Területi Statisztika

Közzététel: 2019. február 12.

A tanulmány címe:

Big Spatial Data: lehetőségek, kihívások és tapasztalatok

Szerző:

Jakobi Ákos ELTE Regionális Tudományi Tanszék**,** E-mail: jakobi@elte.hu

https://doi.org/10.15196/TS590101

TANULMÁNYOK

Big Spatial Data: lehetőségek, kihívások és tapasztalatok*

Big Spatial Data: chances, challenges and experiences

Jakobi Ákos
ELTE Regionális Tudományi Tanszék
E-mail: jakobi@elte.hu

Az utóbbi évek egyértelmű trendje, hogy a kvantitatív elemzések már nemcsak a tradicionális statisztikai adatforrásokra építenek, hanem azokra a nagy mennyiségben, sebességben és változatosságban keletkező új adathalmazokra is, amelyeket a szakma átfogóan csak Big Data állományoknak nevez. Ezeket a forrásokat a területi kutatók egyre intenzívebben használják, ám ennek ellenére még korántsem kiforrott a Big Spatial Data állományok, azaz a térbeli tartalommal bíró és hatalmas mennyiségben keletkező adatok kezelésének és feldolgozásának módszertana. A tanulmány célja a Big Spatial Data állományokkal kapcsolatos fogalmak és lehetőségek rendszerezése, valamint a kapcsolódó kihívások és kritikák áttekintése. Az elméleti részek után a tanulmány egy ismert közösségi oldal adatainak, egy magyarországi bank geolokalizált tranzakciós adatainak, továbbá egy mobilszenzoros nyomkövető-alkalmazás adatainak gyakorlati elemzésével ismerteti a Big Spatial Data állományok analitikai lehetőségeinek és kihívásainak részleteit.

Kulcsszavak:
Big Data,
Big Spatial Data,
Twitter,
területi mintázat,
Budapest

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

Területi Statisztika, 2019, 59(1): 3–26; DOI: 10.15196/TS590101

It's a clear trend in recent years that quantitative analyses are based not only on traditional statistical data sources, but also on new data sets which are being generated in large volumes, speed and variety, and which are comprehensively called by professionals as Big Data. Such resources are increasingly used by the spatial researcher community, although the methodology for handling and processing Big Spatial Data, that is, spatial data which are generated in large volumes, has not been fully established yet. The aim of the study is therefore to systematise the concepts and opportunities related to Big Spatial Data, as well as to overview the parallel existing challenges and critics. After the theoretical parts, the article attempts to point out some details of the possibilities and challenges of Big Spatial Data analysis by introducing practical examples of analysing the data of a well-known social network site, geolocalized transaction data of a Hungarian bank and data of a tracking application.

Keywords:
Big Data,
Big Spatial Data,
Twitter,
spatial pattern,
Budapest

Beküldve: 2018. július 23.
Elfogadva: 2018. szeptember 13.

Adatrobbanás a térelemzésekben

Szakmai körökben a Big Data (szabad fordításban „óriási adathalmaz") néven ismert fogalom arra a hatalmas adatmennyiségre utal, amely információs világunkban gyorsan és folyamatosan keletkezik, s melynek feldolgozása a hagyományos kapacitásokkal és módszerekkel szinte megoldhatatlan kihívást jelent. A Big Data ennek ellenére mégis nagy lehetőségeket kínál. A sokáig csak virtuális melléktermékként számon tartott napi információhalmaz akkor válik értékessé, amikor a különböző adatokat összekötjük, összefüggéseket, felismerhető mintázatokat találunk közöttük, s ebből értékelhető következtetéseket vonunk le. A kormányzati szerveknél, az internetes és telekommunikációs cégeknél vagy egyéb helyeken összegyűlt hatalmas adatmennyiség a társadalom és a gazdaság folyamatainak vizsgálata számára valóságos aranybánya. Közlekedési, vásárlási, szabadidős vagy egyéb mindennapi szokásainkról korlátlanul gyűjtenek adatokat a különféle szervezetek. Mindezek az emberi viselkedés egyedi és csoportos (társadalmi) szintjeibe is betekintést nyújtanak. Az egyre szélesebb körben terjedő, térbeli információkat is használó vagy azokat spontán módon generáló alkalmazások, digitális eszközök és webes szolgáltatások révén a Big Data a társadalom térbeli mű-

ködésének megértéséhez vagy feltérképezéséhez is nagy mennyiségben kínál adatokat, s ezek az információk a területi döntések megalapozását is segíthetik.

A technológiai átalakulás, az új információs és kommunikációs technológiák fejlődése, a digitális eszközök széles körű megfizethetősége egyaránt hozzájárul ahhoz, hogy beindulhasson a Big Data néven ismertté vált adatrobbanás akár a mennyiség, akár a diverzitás vonatkozásában. Egyes értelmezések szerint ez az adatok ipari forradalma (Hellerstein 2008), melynek hatásaként az emberi viselkedés megértésének eddig soha nem látott történelmi korszakába léphetünk (Onnela 2011). Ráadásul ezek az adatok és adatforrások új lehetőséget is hoztak magukkal a társadalmi, a gazdasági vagy az egyéb szakpolitikai döntések támogatásához. A megszámlálhatatlan mennyiségű adat érdemleges információhalmazokba rendezése és a trendek, mintázatok kiszűrése vagy meghatározása jelentős számítási kapacitásokat igényel, melyek csak újabban állnak rendelkezésünkre. A kapott eredmények ugyanakkor új megvilágításba helyezhetik az eddigi hivatalos adatgyűjtési forrásainkat, sőt kiegészíthetik azokat az emberi viselkedés kutatásának mélyebb vagy részletesebb motívumaival.

Napjaink rohamos információtechnológiai fejlődése olyan szolgáltatásokat, alkalmazásokat hívott életre, amelyek a területi kutatás számára is új eszközöket és lehetőségeket kínálnak. Utóbbiak leginkább abból adódnak, hogy az információs és kommunikációs technológiák, illetve a számítógépes megoldások immáron szinte az élet minden részét áthatották. Az angol nyelvű szakirodalomban a „pervasive computing" vagy „ubiquitous computing" fogalmak (Satyanarayanan 2001, Friedewald–Raabe 2011, Weiser 1991) hátterében a térbeli információseszközhasználat terjedése és az információrobbanás is megfigyelhető (Jiang–Yao 2006, Galloway 2004, Zook et al. 2004).

Ezek az adatok manapság úgyszólván rázúdulnak a területi kérdésekkel foglalkozó szakemberekre. A világhálón közzétett strukturált vagy strukturálatlan térbeli információtartalom, illetve a széles körben terjedő, térbeli információkat is használó alkalmazások révén a Big Data a társadalom térbeli működésének megértéséhez is új forrásokat kínál. Terjedőben van a Big Data fogalomkörön belül a külön térbeli tartalommal bíró Big Spatial Data (Ivan et al. 2017) is. Ez a térbeli adatok hatalmasra duzzadt halmaza, ellentétben a Spatial Big Data fogalmával (Thatcher 2014), amely többnyire a nem térinformációs hátterű Big Data források térbeli tartalmi kiegészítését jelenti. E fogalmak között a különbség árnyalatnyi, így akár egymás szinonimái is lehetnek.

Amiket a Big Spatial Data környezetben a területi kutatói és tervezői kör hasznosan vizsgálhat, azok a térbeli tartalommal is rendelkező direkt vagy indirekt digitális nyomok. Az ilyen adatokra épülő adatbázisok közvetlen módon (például az okostelefonok különböző helyalkalmazásaihoz kötődően), vagy egyes honlapok földrajzi azonosító kódokkal (geotag) kibővített közzétételekor keletkeznek. Ennél érdekesebbek a geoinformációkat tartalmazó digitális nyomok indirekt halmazai,

melyek nem szándékosan, de mégis nagy számban jönnek létre. Példaként említhetők azok az elektronikus közlekedési kártyák vagy megfigyelő rendszerek, amelyek rögzítik a közlekedési rendszerbe való belépés és kilépés helyét, illetve idejét, lehetőséget adva – elméletileg – a közlekedési térpályák, szokások stb. vizsgálatára. Digitális nyomokat hagyunk továbbá akkor is, amikor egy-egy weboldalt meglátogatunk. Általában beazonosítható az IP-cím, s ezáltal az a földrajzi hely is, ahonnan a világháló szolgáltatásait igénybe vesszük. A digitális nyomok indirekt felhasználására, elemzésére számos példa említhető (Girardin et al. 2009, Järv et al. 2012, Naaman 2011, Jakobi–Lőcsei 2016), melyek a „melléktermékként" keletkező digitális adatok vizsgálatával hozzák meg a következtetéseiket.

Az új infokommunikációs eszközök jelentős része nemcsak információtovábbításra, de területi adatok gyűjtésére is alkalmas, sőt a felhasználók egyre nagyobb hányada keresi a térbeli információkat is kínáló alkalmazásokat. Az okostelefonok azok számára is mindennapossá tették a térérzékeny adatok használatát, illetve annak lehetőségét, akik korábban nem érdeklődtek iránt uk. Ezek a készülékek általában beépített helymeghatározó alkalmazásaikon keresztül aktív téradathasználatot kínálnak, közelebb hozva ezzel a felhasználókhoz a Globális Helymeghatározó Rendszer *(Global Positioning System* – GPS-) technológiák nyújtotta lehetőségeket. Az okostelefonok és a hasonló eszközök, alkalmazások révén nemcsak a térbeli adatokhoz való hozzáférés lehetősége változott meg. Az alkalmazói kör milliónyi digitális nyomot hagy maga után, melyek jelentős része földrajzi tartalmú. Ezek mellett a területi kutatói és tervezői szakma sem mehet el szótlanul.

Jelen tanulmány először a Big Data fogalomkör sajátosságainak – különösen a térbeli adatokra és azok elemzésére vonatkozó lehetőségeinek és kihívásainak – ismertetésére vállalkozik, majd példák segítségével értékeli az egyes adattípusok használatával kapcsolatos ambivalens tapasztalatokat. Arra számíthatunk, hogy a Big Spatial Data környezet nemcsak pozitív hatásokkal jár, hanem számos új problémát és megoldandó kérdést is felvet.

A Big Data fogalma még nem teljesen tisztázott, bár összefoglaló szakmai kísérletekről már olvashattunk (Szűts–Yoo 2016). Egyesek – főként a társadalomtudósok – gyakorta pusztán hatalmas adatfájlokként értelmezik. Az informatikusok azonban másként vélekednek: ők az adatáramlásra és a folyamatgenerált adathalmazokra gondolnak. Sarkosan fogalmazva, míg korábban csak relatíve kis mennyiségű analóg adat keletkezett és vált elérhetővé a korlátozott számú csatornán keresztül, addig ma rendszeresen nagy mennyiségű digitális adat jön létre különböző csatornákon keresztül, minden percben (King 2011). Egyrészt az adatok keletkezésének, átvitelének sebessége és gyakorisága, másrészt a források száma és sokszínűsége az, ami az adatok özönvízszerű áramlását jellemzi (Giczi–Szőke 2017).

A Big Data forrásokat általában három jellemző alapján különíthetjük el az egyéb adatforrásoktól: ez a nagy adatmennyiség, a nagy változatosság és a nagy sebesség (az angol nyelvű szakirodalomban „3V", azaz volume, variety, velocity)

(Zikopoulos–Eaton 2011, Gartner 2011). Az adatmennyiség nagysága ma már vitathatatlan, eddig fel sem merült nagyságrendek (az exabyte, a zettabyte vagy a yottabyte) kerültek be az informatika mindennapi szóhasználatába. A Big Data e tulajdonságát sok szerző hangsúlyozza, vagy egyszerűsíti le az adatok óriási halmazára (McGuire et al. 2012). Loukides (2010) szerint a Big Data esetében maga a méret válik a probléma lényegévé. Hasonlóan fogalmaz a McKinsey Global Institute (2011) tanulmánya is, mely szerint ezen adathalmazok mérete meghaladja azt, amit a tipikus adatkezelő szoftverek még kezelni tudnak. Ezt emeli ki továbbá Dumbill (2012) is, aki szerint a Big Data túl nagy, túl gyorsan nő és nem illeszthető a korábbi konvencionális adatbázis-architektúrához, azaz feldolgozásához alternatív eljárások kidolgozására van szükség. A Big Data túl nagy ahhoz, hogy mozgatni lehessen (a hálózati kapacitások miatt is), sőt, a Big Data állományokat gyakran nem is tárolják, csak adatfolyamként tekintenek rájuk, melyekből releváns információk nyerhetők ki (Schiller–Burghardt 2015).

Az adatok változatossága is szerteágazóvá vált, ha csak arra gondolunk, hogy hányféle forrásból származnak az újszerű adatok: szenzorokból, közösségi médiából, digitális tartalmakból, tranzakciós adatokból vagy a mobiltelefonokba épített GPS-ek adataiból stb. A Big Data állományait így gyakran hatalmas, többnyire strukturálatlan adathalmaznak is tekintik. A múltban az értékes információknak a strukturálatlan adatokból való kinyerése nagyon munkaigényes volt, a Big Data környezetben viszont mindez automatizáltan és relatíve költséghatékonyan valósulhat meg. Becslések szerint a keletkező információk körülbelül 15%-a strukturált, azaz hagyományos oszlopokból és sorokból álló adattáblákba rendezhető, relációs adatbázisokban tárolható, 85%-a viszont e-mailekben, blogokban, közösségi médiában vagy egyéb helyeken keletkező strukturálatlan adat és információ (TechAmerica Foundation 2012).

Az adatok keletkezésének sebessége is radikálisan felgyorsult. A korábbiakkal ellentétben ma már valós időben, de legalábbis rendkívül hamar elemezhetjük a vizsgált dolgok adatszerű jellemzőit. A sebesség azonban nemcsak az adatok keletkezésére, de azok elérésére, feldolgozására és elemzésére is vonatkozik. Jó példa erre a Massachusettsi Műszaki Egyetem (*Massachusetts Institute of Technology* – MIT) fogyasztóiárindex-becslő kísérlete (Billion Price Project, lásd Cavallo–Rigobon 2016), amelyben a webről gyűjtött napi félmillió online áradatból következtettek a valós idejű árindexek alakulására, sokkal gyorsabban, mint a hivatalos statisztikák.

A korábban említett lényeges jellemzőkön túl a Big Data adatforrások más fontos tulajdonságokkal is rendelkeznek (UN Global Pulse 2012). A Big Data digitálisan keletkezik (nem az analóg adatok manuális digitalizációjával), s csak számítógépes környezetben, digitális formában értékelhető. A Big Data adatok passzív módon termelődnek a mindennapi digitális interakciók melléktermékeként. Adatgyűjtésük automatikusan történik, többnyire nem célzott, vagy direkt adatgyűjtések módján, bár Kitchin (2013) más forrásformákat is említ. Az adatok elemzése folyamatosan,

akár az adatgyűjtéssel párhuzamosan zajlik, és nem határozható meg az adatgenerációs folyamat végső időpontja.

Ezen tulajdonságok azonban alapvetően technológiai jellegűek, melyek az adattárolás és -feldolgozás fejlődésétől függenek. A Big Data egy további jellemzője viszont más karakterű: ez pedig az érték (value), ami ahhoz a növekvő társadalmi-gazdasági hasznossághoz kapcsolódik, amit Big Data források kihasználásával érhetünk el. Nem egy szerző állítja, hogy a Big Data akár új termelési tényezőként is felfogható századunkban (Gentile 2011, Jones 2012). Nem meglepő, hogy a munkaerőpiacon az egyik legnagyobb igény éppen a Big Data lehetőségeit feltáró és kiaknázó adatbányász (data miner) szakértők, továbbá az adattudósok (data scientists) iránt mutatkozik. Ez utóbbi korunk talán legvonzóbb szakmájává kezd válni (Davenport–Patil 2012).

Big Data források, avagy a területi kutatások új lehetőségei

Nincs egyszerű dolgunk a Big Data források csoportosítása során. Általánosságban elkülöníthetők például az adminisztratív eredetűek, melyek forrása lehet állami vagy egyéb intézményi regiszter (elektronikus egészségügyi nyilvántartások, kórházi látogatások, biztosítási nyilvántartások, iskolai adatok, banki adatok). Ezek célja a saját munkafolyamatok támogatása, monitorozása, valamint a munkaprogram végrehajtása. Lehetnek kereskedelmi vagy tranzakciós eredetűek, melyek két entitás közötti tranzakcióból származnak (bankkártya-tranzakciók, online tranzakciók, mobiltelefonos fizetések), s az előző csoporthoz hasonlóan az emberi viselkedés közvetett szenzoraiként hasznosíthatók. Lehetnek azonban közvetlen fizikai szenzoros eredetűek is (műholdképek, forgalomfigyelők, időjárás-figyelők adatai) vagy akár nyomkövető eszközökből származóak (útvonal-követési adatok mobiltelefonokból, GPS eszközökből). Említhetők továbbá az online tartalmakból kiolvasott információk, melyek egyrészt az emberek viselkedéséről tájékoztatnak (honlaplátogatottság, online keresések termékekre, szolgáltatásokra vagy egyéb más jellegű információkra), vagy épp a véleményükről (hozzászólások a közösségi médiában) adnak információkat (Mag 2014). Sőt, hasznos Big Data források lehetnek azok is, amelyekben maguk az emberek az adatszolgáltatók (közösségi adatforrások, felhasználói térképek), ezek az információk bár nem passzív eredetűek, de jó korrekciós forrásai a tartalmaknak.

Az ENSZ Európai Gazdasági Bizottságának (2014) statisztikai munkacsoportja a Big Data források következő osztályozását javasolta:

1. Humáneredetű információk, avagy „people to people" típusú vagy „humán szenzor" adatok, jellemzően a közösségi oldalak adatai. Ezek az adatok lazán strukturáltak és gyakran irányítási kontroll nélküliek.

1100. Közösségi oldalak: Facebook, Twitter, Tumblr stb.

1200. Blogok, hozzászólások

1300. Személyes dokumentumok

1400. Képek: Instagram, Flickr, Picasa stb.
1500. Videók: Youtube stb.
1600. Internetes keresések
1700. Mobiladat-tartalom: szöveges üzenetek
1800. Felhasználó által generált térképek
1900. E-mail

2. Folyamat által közvetített adatok, avagy „people to machine" típusú adatok, jellemzően az üzleti folyamatok adatai, melyeket határozott strukturáltság jellemez, kapcsolati táblákkal, metaadatokkal.
21. Közhivatalok által szolgáltatott adatok
2110. Orvosi/egészségügyi nyilvántartások
22. Kereskedelem által létrehozott adatok
2210. Kereskedelmi tranzakciók
2220. Bank-/készletnyilvántartás
2230. E-kereskedelem
2240. Bankkártya/hitelkártya

3. Automatikus rendszerek adatai, azaz gépek által közvetített „machine to machine" típusú adatok, melyek alapvetően a fizikai világ megfigyeléséből származnak és jól strukturáltak, ám méretük és keletkezési sebességük a tradicionális megközelítéseken is túlmutat.
31. Szenzoradatok
311. Rögzített szenzorok
3111. Otthonautomatizálás
3112. Időjárási, szennyezési szenzorok
3113. Közlekedési szenzorok, webkamerák
3114. Tudományos célú szenzorok
3115. Biztonsági videók, képek
312. Mobilszenzorok (nyomkövetés)
3121. Személyes (mobil-helymeghatározás)
3122. Közúti (autók, teherszállítás)
3123. Vasúti (vonatok)
3124. Légi (repülőgépek)
3125. Vízi (hajók)
313. Műholdas adatok
3131. Topográfiai
3132. Hőmérsékleti
3133. Megfigyelési
3134. Meteorológiai
3135. Egyéb

32. Számítógépes rendszerekből származó adatok
    3210. Naplók (logok)
    3220. Webes naplók (web logok)

Látható, hogy mennyire széles körben jutnak el hozzánk a Big Data korszak adatai, s belegondolhatunk – bár nehéz teljesen átlátni – hogy ezek mögött milyen sokrétű lehetőségek kínálkoznak. Az információs társadalomfejlődés egyik első lépése az 1990-es évek elején az internet közcélú kommunikációs és kereskedelmi lehetőségeinek megjelenése volt, ami egyfajta adat- és információrobbanáshoz vezetett. Ma már szinte bármi hozzákapcsolódhat ehhez a szövevényhez (IoT, internet of things, dolgok internete), ami egyértelműen a bitek és az adatok folyamatos áramlásának bővülését eredményezi és teszi még ezt hosszú ideig.

Természetesen az említett adatforrások, illetve adatgyűjtési formák áttételesen az információs kor területi differenciáltságáról is tájékoztatást adhatnak. Már az is, hogy hol, milyen térségekben keletkezik ez a hatalmas adatmennyiség, informálhat minket az egyes területek virtuális világbeli jelenlétének abszolút súlyáról. A Big Data felhőjében kirajzolódó digitális lábnyomok térségenkénti nagysága, minőségi eltérései az információhasználó lakosság területi jellemzőiről szolgáltatnak adalékokat vagy a különböző térségi szintek folyamatainak megértéséhez adhatnak támpontokat.

Kihívások és kritikák

A lehetőségek ellenére a Big Data források használata rengeteg megoldandó feladatot, továbbá új kérdéseket és kihívásokat von maga után. Ezek az adatok nem készen kapott, letölthető Excel táblázatok, melyeket a jól ismert rutinnal könnyen feldolgozhatunk. Az új kihívások a nagy volumenből, a változatosságból és a nagy keletkezési sebességből adódnak, de egyéb vonatkozásokban is előkerülnek.

A szakértők a Big Data használatával összefüggésben elsők között említik a magánélet, a bizalmasság és az adatvédelem kihívásait. Az egyre szélesebb körben megfigyelt társadalom (surveillance society) nem egy társadalmi ellenreakciót indukált eddig is (Lyon 2001, 2014, Raley 2013), s ezek a problémák az adatszerű megfigyelések kapcsán sem elhanyagolhatók. Bizonyos értelmezés szerint a magánjelleg nem más, mint az egyének joga ahhoz, hogy kontrollálhassák, milyen információk kerülhetnek ki róluk (UN Global Pulse 2012). A Big Data adathalmazokban azonban az egyéneknek kevés esélyük van saját adataik felülvizsgálatára. Megdöbbentő az az individuális szintű információs vagyon, amelyet a Google, a Facebook, a mobilszolgáltatók vagy a hitelkártyarendszereket üzemeltető vállalatok együttesen birtokolnak. A Big Data források használatával kapcsolatos jogszabályi környezet még nincs teljesen stabilizálva, bár az Európai Unió már tett lépéseket ebbe az irányba: az általános adatvédelmi rendelet (General Data Protection Regulation – GDPR) részletesen foglalkozik ezekkel a kérdésekkel.

Más típusú kihívásokat vet fel az efféle forrásokhoz való hozzáférés, az adatok elérésének és megosztásának kérdése. A nyilvánosan hozzáférhető online adatforrások (az „open web" adatai) legyűjthetők, ha a megfelelő technikai feltételek adottak, de hatalmas potenciál rejlik azokban az adatokban is, melyek vállalati kézben vannak (lásd például Facebook belső információk). A vállalatok ezeket saját céljaikra fel is használják (legalábbis részben), nyilvános hozzáférésük viszont általában nem megoldott, vagy üzleti érdekeken alapszik, de ha hozzáférhető is az adat, a mennyiség okán is jelentős költségekbe kerülhet egy Big Data állomány összeállítása.

Technológiai értelemben kihívást jelent maga az adatgyűjtési módszer is. Magas fokú informatikai tudás, nagy háttérkapacitás és sok idő szükséges az adatgyűjtő rendszerek kiépítéséhez. Az adatgyűjtés és az adatbányászat ezért olyan eszközöket és módszereket fejlesztett ki, mint a webscraping, a webharvesting, a crawler robotok, vagy egyéb automatikus adatgyűjtő eljárások használata. A költséges megoldások mellett nagy az olyan relatíve költséghatékony eljárás, mint a közösségi adatgyűjtés (participatory sensing), az alacsony költségű érzékelők (low cost sensors) vagy a másodlagos adatgyűjtési módszerek (webtartalom-elemzés) alkalmazása.

Ha sikerült adatokat gyűjtenünk és a Big Data állomány összeállt, maguknak az adatoknak az elemzése és értékelése is nagy kihívást jelent. Tisztázandó, hogy mit mondhat nekünk valójában az adott adatforrás, milyen következtetések levonására alkalmas és milyenekre nem. Bármennyire is kínálkoznak az újabb és újabb kérdések megválaszolásának esélyei, ezeket a forrásokat sem tekinthetjük csodaszereknek. Gondos vizsgálatok szükségesek annak eldöntésére, hogy a passzív úton előállt és nem direkt módon létrehozott adathalmaz mennyire tükrözi a vizsgált jelenség valódi sajátosságait. Például annak esélye, hogy az idősek is ugyanolyan arányban használják a Google internetes keresőjét, mint a fiatalabb „net-generációhoz" tartozók, vélhetően jóval alacsonyabb, így a keresési eredményekből leszűrt következtetések rájuk nézve nem, vagy csak kevésbé reprezentatívak. A Twitter sem az összes ember véleményét tükrözi, így hiba lenne az „emberek" és a „Twitter-használók" fogalmát szinonimaként kezelni (Boyd–Crawford 2012). Ez a Big Data forrásokkal kapcsolatban megfogalmazott egyik lényeges, általános kritika.

Az adatok valóságtartalmának eldöntése már ettől független kérdés. A beavatkozást nem igénylő, humán szereplő nélküli adatgyűjtés során keletkezett információk tartalmilag kevésbé torzítottak. A közösségi médiában generálódó adatok azonban nem lehetnek teljesen reálisak. Itt számítani kell az információtorzulásra (például ha valaki hamis profillal használ egy online közösségi szolgáltatást), bár a Big Data állományokra jellemző hatalmas elemszám mellett a fals adatok előfordulási valószínűsége kisebb.

A félreértelmezések lehetősége a Big Data állományok (akárcsak más statisztikai adatsorok) esetében továbbra is megmarad, amire nagy nyomást helyez a „hype", azaz a Big Data felkapott és túlértékelt jellege, vagy ahogyan Walsh (2014) fogalmaz: az „automatizált arrogancia avagy big data önhittség". A Google is úgy gondolta, hogy a Big Data módszerek a korábbiaknál is pontosabb predikciós lehetőségeket

kínálnak például a valós idejű influenzatrendek elemzésére. Azóta bebizonyosodott, hogy az adattorzító hatások miatt ez egy téves, bár figyelemre méltó elképzelés (Lazer et al. 2014). A Big Data módszerek Lazer és munkatársai szerint ugyan hasznosak, de csak a tradicionális adatforrásokkal, a „small data" eszközökkel kiegészítve és párba állítva. Attól, hogy valami nagyon nagy, még nem biztos, hogy „Big". Az igazi Big Data állományt nem létrehozzák, hanem keletkezik, a digitális adatáramlás egy pillanatképe, tipikus „flow".

Az adatkör további problematikája a diverzitásból fakad, ami az adatszerkezeti tulajdonságokra vonatkozik. A Big Data állományok a legritkább esetben tiszták a primer formájukban. Sőt, a nagy méretből adódóan a hibák előfordulásának száma is nagy, ami a gondatlan felhasználó számára veszélyt is jelenthet (erre utal a szakirodalom találó „Big Data=Big Errors" szófordulata, lásd Taleb [2013]). Mindez szükségessé teszi, hogy az eddigieknél jóval nagyobb energiákat fordítsunk az adathalmaz megtisztítására, az eklektikus állomány homogenizálására, az outlierek vagy az adathiányok megfelelő kezelésére. Általános kihívás az adatok minősége, pontossága vagy használhatósága érdekében a statisztikai fogalmaknak való megfelelésük (lehetőség szerinti) biztosítása.

Mivel az adatok rengeteg forrásból és tartalommal gyűlnek össze, eddig soha nem látott összefüggések feltárására nyílik lehetőség. Épp ezt tekinthetjük a Big Data források egyik nagy előnyének, ugyanakkor továbbra is körültekintően kell az adatokra vonatkozó, azokból levont következtetéseinket megfogalmazni.

Az újfajta lehetőség a hivatalos statisztika figyelmét is felkeltette, ugyanakkor nem egyszerű a Big Data forrásokat a hivatalos standardoknak megfelelően hasznosítani. Az eddigi gyakorlatban a hivatalos statisztika elsődleges forrásai a teljes körű vagy mintavételes felmérések voltak, amiket a másodlagos (főként kormányzati) adatgyűjtések egészítettek ki. Ezekhez adódtak hozzá harmadikként a Big Data források, melyek lényeges tulajdonsága, hogy nem tervezett, hanem organikus eredetűek. Míg a felmérési (survey-típusú) adatokat egy adott kutatási kérdés támogatására strukturált módon hozzák létre, addig a Big Data spontán módon keletkezik, „csak úgy nő" (Schiller–Burghardt 2015). A Big Data források alkalmazásánál nincs előre megfogalmazott elemzői cél, csak utólag az adatállományból találjuk azt ki. A tudományos közösségnek kell rájönnie arra, hogy ezek az organikus adatok mire és miként használhatók.

Statisztikai értelemben a Big Data mindig minta marad, hiszen a sokaság aránya nem látható. A keletkező adathalmazok inkább az érintőképernyős sorszámkérés úgynevezett „totemoszlop adatállományaira" hasonlítanak, ahol folyamatos az adatgenerálódás (Hajdu 2014). Vizsgálandó a közvetlen statisztikai célra való alkalmasság is. A Big Data alapú statisztikai modellek koefficiensei a nagy méret miatt mindig szignifikánsak lesznek, még rossz determinációs együttható (R^2) mellett is. Ez viszont félreértelmezésekhez vezethet. Az adatok minősége sem feltétlenül megfelelő: nem biztosított a teljes lefedettség (a teljes minta), többszöri előfordulások is lehetnek (többszörös regisztráció miatt), a hagyományos statisztikával ellentétben a vá-

laszmegtagadás kezelése nehéz vagy nem lehetséges (Mag 2014). Más oldalról viszont a Big Data források statisztikai felhasználása igen olcsó, alkalmazásuk helyettesítő, vagy kiegészítő jellegű lehet (az elsődleges statisztikai adatforrások mellett például modelltámogatási vagy validálási céllal). A felmérési adatok és a Big Data források inkább kiegészítői, s nem versenytársai egymásnak (Bethlehem 2015). A hivatalos statisztikai közösségre vár a feladat, hogy megértse, miként hasznosíthatja a Big Data által kínált új forrásokat, gondolatokat és módszereket.

Kísérleti vizsgálati eredmények

Illeszkedve a korábban vázolt és az ENSZ által is javasolt csoportosításhoz, egy-egy példa bepillantást enged a lehetőségek és a kihívások halmazaiba. Az alábbi példák a területi mintázatok értékelési módjaira alapozva szemléltetik a Big Spatial Data eszközök változatosságát.

Twitter, labdarúgás és téradatelemzés

Az *első példa* a people to people típusú adatkörök egy lehetséges felhasználási módját illusztrálja. Látható, hogy ezen adatkörök jellemző forrásai például a közösségi oldalak, amelyekben a strukturálatlanul megosztott adatok és információk utólagos feldolgozásával számos társadalmi viselkedésformát ismerhetünk meg. Jelen példa a Twitter közösségi oldalon megosztott bejegyzések elemzésére vállalkozott.

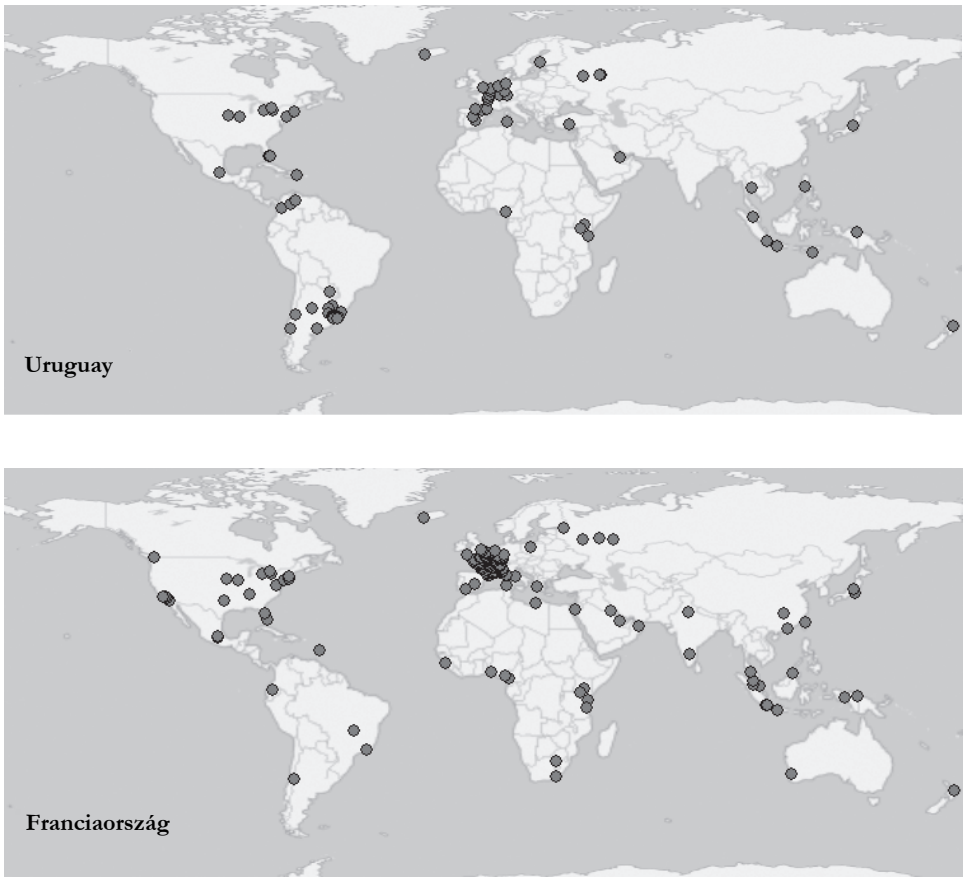
A Twitteren publikált tartalmak földrajzi szempontú elemzésével többen foglalkoztak már, kihasználva, hogy a közzétett információk itt alkalmanként térbeli azonosítókkal együtt is rögzülnek (Graham et al. 2014, Cuevas et al. 2014). Leetaru és munkatársai (2013) tanulmányukban a georeferált (azaz a közzétételi hely földrajzi lokációjához köthető) Twitter-bejegyzések sokszínű elemzési lehetőségeit ismertették, minden esetben a bejegyzésekhez kapcsolt földrajzi metaadatokra építve a megállapításaikat. Az elérhető geoadatok köre kétféle lehet: egyrészt településekre vonatkozó (a Twitter-felhasználók ezt manuálisan állíthatják be egy menürendszer segítségével), másrészt pontos földrajzi lokációt jelölő koordinátapár (amit általában a GPS és egyéb celluláris helymeghatározó alkalmazások szolgáltatnak). A település megjelölését a felhasználók a Twitter által felkínált listából választják ki, főleg akkor, ha a közösségi portált asztali vagy fix helyzetű eszközön keresztül használják. Ezt a helymegjelölést a felhasználó manuálisan frissíti, így a helyváltoztatáskor (például utazáskor) küldött bejegyzések csak a legutóbb választott lokáció szerint rögzülnek. Ezzel ellentétben a pontos helykoordinátákat közlő mobilalkalmazásokat használó változatban a felhasználónak semmi dolga sincs, hogy a rá vonatkozó helyinformációkat frissítse, mivel ez automatikusan megtörténik. A felhasználók aktuális helyzetét a bejegyzések közzétételekor négytizedes pontosságú koordinátaértékekkel rögzítik, ami lehetővé teszi a felhasználók helyzetének pontos utca, házszám, vagy épület szintű beazonosítását is (személyiségi jogi kockázatok miatt a felhasználóknak enge-

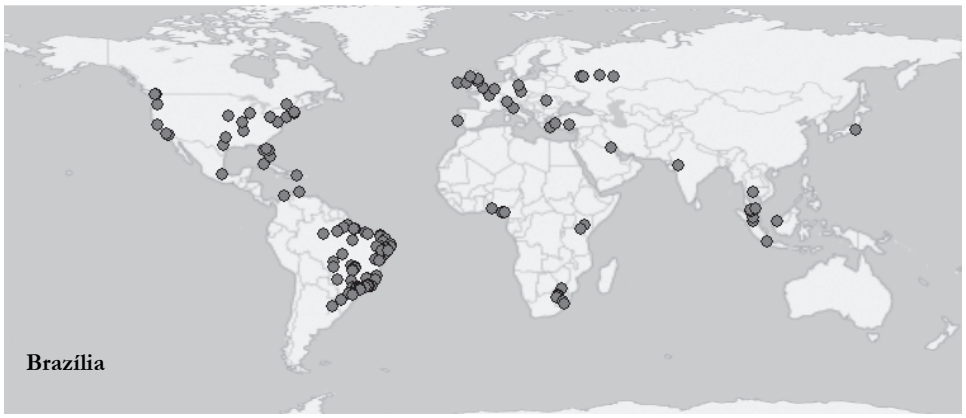
délyezniük kell az ilyen pontosságú térbeli azonosítást). Egy átlagos napon, Leetaru és munkatársainak (2013) kutatásai szerint, a bejegyzések 2,0%-a tartalmazott földrajzi metaadatokat, 1,8%-a települési megjelöléssel, 1,6%-a pontos helykoordinátával, de előfordult, hogy egy bejegyzés mindkettővel rendelkezett. Az adatkör így is elégségesnek bizonyult arra, hogy a társadalom térbeli működésének sajátosságait, esetleg részleteit is megismerhessük (lásd Graham et al. 2014).

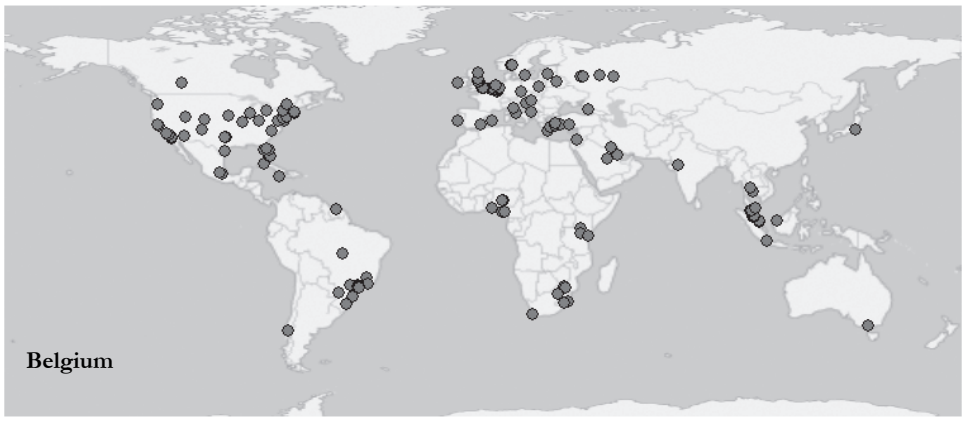
1. ábra

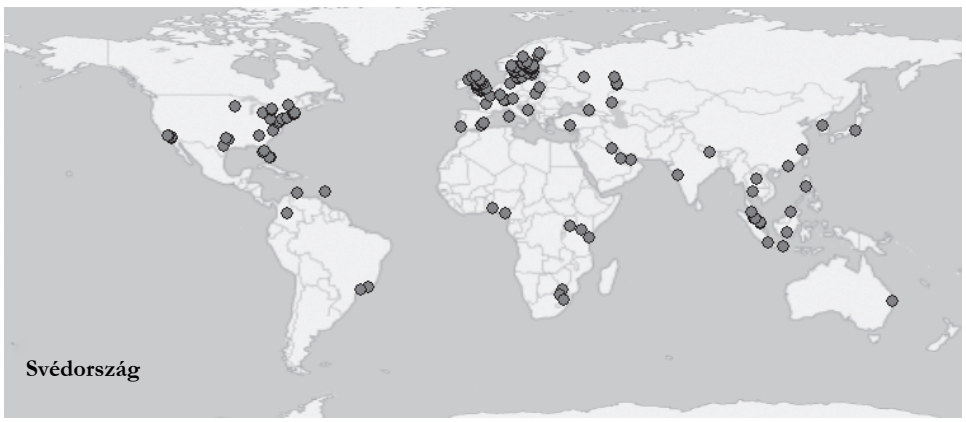
A 2018-as labdarúgó világbajnokság negyeddöntőt játszó országaira hivatkozó geolokalizált Twitter-bejegyzések elhelyezkedése a Földön a mérkőzések időpontja körül gyűjtött bejegyzések alapján

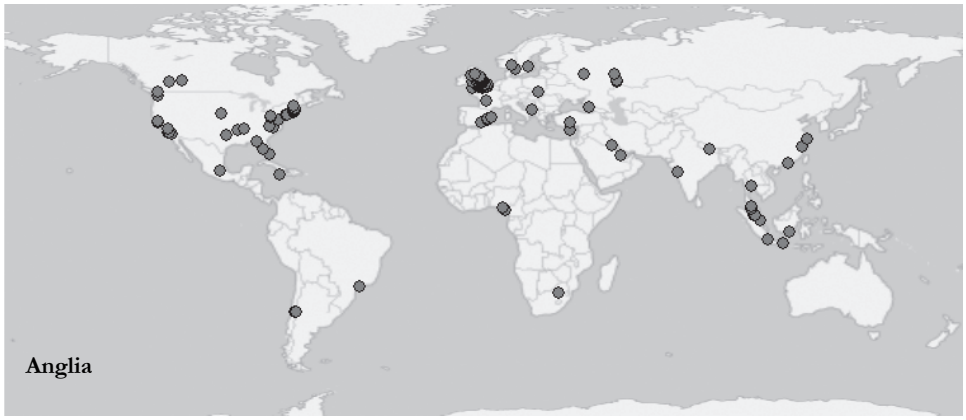
Location of geolocalized Twitter posts referring to the 2018 FIFA World Cup quarter-finalist countries based on entries collected around the time-period of the quarter-final matches

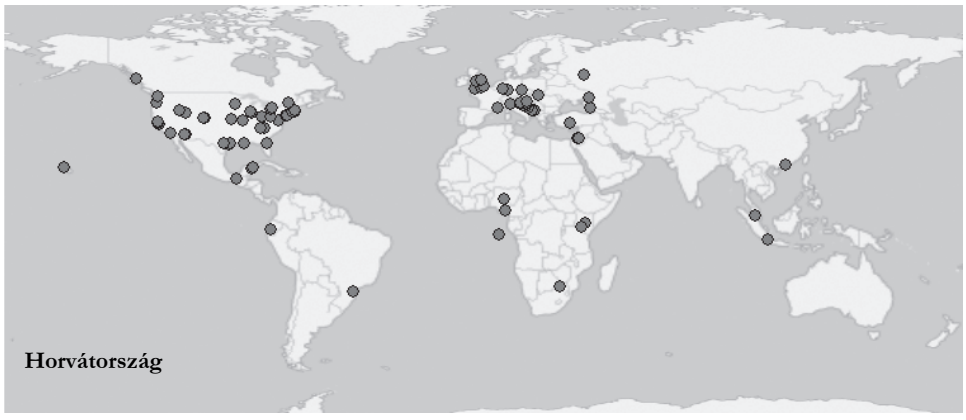



Brazília


Belgium


Svédország





Forrás: A Twitter Streaming API felhasználásával a szerző saját adatgyűjtése alapján.

A példaként említett vizsgálat során a Twitter Streaming API-ja és némi Python programozás segítségével összegyűjtöttük azokat a bejegyzéseket, amelyek a 2018-as labdarúgó világbajnokság negyeddöntős mérkőzéseinek időpontjai körül keletkeztek, és egyik vagy másik mérkőző fél országára hivatkoztak. Más szóval 2018. július 6-án és július 7-én a labdarúgó mérkőzések időpontjaiban, valamint a mérkőzéseket követő másfél-két órában a Twitter-bejegyzésekből leszűrtük azokat az eseteket, amelyek az aktuálisan játszó egyik vagy másik ország nevét szövegszerűen tartalmazták. Az 1. ábra az egyes országokra kapott találati listából azoknak a bejegyzéseknek a földrajzi elhelyezkedését mutatja, amelyek geolokalizációval is rendelkeztek.

A relatíve rövid, de célzott vizsgálati időszakban összesen egymillió tweetet sikerült regisztrálni, melyekből körülbelül kétezer bejegyzés geolokációs információval is rendelkezett, a teljes állomány 0,2%-át adva. Ez egyértelműen elmarad a Leetaru és szerzőtársai (2013) által jelzett tapasztalati arányoktól, bár a halmaz a mintázatok jellegzetességének kirajzolására már így is alkalmas. A szigorodó adatpolitika, s főleg a privát adatok kezelésének szabályozása (lásd GDPR) miatt, úgy tűnik, egyre több felhasználó választja a geopozíció közzétételének letiltását. Az IP-cím-alapú geolokalizációt a Twitter nem támogatja, mivel az erősen sértené a magánélet biztonságát. A rendelkezésre álló egymillió tweet ugyanakkor utófeldolgozással lehetőséget kínál a legalább nagy léptékű földrajzi azonosításra a bejegyzések egy további részénél, amivel a lokalizációs arány akár többszörösére is növelhető (ezt a munkafázist a vizsgálat már nem érintette). A geolokalizált tweetek relatíve alacsony arányát a nem lokalizált Twitter-bejegyzések hirtelen (az aktuális futballesemény kapcsán) megugró számával is magyarázhatjuk. E feltételezés tesztelésére a korábbi kulcsszavak előfordulási gyakoriságát egy, a mérkőzésektől független, vagy semleges időpontban is megvizsgáltuk. A geolokalizált találatok aránya – ugyanazon kulcsszavakkal – ekkor 1,95% körül alakult, ami egybeesett a Leetaru és munkatársai által megfigyelt tapasztalati arányokkal. A Twitter-bejegyzések számának statisztikai szempontú alakulására úgy tűnik, egyértelmű hatással vannak a jelentősebb (főleg globális érdeklődést keltő) események.

Az 1. ábra résztérképei az egyes keresési kulcsszavakkal szűrt találatok földrajzi differenciáltságát mutatják, amellett, hogy a globális Twitter-aktivitás általános trendjeihez is igazodnak. Az „Uruguay" keresési kulcsszóval szűrt geolokalizált bejegyzések térképén egyrészt nagyobb koncentráció látszik Uruguay környezetében, másrészt az aktuális ellenfél, Franciaország területén is sok lokalizált tweet fordult elő, ami ebben az esetben valószínűsíti a mérkőzés kapcsán kialakult összefüggő kapcsolatot a két ország között. Fordított irányban ez azonban nem figyelhető meg. Míg a „France" kulcsszóval szűrt Twitter-bejegyzések markáns koncentrációt mutattak Franciaország környékén, addig ez esetben Uruguay nem is jelent meg a mintában. A magyarázat hátterében az eltérő nyelvi környezetet véljük felfedezni, azaz az uruguayi felhasználók nem a francia vagy angol nyelvű „France" kifejezést, hanem a spanyol nyelvű „Francia" szót használták tweetjeikben az ország megnevezésére.

A Twitter analitikai vizsgálatának egyik sarkalatos pontja a nyelv. A vizsgálatban az egyes országokra referáló kulcsszavakat az angol írásmód szerint választottuk ki. Ezek alkalmanként egybeestek a nemzetközi szinten gyakran használt országmegnevezésekkel, sőt esetenként az ország saját hivatalos nyelv szerinti megnevezésével is (lásd France), máskor viszont eltértek a helyileg használt vagy általános elnevezésektől. A felmerülő problémák orvosolhatók, ha egyidejűleg több nyelv szerint is lekérdezzük az eredményeket. Az említett angol írásváltozatú keresési kulcsszavak ráadásul felülreprezentálták az angolszász országokat vagy az angolul beszélőket a mintában. Ezzel magyarázható, hogy a térképeken jelentős találati sűrűség jellemzi az Egyesült Államokat, illetve Nagy-Britanniát. A felülreprezentáltság egy másik oka lehet a Twitter helyi ismertségének és felhasználói elterjedtségének foka is. Ez alapján ismét az Egyesült Államok és Nagy-Britannia emelhető ki az országok közül, de elterjedt még a Twitter használata a nyugat-európai országokban, Törökországban, valamint Indonéziában is. Nem feledkezhetünk meg az időzónahatásokról sem. Az adatfelvételek időpontjában Ázsia keleti részén éjjel volt, a várható Twitter-aktivitás sem volt ott túl magas, ami a térképeken is megjelent.

Tisztában kell lennünk azzal is, hogy a keresési kulcsszó valójában mire vonatkozott. Esetünkben e rövid időintervallumban valószínűsíthető volt, hogy az egyes országok említései leginkább a labdarúgó mérkőzésekhez kötődően jelentek meg, ám ez nem lehetett ebben az időszakban sem kizárólagos. Az időintervallum maga is erős korlátokat szab, tehát az eredményeink csak rövid, aktuális helyzetkép felvázolására alkalmasak. Komplex, hosszabb távú lekérdezésekkel ez a Big Data állomány tovább javítható, esetleg relevánssá tehető. További kontextuális elemzés vagy szöveganalitika nagyban segítheti a megértést (ami igazolja a Big Data állományok költséges adattisztítási és harmonizációs igényeit).

A mintavételi torzulás magából a Twitter-közösségi háló adottságaiból is eredhet. Itt is nyilvánvaló a populációs torzulás, csak a Twitter-használó (ráadásul aktív) lakosságra vonatkozóan gyűjthetünk információkat (ez esetben főként a futball iránt érdeklődő csoportról). A Twitter Streaming API-val gyűjtött adatokban nincs reprezentativitási szűrés, sőt a geolokalizált adatok között duplikátumok is előfordulhatnak (egy felhasználó adott helyről több bejegyzést is közzétehet). Mindezeket figyelembe kell venni az adatok értékelése során.

A korábban említett sok kérdőjel, illetve tartalmi vagy módszertani kihívás ellenére egyértelműen megállapíthatjuk, hogy nem véletlenszerű a térképeken kirajzolódó mintázat. Minden keresési kulcsszó esetében világos sűrűsödés fedezhető fel a kulcsszóban jelzett országok körül, ugyanakkor szembetűnő a mérkőző felek párhuzamos (vélhetően összekapcsolódó) megjelenése is az egyes ábrákon (lásd Svédország és Anglia találati térképeit). A további szövegelemző vizsgálatok a feltevések összefüggéseit igazolhatják.

Vásárlási adatok térbeli mintázata

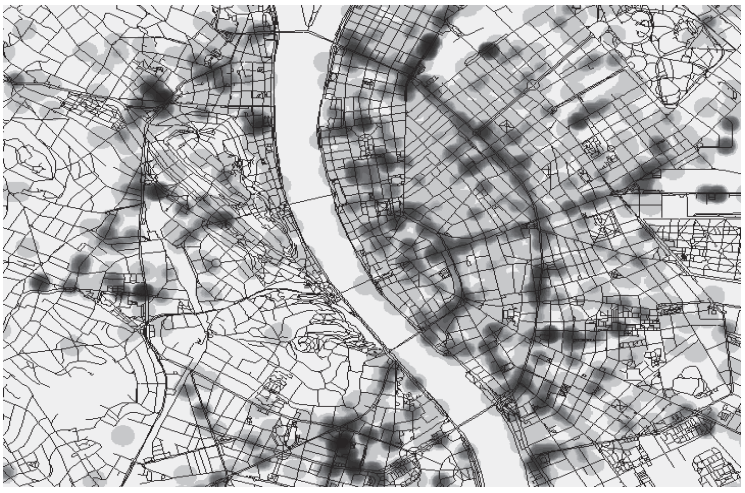
A *második példa* a people to machine adatkörből kiválasztott lokalizált kereskedelmi tranzakciós adatok alkalmazási lehetőségét ismerteti. Az adatkört maguk az emberek, tehát nem automatikus (szenzoros) rendszerek generálták, bár az adatgyűjtéshez, az adatok regisztrációjához gépekre volt szükség. A példa egy magyarországi bank ügyfeleinek POS- (Point-of-Sale) terminálokon feljegyzett vásárlásainak aktivitási adatai alapján készült. A terminálok az adatrögzítő eszközök, míg az ügyfelek azok, akiknek a viselkedéséről információ szerezhető.

A bankok üzleti érdekeik miatt és saját tevékenységükből adódóan folyamatosan rögzítik ügyfeleik tranzakcióit. Leszámítva az online, a fióki vagy az ATM tranzakciókat a banki ügyfelek legtöbbször a POS-terminálokon keresztül (a helyszíni bankkártyás fizetés során) hagynak digitális nyomokat maguk után. Ezek a digitális nyomok nemcsak azt jegyzik fel, hogy az egyes bankkártya-tulajdonosok mikor és milyen összegben vásároltak, de a POS-terminálok lokalizációja révén azt is, hogy mindez hol történt. Egy magyarországi bank korábbi mintaállományát felhasználva, s az adatokat aggregáltan térképezve látható, hogy a bank ügyfélköre általában hol költi el a pénzét. Ha ez az ügyfélkör, avagy a vizsgált minta kellően nagy, akkor a bolti kereskedelmi aktivitás társadalmi szintjeibe is bepillantást nyerhetünk (a mintaelemszámot, illetve a bank nevét nem közölhetjük).

2. ábra

A POS-terminálokon jegyzett vásárlások sűrűségi térképe egy magyarországi bank ügyfeleinek adatai alapján (Budapest belső területei, részlet)

Density map of the purchases made on POS terminals based on customer data of a Hungarian bank (central areas of Budapest, detail)



Megjegyzés: a sötétebb területek a nagyobb sűrűségű zónákat, a fekete vonalak az úthálózatot jelölik.
Forrás: a szerző saját szerkesztése egy magyarországi bank mintaadatbázisának felhasználásával.

A bankok egyes boltokba vagy értékesítési pontokra kihelyezett POS-termináljainak lokációi általában ismertek (legalábbis az üzemeltető bank számára). Ezek térképezése nem jelent különösebb kihívást, s nem nyújt érdemi információkat sem (hacsak a terminálok ponteloszlása nem érdekel bennünket). Az egyes terminálok tranzakciószámmal (esetleg tranzakciós összegekkel) súlyozott értékei viszont már érdemlegesen kirajzolhatják a kereskedelmileg aktív körzeteket a vizsgált térben. A 2. ábra a banki ügyfelek POS-tranzakciós (bolti vásárlási) aktivitásának térbeli eloszlásviszonyait mutatják Budapest belső területén. Jól láthatók a főváros frekventált kereskedelmi zónái, amelyek főként a Nagykörút, illetve a Kiskörút vonalaihoz igazodnak. Sűrűsödés jellemző a pesti belváros környékén, valamint egyes sugárirányú utak (Rákóczi út) vonalában. Budán a vonalas szerkezetek kevésbé jellemzőek, inkább lokális kereskedelmi gócok láthatók például a Széna tér, a dél-budai Móricz Zsigmond körtér és Újbuda-Központ környezetében. E gócterületek utólag is jól beazonosítják a bevásárlóközpontokat (Mammut, Allée, Westend, Aréna Pláza, Corvin Pláza stb.), amelyek a kiskereskedelmi fogyasztás urbánus centrumai. A modell megerősíti a városi kereskedelmi térszerkezettel kapcsolatos eddigi ismereteinket (Sikos T. 2013), de ki is egészíti őket azzal, hogy bizonyítékot szolgáltat a valós kereskedelmi térhasználat alakulásáról. Ilyen léptékű lokalizált kérdőíves vagy statisztikai felmérés más módon amúgy nem lenne elérhető.

E lehetőség a pozitívumai ellenére számos kihívással is jellemezhető. A legfőbb gondot maguknak az adatoknak a hozzáférése és felhasználási korlátai jelentik. A banki vagy általánosan fogalmazva business-típusú adatok a legérzékenyebb források közé tartoznak. Nyilvános felhasználásuk erősen korlátozott, az adatok (bizonyos keretek között) üzleti titkot képeznek, vagy ha nem, akkor is individuális szinten a lakosság közvetlen magánérdekű adatainak tekinthetők. Legtöbb esetben csak aggregátumok szintjén lehet – külső szakértőként – ezekkel az adatokkal dolgozni. Az adatok beszerzése is kihívást jelent, hiszen az ilyen adatkörök tulajdonosai üzleti szereplők, az adatelemzésekhez inkább piaci igények párosulnak, a nyilvános adatközlés pedig igen ritka.

A POS-terminálokon regisztrált tranzakciós adatokból felvázolt példa statisztikai értelemben is felvet kérdéseket. Big Data állományról van szó, így a vizsgálat elemszáma nagy, de nem biztos, hogy reprezentatív is a fogyasztóképes lakosság teljes egészére nézve. Az adatbázisban csak egyetlen bank és azon ügyfeleinek adatai szerepeltek, akik egyben bankkártya-tulajdonosok is és valamely POS-terminál által rögzített vásárlási rekorddal rendelkeztek. Ez egy szűkített állomány (a mintaválasztás statisztikailag sem megalapozott), ám mégis alkalmas a nagyvonalú területi mintázati sajátosságok meghatározására. Az alkalmazott modell javítható és finomítható más Big Data források (például más bankok adatainak) bevonásával, illetve kellően hosszú időre vonatkozó adatsorok alkalmazásával.

A sport és szabadidős térhasználat digitális nyomai

A harmadik példa a machine to machine típusú adatkörök lehetőségeit illusztrálja. Az ilyen adatokat általában technikai szenzorok, mérőműszerek, automatikus adatgyűjtők szolgáltatják, melyek között találhatók téradatokkal dolgozók is. Az alábbi példa egy mobilszenzoros nyomkövető adatgyűjtő forrás adataira épült.

A 3. ábra az egyre divatosabbá váló sport és szabadidős applikációk egyike alapján készült. Az adatbázis alapját képező okoseszközökre (leginkább okostelefonokra) telepíthető alkalmazás az alapkészülékbe épített helymeghatározó rendszer segítségével megállapított helyadatokat gyűjti össze a felhasználó számára a szabadidős sporttevékenység (futás, kerékpározás) közben. A sportteljesítmény monitorozása alatt az eszköz rögzíti a sportoló által bejárt útvonalat is. Ha nagyszámú felhasználó rögzíti az adatait, akkor a bejárt útvonalak egyidejű térképi ábrázolásával lehetővé válik a sport és szabadidős tevékenységek városi térpályáinak és frekventált helyeinek meghatározása is.

3. ábra

A szabadidős sporttevékenységek (futás, kerékpározás) városi térpályái a főváros budai oldalán és a belvárosban (részlet)

Urban trajectories of leisure sports activities (running, cycling) in central and western parts of Budapest (detail)

Megjegyzés: a vastag vonalak sűrűbben, a vékony vonalak ritkábban használt útszakaszokat jelölnek.
Forrás: www.strava.com (átdolgozott részlettérkép).

Az efféle alkalmazások helykoordináták időbélyeggel ellátott sorozatát jegyzik fel a háttéradatbázisokban, mely adatok bármilyen (nagy adatmennyiséget kezelő) térinformatikai rendszer segítségével vizuálisan is megjeleníthetők. A példaként bemutatott 3. ábra (háttértérkép hiányában is) felismerhetően kirajzolja a fővárosi utak hálózatát, illetve azokat az utakat is, amelyek csak a szabadidős sporttevékenység (futás, kerékpározás) szempontjából fontosak. Ilyen útszakaszokat főként a budai hegyvidéki részeken találtunk. A belvárosi és a Dunához közeli területek frekventáltabb útszakaszokkal jellemezhetők, főleg ott, ahol kerékpársávok és sportolási területek is találhatók. A térképen egy-egy sportpálya futóköre is kirajzolódott.

A 3. ábra egyfajta közösségi adatgyűjtés terméke. Alaphelyzetben a felhasználó kizárólag saját magának rögzíti és gyűjti a sporttevékenységéhez kapcsolódó útvonaladatokat. Gyakori azonban, hogy a teljesítményt rögzítő útinaplókat és a bejárt útszakaszok térképét a felhasználók megosztják egymással vagy a közösséggel. A térképek ezek alapján készülnek el. Mivel az adatok közzétételéhez az alkalmazásban a felhasználóknak hozzá kell járulniuk, így a publikált eredménytérképek már kevésbé sértik a magánéleti jogokat. Az aggregált térképek esetében kevésbé jelentkeznek a „privacy-típusú" kérdések, bár ha valaki nagy számban tölt fel adatokat a lakóhelyéről kiinduló útvonalakról, akkor esetenként még így is beazonosítható a felhasználó amúgy nem közzétenni kívánt lakcíme (Liptak 2018).

A legtöbb Big Data forrás csak egy adott tematika adott szeletét képes részletesen megjeleníteni. Bár kérdőíves és egyéb felmérési forrásokból eddig nem volt lehetőség a sport és szabadidős tevékenység ilyen részletes területi képének megjelenítésére, a kapott eredmény még így sem tekinthető minden szempontból problémamentesnek. Az adatok vonatkozásában például itt sem lehet meghatározni a reprezentativitás mértékét. Nem ismert az adatgyűjtő populáció, de feltételezhető, hogy nemcsak a sporttevékenységben aktív, hanem az infokommunikációs eszközök használatában is jártas (azzal egyáltalán rendelkező) csoportról van szó. Az analízisből levont következtetések természetesen csak erre a populációra lehetnek relevánsak.

Bár folyamatosan bővülő adathalmaz áll az elemzők rendelkezésére, az adatkör akkor lesz igazán használható, ha nagyon sok felhasználó, hosszú időtávra visszanyúló téradatai rögzülnek a rendszerben. Amíg ez (egyes lokális körzetekben) nem valósul meg, addig érdemleges torzító hatások jelenhetnek meg például az átlagnál sokkal aktívabb felhasználókkal összefüggésben.

Összegzés

Lehetőségek és kihívások. Ez a két szó kiválóan jellemzi a Big Spatial Data állományokkal kapcsolatos szakértői véleményeket. Az elmúlt évtized jelentős szakmai fordulatának tekinthetjük a Big Data témakör intenzív előretörését, s ezzel egyidejűleg a hozzá kapcsolódó térbeli kontextusú analitikai eredményeket is. A trendeket nézve abban bízhatunk, hogy a Big Data eszközöket hasznosító területi kutatások szerepe és aránya szignifikánsan növekedni fog a jövőben. Lassan már ma is termé-

szetesnek tűnik, de legalábbis nem feltétlenül különleges a Big Data állományokkal való foglalkozás. A hangsúly a lehetőségek felől inkább a kihívások és a jó minőségű Big Spatial Data elemzések felé tolódik el.

A példák rávilágítottak arra, hogy a Big Data állományok segítségével a társadalmi viselkedésre vonatkozó, s eddig csak vélt összefüggések akár bizonyítást is nyerhetnek, vagy eddig nem ismert új megvilágításba is kerülhetnek. A labdarúgó világbajnokság idején közzétett Twitter-adatok a lokalizált földrajzi események digitális térbeli hatását igazolták, a banki tranzakciós adatok nagyobb mintán is bizonyították a kereskedelmi aktivitás városi térszerkezeti sajátosságait, a sport és szabadidős adatok az eddig nemigen értékelt szabadidős térhasználat jellemzőiről nyújtottak hasznos információkat. A Big Spatial Data típusú adatkörök eközben új problémákat is felvetettek, például a jelentős méretű, ugyanakkor definiáltan torzított minták kapcsán. Mindez korlátja a megállapítások teljes sokaságra történő kiterjesztésének, de egyben inspirálója is az új kutatói gondolatok meghozatalának.

Köszönetnyilvánítás

A tanulmány a Felsőbbfokú Tanulmányok Intézete (Institute of Advanced Studies Kőszeg) támogatásával készült.

IRODALOM

BOYD, D.–CRAWFORD, K. (2012): Critical questions for Big Data: Provocations for a cultural, technological and scholarly phenomenon *Information, Communication and Society* 15 (5): 662–679. https://doi.org/10.1080/1369118X.2012.678878

CAVALLO, A.–RIGOBON, R. (2016): The Billion Prices Project: Using Online Prices for Measurement and Research *Journal of Economic Perspectives* 30 (2): 151–178. https://doi.org/10.1257/jep.30.2.151

CUEVAS, R.–GONZALEZ, R.–CUEVAS, A.–GUERRERO, C. (2014): Understanding the locality effect in Twitter: measurement and analysis *Personal and Ubiquitous Computing* 18 (2): 397–411.

DAVENPORT, T. H.–PATIL, D. J. (2012): Data Scientist: The Sexiest Job of the 21st Century *Harvard Business Review* 2012 October.

FRIEDEWALD, M.–RAABE, O. (2011): Ubiquitous computing: An overview of technology impacts *Telematics and Informatics* 28 (2): 55–65. https://doi.org/10.1016/j.tele.2010.09.001

GALLOWAY, A. (2004): Intimations of everyday life: Ubiquitous computing and the city *Cultural Studies* 18 (2): 384–408. https://doi.org/10.1080/0950238042000201572

GARTNER (2011): *Gartner Says Solving 'Big Data' Challenge Involves More Than Just Managing Volumes of Data* Stamford, Connecticut, USA http://www.gartner.com/newsroom/id/1731916.

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

GIRARDIN, F.–VACCARI, A.–GERBER, A.–BIDERMAN, A.–RATTI, C. (2009): Quantifying urban attractiveness from the distribution and density of digital footprints *International Journal of Spatial Data Infrastructures Research* 4: 175–200. https://doi.org/10.2902/1725-0463.2009.04.art10

GRAHAM, M.–HALE, S. A.–GAFFNEY, D. (2014): Where in the World are You? Geolocation and Language Identification in Twitter *The Professional Geographer* 66 (4): 568–578. https://doi.org/10.1080/00330124.2014.907699

HAJDU, O. (2014): *Big Data, adatbányászat, statisztika: terminológia, adatstruktúra, módszertan* Konferencia-előadás. Big data - forradalmasítja mindennapjainkat? Az MTA IX. Osztály Statisztikai és Jövőkutatási Tudományos Bizottságának tudományos ülése, 2014. november 20., MTA, Budapest.

IVAN, I.–SINGLETON, A.–HORÁK, J.–INSPEKTOR, T. (eds.) (2017): *The Rise of Big Spatial Data* Springer International Publishing, Cham, Switzerland.

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

JÄRV, O.–AHAS, R.–SALUVEER, E.–DERUDDER, B.–WITLOX, F. (2012): Mobile Phones in a Traffic Flow: A Geographical Perspective to Evening Rush Hour Traffic Analysis Using Call Detail Records *Plos One* 7 (11): e49171. https://doi.org/10.1371/journal.pone.0049171

JIANG, B.–YAO, X. (2006): Location-based services and GIS in perspective *Computers, Environment and Urban Systems* 30 (6): 712–725. https://doi.org/10.1016/j.compenvurbsys. 2006.02.003

KING, G. (2011): Ensuring the Data-Rich Future of Social Science *Science* 331 (6018): 719–21. https://doi.org/10.1126/science.1197872

KITCHIN, R. (2013): Big data and human geography: Opportunities, challenges and risks *Dialogues in Human Geography* 3 (3): 262–267. https://doi.org/10.1177/2043820613513388

LAZER, D.–KENNEDY, R.–KING, G.–VESPIGNANI, A. (2014): The Parable of Google Flu: Traps in Big Data Analysis *Science* 343 (6176): 1203–1205. https://doi.org/10.1126/science.1248506

LYON, D (2001): *Surveillance Society: Monitoring Everyday Life* Open University Press, Buckingham – Philadelphia, USA.

LYON, D. (2014): Surveillance, Snowden, and Big Data: Capacities, consequences, critique *Big Data & Society* 2014: 1–13. https://doi.org/10.1177/2053951714541861

MAG, K. (2014): *Big data a hivatalos statisztikában kihívások és lehetőségek* Konferencia-előadás. Big data - forradalmasítja mindennapjainkat? Az MTA IX. Osztály Statisztikai és Jövőkutatási Tudományos Bizottságának tudományos ülése, 2014. november 20., MTA, Budapest.

NAAMAN, M. (2011): Geographic information from georeferenced social media data *SIGSPATIAL* 3 (2): 54–61. https://doi.org/10.1145/2047296.2047308

RALEY, R. (2013): Dataveillance and countervailance In: GITELMAN. L. (ed.): *Raw Data Is an Oxymoron* pp. 121–145., MIT Press, Cambridge, MA.

SATYANARAYANAN, M. (2001): Pervasive computing: vision and challenges. Personal Communications *IEEE* 8 (4): 10–17. https://doi.org/10.1109/98.943998

SIKOS, T. T. (2013): Kereskedelmi földrajz In: JENEY, L.–KULCSÁR, D.–TÓZSA, I. (szerk.) *Gazdaságföldrajzi tanulmányok közgazdászoknak* pp. 239–252., BCE Gazdaságföldrajz és Jövőkutatás Tanszék, Budapest.

SZŰTS, Z.–YOO, J. (2016): Big Data, az információs társadalom új paradigmája *Információs Társadalom* 16 (1): 8–28. http://dx.doi.org/10.22503/inftars.XVI.2016.1.1

TECHAMERICA FOUNDATION (2012): *Demystifying Big Data. A Practical Guide To Transforming* The Business of Government. TechAmerica Foundation's Federal Big Data Commission, Washington D.C, USA.

THATCHER, J. (2014): Living on Fumes: Digital Footprints, Data Fumes, and the Limitations of Spatial Big Data *International Journal of Communication* 8: 1765–1783.

UN GLOBAL PULSE (2012): *Big Data for Development: Opportunities & Challenges* Global Pulse White Paper, May 2012, New York, USA.

WEISER, M. (1991) The computer for the 21st century *Scientific American* 265 (3): 94–104.

ZIKOPOULOS, P.–EATON, C. (2011): *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data* McGraw-Hill Osborne Media, New York, USA.

ZOOK, M.–DODGE, M.–AOYAMA, Y.–TOWNSEND A. (2004) New Digital Geographies: Information, Communication, and Place In: BRUNN, S. D.–CUTTER, S. L. – HARRINGTON, J. W. (eds.): *Geography and Technology* pp. 155–176., Kluwer Academic Publishers, Norwell, MA.

INTERNETES HIVATKOZÁSOK

BETHLEHEM, J. (2015): *The ever changing landscape of official statistics* New Techniques and Technologies for Statistics (NTTS) Conference, 2015.03.09.-13., Brussels, Belgium. http://www.cros-portal.eu/sites/default/files//Presentation%20S20AP1-ntts-2015.pdf (letöltve: 2018. július)

DUMBILL, E. (2012): *What is big data? An introduction to the big data landscape* O'Reilly Radar, 11 January. O'Reilly Media Inc., Sebastopol, California, USA. http://radar.oreilly.com/2012/01/what-is-big-data.html (letöltve: 2018. július)

ENSZ (2014): *How big is Big Data?* United Nations Economic Commission for Europe, Genf, Svájc. http://www1.unece.org/stat/platform/pages/viewpage.action?pageId=99484307 (letöltve: 2018. július)

GENTILE, B. (2011): The New Factors of Production and the Rise of Data-Driven Applications *Forbes* 31 October. http://www.forbes.com/sites/ciocentral/2011/10/31/the-new-factors-of-production-and-the-rise-of-data-driven-applications/ (letöltve: 2018. július)

HELLERSTEIN, J. (2008): *The Commoditization of Massive Data Analysis* University of California, Berkeley, USA. http://radar.oreilly.com/2008/11/the-commoditization-of-massive.html (letöltve 2018. július)

JONES, S. (2012): Why 'Big Data' is the fourth factor of production *Financial Times* 27 December. www.ft.com/intl/cms/s/0/5086d700-504a-11e2-9b66-00144feab49a.html (letöltve: 2018. július)

LEETARU, K. H.–WANG, S.–CAO, G.–PADMANABHAM, A.–SHOOK, E. (2013): Mapping the Global Twitter Heartbeat: The Geography of Twitter *First Monday* 18: (5–6).

https://firstmonday.org/ojs/index.php/fm/article/view/4366/3654 (letöltve: 2018. július) https://doi.org/10.5210/fm.v18i5.4366

LIPTAK, A. (2018): Polar Fitness suspends its global activity map after privacy concerns *The Verge* Jul. 8. https://www.theverge.com/2018/7/8/17546224/polar-flow-smart-fitness-company-privacy-tracking-security (letöltve: 2018.07.12)

LOUKIDES, M. (2010): *What is data science? The future belongs to the companies and people that turn data into products* O'Reilly Radar, 2 June, O'Reilly Media Inc., Sebastopol, California, USA. http://cdn.oreilly.com/radar/2010/06/What_is_Data_Science.pdf (letöltve: 2018. július)

MCGUIRE, T.–MANYIKA, J.–CHUI, M. (2012): Why Big Data is the New Competitive Advantage *Ivey Business Journal* July/August, http://iveybusinessjournal.com/publication/why-big-data-is-the-new-competitive-advantage/ (letöltve: 2018. július)

MCKINSEY GLOBAL INSTITUTE (2011): *Big data: The next frontier for innovation, competition and productivity* McKinsey & Company, San Francisco, California, USA. http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation (letöltve: 2018. július)

ONNELA, J. P. (2011): *Social Networks and Collective Human Behavior* UN Global Pulse. http://www.unglobalpulse.org/node/14539 (letöltve 2018. július)

SCHILLER, D.–BURGHARDT, A. (2015): *Using Research Data Centres (RDCs) to access Big Data* New Techniques and Technologies for Statistics (NTTS) Conference, 2015.03.09.-13., Brussels, Belgium. http://www.cros-portal.eu/sites/default/files//Schiller-etal_RDCs_to_access_Big_Data_0.pdf (letöltve: 2018. július)

TALEB, N. (2013): Beware the big errors of big data *Wired* February 8, 2013. http://www.wired.com/2013/02/big-data-means-big-errors-people/ (letöltve: 2018. július)

WALSH, B. (2014): Google's Flu Project Shows the Failings of Big Data *Time Online* 13. March 2014. http://time.com/23782/google-flu-trends-big-Data-problems/ (letöltve: 2018. július)