
E-mail:
sagvari.bence@tk.mta.hu

A társadalomtudományi kutatásokban használt Big Data források jelentős része térben is értelmezhető információ. Így a szociológia és a társadalomföldrajz határterületein újfajta értelmezési keretek, módszertanok, továbbá gyors és olcsó adatforrások alkalmazására nyílik lehetőség. Az adathiányok, a reprezentativitás, a longitudinális összehasonlítás lehetősége, vagy csak az adatokhoz való hozzájutás folyamata olyan kihívás, amelyekkel mindig számolni kell. A Big Data távolról sem olyan varázsszer a társadalomtudományok és ezen belül a szociológiai fókuszú társadalomföldrajzi kutatások számára, amely egyik napról a másikra korábban nem ismert összefüggések feltárását teszi lehetővé, továbbá a hagyományos módszertanok alkalmazását feleslegessé és meghaladottá. Inkább egy olyan organikus fejlődésnek vagyunk részesei, amelynek számtalan zsákutcája van (és lesz), továbbá a kutatási irányok és módszerek sok dilemmát, megválaszolandó elméleti, módszertani, kutatásetikai, jogi, szervezeti stb. kérdést vetnek fel.

Kulcsszavak:
Big Data,
szociológia,
társadalomföldrajz,
kutatásmódszertan,
területi egyenlőtlenségek

* A „Big Data a területi kutatásokban” címmel rendezett konferencián, 2018. február 27-én tartott előadás szerkesztett változata.

Területi Statisztika, 2019, 59(1): 27–48; DOI: 10.15196/TS590102

Majority of Big Data sources used in social science research comprise spatial information. Therefore, at the crossroads of sociology and social geography new frames for interpretation, new methods, as well as fast and inexpensive data sources are emerging. However, missing data, the issue of representativeness, the possibility of longitudinal comparison, or simply the process of obtaining data are all challenges that researchers have to face with. Big Data are far from being such a 'magic sauce' for social sciences and especially for geographical research with sociological focus that helps to reveal previously unknown facts from one day to the other, or makes the use of traditional research methodology unnecessary and obsolete. Rather, we are witnessing an organic development with deadlocks and dilemmas, and the research approaches and methods raise a great deal of dilemmas and issues related to theory, methodology, research ethics, legal and organisational issues, etc. which need to be addressed.

Keywords:
Big Data,
sociology,
social geography,
research methodology,
spatial disparities

Beküldve: 2018. június 26.
Elfogadva: 2018. október 5.

Bevezetés

A közel másfél évtizede tartó folyamat során a szociológia, a közgazdaságtan és a társadalomföldrajz határterületein egyre több olyan kutatási eredmény jelent meg, amelynek empirikus anyagát valamilyen Big Data forrás képezte (Miller 2010). Ezek az elemzések többségében nehezen sorolhatók be a hagyományos tudományterületek által kijelölt elméleti és módszertani keretek közé. Ehelyett a *computational social science* (számítógépes társadalomtudományként fordítható) tudományos köztudatban egyre inkább polgárjogot nyert ernyője alá rendezhetjük őket (Conte et al. 2012, Lazer et al. 2009, Salganik 2017).

A Big Data tudományos hasznosítása ma már korántsem kuriózum. A világ számos egyetemén alakultak olyan interdiszciplináris kutatóközpontok, amelyek különböző témákban végeznek kutatásokat; évek óta megrendezik a téma rangos nemzetközi konferenciáit, és megjelentek a terület dedikált nemzetközi folyóiratai is. Ez a folyamat már Magyarországon is megfigyelhető, még akkor is, ha összességében alacsony a publikált empirikus kutatási eredmények száma, és viszonylag kevesen vannak (néhány intézményhez kötődnek) azok, akik ezzel a területtel foglalkoznak.

A tanulmánynak nem feladata annak kifejtése, hogy mit is értünk Big Data alatt, hiszen ezt korábban már magyar nyelven is többen megtették az elmúlt években (Baji 2017, Dessewffy–Láng 2015, Giczi–Szőke 2017, Jakobi 2014, Kmetty 2018, Nagy–Veroszta 2018, Németh 2015, Ságvári 2017a, 2017b, 2017c, Székely 2015, Vukovich 2015, Jakobi–Lőcsei 2016, Z. Karvalics 2018). Figyelmünket inkább arra összpontosítottuk, hogy a társadalomtudományi kutatásokban használt Big Data források jelentős része térben is értelmezhető információkhoz (települések nevei, postai címek, GPS-koordináták, mobilhálózatok cellainformációi, vagy IP-címek) jutasson bennünket (Leszczynski–Crampton 2016). A szociológia és a társadalomföldrajz határterületein így egy újfajta értelmezési keretre, módszertanra, továbbá gyors és olcsó adatforrások alkalmazására nyílt lehetőség. A tanulmány célja, hogy a kvantitatív szociológiai kutatások módszertani hagyományaiból kiindulva ismertesse a Big Data alapú kutatások néhány fontos, nem megkerülhető módszertani kihívását. Az alábbiakban (1) a reprezentativitás és általánosíthatóság, (2) a rendszerszintű torzítások és adathiányok, illetve (3) a hozzáférés és személyes adatvédelem legfontosabb kérdéseit tekintjük át elméleti megfontolások és konkrét kutatásokban alkalmazott módszerek alapján.

Méret és reprezentativitás

A Big Data kutatások folyamatát alapvetően az határozza meg, hogy a kutatást végzők többnyire „fordítva ülnek a lovon". Elméleti keretek és konkrét hipotézisek nélkül természetesen ezek a kutatások sem létezhetnek, az elérhető adatok köre azonban az elmélet és az empíria terén is korlátozhatja a lehetőségeket. A „klasszikus" kutatási folyamat szerint a kijelölt elméleti kereteken belül fogalmazódnak meg a kutatás konkrét kérdései és hipotézisei, majd ehhez kapcsolódva – különböző adatfelvételi módszerek segítségével – előáll az adat, amelyből kiolvashatók a válaszok, megerősítve vagy cáfolva az eredeti hipotéziseket. A Big Data esetében többnyire a „gombhoz kell varrni a kabátot". Olyan kutatási kérdéseket, hipotéziseket lehet csak megfogalmazni, amelyekre a válasz vélhetően megtalálható az adatokban. Ez arra is lehetőséget adhat, hogy csak olyan felfedező attitűddel közelítsünk az adatokhoz, amelynek az eredménye akár új elméleti keretek vagy korábban elképzelhetetlen kutatási kérdések megfogalmazásához vezethet (Miller–Goodchild 2014). Ez a jelenség vezet el bennünket az adatok méretének kérdéséhez.

Ma már közhely, hogy a nagy adatbázisok előnye a belőlük nyerhető részletgazdagság, a korábban láthatatlan mikroösszefüggések tudatos vagy akár véletlenszerű feltárásának lehetősége. A területi fókuszú elemzések esetében ez nagyon hatékony „zoomként" működhet, hiszen minél nagyobb mennyiségű, pontszerű adat áll rendelkezésre, annál élesebb és megbízhatóbb kép tárul elénk. Ez persze nem jelenti azt, hogy a nagyobb adatmennyiség szükségszerűen jobb is. Így az a törekvés, hogy minél több adatot begyűjtve, minél nagyobb adatbázisokon lehessen elemzéseket végezni,

bizonyos méret felett könnyen öncélúvá és értelmetlenné válhat. Ugyanakkor azt is nehéz meghatározni, hogy mekkora az elégséges adatmennyiség, hiszen a korlátokat sokszor csak a technikai háttér (a tároló és adatfeldolgozási kapacitás) jelöli ki. Ezért érdemes mindig szem előtt tartani, hogy vajon a kutatási kérdések tudományos értelemben megbízható megválaszolásához mekkora mennyiségű adatra van szükség.

Eagle és szerzőtársai (2010) a telefonhívások adataiból megfigyelhető hálózati jellemzők és a népszámlálásból származó szocioökonómiai „jóllétadatok" közötti kapcsolatokat elemezték a Big Data alapú kutatások szempontjából. Úttörőnek tekinthető tanulmányukban elsődleges elméleti alapvetésük az volt, hogy a közösségeken belüli társadalmi kapcsolatok összefüggésbe hozhatók a gazdasági fejlődéssel. Másképpen fogalmazva: a gazdasági növekedést biztosító lehetőségek a helyi közösségektől függetlenül, kívülről érkeznek. Ahol viszont a társas kapcsolatok ritkák és egymástól elszigetelt kis csoportok élnek együtt, ott korlátozottak a külső kapcsolódások. Ezzel ellentétben a társas kapcsolatok heterogenitása (természetesen más tényezők megléte esetén) pozitív hatással van egy közösség gazdasági értelemben vett sikerességére. A szerzők a társadalmi-gazdasági fejlettséget egy kompozit deprivációs index segítségével mérték. A kutatás empirikus alapját egy olyan adatbázis képezte, amely míg az Egyesült Királyság területén belül 2005 augusztusában kezdeményezett mobilhívások 90, addig a lakossági és üzleti vezetékes hívások közel 100%-át tartalmazta. Az így létrejött gráf 65 millió csomópontot és ezek között 368 millió élt (azaz hívást) tartalmazott. Figyelembe véve az említett ország méretét és lakosságszámát, ez óriási adatbázisnak tekinthető, amelynek segítségével vélhetően először sikerült Granovetter (1985) gyenge kötések elméletét egy ország teljes területére kivetítve, Big Data alapokon bizonyítani. A hatalmas mennyiségű adat ellenére is feltehető a kérdés, hogy egyetlen nyári hónap adatai (amikor vélhetően a munkajellegű hívások aránya volt alacsonyabb) nem torzíthatták-e az eredményeket, továbbá az sem ismert, hogy az összes mobilhívás adatának melyik 10%-a nem volt elérhető a kutatók számára. Ez a tény pedig a reprezentativitás és az eredmények általánosíthatóságának kérdését veti fel.

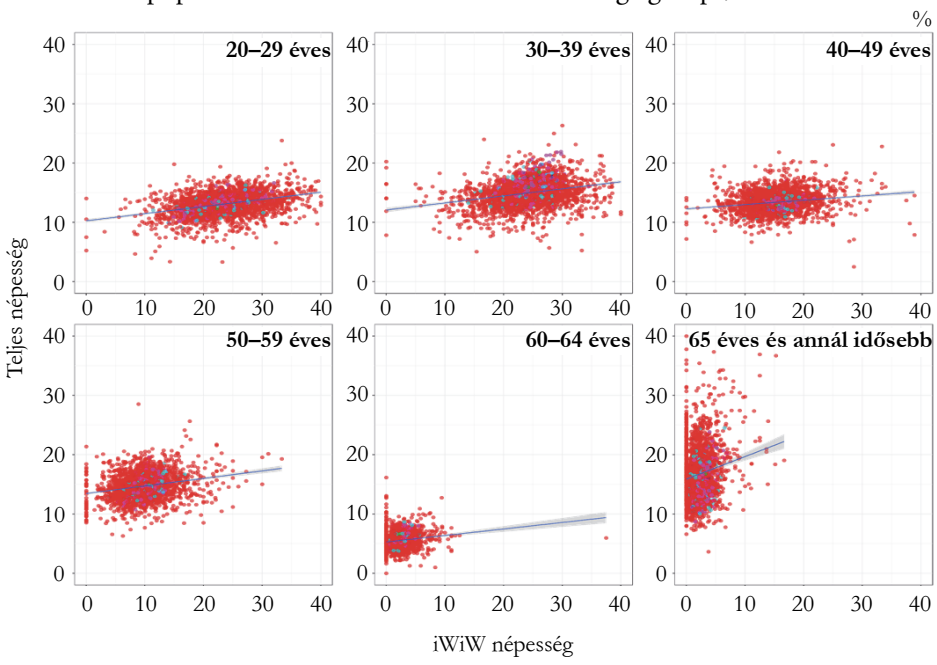
A reprezentativitás elvén szocializálódott társadalomtudományos gondolkodás a kvantitatív, Big Picture típusú kutatásokban általában nehezen fogadja el minőségi adatként az olyan elemzéseket, amelyek nem valószínűségi mintavételen vagy a teljes sokaság vizsgálatán alapulnak. A szélsőségesen kételkedő vélemény szerint az ilyen kutatásokból lényegében semmit sem tudhatunk meg. Az empirikus szociológia tankönyvi példája az egyesült államokbeli Literary Digest 1936-os fiaskója az ottani elnökválasztás eredményének előrejelzésében (Squire 1988), továbbá a Gallup kvótaalapú mintavételének közvélemény-kutatási kudarca 1948-ban (Frankel–Frankel 1987). Ezek mind azt mutatták, hogy miként lehet hatalmas méretű adatbázisokon, vagy a kiválasztásban szubjektív szempontokat (is) megengedő módszerekkel „felsülni" az előrejelzésekben. A Big Data vonatkozásában gyakran emlegetett „*n=all*" „bonmot"-ja (Mayer-Schönberger–Cukier 2013), „messziről" nézve igaz ugyan, de

a valóságban egy adatbázis szinte mindig hiányos vagy hibás. A mintavételen alapuló reprezentativitás vagy a teljeskörűség még a legjobb minőségű Big Data esetében sem garantált. Az információk töredékesek, „zajosak" vagy éppen szisztematikusan torzítottak, és ezekkel a korlátokkal tudatosan foglalkozni kell(ene) a kutatást végzőknek. Egy közösségi hálózat teljes adatbázisa, a felhasználók által megosztott üzenetek, vagy egy szolgáltatónál regisztrált éves hívásadat rengeteg információt tartalmaz, de az ezekből levont, teljes társadalomra vonatkozó, reprezentatív megállapításokkal nyilvánvalóan óvatosan kell bánni.

1. ábra

A magyarországi települések állandó lakosságának és a településen regisztrált iWiW-felhasználók egymáshoz viszonyított aránya az egyes korcsoportokban, 2013

Proportion of iWiW users registered in Hungarian settlements to the resident population of the settlements in various age groups, 2013



Forrás: iwiw, KSH-TeIR (2013) alapján saját szerkesztés.

Kutatócsoportunk[1] évekkel ezelőtt hozzáférést kapott a 2002 és 2014 között működő iWiW online közösségi oldal anonimizált adatbázisához, amelyben megta-

[1] „Egy online közösségi hálózat életciklusa: big data elemzés" című OTKA (K112713) kutatás. A kutatócsoport tagjai: Jakobi Ákos, Kertész János, Koltai Júlia, Lengyel Balázs, Lőrincz László, Török János és Ságvári Bence.

lálhatók voltak a felhasználók publikusan elérhető demográfiai adatai, illetve a felhasználók között létrejött kapcsolatok (Lengyel et al. 2016, 2015). Ritkaság, hogy egy 10 millió fős ország több mint 4 millió lakójáról rendelkezésre álljon ilyen részletezettségű adatforrás. Ez az adatbázis is, a teljes magyar társadalomra nézve hiányos volt, hiszen nem volt véletlenszerű, hogy ki volt és ki nem a szolgáltatás regisztrált felhasználója. Az sem volt mellékes probléma, hogy a felhasználók jelentős része állította bizonyos demográfiai adatait (például a korát) nem nyilvánossá, így számunkra elérhetetlenné. A hiányzó demográfiai adatokat – ahol ennek volt jelentősége – különböző hálózati jellemzőket is figyelembe vevő statisztikai modellekkel pótoltuk, de ezek pontossága nem mérhető össze azokéval, akik ezt az információt nem tagadták meg. Az 1. ábra szemlélteti, hogy településenként hogyan tért el egymástól a valós, illetve az iWiW-népesség százalékos aránya az egyes (nem becslésen alapuló) korcsoportokban, azaz milyen megkötésekkel kellett dolgozni azokban az esetekben, amikor a magyar társadalom egészének működéséről akartunk tudományosan legitim állításokat tenni.

Az iWiW-hez hasonló adatforrást használt Norbutas és Corten (2018) is, akik egy holland online közösségi hálózat teljes adatbázisát felhasználva vizsgálták a különböző hálózati strukturális jellemzők, illetve 441 önkormányzati közigazgatási egység (*gemeenten)* gazdasági fejlettségének kapcsolatát. Az elemzés kereteit a Granovetter–Coleman–Putnam-féle társadalmi tőke elmélete jelölte ki, és ennek megfelelően három jól operacionalizálható hálózati jellemzőt vizsgáltak (Coleman 1990, Granovetter 1985, Putnam 2000). Az első ezek közül a hálózat sűrűsége volt, amely mögött az a hipotézis húzódott meg, mely szerint önmagában a sűrű(bb) hálózatok hozzájárulnak a – gazdasági jellegű együttműködésekhez nélkülözhetetlen – bizalom magas(abb) szintjéhez (Knack–Keefer 1997). A másik két dimenzió a társadalmi tőke összekötő (*bridging*) és összetartó (*bonding*) típusát próbálta mérhetővé tenni. Az előbbi két település közösségei közötti hálózati kapcsolatok sűrűségére utal. Ezek azok a gyenge kötések, amelyeken keresztül új információ, tudás, korábban nem ismert nézetek, gondolatok jelenhetnek meg egy közösségben. Gazdasági szempontból elsősorban az információk terjedésének van jelentősége, amelyek új munkalehetőségeket, új termékek és szolgáltatások innovációját segíthetik elő.

A társadalmi tőkének ez a típusa elsősorban az egyén szintjén – mint az egyéni versenyképesség, sikeresség egyik alapvető kritériuma – értelmezhető. Ezeknek a kapcsolatoknak a települési szinten aggregált mutatója azonban fontos indikátora lehet egy-egy térség gazdasági fejlettségének és fejlődési potenciáljának. Az említett szerzők azt is feltételezték, hogy a távolabbi településekkel fenntartott sűrűbb kapcsolathálózat (a közeli, szomszédos településekkel összehasonlítva) még inkább kapcsolatba hozható a gazdasági fellendüléssel. Ezzel szemben a településeken belüli kapcsolatok sűrűsége és modularitása (töredezettsége) az összetartó társadalmi tőke mértékének közvetett mérőszáma. A kiinduló hipotézis szerint a nagyobb modulari-

tás, azaz a kapcsolathálózat töredezettsége, az egymástól elszigetelt közösségek jelenléte negatívan befolyásolja a gazdasági folyamatokat.

Az elemzés empirikus alapjául szolgáló holland online közösségi hálózat (Hyves) története nagyban hasonlít a magyar iWiW gyors sikerére, majd látványos bukására. A Hyves 2004-es indulását követően 2010-ben érte el csúcspontját, amikor több mint 10 millió regisztrált felhasználója volt, többségében Hollandiából. Az adatok tisztítását és geokódolását követően 6,3 millió felhasználó adata állt rendelkezésre, akik országos szinten a teljes lakosság 34%-át reprezentálták. Az arányok településenként eltérők voltak, de szinte mindenhol minimálisan elérték a 20%-ot. Nem meglepő módon a felhasználók átlagos életkora a teljes lakossághoz képest alacsonyabb volt. Ez a tény kapcsolódik a Big Data előzőekben említett *„n=all"* és reprezentatitivitás problémához, azaz joggal merül fel a kérdés, hogy milyen kompromisszumokkal használható egy ilyen jellegű adatforrás térségi szintű makrogazdasági folyamatok magyarázatához? A különböző hipotézisekre felállított modellek eredményei azt mutatták, hogy a távoli térségekkel fenntartott sűrű kapcsolathálózat szoros összefüggést mutatott a gazdasági fejlettséggel, ezzel párhuzamosan azok a térségek voltak gazdaságilag elmaradottabbak, ahol a belső kapcsolathálózat töredezett volt, azaz a közösségek egymástól elszigetelten léteztek. További érdekesség, hogy az eredmények alapján az egyes területi egységek belső hálózati sűrűsége fordított összefüggést mutatott a gazdasági fejlettséggel. Nem sikerült empirikusan bizonyítani, hogy a sűrű hálózatok elméletben magasabb bizalomszintje a gazdasági folyamatokat is pozitívan befolyásolja. Ezeket a megállapításokat azonban érdemes aszerint is átgondolni, hogy vajon mennyiben befolyásolta az eredményeket az a tény, hogy a teljes népességnek csak egy kisebb részére vonatkozóan ismert a kapcsolati háló, és ezeknek a kapcsolatoknak (éleknek) a súlya egységnyi, azaz az adatok alapján nem dönthető el, hogy milyen intenzitásúak a társas kapcsolatok.

A bárki által használható adatexport lehetősége miatt az elmúlt néhány évben az első számú társadalomtudományos Big Data kutatási nyersanyaggá a Twitter vált. Ingyenessége mellett a másik fontos vonzereje, hogy a tweet-ek geolokációs adatai is elérhetők, így a társadalomföldrajz is intenzíven használhatja ezt a lehetőséget. Az emberek térbeli mobilitásának vizsgálatára Jurdak (2015) helymeghatározáson alapuló (geolokált) Twitter-bejegyzéseket használt. Kutatásai során eltérő mobilitásmintázatokat azonosítottak ausztráliai Twitter-felhasználók körében. Eredményeik fő üzenete az volt, hogy a tweet-ek alapján jól rekonstruálható és (ami még fontosabb) előre jelezhető az emberek mozgása.

Egy másik kutatásban Llorente és szerzőtársai (2015) a spanyolországi munkanélküliségi adatokat és a Twitter-használat összefüggéseit vizsgálták. Kimutatták, hogy azok a térségek, ahol az emberek az általuk megosztott tweet-ek lokációs adatai alapján nagyobb földrajzi mobilitást mutattak, alacsonyabb munkanélküliségi rátával rendelkeztek, továbbá a tweet-ek dinamikája alapján az itt élők korábban kezdték a napot, és a tweet-jeik szövege kevesebb nyelvhelyességi hibát tartalmazott, ami indirekt módon a magasabb iskolai végzettségre utalt.

Gonclaves és szerzőtársai (2018) az amerikai angol nyelv térnyerését mutatták be térben és időben, Twitter-adatok segítségével. Fő megállapításuk az volt, hogy az Egyesült Királyság területén kívül mindenhol kimutatható volt az amerikai angolnyelv-használat előtérbe kerülése (annak hatását még az ország határain belül is meg lehetett figyelni).

Nyilvánvaló, hogy mindhárom példa esetében felmerül, vajon mennyiben legitimek az eredmények, mennyiben lehet ezekre valódi döntéseket alapozni? Itt legalább négy szempontot érdemes megemlíteni. (Ezeket a problémákat általában a szerzők maguk is említik.) (1) a Twitter aktív felhasználói az internetet használó, fiatalabb és „technológiabarát" emberek csoportjaiból kerülnek ki; (2) a Twitter API-n keresztül letölthető adatok egy olyan 1%-os (a szolgáltató ígérete szerint véletlen) mintavételen alapszik, aminek a pontos algoritmusáról semmilyen nyilvános dokumentáció nem érhető el (Morstatter et al. 2014); (3) a geolokált tweet-ek az összes tweet 1%-át adják, és az ezzel foglalkozó kutatások szerint demográfiailag – vélhetően mobilitásában is – eltér egymástól azok csoportja, akik engedélyezik és akik nem engedélyezik az app-ban a tartózkodási helyük megosztását; (4) végül arról sincsenek megbízható információink, hogy egy adott felhasználó aktuális tartózkodási helye mennyiben befolyásolja a „tweet-elési" szokásait. Feltételezhetjük, hogy vannak olyan pontok, amelyek alul- vagy felülreprezentáltak, attól függően, hogy ott éppen milyen külső ingerek érik a felhasználót.

A Twitter-alapú térbeli dimenziót is vizsgáló kutatások módszertani legitimitásának kérdése ennél természetesen jóval összetettebb, de annyi bizonyos, hogy a reprezentativitás szempontjából jelentős potenciális torzításokat figyelembe kell venni a kutatások megtervezésekor és az eredmények értelmezésekor.

Rendszerszintű torzítások az adatokban

A hagyományos, emberek személyes megkérdezésén és megfigyelésen alapuló kutatásokkal szemben az egyik alapvető kritika, hogy a résztvevők tudatában vannak annak, hogy egy kutatási folyamat részesei, ezért hajlamosak a helyzethez „igazítani" a viselkedésüket és a véleményüket. A hagyományos survey-típusú kutatások esetében ennek része például a kérdezőbiztos hatása. A kérdezőbiztos demográfiai jellemzői (nem, életkor), illetve az adott témával kapcsolatos vagy az eredmények megoszlására vonatkozó saját prekoncepciói, vagy a kérdezett fejében lévő társadalmi elvárások valamilyen irányba befolyásolják a válaszokat is (Groves 1989).

Ebből a szempontból a nem kutatási célból létrejövő Big Data mint nyersanyag alapértelmezésben megbízhatóbb, akár területi, akár nem területi jellegű kutatások esetében, ezek tulajdonképpen *noninvazív* beavatkozásnak tekinthetők. Azonban esetükben is ajánlott az óvatosság. Egy egyszerű példával szemléltetve: a Facebook- és a Twitter-bejegyzéseket számos kutatás használta arra, hogy belőlük az emberek véleményére, hangulatára, érzelmi állapotára következtessen (Coviello et al. 2014,

Eichstaedt et al. 2015, Šćepanović et al. 2017, Weller et al. 2013). Ezek a közösségimédia-platformok nem feltétlenül az emberek őszinte gondolatainak lenyomatai, hanem a Goffman-féle homlokzatépítés eszközei (Goffman 1981). Az emberek általában saját helyzetüket, életüket inkább pozitívnak szeretnék láttatni, a véleménynyilvánításokban pedig nagyobb arányban találjuk meg az extrém álláspontokat. Ezek a hatások a hagyományos adatgyűjtések esetében is megfigyelhetők, a közösségi média torzító hatásainak összetett mechanizmusait azonban még nem ismerjük annyira, hogy ezeket rutinszerűen korrigálhassuk. (Ha egyáltalán lehetséges, és feltételezve, hogy nem ennek az idealizált képnek a feltárása a cél.)

A hagyományos adatfelvételi módszerekkel ellentétben – ahol a résztvevők tudatában vannak annak, hogy részesei az adott kutatásnak – a Big Data lehetőséget ad arra, hogy az emberekről „megzavarásuk" nélkül gyűjtsünk adatokat. A közösségi média szolgáltatások (Facebook, Twitter, Instagram) mint leggyakrabban használt empirikus kutatási források mögött lévő rendszerek működését azonban bonyolult, a külvilág számára többnyire visszafejthetetlen, összetett algoritmusok irányítják. A szolgáltatók számára ezek elsődleges célja, hogy minél hatékonyabban szolgálják ki a felhasználók igényeit. Másképp fogalmazva: növeljék az emberek lojalitását az adott szolgáltatás iránt; a még hatékonyabb, célzott hirdetések érdekében egyre pontosabb személyiségprofilt alakítsanak ki róluk. Ezek az algoritmusok tudományos értelemben véve torzítják az adatokat, hiszen képesek az emberi viselkedés bizonyos aspektusait felül-, illetve alulsúlyozni. Ugander és szerzőtársai (2011) például egy 2011-es adatokon alapuló elemzésükben kimutatták, hogy feltűnően magas a 20 ismerőssel rendelkezők száma a Facebook-on. Az emberek kapcsolathálózatában azonban nem nagyon vannak ilyen „mágikus" számok. A magyarázat a Facebook azon algoritmusában rejlett, amely megpróbálta mesterségesen „feltornászni" a felhasználók számát abban az esetben, amikor az 20 alatt volt. A háttérben pedig az a kapcsolati triádokról szóló klasszikus szociológia- és hálózatelméleti megfontolás húzódott meg, mely szerint, „*ha nekem barátom Péter és Pál is, akkor jó eséllyel Péter és Pál is ismerik egymást*". A rendszer így a „Kit ismerhetek?" funkció keretében Péternek és Pálnak is felajánlotta, hogy legyenek ismerősök a Facebook-on. A Facebook ezeket az ajánlásokat a felületén intenzíven kommunikálta mindaddig, amíg az ismerősök száma el nem érte a 20 főt. Miután megszűntek ezek a jól látható ajánlások, a felhasználók egy része nem lépett már mással kapcsolatba, így az ismerőseinek a száma maradt 20 (idézi Salganik 2017, 35. old.).

Az algoritmusalapú torzítások lehetőségét nem lehet eléggé hangsúlyozni. Ugyanakkor fontos a feltételes mód használata, hiszen egyáltalán nem biztos, hogy a probléma a valóságban is megváltoztatja a kutatás eredményeit. A kihívás azonban abban rejlik, hogy ezek az algoritmusok folyamatosan változnak, a változások tartalmát azonban nem hozzák nyilvánosságra, a visszafejtésük lényegében önálló, kevesek által művelt kutatási terület (Burrell 2016, Pasquale 2015, Sandvig et al. 2014).

A Big Data alapú társadalomtudományi elemzések az alapvetően pozitivista tudományelméleti elvből kiindulva az adatra többnyire értéksemleges nyersanyagként tekintenek. Ez nem azt jelenti, hogy ennek segítségével ne lehetne rámutatni a társadalmi egyenlőtlenségek széles spektrumára. Létezik azonban egy olyan nézőpont is, amely arra fókuszál, hogy vajon maga az adat milyen körülmények között keletkezik, az adatok szerkezetét befolyásolhatják-e olyan kialakult társadalmi gyakorlatok, amelyek a későbbi kutatások eredményeire is hatással lehetnek. Az adatok keletkezése, tárolása, hozzáférhetősége ugyanis minden esetben része a társadalom kialakult hatalmi erőviszonyainak. A kormányzati, a vállalati adatgyűjtések az adott szervezet működési logikájából adódnak, és annak mindenkori érdekeit kell hogy szolgálják. Létezik egy olyan speciális terület is, ahol empirikusan kevésbé tetten érhetők ezek az erőviszonyok: a közösségi, alulról építkező adatbázisok és információs források esetében izgalmas kérdés, hogy ezek milyen módon tükrözik vissza a társadalmi normákat, kulturális különbségeket, beidegződéseket. A területi elemzések szempontjából ennek egyik jó példája a felhasználók által rögzített információk az olyan közösségitérkép-platformokon, mint az OpenStreetMap vagy a Google MapMaker. (Ez utóbbi 2017 óta a Google Maps részévé vált.) Vannak arra vonatkozó kutatási eredmények, melyek szerint ezek akarva-akaratlanul inkább férfias világképet közvetítenek. Egy 2011-es online kérdőíves vizsgálat azt találta, hogy ezekben a rendszerekben demokratikusan, közösségi döntések alapján jelennek meg az új információk, azonban a tény, hogy ezzel elsősorban férfiak foglalkoznak, ők kerülnek az információs kapuőr (*gatekeeper*) szerepbe. Emiatt például a jellemzően férfias (kocsma, bár, night bar) és nőies (bölcsődék, óvodák) helyek kategóriáinak részletessége, a kategorizálások „finomsága" eltérő volt, azaz nagyobb mértékben tükrözte a férfias látásmódot (Stephens 2013). Ehhez hasonló az a jelenség is, hogy a magasabb státusú társadalmi rétegek által lakott, vagy más okok miatt frekventált területek kidolgozottsága, információtartalma gazdagabb, mint azoké, ahol szegény, hátrányos helyzetű csoportok élnek.

A Big Data egyik fontos ígérete lehet(ne), hogy mivel az adatok folyamatosan keletkeznek, ezért trendek megfigyelésére, így longitudinális elemzésre is alkalmasak. Ez azonban korántsem biztos, hogy így van. Módszertanilag két (vagy több) időpont közötti változás méréséhez azonban alapfeltétel, hogy sem az alapsokaság, sem pedig az alkalmazott módszer ne változzon. Ez a Big Data esetében nem magától értetődő. Az előzőekben bemutatott konkrét kutatások módszertanának egyike sem alkalmas arra, hogy azt megismételve az időbeli változásokat is vizsgálni lehessen. Salganik (2017, 33. old.) három olyan tényezőt sorol fel, amelyek nem teljesülnek abban az esetben, ha két vagy több időpont vagy időszak változásait szeretnénk összehasonlítani. A Big Data viszonylagos „fiatalsága" miatt ez ma még inkább kuriózum, nem pedig általános gyakorlat. Láthattuk, hogy a Big Data általában nem teljes körű és nem reprezentatív. Ez azt is jelenti, hogy két különböző időpontra vonatkozó adatok különböző alapsokaságot is jelentenek, az eltérések pedig vélhe-

tően nem lesznek véletlenszerűek. Abban az esetben, ha nem közvetlenül egy adott szolgáltatást használók összetételének változása a kutatás témája, hanem valamilyen közvetett megfigyelés, az összehasonlítás még kevésbé megvalósítható (Ezt nevezi Salganik „*population drift*"-nek). A Twitter példájánál maradva: két időpontban számosságában és összetételében is eltérő lesz az alapsokaság, így nem tudjuk megállapítani, hogy összességében az emberek mobilitási szokásai változtak-e meg, vagy csak olyan új felhasználók csatlakoztak (vagy mások hagyták ott a szolgáltatást), akik újfajta mobilitási mintázatokat „hoztak be" a rendszerbe.

Egy másik példa: az említett, 2002 és 2014 között létező iWiW online közösségi hálózat különböző életciklusokra osztható aszerint, hogy milyen társadalmi csoportokból érkeztek a felhasználók, és kik tartoztak az új belépők jellegzetes csoportjaiba. Az innovációk terjedésének alapvető törvényszerűségei szerint míg az első pár évben felülreprezentáltak voltak a többségükben városi, iskolázottabb fiatalok és fiatal felnőttek, addig az utolsó pár évben már jóval nagyobb arányban csatlakoztak az alacsonyabb státusú, idősebb, kistelepüléseken élő emberek. Ez azt is jelenti, hogy a felhasználók száma is különböző volt az egyes időszakokban. Emiatt tetszőleges két időpontot kiválasztva aligha készíthetünk olyan elemzést, amely függetlenül más tényezőktől, a magyar társadalom kapcsolathálózatának változását vizsgálná, hiszen lényegében mindig más volt az alapsokaság, akikről rendelkezésre álltak információk.

Ehhez hasonló probléma, amikor ugyanezt a kérdést vagy akár az emberek földrajzi mobilitását két eltérő időszakra vonatkozó mobiltelefon-hálózati hívásadatok alapján vizsgáljuk. Míg az előbbi esetben az alapsokaság az eltérő, addig utóbbiban a kommunikációs szokások megváltozása vihet(ne) tévútra bennünket (ezt nevezhetjük „*behavioral drift*"-nek). Az a tény, hogy az emberek kevesebbet hívják egymást, nem feltétlenül jelenti azt, hogy a társadalmi kapcsolatok visszaszorulóban vannak. Inkább arról van szó, hogy a korábban telefonon történő kommunikáció más platformokra (például szöveges üzenetek, hanghívásra is alkalmas „app-ok" stb.) terelődött át. A kevesebb hívásadat a földrajzi mobilitással kapcsolatos információkat is töredékesebbé, megbízhatatlanabbá teszi, megnehezítve vagy lehetetlenné téve azt, hogy két időpontra vonatkozó trendeket legitim módon hasonlíthassunk össze.

A longitudinális vizsgálatoknak létezik egy harmadik korlátja is, amely közvetlenül az adatot szolgáltató rendszer időszakos változásaiból következik („*system drift*"). Például az iWiW esetében egyik napról a másikra megszűnt a korlátozás, mely szerint csak meghívóval lehetett regisztrálni, vagy akkor, amikor a Twitter és a Facebook megváltoztatta a bejegyzések maximális karakterszámát, jelentős mértékben befolyásolta a felhasználók lehetőségeit és rövid időn belül átalakította a használati szokásokat, azaz a kutatásokhoz használt nyersanyagot. Összességében tehát két azonos forrásból, azonos módon, de két különböző időpontban létrehozott adatbázis nagyban eltérhet egymástól abban, hogy kik voltak a felhasználók, és ők hogyan használtak egy olyan rendszert, amelynek működési elveit időről időre a tulajdonosok is megváltoztatják. Mindezekből adódóan ma még bizonytalan, hogy a Big Data

források hogyan alkalmazhatók longitudinális jelleggel, módszertanilag legitim módon a hosszú távú társadalmi változások vizsgálatára.

Hiányzó információk és „koszos" adatbázisok

A minden kérdésünkre választ adó tökéletes adat természetesen nem létezik. Az empirikus szociológia hagyományos módszereit tekintve sincsen olyan hosszú és részletes kérdőív vagy interjú, amely ne lenne valamilyen szempontból hiányos. Még inkább ez a helyzet azokkal a szociológiai másodelemzésekkel, amelyek a nyilvánosan elérhető longitudinális és/vagy nemzetközi összehasonlításra is alkalmas adatbázisokra (ESS, EVS, SHARE stb.) épülnek. Ezek esetében még több kompromisszumot kell megkötni, hiszen az adatokat előállítók (az eredeti kérdéseket megtervezők) és az azokból elemzést készítők csoportja nem azonos, így a kutatási folyamat első és második része elválik egymástól.

A Big Data források esetében ezek a hiányosságok még nagyobbak lehetnek. Köztudott, hogy az adatok többnyire nem kutatási céllal keletkeznek, így a kutatás szempontjából alapvető fontosságú információk hiányozhatnak belőlük. A „klasszikus" szociológiai kutatások esetében használt alapvető demográfiai változók is valószínűleg hiányosak, vagy ha rendelkezésre is áll az információ, még akkor sem lehetünk teljesen biztosak annak megbízhatóságában. Egy példa: mobilhálózatok hívásadataiból sokféle kutatási kérdés válaszolható meg az emberek kapcsolathálózatával, rövid és hosszú távú földrajzi mobilitásával, napi rutinjaival, az általuk használt terekkel kapcsolatban. Nagy valószínűséggel azonban még közelítő adataink sem lesznek arról, hogy milyen demográfiai jellemzői vannak az adott készülék használójának. A szolgáltatók rögzítenek ugyan előfizetői adatot a szerződés megkötésekor, de ez egyáltalán nem biztosíték arra, hogy a későbbi használó is ugyanaz a személy. Flottás előfizetések esetében még ezek az információk sem fognak rendelkezésre állni.

A hívásadatok példájánál maradva, amennyiben az a cél, hogy az emberek kommunikációs szokásait, vagy kapcsolathálózatait vizsgáljuk – az ilyen típusú adatokból nagyon sok kérdésre választ kaphatunk – csak a valóság egy szelete áll rendelkezésre. Kicsi az esély arra, hogy azonos formátumban az összes szolgáltató adataihoz hozzáférhessünk. Ez esetben pedig gondolnunk kell az adatok rendszerszintű torzítására is, hiszen szolgáltatónként eltérhet az ügyfélbázis, bizonyos társadalmi csoportok alul-, míg mások felülreprezentáltak lehetnek. Egy ember térbeli mozgása többékevésbé rekonstruálható, de a kommunikációs mintázatoknak egyszerre csak kisebb részét tudjuk vizsgálni. Például hiába ismerjük valakinek a mobiltelefonálási szokásait, ehhez nem tudjuk párosítani a különböző üzenetküldő alkalmazásokon (appokon) keresztül küldött üzeneteket, illetve a hangalapú kommunikációt. Sőt, ma még a mobilszolgáltatók sem feltétlenül tudják, hogy a hálózatukon mobilinternetet használó előfizető valamilyen tartalmat fogyaszt, vagy éppen valakivel kommunikál. Léteznek olyan megoldások, amelyekkel ezeket a problémákat (részben) orvosolni tudják, de ezek nagyon gondos előkészítést igényelnek, a siker sem feltétlenül garan-

tált, és természetesen jóval drágábbak. Például a magyarra csak nehézkesen lefordítható *enriched asking* lényege, hogy a Big Data forrást hagyományos adatfelvétellel egészítik ki, így pótolva a hiányzó információkat. Ennek a módszernek egy további alkalmazása, amikor a kis elemszámú mintán *„big"* és survey-alapú adatforrások összekapcsolása után prediktív modellek segítségével a teljes adatbázisra készülnek becslések (*amplified asking*).

Ez utóbbira példa az a néhány évvel ezelőtti kutatás, amelyben jövedelmi helyzet és a szegénység területi szintű becslését végezték el. A fejlett világ országaiban ma már hosszú évtizedekre visszanyúló és egyre bővülő hivatalos statisztikai adatok állnak rendelkezésre a társadalom, a gazdaság aktuális állapotáról. Azokban az államokban azonban, ahol fejletlen az intézményrendszer, a gazdaság pedig elmaradott, ott jóval kevesebb ilyen típusú adatforrás érhető el. Pedig ezek éppen a sikeres gazdaságfejlesztési programok és a társadalmi reformok pontos tervezéséhez lennének elengedhetetlenek. Blumenstock és szerzőtársai (2015) egy ruandai (közép-afrikai) mobilszolgáltató hívásadatait kombinálták hagyományos survey-technikával, és ennek alapján területi bontásban, a teljes lakosság vagyoni helyzetére próbáltak becslést adni. A módszer röviden a következő volt: a rendelkezésre álló milliárdos nagyságrendű hívásadatból (hanghívások és szöveges üzenetek) egyéni szinten komplex indikátorokat hoztak létre, amelyek a hívások időpontját, időtartamát, a hívott fél hasonló jellegzetességeit, illetve a hívásokból kiolvasható földrajzi mobilitás mintázatait vették figyelembe. Ezzel párhuzamosan egy közel ezer fős, területi szinten reprezentatív mintán, telefonos kérdőíves kutatással vizsgálták az előfizetők társadalmi és vagyoni helyzetét, infrastrukturális ellátottságát. A létrehozott személyes profilokhoz hozzárendelték a mobilhasználati szokásokból kiolvasható információkat, majd a kettő adatforrás között statisztikai összefüggéseket kerestek. Az utolsó lépésben a kérdőíves kutatásból kimaradt mobiltelefon-előfizetőkre lefuttatták ezeket a prediktív modelleket, azaz telefonhasználati szokásaik alapján becslést adtak az adott személy vagyoni helyzetére. Ez a módszer természetesen nem eredményezhet 100%-ban pontos adatokat, de összességében a modell alkalmas volt arra, hogy viszonylag jól becsülje, hogy egy adott háztartás használ-e elektromosáram-szolgáltatást, rendelkezik-e motorkerékpárral, televízióval, hűtőszekrénnyel. A legfontosabb, hogy mindezt a hagyományos adatgyűjtéshez képest tízszer gyorsabban és ötvenszer olcsóbban tudták elvégezni (Salganik 2017, 359. old.). Gyakorlati szempontból ennél is fontosabb, hogy az ilyen eljárások segítségével a kevésbé fejlett országokban hivatalos statisztikák hiánya miatt meglévő információs űrt Big Data alapú közvetett információk és algoritmusok segítségével lehet kitölteni, és becslést adni arról, hogy az országban hol élnek a leginkább rászoruló, kiszolgáltatott társadalmi csoportok.

A korábban említett probléma leküzdése, mely szerint nagyon nehéz ugyanazon időszakra az összes mobilszolgáltató hívásadataihoz hozzájutni, ezeket megfelelő formátumban egyesíteni, és így a teljes gráfot létrehozni nem lehetetlen. Erre példa Coscia és szerzőtársainak (2017) tanulmánya, amelyben a gazdaságfejlődés azon

klasszikus elméleti kérdésére keresték a választ, hogy milyen tényezők húzódnak meg bizonyos térségek gazdasági elmaradottsága mögött. Vajon a rosszul működő intézményrendszer az elsődleges ok, vagy pedig az, hogy el vannak zárva azoktól a társadalmi közvetítőcsatornáktól, amelyek elősegítenék a technológia terjedését, majd az abból következő gazdasági fellendülést. Az elemzés tárgya Kolumbia volt, ami talán nem véletlen, mivel sokszor olyan nem észak-amerikai és európai országokban van lehetőség ezeknek az adatoknak a megszerzésére, ahol kevésbé szigorúak az adatvédelmi szabályozások. A kutatók területi szintű foglalkoztatási és jövedelemadatokat használtak, illetve (ugyanezen a területi szinten) aggregálták a rendelkezésükre álló mobiltelefon-hívások metaadatait. Hat hónapos időszak alatt összesen 2,2 milliárd (!) hívás metaadatával dolgoztak. Az adatbázisban 40 millió egyedi telefonszám volt, ami (figyelembe véve, hogy Kolumbia lakossága 46 millió fő) hatalmas méretű információforrás. Az elemzés kuriózuma az volt, hogy nemcsak egy szolgáltató adatait használták, hanem képesek voltak – bizonyos megkötésekkel ugyan, de – egyesíteni több szolgáltató hívásait. A „tökéletes" adatbázis nyilvánvalóan az lenne, ahol az összes hívó és hívott fél pontos földrajzi koordinátája ismert, az adott hívás egyéb metaadataival együtt. A mobiltelefonos hívásadatoknak azonban sajátossága, hogy amennyiben két különböző szolgáltató ügyfelei hívják egymást, adott szolgáltatónál csak a hívó fél tartózkodási helye ismert. Ez a korlát ebben a kutatásban is jelen volt, ezért ezekben az esetekben a hívott félnek a leggyakoribb tartózkodási helyét (ahonnan a másik szolgáltató adatbázisa alapján a leggyakrabban telefonált) használták. A kutatás eredményei azt mutatták, hogy azokban a térségekben volt erőteljes a gazdasági fejlődés dinamikája, amelyek sűrű kommunikációs kapcsolathálózattal rendelkeztek. Bár a kutatás csak egy rövid időszakra fókuszált, mégis közvetett bizonyítékokkal szolgált arra, hogy egy homogén intézményrendszer keretein belül azok az elmaradott térségek tudtak felzárkózni, amelyek intenzív kommunikációs kapcsolatokat alakítottak ki a fejlettebb területekkel, így az információk, a tudás és a technológia terjedése hatékony volt.

Az adatelemzéssel foglalkozók körében gyakran hangoztatott a 90/10 hüvelykujj szabály, amely még a Pareto-elv ismert 80/20 arányainál is szélsőségesebben határozza meg az adatelemzés teljes folyamata során az adatok tisztítására, megfelelő struktúrába rendezésére, az elemzésre való előkészítésére szánt időt (90%), illetve magát a tényleges elemzést (10%). Ez az elv a Big Data esetében is helytálló. Sőt, itt azzal is meg kell küzdenie a kutatónak, hogy mivel ezeket az adatokat nem kutatási céllal hozták létre, az adatstruktúrák kutatási szempontból szokatlanok lehetnek, a helyzetet pedig tovább nehezítik a tudományos(abb) világban megszokott dokumentáció és a metaadatok hiányosságai. Gondolhatnánk, hogy az automatikusan generált adatok „tisztábbak" mint a hagyományos, számtalan emberi (kutatói, adatfelvevői) manuális beavatkozást is igénylő adatforrások. A helyzet azonban nem feltétlenül ez: a Big Data is „koszos". Az általunk elemzett iWiW adatbázisában például sok százezer olyan felhasználó volt, akik vélhetően valamilyen hiba, vagy a szolgáltatás lényegét szándékosan meghekkelő „spammerek" és robotok regisztrációja során ke-

rültek az adatbázisba. Egy ilyen probléma összességében könnyen kezelhető. Ugyanakkor a Twitter esetében egyre nehezebb szétválasztani az automatizált „botok" által generált üzeneteket, illetve ezek automatizált megosztóit a valódi felhasználóktól (Vosoughi et al. 2018). Ez a jelenség a Twitter-adatokat használó kutatások számára összességében inkább egy „bosszantó" probléma, mindeközben maga a jelenség az információs propaganda egyik fő eszköze lett (Lazer et al. 2018). Az ilyen típusú kockázatok kiküszöbölésére egyelőre nincsenek általános receptek, a kutatók felelőssége így leginkább abban rejlik, hogy a lehető legtöbb információt begyűjtsék az adatok keletkezésének körülményeiről, továbbá tisztában legyenek a vonatkozó adatforrás esetében szükséges adattisztítási eljárásokkal.

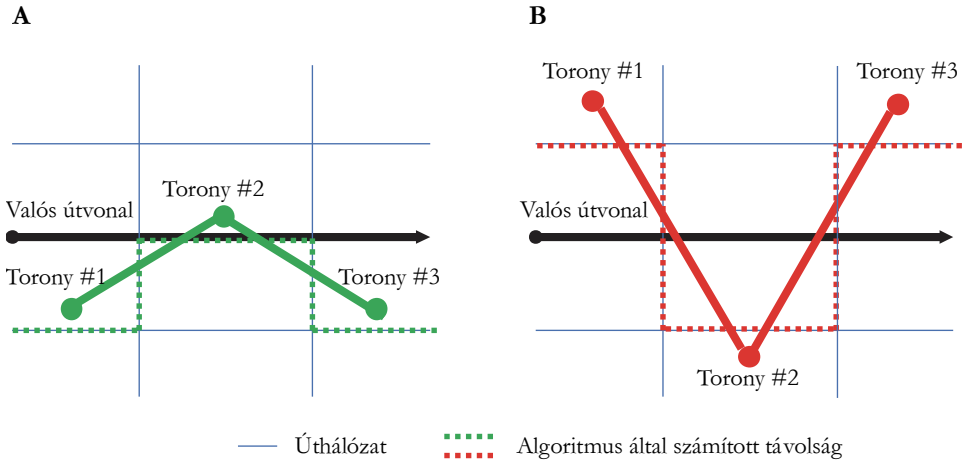
A térbeli adatok pontatlanságának speciális esetét jelentik a mobilhálózatokból származó cellainformációk. A tornyok pontos földrajzi koordinátái ismertek ugyan (ezeket a mobilszolgáltatók nem feltétlenül adják ki), de azt, hogy egy-egy mobilkészülék pontosan hol tartózkodik, csak tág határok között lehet meghatározni. Ez pedig számos problémát vet fel, ugyanis az algoritmusok által számolt utak nyilvánvalóan eltérnek a valóságostól. Túl- és alulbecslés éppúgy előfordulhat, attól függően, hogy milyen a tornyok egymáshoz viszonyított elhelyezkedése, és milyen algoritmus alapján jön létre a pontszerű adatokból számolt útvonal (Kwan 2016).

A 2. ábrán ez a bizonytalansági tényező látható: A és B esetben is ugyanaz a valóságban megtett út, mégis, a tornyok elhelyezkedéséből adódóan jelentősen eltér a zöld, illetve a piros pontozott vonallal jelzett kalkulált távolság.

2. ábra

A megtett távolság algoritmusalapú mérésének torzítása cellainformációk alapján, 2016

Distortion of algorithm-based measurements of the distance taken, based on cell information, 2016



Forrás: Kwan (2016) alapján saját szerkesztés.

Hozzáférés az adatokhoz és a személyes adatok védelme

A Big Data vonatkozásában az egyik legfontosabb, sokszor említett alapvetés, hogy a keletkezése általában független a későbbi tudományos kutatástól. A kormányzatok és állami intézmények saját működésükhöz kapcsolódóan generálják és gyűjtik azokat, a vállalatok esetében pedig az adat az üzleti tevékenységük része, annak stratégiai alapja. Ezeknek az adatforrásoknak jelenleg nincsen és várhatóan nem is lesz formális, egyértelmű szabályok alapján működő piaca. Ez a kapcsolatrendszer nem hasonlítható ahhoz az egyértelmű megrendelő/vállalkozó viszonyhoz, amikor például egy kutató megbízást ad az adatfelvételt végző cégnek survey-típusú terepmunkára, vagy interjúk, fókuszcsoportos beszélgetések lebonyolítására. A Big Data esetében a kifejezetten tudományos célú hozzájutás az egyéni tudományos reputáción, az intézményi beágyazottságon, az akadémiai világon kívüli kapcsolatrendszeren és leginkább a szerencsén múlik. Ajánlatos a nagyvállalati és a bürokratikus szervezeti kultúrákban (annak sajátos nyelvhasználatában) való jártasság, a megfelelő tárgyalási készségek, illetve annak felismerése, hogy mik lehetnek ezeknek a szervezeteknek az igényei, érdekei és működésük külső korlátjai (például azok a keretek, amelyeket az anyavállalat nemzetközileg kijelöl). Ezen a területen a társadalomtudományokat művelőknek sokat kell még tanulniuk, fejlődniük annak érdekében, hogy képesek legyenek meggyőzni az adatok tulajdonosait arról, hogy a tervezett kutatásnak van értelme, az adatokat megfelelő módon használják fel, nem sérülnek az üzleti érdekek, tehát az együttműködés mindkét fél számára előnyös lehet.

Egy adatgazda vállalat vagy kormányzati intézmény számára adataik kiadása inkább kockázatot, illetve többletmunkát jelent, mint a tudományos nyilvánosságban számukra nem feltétlenül vágyott megdicsőülés vagy közvetlen anyagi bevétel. Az „*osszátok meg az adataitokat velünk, mert abból mi nagyon jó kutatásokat tudunk csinálni*" típusú kutatói hozzáállás mellett/helyett leginkább eredményesnek az a stratégia tűnik, amelyben a kutatók megpróbálnak valamilyen formában *know-how*-t, konkrét elemzéseket, üzletileg közvetlenül is hasznosítható tudást visszaadni az adatok tulajdonosainak. Az ilyen típusú bartermegállapodások adminisztratív terhei is általában kisebbek, hiszen konkrét ki- vagy befizetések nélkül mindkét fél a saját szféráján belül képes eredményt felmutatni.

Mindezek alapján talán nem túlzás azt állítani, hogy a Big Data újfajta választóvonalakat hoz létre „adatgazdagok" és „adatszegények" között (Boyd–Crawford 2012, Ságvári 2017a, 2017c). Itt nem csak arról az aszimmetrikus viszonyról van szó, amely az adatok tulajdonosai (akik gyűjtik, tárolják, „kibányásszák" és elemzik azt) és az adatgyűjtés tárgyai (például a felhasználók) között létezik (Andrejevic 2014), hanem azokról a régi-új választóvonalakról, amelyek kutatók és kutatócsoportok, a vállalatok és az akadémiai világ vagy éppen országok között jönnek létre, attól függően, hogy fizikailag mennyire közel, illetve távol vannak az adatoktól, pontosabban annak tulajdonosaitól. Ez az új helyzethez való alkalmazkodást éppúgy igényli, mint

a kreatív stratégiákat arra vonatkozóan, hogy miként lehet az „adatszegények csoportjából átlépni az adatgazdagok közé".

A Big Data alapvetően két forrásból származhat: valamilyen üzleti tevékenység, vagy valamilyen állami intézmény működése során keletkezik. Emiatt jó eséllyel érzékeny, személyes adatokat fog tartalmazni. Az adatvédelemmel kapcsolatos jogi szabályozás, a kutatások etikai elvei és az adatgazdák jól felfogott érdeke, hogy az adatok csak anonimizált, konkrét személyek azonosítására nem alkalmas formában kerüljenek a kutatókhoz. Az engedélyezési folyamat egy nagyobb szervezet esetében olyan soklépcsős és sokszereplős, melynek eredményeként ez a feltétel közvetlenül nagy valószínűséggel teljesülni fog. Azt azonban már nagyon nehéz teljes biztonsággal garantálni, hogy más, nyilvánosan elérhető adatok és kifinomult statisztikai eljárások segítségével ne lehessen ezeket részben vagy teljesen visszafejteni. Salganik (2017) a Netflix Prize 2006-os esetét említi, amikor 500 ezer felhasználó 100 millió mozifilmre adott értékelésének előzetesen gondosan anonimizált adatbázisát két héten belül sikerült úgy visszaalakítani, hogy abból konkrét személyekre vonatkozó adatokhoz jutottak (Narayanan–Shmatikov 2008). A 2018 májusában életbe lépett uniós adatvédelmi rendelet, a Cambridge Analytica személyes Facebook-adatokkal kapcsolatos botránya pedig világszerte még inkább óvatossá tette az adatgazdákat a tudományos kutatókkal való együttműködésekben.

Az adatok megszerzésének mindezen körülményei egy további problémát is felvetnek. Az empirikus társadalomtudományok művelőinek gondolkodásában talán ma még kevéssé van jelen a kutatások azonos adatokon való reprodukálhatóságának elve. Az egyéni alkuk eredményeként létrejött, titoktartási szerződésekkel szabályozott adatexport esetében ez lényegében lehetetlen. Egy tudományos publikációnál nem lehet elvárás, hogy a szerzők az eredmények kontrollálásának és reprodukálhatóságának érdekében harmadik féllel eredeti formában megosszák nyers adataikat. Egy ilyen lépéssel vélhetően megszegnék az adatok eredeti tulajdonosaival kötött szerződésüket.

Innováció a kutatásban

Minden Big Data adatforrás társadalomtudományos kutatási célú felhasználása kreativitást és módszertani innovációt igényel. Ezek a kutatások jellemzően ma még a felfedező, kísérletező szakaszban vannak, így nincsenek univerzális receptek, „dobozos megoldások". Ezeken belül is vannak olyan próbálkozások, amelyek kifejezetten egyedinek tekinthetők. A következő két kutatás kifinomultabb képfelismerési algoritmusokat alkalmazott a területi szintű társadalmi-gazdasági indikátorok verifikálására és/vagy előállítására, és egyben jól mutatja, hogy melyek lehetnek a kutatások innovatív irányai. Jean és szerzőtársai (2016) kutatását szintén az a tény inspirálta, hogy az afrikai országok társadalmaira vonatkozóan kevés megbízható statisztikai adat áll rendelkezésre. Módszerük lényege egy olyan gépi tanuláson alapuló modell létrehozása volt, amely nagy felbontású éjszakai és nappali műholdfelvételek, illetve

a szűkösen rendelkezésre álló területi statisztikák alapján képes szó szerint feltérképezni egy adott térségben élők fogyasztását és vagyoni helyzetét. A modell a konvolúciós neurális hálózatok (convolutional neural network – CNN) módszerére épült, és egyik fő eredménye, hogy az egyik országon „betanított" modell hatásfoka egy másik országra alkalmazva sem csökken számottevően. Összességében illúzió azt gondolni, hogy az ilyen vagy hasonló eljárások rövid időn belül pontos adatokkal szolgálhatnak a hivatalos statisztika számára. Figyelembe véve, hogy az elemzéshez „csak" az egyre nagyobb felbontásban, a világ minden részéről rendelkezésre álló műholdképek, illetve a megfelelő számítási kapacitás szükségesek, tehát a hagyományos adatgyűjtésekkel összehasonlítva nagyságrendekkel olcsóbb módszerről van szó, igen ígéretes kutatási irányzatnak tekinthető.

Szintén a képfelismerés és gépi tanulás, illetve a rendelkezésre álló hivatalos adatforrások összekapcsolásából jutott eredményekre az a kutatás, amely a Google Street View egyesült államokbeli városokban készített képein felismerhető járművek és számos társadalmi jellemző között keresett kapcsolatot. Ismerve az ottani társadalom és az autó, mint tárgy közötti „bensőséges" viszonyt, ez utóbbi egy olyan proxyváltozó, amely nagyon jól kifejezi a társadalom különböző csoportjainak jövedelmi helyzetét, fogyasztói szokásait és értékválasztásait. Ennek a kutatásnak a hátterében a cél szintén a hatékonyság növelése volt, azaz a drága és időigényes hagyományos, személyes megkérdezésen alapuló adatfelvételek vagy más hivatalos statisztikai adatgyűjtések kiváltása valamilyen kreatív eljárással. Gebru és kollégái (2017) az Egyesült Államok 200 városában készült 50 millió képfelvételt elemeztek, amelyeken összesen 22 millió jármű volt azonosítható (Ez az ország teljes járműállományának 8%-a.). A kutatók létrehoztak egy olyan kiinduló adatbázist, amely több mint 2600 különféle gyártmányú, évjáratú és típusú jármű képeit és pontos adatait tartalmazta. A képfelismerő algoritmus (amely szintén neurális hálózatokon alapult) 0,2 másodperc alatt volt képes egy képen szereplő járművet beazonosítani. (A folyamat alig 2 hétig tartott. Ugyanezen munka elvégzése egy ember munkaidejét és 10 másodperces azonosítási időt alapul véve több mint 15 évig tartott volna.) Az eredmények néhány széles körben ismert sztereotípiát is visszaigazoltak. Például azokban a választási körzetekben, ahol a szedán-típusú járművek voltak többségben, az ott lakók 88%-kal nagyobb valószínűséggel szavaztak demokrata jelöltre. Ezzel ellentétben, ahol a pick-up truckok voltak többségben, ott a republikánus jelölt megválasztására volt nagyobb esély. A különböző autómárkák a lakónegyedek etnikai összetételével is szoros kapcsolatban vannak: az ázsiaiak inkább az ázsiai márkákat (Honda, Toyota), az afroamerikaiak a klasszikus hazai márkákat (Chrysler, Buick, Oldsmobile) részesítik előnyben.

Konklúzió

A tanulmányban bemutatott módszertani problémák és a konkrét kutatási eredmények alapján megfogalmazható dilemmák jól mutatják, hogy a Big Data távolról sem

olyan varázsszer a társadalomtudományok és ezen belül a szociológiai fókuszú társadalomföldrajz számára, amely egyik napról a másikra korábban nem ismert összefüggések feltárását teszi lehetővé, továbbá hagyományos módszertanok alkalmazását feleslegessé és meghaladottá. Ehelyett inkább egy olyan organikus fejlődésnek vagyunk részesei, amelynek számtalan zsákutcája van (és vélhetően lesz is), és az ígéretes kutatási irányok és módszerek is megannyi dilemmát és megválaszolandó elméleti, módszertani, kutatásetikai, jogi, szervezeti stb. kérdést vetnek fel. Nincs „régi" és „új" társadalomtudomány, és ezen belül szociológia. Sőt, úgy tűnik, hogy a két megközelítés egyesítése, a hibrid kutatási módszerek alkalmazása lehet az az út, amely elvezethet az igazi tudományos legitimáció felé. Az ebben részt venni kívánó kutatóktól ez folyamatos és intenzív megújulási készséget, kritikai érzéket, az adatok érzékeny és sokakat érintő személyes mivolta miatt pedig etikus és felelősségteljes hozzáállást kíván. Eközben pedig a szociológia évszázados elméleti hagyományaira gondolva azt sem szabadna elfelejteni, hogy a hatalmas méretű adatbázisok elemi részecskéi – bár egyre kevésbé látszanak – továbbra is emberek, akiknek az érzéseit, motivációit és cselekedeteit meg kívánjuk érteni, és akiknek végső soron, áttételesen segíteni kellene a munkánk eredményeivel.

IRODALOM

ANDREJEVIC, M. (2014): The Big Data Divide *International Journal of Communication* 8 (8): 1673–1689.

BAJI, P. (2017): Okos városok és alrendszereik – Kihívások a jövő városkutatói számára? *Tér és Társadalom* 31 (1): 89–105. https://doi.org/10.17649/TET.31.1.2807

BLUMENSTOCK, J.–CADAMURO, G.–ON, R. (2015): Predicting poverty and wealth from mobile phone metadata *Science* 350 (6264): 1073–1076. https://doi.org/10.1126/science.aac4420

BOYD, D.–CRAWFORD, K. (2012): Critical Questions for Big Data *Information, Communication & Society* 15 (5): 662–679. https://doi.org/10.1080/1369118x.2012.678878

BURRELL, J. (2016): How the machine 'thinks': Understanding opacity in machine learning algorithms *Big Data & Society* 3 (1):1–12. https://doi.org/10.1177/2053951715622512

COLEMAN, J. (1990): *Foundations of Social Theory* Belnap Press, Cambridge, MA.

CONTE, R.–GILBERT, N.–BONELLI, G.–CIOFFI-REVILLA, C.–DEFFUANT, G.–KERTESZ, J.–HELBING, D. (2012): Manifesto of computational social science *The European Physical Journal Special Topics* 214 (1): 325–346. https://doi.org/10.1140/epjst/e2012-01697-8

COSCIA, M.–CHESTON, T.–HASUMANN, R. (2017): Institutions vs. Social Interactions in Driving Economic Convergence: Evidence from Colombia *Working Paper - Center for International Development at Harvard University*(No.331), Cambridge, MA.

COVIELLO, L.–SOHN, Y.–KRAMER, A. D.–MARLOW, C.–FRANCESCHETTI, M.–CHRISTAKIS, N. A.–FOWLER, J. H. (2014): Detecting emotional contagion in massive social networks *PLoS One* 9 (3): e90315. https://doi.org/10.1371/journal.pone.0090315

DESSEWFFY, T.–LÁNG, L. (2015): Big Data és a társadalomtudományok véletlen találkozása a műtőasztalon *Replika* 92-93: 13.

EAGLE, N.–MACY, M.–CLAXTON, R. (2010): Network diversity and economic development *Science* 328 (5981): 1029–1031. https://doi.org/10.1126/science.1186605

EICHSTAEDT, J. C.–SCHWARTZ, H. A.–KERN, M. L.–PARK, G.–LABARTHE, D. R.–MERCHANT, R. M.–SELIGMAN, M. E. (2015): Psychological language on Twitter predicts county-level heart disease mortality *Psychol Sci* 26 (2): 159–169. https://doi.org/10.1177/0956797614557867

FRANKEL, M. R.–FRANKEL, L. R. (1987): Fifty Years of Survey Sampling in the United States *The Public Opinion Quarterly* 51: S127-S138.

GEBRU, T.–KRAUSE, J.–WANG, Y.– CHEN , D.–DENG, J.–LIEBERMAN AIDEN, E.–FEI-FEI, L. (2017): "Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States." PNAS 114(50): 13108-13113.

GICZI, J.–SZŐKE, K. (2017): Hivatalos statisztika és a Big Data *Statisztikai Szemle* 95 (5): 461–490. https://doi.org/10.20311/stat2017.05.hu0461

GOFFMAN, E. (1981): *A hétköznapi élet szociálpszichológiája* Gondolat Könyvkiadó, Budapest.

GONCALVES, B.–LOUREIRO-PORTO, L.–RAMASCO, J. J.–SANCHEZ, D. (2018): Mapping the Americanization of English in space and time *PLoS One* 13 (5): e0197741. https://doi.org/10.1371/journal.pone.0197741

GRANOVETTER, M. (1985): Economic Action and Social Structure: the Problem of Embeddedness *American Journal of Sociology* 91 (3): 481–493.

GROVES, R. M. (1989): *Survey errors and survey costs* Wiley, New York.

JAKOBI, Á. (2014): Újszerű területi statisztikai adatgyűjtési lehetőségek az információs világ egyenlőtlenségeinek kutatásában *Területi Statisztika* 54 (1): 35–52.

JAKOBI, Á.–LŐCSEI, H. (2016): Brand wars in cyberspace: a GIS solution *Regional Statistics* 6 (2): 173–176. https://doi.org/10.15196/RS06209

JEAN, N.–BURKE, M.–XIE, M.–DAVIS, W. M.–LOBELL, D. B.–ERMON, S. (2016): Combining satellite imagery and machine learning to predict poverty *Science* 353 (6301): 790–794. https://doi.org/10.1126/science.aaf7894

JURDAK, R. (2015): "Understanding Human Mobility from Twitter." PLoS One 10(7): e0131469.

KMETTY, Z. (2018): A szociológia helye a Big Data-paradigmában és a Big Data helye a szociológiában *Magyar Tudomány* 179 (5): 683–692. https://doi.org/10.1556/2065.179.2018.5.11

KNACK, S.–KEEFER, P. (1997): Does Social Capital Have an Economic Payoff? A Cross-Country Investigation *The Quarterly Journal of Economics* 112 (4): 1251–1288. https://doi.org/10.1162/003355300555475

KWAN, M.-P. (2016): Algorithmic Geographies: Big Data, Algorithmic Uncertainty, and the Production of Geographic Knowledge *Annals of the American Association of Geographers* 106 (2): 274–282. https://doi.org/10.1080/00045608.2015.1117937

LAZER, D.–BAUM, M. A.–BENKLER, Y.–BERINSKY, A. J.–GREENHILL, K. M.–MENCZER, F.–ZITTRAIN, J. L. (2018): The science of fake news *Science* 359 (6380): 1094–1096. https://doi.org/10.1126/science.aao2998

LAZER, D.–PENTLAND, A.–ADAMIC, L.–ARAL, S.–BARABASI, A. L.–BREWER, D.–VAN ALSTYNE, M. (2009): Social science. Computational social science *Science* 323 (5915): 721–723. https://doi.org/10.1126/science.1167742

LENGYEL, B.–SÁGVÁRI, A. V. B.–JAKOBI, Á.–KERTÉSZ, J. (2016): The geography of the iWiW *Területi Statisztika* 56 (1): 30–45. https://doi.org/10.15196/TS560103

LENGYEL, B.–VARGA, A.–SÁGVÁRI, B.–JAKOBI, A.–KERTÉSZ, J. (2015): Geographies of an Online Social Network *PLoS One* 10 (9): e0137248. https://doi.org/10.1371/journal.pone.0137248

LESZCZYNSKI, A.–CRAMPTON, J. (2016): Introduction: Spatial Big Data and everyday life *Big Data & Society* 3 (2): 1–6. https://doi.org/10.1177/2053951716661366

LLORENTE, A.–GARCIA-HERRANZ, M.–CEBRIAN, M.–MORO, E. (2015): Social media fingerprints of unemployment *PLoS One* 10 (5): e0128692. https://doi.org/10.1371/journal.pone.0128692

MAYER-SCHÖNBERGER, V.–CUKIER, K. (2013): *Big Data: A revolution that will transform how we live, work, and think* Houghton Mifflin, Harcourt.

MILLER, H. J. (2010): The Data Avalanche Is Here. Shouldn't We Be Digging? *Journal of Regional Science* 50 (1): 181–201. https://doi.org/10.1111/j.1467-9787.2009.00641.x

MILLER, H. J.–GOODCHILD, M. F. (2014): Data-driven geography *GeoJournal* 80 (4): 449–461. https://doi.org/10.1007/s10708-014-9602-6

MORSTATTER, F.–PFEFFER, J.–LIU, H. (2014): *When is it biased?: assessing the representativeness of twitter's streaming API* Paper presented at the Proceedings of the 23rd International Conference on World Wide Web, Seoul, Korea.

NAGY, P. T.–VEROSZTA, Z. (2018): Az új adatkezelés lehetőségei és kockázatai a társadalomkutatásban *Magyar Tudomány* 179 (5): 651–652. https://doi.org/10.1556/2065.179.2018.5.9

NARAYANAN, A.–SHMATIKOV, V. (2008): *Robust de-anonymization of large sparse datasets* Paper presented at the Security and Privacy, 2008. SP 2008. IEEE Symposium, Oakland, CA, USA. https://doi.org/10.1109/SP.2008.33

NÉMETH, R. (2015): A számok tényleg magukért beszélnek? *Replika 92–93 (2015/3-4.)*: 203–208.

NORBUTAS, L.–CORTEN, R. (2018): Network structure and economic prosperity in municipalities: A large-scale test of social capital theory using social media data *Social Networks* 52: 120–134. https://doi.org/10.1016/j.socnet.2017.06.002

PASQUALE, F. (2015): *The black box society : the secret algorithms that control money and information* Harvard University Press, Cambridge.

PUTNAM, R. D. (2000): *Bowling alone : the collapse and revival of American community* Simon & Schuster, New York.

SÁGVÁRI, B. (2017a): The Computational Turn in Social Sciences. Challenges of the New Empiricism in the Age of Big Data *Intersections. East European Journal of Society and Politics* 3 (1): 5–14. https://doi.org/10.17356/ieejsp.v3i1.348

SÁGVÁRI, B. (2017b): Diszkrimináció, átláthatóság és ellenőrizhetőség. Bevezetés az algoritmus-etikába *Replika 103 (*2017/3): 61–79.

SÁGVÁRI, B. (2017c): Társadalomtudomány a Big Data korában *Statisztikai Szemle* 95 (2017/5): 491–504. https://doi.org/10.20311/stat2017.05.hu0491

SALGANIK, M. J. (2017): *Bit by Bit: Social Research in the Digital Age* Princeton University Press, Princeton, New Jersey.

SANDVIG, C.–HAMILTON, K.–KARAHALIOS, K.–LANGBORT, C. (2014): *Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms* Paper presented at the 64th Annual Meeting of the International Communication Association, Seattle, WA, USA.

ŠĆEPANOVIĆ, S.–MISHKOVSKI, I.–GONÇALVES, B.–NGUYEN, T. H.–HUI, P. (2017): Semantic homophily in online communication: Evidence from Twitter *Online Social Networks and Media 2*: 1–18. https://doi.org/10.1016/j.osnem.2017.06.001

SQUIRE, P. (1988): Why the 1936 Literary Digest Poll Failed? *Public Opinion Quarterly* 52 (1): 125–133. https://doi.org/10.1086/269085

STEPHENS, M. (2013): Gender and the GeoWeb: divisions in the production of user-generated cartographic information *GeoJournal* 78 (6): 981–996. https://doi.org/10.1007/s10708-013-9492-z

SZÉKELY, I. (2015): Az adatmentes zónák szükségessége és esélye *Replika* 92–93 *(2016/3–4):* 209–225.

UGANDER, J.–KARRER, B.–BACKSTROM, L.–MARLOW, C. (2011): The anatomy of the Facebook social graph *arXiv preprint arXiv:*1111: 4503.

VOSOUGHI, S.–ROY, D.–ARAL, S. (2018). The spread of true and false news online *Science* 359 (6380): 1146–1151. https://doi.org/10.1126/science.aap9559

VUKOVICH, G. (2015): Adatforradalom és hivatalos statisztika *Statisztikai Szemle* 93 (2015/8–9): 745–758.

WELLER, K.–BRUNS, A.–BURGESS, J.–MAHRT, M.–PUSCHMANN, C. (Eds.). (2013): *Twitter and Society* Peter Lang, New York.

Z. KARVALICS, L. (2018): Nagy adat és digitális történelem: egy izgalmas házasság múltja, jelene és jövője *Magyar Tudomány* 179 (5): 668–682. https://doi.org/10.1556/2065.179.2018.5.10