Németh Renáta

A számok tényleg magukért beszélnek?

(Hozzászólás Dessewffy Tibor és Láng László írásához)

Mottó:

With enough data, the numbers speak for themselves

(Részlet Chris Anderson Dessewffy által is idézett programadó cikkéből.)

Dessewffy Tibor és Láng László tanulmányához a társadalomtudományok kvantitatív módszereinek oktatójaként szeretnék hozzászólni, a címben szereplő kérdőjel által jól jelzett pozícióból. Hozzá kell tennem, kényelmetlen érzéssel teszem ezt. A Big Data-t képviselő adattudományok és az adatgyűjtés klasszikus módjait képviselő klasszikus tudományok (mint a survey-t használó szociológia vagy a kísérleteket végző biológia) viszonya a témában megjelent írásokat tekintve a legkevésbé sem tűnik harmonikusnak. Az új diszciplína provokatív cikkekben kopogtat a „régi" tudományoknál, s azok önnön tekintélyükről alkotott meggyőződésük függvényében vagy dühösen utasítják vissza a kopogtatást (biológia) vagy kétségbeesetten próbálják az ablak helyett az ajtót ajánlani a bejövetelhez (szociológia). Elég beütni a keresőbe Andersonnak a mottóban idézett cikkének címét. Az írásra született reakciók között van a molekuláris biológus Pigliuccié, aki szerint „információk petabyte-jai használhatók tehát ezen a területen. De kérem, ne nevezzék ezt tudománynak" (White 2009). A biokémikus White szerint

> ha a komplex rendszerek kutatói nem veszik komolyan a tudományos módszereket, akkor tudományuk lassan szétesik […] A számítógépes modellek lehet, hogy káprázatatóak, de hacsak nem produkálnak olyan sikereket, amik végül megváltoztatják az adott terület (molekuláris biológia, szociológia, közgazdaságtan) képviselőinek gondolkodását, a komplexitás tudományai terméketlenként halnak el (White 2009).

A társadalomtudósok saját területiumuk védelmekor kevésbé magabiztosak, talán a természettudományokkal szembeni régi kisebbségi érzés okán. Ahogyan egy Big Data hírportálon olvasható, „a szociológusok ott maradtak fel-le ugrálva, karjukkal integetve, azt kérdezve, miért nem konzultált velük senki erről az egészről" (Foster 2012). Ebben a viszonyban az óvatoskodó társadalomkutatók oldaláról való megszólalás részemről tehát nem tűnhet túl progresszív attitűdnek. De álláspontomban nem vagyok egyedül: Dessewffy is Anderson kritikusaként lép fel, az adatok önmagukban való értelmezhetetlenségét, a szociológiai elméletek megkerülhetetlenségét állítva. Érdemes itt felidézni még Szántó és SYI kitűnő, 2010-es BUKSZ-beli cikkét, ahol Barabási Albert-László kapcsán találkozhatunk hasonló érvekkel.[1] Hozzászólásom ezeket a gondolatokat támasztja alá és egészíti ki, a Big Data valamennyi válfaja felvet ugyanis kérdéseket a módszertan és az adatminőség terén, s mint látni, fogjuk, ezek izgalmas elméleti kérdésekhez vezetnek.

N=all

Dessewffy és Láng a szociológiában klasszikusan használt survey-ekkel szemben a Big Data előnyét abban látják, hogy az utóbbi esetén a „minta" a teljes populációt fedi. Én ezzel

[1] A szociológia nemzetközi szinten is ritkán reflektál a kihívásra, kivételként említhető Chris Snijders és szerzőtársai 2012-es *International Journal of Internet Science*-beli cikke, ahol a szociológia Big Data-iparhoz való potenciális hozzájárulásaként elsősorban a makroszintű jelenségek mikroszintű megközelítéssel való kiegészítését ajánlják. Szántó és SYI is említi ezt, de szélesebb kontextusban.

szemben azt gondolom, hogy ez utóbbi feltétel ritkán teljesül. Jó példa a boldogságmérő, mellyel (lásd hedonometer.org) Twitter-szövegek milliárdjainak boldogságszintjét mérik alkotószavaik boldogságtartalmának átlaga alapján,[2] majd például az USA államainak boldogságszintjét hasonlítják össze ezen a módon. Mivel a Twittert nem mindenki használja, kérdés, reprezentatívak-e az adatok, ténylegesen jellemzik-e az adott államot.[3] Választ kell tudni adni arra a kérdésre, hogy milyen mechanizmus áll a lefedettségi probléma mögött. Nyilván nem feltételezhetjük, hogy a Twitter-használók random mintáját adnák a populációnak, sőt azt sem, hogy minden államban ugyanolyan súlyú és ugyanolyan demográfiai összetételű a felhasználók csoportja. Mindez az adatbázisok (méretüktől független) tulajdonságainak és korlátainak a megértését igényli.[4] A társadalomtudományban évszázados reflexió létezik erre a problémára, a survey-alapú adatgyűjtés- és elemzés kidolgozott módszereivel. Ugyanakkor a survey-ek válaszadási aránya, így megbízhatósága is meredeken csökkent az utóbbi években. Illusztrációként: a Budapesten történő személyes kérdezések esetén a válaszadási arány manapság sokszor még a 20%-ot sem éri el.

Amikor azt látjuk, hogy a hedonométer szerint Louisiana a legboldogtalanabb amerikai állam, míg ugyanarra az időszakra vonatkozóan egy nagymintás survey Lousiana-t hozza ki legboldogabbnak (Oswald és Wu 2011), be kell látnunk, hogy megkerülhetetlen probléma az adatbázisok létrejöttének megértése. Az ellentmondás egyik lehetséges feloldása ugyanis a lefedett populációk eltérése lehet.

Adat és „valóság”

Az ellentmondás egy másik lehetséges magyarázata a két mérőeszköz eltérésén alapulhat. Míg a hedonométer az államok területén létrejött tweetekből indul ki, a survey a „Mennyire érzi boldognak magát egy 1–4-es skálán?” kérdésre adott válaszokon alapult. Ezt az eltérést is említik Dessewffy és Láng, amikor így írnak:

> Kérdőívekből véleményeket ismerünk meg. Ezen vélemények többnyire ismeretelméletileg „maszatosak”, a belőlük kirajzolódó kép a valósághoz fűződő viszonya esetleges […] Nem azért, mert félre akarjuk vezetni a kutatókat, hanem mert legtöbbször nincs kikristályosodott és tiszta álláspontunk az adott kérdésekről. […] Ezzel szemben a Big Data alapú kutatás alapvetően a digitális lábnyomokat, azaz a valódi viselkedés nyomait használja (Dessewffy és Láng 2015: oldalszám).

Tehát a Big Data objektív lenne?

A digitális lábnyomoknak önmagukban nincsen jelentésük, a jelentést a kutató alkotja meg. A legjobb algoritmus sem képes értelmes kérdésfeltevésre, a mérési adatok interpretálására vagy helyes konklúziók levonására. A hedonométer a tweetekből az államok boldogságszintjére következtetve egy többlépcsős interpretációs folyamatot jár végig. Amikor a tweet boldogságát alkotószavainak átlagos boldogságával, amikor egy állam boldogságát a területén írt tweetek boldogságával azonosítja, az adatok legkevésbé sem objektív értelmezését végzi el. És érdemes azt is tudni, honnét ismerik a kutatók az egyes angol szavak boldogságszintjét: egy nagymintás survey-ben (!) kértek meg egyéneket, hogy

---

[2] A cikkben nem tárgyalom azt a problémát, hogy az idézett kutatásokban a szövegeket nagyon leegyszerűsített, ún. szózsákmodellel közelítették meg, tehát alkotószavaik halmazaként kezelték őket, a potenciálisan jelentésmódosító kontextusra érzéketlenül.

[3] A boldogságmérő korpusza a Twitter teljes adatbázisának csupán 10%-os random mintáját adja, tehát ilyen szempontból sem teljes körű. Problémát jelent az is, hogy a tweetek populációján kapott információkat az egyének populációjára vonatkoztatják. Ez azért hoz torzítást, mert a gyakrabban twittelők nagyobb súllyal szerepelnek. Ezek ugyanakkor nem elvi problémák, technikailag kezelhetők lennének.

[4] A hedonometer.org-on a FAQ alatt szerepel a reprezentativitás problémája, de a publikációk, jelentések gyakran csak zárójelben említik meg, és elcsúsztatják az interpretációt a teljeskörűség felé, pl. a teljes populációra vonatkozó, más forrásból származó információk (fegyveres gyilkosságok száma) és a twitterezőkre vonatkozó információk együttes használatával.

boldogságtartalmuk szerint értékeljék (!) az angol korpusz egy részét. A tesztalanyok pl. a *hope, hero, to win* szavakat magas pontszámmal értékelték, a tagadószavakat alacsonnyal.

Az adatok méretétől (big vagy small) függetlenül, az adatelemzés középpontjában az interpretáció áll. Ám az adat és a „valóság" közötti távolság, így az őket összekötő interpretáció szerepe a Big Data esetén nem kisebb, mint a survey-nél. Igaz, hogy a survey-nél a válaszadó reflexiója áll az adat és a valóság közé, ugyanakkor a Big Data kutató (1) szemben a survey kérdőívkészítőjével, nem határozhatja meg a rendelkezésére álló adatok körét, azok mintegy melléktermékként jönnek létre. Továbbá, (2) az adat csak közvetett információ, „lábnyom". E távolság problémái a tweet és a tényleges boldogságérzet azonosításakor, azt hiszem, nyilvánvalóak – elég csak a megformált szöveg kulturális beágyazottságára utalni. Éppen a kulturális kontextus (a stílus, a nyelvi normák) figyelmen kívül hagyása, s a szövegnek egyfajta információközlő viselkedéssel való azonosítása miatt nehéz a hedonométer eredményeinek (idősebb bloggerek boldogtalansága, dalszövegek időben csökkenő boldogsága stb.) interpretálása.

Mindez tehát nem növeli, hanem éppen csökkenti a kutatói objektivitást. A jelentés nélküli „lábnyomok" miatt a Big Data esetében nagyobb a jelentéskonstrukció szerepe, az adatokban a kutató is benne van. Így az ideális tudományos eljárás – empirikus adatok ismerete nélkül felállított hipotézisek falszifikációs próbája utólag megszerzett adatokon – is ellehetetlenül, a hipotézis felállítása és próbája közötti válaszvonal elmosódik, a falszifikáció esélye csökken.

Kontextus

Az adatok kontextusból való kiragadása a Big Data egy másik jellemzője. A kontextus egyedi jelenség, ezért nehezen fogható meg nagy volumenek szintjén. A hedonométer esetén is triviálisan felmerül a probléma: szavak boldogságát mérjük a kontextusuk (tweet) és tweetek boldogságát a saját kontextusuk (beszédszituáció) nélkül. A kontextusból való kiragadás másik ismert Big Data példája a társadalmi hálózatok esete. Ezek esetén gyakran csak a hálózat, mint matematikai objektum kerül elemzésre, már a hálózat összekötő vonalainak (az éleknek) a valódi tartalma nélkül. Nézzük pl. Barabási Albert-László és szerzőtársai 2011-es Nature-beli cikkét a komplex rendszerek kontrollálhatóságáról. A cikk olyan rendszereket vizsgál, amelyeket gráffal reprezentálhatunk, csúcsaikhoz számértékű állapotok vannak rendelve, az állapot megváltozása lineárisan függ a szomszédok állapotától és egy külső vezérlés által bizonyos csúcsokhoz az adott pillanatban adott inputtól. A cikk fő tétele belátja, hogy a rendszer kontrollálhatóságához szükséges csúcsok (driverek) minimális száma csak a gráf topológiájától függ. Ezután társadalmi hálózatokra (börtönbeli bizalomháló, www, kommunikációs hálók) is kiszámolja a driver csúcsok minimális számát – ez úgy interpretálható, mint az a legkisebb emberhalmaz, amely az adott társadalmi hálózatot érdemben képes befolyásolni. A problémát az jelenti, hogy a társadalmi hálózatok, illetve a cikkben leírt dinamikus rendszerek analógiája nem terjed túl a hálózatok topológiáján. Milyen valós szám lenne rendelve a börtönbeli barátsággráf (ember)csúcsaihoz, amely jelentéssel bír a börtöngráf kontextusában? Mi az az interpretálható lineáris függvény, amely az (ember)csúcsok állapotváltozásait a szomszédai állapotához köti?

Korreláció és okság

Dessewffy és Láng: „A digitalizált, Big Data által uralt valóságban az elméletre új szerep vár. Bár valóban vannak esetek, amikor az adatbányászat meglepő korrelációkat mutat meg, ám a legtöbb problémát nem lehet megoldani kizárólag a korrelációs mintázatok bemutatásával" (2015: oldalszám).

A Big Data-információkat elsősorban (gazdasági célú) előrejelzésekre használják, s ritkán merül fel a megfigyelt mintázatok magyarázatának kérdése. Ahogyan egy 2013-as Big Data-bestseller fogalmaz: a miértre vonatkozó kérdéseket a mire vonatkozókra kell cserélni. A korrelációelemzés a kísérleteken alapuló oksági elemzéssel szemben gyors és olcsó, ráadásul „világos betekintést ad, ami újra zavarossá válhat, amint az okságot visszahozzuk a képbe" (Mayer-Schönberger és Cukier 2013: 66-68). Hogy e gondolatok kontextusát lássuk: példája szerint az, hogy a narancssárga autók adatbányászati eredmények szerint ritkábban mennek tönkre, a használtautó-kereskedők számára előrejelzésként értékes információt jelent, mégsem érdemes sem a miérten gondolkodni, sem autónkat narancssárgára festeni. Chris Andersonnak már idézett cikke ugyanerről:

> A hatalmas adattömegek elérhetősége és az elemzésükre alkalmas statisztikai eszközök a világ megértésének új útját kínálják. A korreláció fontosabbá válik az okságnál, és a tudomány anélkül is haladást tud elérni, hogy koherens modellekre, egységes elméletekre vagy mechanisztikus magyarázatokra támaszkodna (Anderson 2007: 16).

Úgy tűnik, a tudománytörténet fejezetei ismétlik önmagukat. Az okságot ismét a korreláció felől éri kihívás, mint száz évvel ezelőtt Karl Pearson (a korrelációs mérőszám megalkotója és az okság általa misztikusnak tekintett fogalmának legnagyobb kritikusa) és iskolája felől. A kihívásra felhozott, kauzalitás melletti érvek is ismerősek lehetnek, akár egy szociológia alapszakos diáknak is. Például: a korreláció megtalálása nem a kutatás végpontjaként, hanem inkább hipotézisek kiindulópontjaként lehet fontos. A Big Data olyan rendszerek esetén használható előrejelzésre, amelyek időben konzisztensek, nem változnak megjósolhatatlanul, kis komplexitásúak – a narancssárga autókat termelő autóipar ilyen lehet, a társadalmi rendszerek általában nem ilyenek. A tudományos elméleteknek nem az előrejelzés a végső célja, hanem a magyarázat, ami viszont oksági elméletet kíván. A fontos korrelációk megkülönböztetése a jelentéktelenektől oksági modellt kíván. És így tovább.

A hedonométerre visszatérve: a korrelációs elemzésekben gyakoriak a szociológia számára ismerős interpretációs csapdák. Ilyen a hamis korreláció esete: regionális szinten a tweetek átlagos boldogsága és a házasságok aránya korrelál, de ebből nem következik feltétlenül, hogy érdemes megházasodni. A két változónak közös oka is lehet, mint pl. a nyelvhasználatban is megmutatkozó kultúra – történetesen az USA konzervatív hagyományokkal rendelkező déli államai azok, ahol magasabb a házasságok aránya, és az angolszásszal érintkező kultúrák hatására ugyanitt jellemzőbb lehet az emocionális nyelvhasználat is. Ide tartozik a hatásirány kérdése is: Twitter-hálózatokban a csúcsok fokszáma és tweetjeinek boldogsága összefügg, de nem csak az előbbi lehet az ok és utóbbi az okozat. Felmerül még az elemzési szintek meg nem különböztetéséből fakadó ökológiai tévkövetkeztetés lehetősége – az átlagosan iskolázottabb államokban magasabb a boldogság átlagos szintje, tehát az iskolázottabbak boldogabbak? Sőt továbblépve, ahogy Szántó és SYI írják már idézett cikkükben, a megfigyelt mintázatok megértésének, az érdekek, szándékok és normák figyelembevételének szükségessége is felmerül. Az USA államaiban a boldogságmérő karácsony körül a boldogság rendszerszerű emelkedését mutatja. Az ünnep által kiváltott tényleges öröm vagy a jókívánságokra vonatkozó, tweetekre is érvényes társadalmi konvenciók állnak-e emögött?

Konklúzió

A Big Data és a survey előnyeinek/hátrányainak számbavétele után elmondhatjuk, hogy nem szembeállításuk, hanem egymást kiegészítő alkalmazásuk ad gyümölcsöző megközelítést. A Big Data és survey-adatok összefésülhetők egymással. A Big Data interpretációját és lefedettségi kérdését kiegészítő survey-jel támogathatjuk. És fordítva: a survey is támogatható

Big Data-val, pl. a mintavétel tervezését vagy a válaszmegtagadás kezelését (hiányzó jövedelemadatok imputációja adminisztrációs adatbázisokból) tekintve.

Bár a fenti példák némelyike a hedonométer Twitter-elemzéseivel kapcsolatban intett óvatosságra, nem állítom, hogy nem érdemes digitális szövegeket elemezni szociológiailag, még ilyen egyszerű szózsákmodellel sem. Fontos probléma például, hogyan tulajdonítunk a társadalmi diskurzusban jelentést egyes jelenségeknek. Hatékony szövegbányászati módszerek léteznek szövegek tematikus struktúrájának azonosítására, ezekkel vizsgálható például, hogyan szabadult el a szélsőjobboldali politikai szerveződések cigányellenes, rasszista retorikája a közbeszédben a média legitimizáló közvetítésén keresztül. Vagy ugyanezekkel a módszerekkel olyan új jelenségek, mint a bevándorlás jelentéssel való felruházásának dinamikája is követhető – hogyan alakítja pl. a félelem témája ezt a jelentést, hogyan, milyen aktorok által, milyen dinamikában foglalja el a diszkurzív mezőt az új jelentés.

A mottóra visszatérve: a számok, bármilyen sokan vannak is, nem beszélnek önmagukért. A társadalmi Big Data kutatói a szociológia pionírjaihoz hasonlóan, ismét a természet- és bölcsészettudományok közötti határvonalon, a tudományos objektivitást érintő episztemológiai kérdések kereszttűzében találják magukat. A Big Data kétségtelenül korszakváltást jelent a szociológia számára. A hatalmas adattömegeken és hatékony algoritmusokon kívül azonban, ahogy a felmerülő kérdések kapcsán látható, szükség van a szociológia és a szociológiai módszertan (újra)felfedezésére is.

Hivatkozott irodalom

Anderson, Chris (2007): The End of Theory. The Data Deluge Makes the Scientific Method Obsolete. *Wired* (16). Interneten: http://www.wired.com/science/discoveries/magazine/16-07/pb_theory/.

Foster, Ian Armas (2012): What Sociologists Say About Big Data. *Datanami.* Interneten: http://www.datanami.com/2012/09/03/what_sociologists_say_about_big_data.

Liu, Yang-Yu, Jean-Jacques Slotine és Barabási Albert László (2011): Controllability of Complex Networks. *Nature* (473): 167–173.

Mayer-Schönberger, Viktor és Kenneth Cukier: (2013): *Big Data. A Revolution That Will Transform How We Live, Work, and Think*. Boston: Houghton Mifflin Harcourt.

Oswald Andrew J. és Stephen Wu (2011): Well-Being across America. *Review of Economics and Statistics* 93(4): 1118–1134.

Pigliucci, Massimo (2009): The End of Theory in Science? *EMBO Reports* 10(6): 534.

Snijders, Chris, Uwe Matzat és Ulf-Dietrich Reips (2012): „Big Data" – Big Gaps of Knowledge in the Field of Internet Science. *International Journal of Internet Science* 7(1): 1–5. Interneten: http://www.ijis.net/ijis7_1/ijis7_1_editorial.pdf.

Szántó Zoltán, SYI (Szakadát László) (2010): Fizikusok, bélyeggyűjtők, emberjárás-jelentők. BUKSZ 3., 201-213.

White, Michael (2009): *Networks Are Killing Science*. (Blogbejegyzés a Science 2.0 tudományos közösségi oldalon.) Interneten: http://www.science20.com/adaptive_complexity/networks_are_killing_science.