Magyar Tudomány 179(2018)5, 683–692
DOI: 10.1556/2065.179.2018.5.11

A SZOCIOLÓGIA HELYE A BIG DATA-PARADIGMÁBAN ÉS A BIG DATA HELYE A SZOCIOLÓGIÁBAN

ROLE OF SOCIOLOGY IN BIG DATA PARADIGM, AND ROLE OF BIG DATA IN SOCIOLOGY

Kmetty Zoltán

PhD, egyetemi adjunktus, Eötvös Loránd Tudományegyetem Társadalomtudományi Kar, Szociológia Intézet
zkmetty@tatk.elte.hu

ÖSSZEFOGLALÁS

A digitális adatrögzítés, a közösségi média, a dolgok internete (Internet of Things, IoT) és összességében a digitalizáció olyan mennyiségű adatot generált, ami korábban elképzelhetetlen lett volna. Az adattudomány teljesen új perspektívája nyílt meg, amelynek a lehetőségei és a határai sem láthatóak még. Mivel a nagy adatbázisok (Big Data) feldolgozásának képessége nem része általában a szociológusok standard eszköztárának, más tudományterületek képviselői (például a fizikusok), aktív résztvevőivé váltak a társadalomtudományi Big Data-kutatásoknak, továbbá olyan új szakmák is kialakultak, mint az adattudósé. De az adatok keletkezésének megértése és a jó kutatási kérdések feltevése a Big Data-paradigmán belül is elengedhetetlen ahhoz, hogy új tudományos eredményekhez jussunk. A társadalomtudósok rengeteg tudást felhalmoztak a világról, tudják mik a társadalom strukturálódásának mozgatórugói, sokat tudnak az emberi viselkedésről is. Ezt a tudást fel kell használni a kutatási kérdéseink megfogalmazásához, annak érdekében, hogy a társadalom Big Data-alapú vizsgálatában előre tudjunk lépni.

ABSTRACT

Digital data recording, social media, IoT (Internet of Things) and the entire digitalisation produce such amount of data that was unimaginable in the past. It opens entire new perspectives for data science, but its potential and limitations are hard to see know. As the techniques of data processing at large scale databases (so-called 'Big Data') are not always the part of the basic inventory of sociologists, scientists of other fields (like physicists) became active participants of Big Data analysis in social researches and new professions like data scientists have emerged. But understanding the data generating mechanisms and asking the good questions is essential, if we would like to create new scientific knowledge based on this kind of data. Social Scientist have gathered a lot of knowledge about the world, they know how societies structure, and they know a lot about human behaviour. This knowledge has to be used to shape our research questions in order to reach further in studying the society based on Big Data.

Kulcsszavak: Big Data, szociológia, adattudomány, érvényesség, konfirmatív logika

Keywords: Big Data, sociology, data science, validity, confirmative logic

BEVEZETÉS

A digitális adatrögzítés és adattárolás felfutása és a digitalizáció általános elterjedése önmagában olyan adatmennyiséget generált az elmúlt években, ami korábban elképzelhetetlen lett volna. Az adatokkal foglalkozó és dolgozó tudományokban ez egy olyan új perspektívát nyitott meg, amelynek a lehetőségeivel és korlátaival csak most kezdenek el ismerkedni az ezzel foglalkozó kutatók. Az adatmennyiség nemcsak a nagysága miatt érdekes (persze ez is számos következménnyel jár), hanem a keletkezés módja is újszerű, főleg a társadalomtudományok számára. A számokkal és kvantitatív elemzéssel foglalkozó szociológusok többsége abban a paradigmában élt eddig, hogy hivatalos adatgyűjtésekre vagy speciális *survey* kutatásokra támaszkodva próbáljon eredményekre jutni. Ezekben a típusú adatgyűjtésekben alapvetően az volt a közös, hogy tervszerűen elemzési célokra készítették ezeket az adatokat, legalábbis az adatok magas strukturáltsággal rendelkeztek. Ezzel szemben a digitális térben keletkező adatok nagy része nem adatgyűjtési célból készül, és emiatt csak részben vagy egyáltalán nem strukturált (Dessewffy–Láng, 2015). Itt tehát rögtön két kihívással kell megküzdeni a kutatónak, valahogy meg kell birkózni azzal az adatmennyiséggel, amin a korábban használt programok és rutinok már nem működnek, és valahogy „rendet is kell vágni" a strukturálatlanságban. A rendteremtésre sokan bejelentkeztek, a természettudósok egy része kimondva-kimondatlanul helyet követel magának az ezen adatok értelmezéséről szóló vitákban. A társadalomtudományoknak és ezen belül a szociológiának azonban nem szabad ijedten tekintenie erre az adatforradalomra, sőt inkább lehetőségként kell felfognia, mert számos olyan kérdés merül fel a rendteremtés és értelmezés közben, amelyre elsősorban a szociológia felől tud adekvát válasz érkezni. A szociológiának megvan a helye a Big Data-paradigmán belül, és a Big Datának is megvan a helye a szociológián belül.

A SZOCIOLÓGIA HELYE A BIG DATA-KUTATÁSOKBAN

Mivel foglalkozik egy szociológus? Akár furcsának is tűnhet ez a kérdés, egy Big Data-témájú tanulmány alcímében, de ha a Big Data és a szociológia kapcsolatát meg szeretnénk érteni, akkor nem hagyhatjuk ezt az aspektust figyelmen kívül. Arra, hogy mi a szociológia, és mit csinál, egy szociológus biztos nem tudna egyszerűen egy konszenzusos definíciót találni, abban azonban valószínűleg a

legtöbben egyetértenének, hogy ezek a definíciók időben és valamennyire térben is (utóbbi talán most már kevésbé releváns), változtak, változnak. A szociológia változását a survey kutatási módszer megjelenésén keresztül mutatom be.

A survey megjelenése és tömegessé válása egészen biztosan jelentősen átrendezte a szociológiai mezőt, hiszen olyan, korábban nem használt tudáselemeket kellett, hogy beépítsen a szakma, amelyek addig kívül voltak a határain. A survey megjelenése jelentősen felértékelte a statisztika szerepét, hiszen mintákat kellett tervezni, hibákat kellett számolni, statisztikai modelleket kellett alkotni. A *kellett* persze erős szó, hiszen az elméleti és történeti szociológusok vagy a kvalitatív kutatásokra koncentráló kutatók ideig-óráig elkerülhették ezt a kérdést, de ma már nem nagyon lehet anélkül a szociológia szakot elvégezni, hogy valaki legalább minimális mértékben ne sajátítsa el a statisztikai alapokat, és a kvantitatív adatelemzési technikákat. Az 1930–60-as években a survey-jellegű kutatások felfutása (Groves, 2011) valószínűleg megrémítette az akkori statisztikailag még nem képzett társadalomtudós-generációt, hiszen olyan tudásra lett volna szükségük, amellyel általánosságban nem rendelkeztek. Sokakban akkor fel is merülhetett, hogy az ezzel a tudással viszont rendelkező statisztikusok átvehetik majd a „hatalmat" a társadalomértelmezés addigi szociológia dominanciája fölött. Több mint ötven évvel később, bátran kijelenthetjük, hogy ez nem történt meg, csak azok a statisztikusok, „módszertanászok" lettek fontosak a társadalomértelmezés szempontjából, akik maguk is releváns szociológia tudást vettek magukra. A tanulság az, hogy külön kell választani a módszert, és külön az értelmezést, ha valaki jó az elsőben, még nem lesz jó a másodikban. Ezt az előbb bemutatott példát analógiaként használhatjuk a Big Data esetére. A hatalmas mennyiségű digitális tartalom kezelése olyan tudást igényel, amelylyel a legtöbb szociológus nem rendelkezik. Az adatok „megszerzése", mozgatása, elemzése olyan programokat, elemzői rutinokat kíván meg, amit nem egyszerű elsajátítani. A korai statisztikusok helyére ebben a mostani helyzetben az „adattudósok" érkeznek, statisztikai tudás helyett különböző programozási nyelvek és adattárolási technikák ismerői. De ők továbbra is csak a módszert tudják nyújtani. És ez igaz azokra a természettudósokra, akik sokéves Big Data kezelési tapasztalattal a hátuk mögött kiváltják az adattudósokat, és a maguk útját járva elkezdik a társadalmi folyamatokat értelmezni. Ha nincs meg a jó kérdés, és nincs meg az értelmezési keret, akkor nem lesz meg a jó válasz sem. A számok bármekkora sokaságból, és bármilyen bonyolult módon is keletkezzenek, magukban nem fognak sok mindent jelenteni, nem lesz társadalomtudományi relevanciájuk. Az adatok pusztán induktív feldolgozásából nem fognak szociológiai szempontból izgalmas eredmények kijönni, pusztán olyan állításokig fogunk eljutni, hogy az emberek idejük nagy részét otthon és a munkahelyükön töltik, ezért egyszerűen megjósolható a térbeli helyzetük (Barabási, 2010). Nem akarom elvetni (és nyilván a szociológusok többsége sem veti el)

az induktív tudásfelhalmozást (lásd ezt támogatva a Big Data-paradigmán belül: Lee–Martin, 2015), de amellett mindenképpen érvelnék, hogy a Big Data-elemzésekben is fontos lenne érvényesíteni a konfirmatív logikát. A konfirmatív logikai munkákhoz viszont kérdések és társadalomtudományi értelmezések kellenek, és ez már a szociológusok terepe. És itt vissza is kanyarodhatunk a fejezet címéhez – mivel foglalkozik a szociológus? Vagyis pontosabban a Big Datával foglalkozó szociológusok (hiszen nyilván lehet majd e nélkül is szociológiát művelni, akárcsak statisztika nélkül), milyen tudással kell, hogy rendelkezzenek?

Az adattudósokra jellemző tudás hasznos lehet, de nem előfeltétele a Big Datával való foglalkozásnak. Sok szociológus dolgozik úgy (a kutatóintézetek is így épülnek fel), hogy vannak a társszerző kollégák között survey-statisztikusok, akik speciális módszertani és adatelemzési tudással rendelkeznek. A Big Data-paradigmán belül ezt a pozíciót veszi át, vagy ezen pozíció mellé épül be egy adattudós is, aki képes nagy mennyiségű adathalmazokkal jól és hatásosan bánni. A szociológus szerepe azonban egy ilyen munkában nemcsak az, hogy a releváns társadalomtudományi kérdést megfogalmazza (és az adatok alapján megválaszolja), hanem az is, hogy eldöntse, hogy a nagy mennyiségű, sokszor strukturálatlan adatból megválaszolható-e egyáltalán az a kérdés, és ha igen, akkor milyen érvényességgel. Az érvényesség itt minimum három szempont mentén érdekes.

Az érvényesség kérdése

A strukturálatlanság bár nem szinonimája a „zajnak", de gyakran együtt jár vele. Legyen szinte bármilyen adatunk, mindig valamennyire „zajos" lesz, még egy jól kivitelezett surveyben is van bizonytalanság. Ugyanaz a válaszadó már egy órával később máshogy töltené ki a kérdőívet, kérdéseket félreértenek a válaszadók, lankadhat a figyelmük a kitöltés közben. Ez a zajszint egy strukturálatlan adathalmazon még magasabb, nehezebb kihámozni a „valóságot" az adathalmaz mögül. A szociológus egyik feladata annak az eldöntése, hogy vajon a zaj nem nyomja-e el teljesen a vizsgálni kívánt kérdést, lehet-e érvényes válaszhoz jutni? A zaj szűrése pedig nem egyszerű, akár néhány kezdeti kis (de rossz) döntés is teljesen romba döntheti az elemzésünket (Diesner, 2015).

A zajnál picit komplikáltabb a második érvényességet érintő aspektusunk, ami a Big Data-adatok keletkezési körülményeit érinti. Akárcsak egy kérdőívre adott válaszok, a Big Data-adatok is valamilyen kontextusban keletkeznek, nem semlegesek (Lewis, 2015). Kontextuson nemcsak a dinamikai kontextust értjük (például egy bejegyzés milyen más bejegyzésre reagál), hanem legalább ennyire fontosnak gondoljuk a makrokontextust, vagyis azokat a kereteket, normákat és szabályokat, amelyek például egy adott oldal használóit szabályozzák. Ilyen normatív keret például az, hogy mik az elvárt viselkedési minták egy oldalon. Megmondhatjuk-e Facebook-használók posztjai alapján, hogy az emberek hány

százaléka boldog, ha feltételezzük azt, hogy a boldogságot tökéletesen tudjuk definiálni a felhasználói aktivitás alapján? A válaszunk természetesen nem. Még csak a Facebook-felhasználókra sem tudunk ilyen állítást megfogalmazni. A közösségi oldalakon a „jó" oldalunkat szeretjük megmutatni, ebből következően inkább a boldogságot jelző üzenetek fognak dominálni. A normatív mozzanatoknál még erősebb cezúrát jelölnek ki, a letiltási, kitiltási, moderálási szabályok. Mivel a legtöbb oldal igyekszik valahogy meggátolni, csökkenteni az uszító tartalmak arányát, megjelenését, ezért ilyen jellegű téma elterjedtségének vizsgálata valószínűleg nem tud érvényes tudáshoz vezetni minket.

Dinamikai folyamatok vizsgálatánál még inkább fontos az adatkeletkezési környezet ismerete, mivel ezekben az esetekben azzal is tisztában kell lennünk, hogy történt-e változás az adatkeletkezést befolyásoló szabályrendszerben. A keletkezési körülmények kapcsán a normán és a szabályrendszeren felül az oldalakat működtető algoritmusokban is lehetnek olyan pontok, amelyek torzíthatják az adatstruktúrát (Shaw, 2015). A Facebook-falat (feed) generáló algoritmus például ilyen, hiszen ismerőseink posztjai nem véletlenül jelennek meg, hanem azokat látjuk gyakran, akik/amik az algoritmus szerint érdekesek nekünk. Azt, hogy az „érdekes" milyen számítások alapján generálódik, nem lehet tudni, és igazából most számunkra nem is fontos. Azt viszont szem előtt kell tartanunk, hogy ha Facebook interakciós gráfot szeretnénk elemezni, akkor ez a típusú torzító mechanizmus ott lesz az adatok mögött. Ha megváltozik az algoritmus, megváltozik az interakciós gráf is. Tehát nem a „valóságról" tudunk meg valamit, hanem az adatfolyam-generáló mechanizmus természetéről vagy legalábbis a kettő valamilyen kihámozhatatlan egyvelegéről.

A problémát szemléltetem egy telefonhálózatos példán is. „A" körzet lakói gyakran telefonálnak településen belül, „B" körzet lakói kevésbé gyakran. Ebből vajon következtethetünk arra, hogy „A" körzetben erősebb a lakóhelyi közösségi háló? Talán. Ha azonban tudom azt, hogy az „A" körzetben működő telefontársaságnál már hosszú évek óta ingyen van a helyi beszélgetés, míg „B" körzetben lévő másik telefontársaságnál ez a kedvezmény nem létezik, akkor fennáll annak a veszélye, hogy a nagyobb helyi interakciós sűrűség csak a kedvezményre vezethető vissza. Itt is az adatgeneráló mechanizmusok torzítási hatását kell megértenünk és mérlegelnünk, különben búcsút inthetünk a szociológiai szempontból is érvényes állításoknak.

Az érvényesség harmadik sarokköve a mintaszelekció, illetve az önszelekció kérdése. A „kikről" van adatunk ezen paradigmán belül még égetőbb kérdés, mint általában a survey kutatások esetében, mivel a mintákat nem mi tervezzük, hanem készen kapjuk őket. Kik tweetelnek, kik facebookoznak rendszeresen, kik használnak helymegosztó szolgáltatásokat? Ha ezekre a kérdésekre nem tudunk megnyugtató választ adni, akkor nem beszélhetünk érvényes eredményekről, legyen bármennyi adatunk is.

Tehát a szociológusnak nemcsak az a feladata ezen paradigmán belül, hogy társadalmi szempontból releváns kérdéseket tegyen fel, és válaszokat fogalmazzon meg, hanem az is, hogy az eredmények érvényességéről is meggyőződjön. Ehhez pedig értenie kell az adatgeneráló mechanizmusok természetét, és tudnia kell, hogy milyen adatforrásnak hol vannak az értelmezési határai. Egy Big Data-szociológusnak ezzel is kell foglalkoznia…

A BIG DATA HELYE A SZOCIOLÓGIÁBAN

Az előző rész azzal foglalkozott, hogy hol van a szociológusok helye a Big Data-kutatásokban. Ezen most fordítunk egyet, és arra próbálunk választ adni, hogy hol van a helye a Big Data-paradigmának a szociológiai kutatásokban, és részben arra is, hogy mit változtathat meg a Big Data a szociológiában. A következő felsorolás nyilvánvalóan nem lesz teljes, de igyekszem azokat a területeket megvilágítani, amelyekben a Big Data nagy szerepet kaphat, és azok közül is néhányat, ahol nem várható nagy változás.

Az idő szerepe

A szociológiai kutatásokban az idő általában nem játszik szerepet, vagyis pontosabban fogalmazva inherens a szerepe. Ez abból adódik, hogy a survey kutatások nagy része keresztmetszeti kutatás, ritka, amikor panel és/vagy trend-adatok állnak rendelkezésre. Összességben ritka a több időpontot átfogó kutatás, ha van is ilyen, az időpontok száma véges, ritkán haladja meg a kettő-hármat. A Big Data átformálja az időérzékelést (Nooy, 2015), hiszen folyamatos adatgenerálás működik. Az idő, a dinamika így kulcskérdéssé válik. Nem a mi és a mennyi, hanem a mennyit változott, milyen gyorsan változott kérdések lesznek érdekesek. Ez az új adatfolyam megnyithat korábban már lezárt témákat, új aspektusba helyezve őket. Az idő az adatok „frissességét" is érinti. Nem kell feltétlenül kettő–öt–tízéves survey kutatásokhoz nyúlnunk egy adott probléma kapcsán, hanem akár az elmúlt órák, napok, hetek, hónapok eseményeinek elemzése is lehetővé válik (Csepeli, 2015). Lerövidül az adatgenerálás és az adatelemzés közötti szakasz, a mai társadalomról a mai adatok alapján tudunk állításokat megfogalmazni. A szociológia sosem tudott olyan friss és mai lenni, mint amilyen a Big Data által lehet.

Kiscsoportok, ritka események

A kvantitatív szociológusok nagy része a survey kutatások által teremtett kereten belül tud működni. 1000–2000 fős kutatásokból próbáljuk megérteni a társadalmat. Ez óhatatlanul oda vezet, hogy a „nagyon pici" csoportokról vagy a nagyon

ritka eseményekről a szociológusok nehezen tudnak állításokat megfogalmazni. A nagyon pici és a nagyon ritka itt survey mintaméret nagyságrendben érthető. Egy 1 százalékos csoport egy standard 1000 fős surveyben 10 fő – ez nehezen elemezhető. A rétegződéssel foglalkozó szociológusok előszeretettel hivatkozzák David B. Gruskyt (Grusky–Wedden, 2005), aki a mikroosztály kutatások mellett tört lándzsát, szembemenve az 5–20 osztályos modellekkel, ő 100 feletti társadalmi/foglalkozási osztályt is relevánsnak gondolt. Ilyet persze lehet csinálni egy népszámlálásból, de ha kisebb adatforrás áll a rendelkezésünkre, akkor nem fog sikerülni egy ilyen vállalkozás. Ha Big Data-jellegű adatforrásunk van, más a helyzet, szinte bármilyen kis csoportot vizsgálhatunk (ha megjelenik a populációs keretben), ezeket tetszőlegesen kombinálhatjuk, összehasonlíthatjuk. És ehhez nem is kell nagyon speciális mintákat tervezni, elég csak a megfelelő csoportokat leszűrni. A méretnövekedés magával hozza azt is, hogy a megértés is mélyebb lehet. Egy 50 fős mintaszeletet egy maximum 5 változós modellel célszerű/illik vizsgálni, elkerülve a túlillesztés problémáját. Ha 5000 fős csoportunk van, ilyen problémák már nem jelentkeznek.

A kis csoportok mellett a ritka események is jobban vizsgálhatóvá válnak. Meg lehet-e válaszolni azt a kérdést egy surveyből, hogy a szülinaposok boldogabbak az életükkel? Aligha. Egy Big Data-adatforrásból ez a kérdés viszont akár meg is válaszolható.

Kétségkívül ezek a lehetőségek megnyitnak számos olyan területet a kvantitatív szociológia előtt, ahova eddig csak a kvalitatív szociológia tudott bejutni. Ez olyan lehetőség, amit nem szabad kihagyni.

Régi módszerek, új lehetőségek, új dilemmák

Az adatelemzési módszerek mindegyike többé-kevésbé érzékeny az elemszámra. A regressziós modellekben ez például a bevonható változók számában és a becslések standard hibáiban manifesztálódik. A regressziós módszereket mégsem gondoljuk kifejezetten elemszám érzékenynek, akár egy 500 fős mintából is lehet viszonylag alacsony standard hibájú becsléseket adni, akár sok változó bevonásával. Ezzel szemben például a kereszttábla-elemzés kifejezetten elemszám érzékeny. Nem is feltétlenül a kétdimenziós kereszttáblákra kell gondolni (bár ott is jelentkezhet a probléma), hanem a három vagy több dimenziót magában foglaló modellek esetére. Általában ezek elemzése valamilyen loglineáris modell segítségével történik, legtöbbször egy sokdimenziós interakciós térben leképezve az összefüggésrendszert. Ez három-négyre korlátozza a bevonható változók és kategóriák számát, ami óhatatlanul szegényesebb elemzési keretet tud biztosítani, mint egy regressziós modell (legalábbis a változók sokszínűsége szerint). A Big Data-paradigmán belül ez a probléma is feloldódik, több dimenziót lehet bevonni, nem kell félni az esetek elfogyásától.

Ez várhatóan oda fog vezetni, hogy a loglineráis modellek használata gyakoribb lesz, és olyan területeken is elkezdik ezt a modellosztályt használni, ahol eddig háttérbe szorult.

A kérdést a másik oldalról is megközelíthetjük – megnehezíti-e egyes módszerek használatát az, ha túl nagy az adathalmaz? A válasz itt is egyértelműen igen. Az adatkezelés nehézségeiről és gépigényéről nem kell beszélnünk, az az adattudósok dolga. Ebben az esetben inkább az érdekel bennünket, hogy a statisztikai módszerek hogyan reagálnak a nagyobb adathalmazokra, vagyis pontosabban fogalmazva, vannak-e olyan módszerek, amelyek rosszul reagálnak a nagy adathalmazokra? Két példát hoznék ezzel kapcsolatban. A hierarchikus klaszteranalízis tipikusan olyan módszer, amelyet viszonylag kis adathalmazon szoktak csak futtatni, mivel a futási idő nem polinomiálisan függ össze az esetszámmal, a polinomiálisnál nagyobb a futási idő növekedése. Több tíz- vagy százmilliós adathalmaznál a hierarchikus klaszterezés futtatása ellehetetlenül. És ez igaz minden olyan módszerre, ahol a futási idő a polinominálisnál gyorsabban nő. A másik példa szintén egy csoportosító eljárás, a látens profilelemzés, ami tulajdonképpen egy modellalapú klaszterezési módszer, ami a kevert eloszlások szétválasztásának logikájára épül (Vermunt–Magidson, 2002). Itt a futási idő nem feltétlen jelent gondot nagy adathalmaz esetében sem, viszont a csoportszám meghatározása annál inkább. Mivel modellalapú klaszterezésről van szó, a modellek illeszkedését különféle mutatók segítségével tudjuk vizsgálni (ilyen lehet, többek között, az AIC- vagy a BIC-mutató). A probléma csak az, hogy minél nagyobb az adathalmazunk, annál több a kategóriás szegmentáció, ami a legjobban illeszkedik (Fraley–Raftery, 1998). Ebben az esetben tehát a megszokotthoz képest új kapaszkodókat kell keresnünk az ideális csoportszám megtalálásához.

A szignifikancia rabságában

Az előző két alfejezet bár más-más hangsúllyal, de a mintaméret okozta lehetőségeket, problémákat taglalta, ez a mostani fejezet sem kivétel ebből a szempontból. A kvantitatív szociológusok tipikus kérdése arra vonatkozik, hogy szignifikáns-e az összefüggés, azaz a látott mintázatok a teljes sokaságban is megtalálhatók-e, nem a mintavétel okozza-e az egyes dimenziók együttmozgását. Ha szignifikáns az összefüggés, akkor gyakorlatilag sokan már hátra is dőlnek, a hatáserősségről kevés szó esik. Egy Big Data-projektben a szignifikancia-teszteknek nincs általában értelmük, egy többmilliós adathalmazon minden összefüggés szignifikánsnak fog bizonyulni. Az eddig sokszor másodlagos kérdés a hatáserősségről felértékelődik, hiszen ez tudja a mondanivaló lényegét adni.

Nyertesek és vesztesek

A Big Data-paradigma eltérő módon fogja érinteni a szociológia egyes ágait, vannak olyan szegmensek, amelyek sokat profitálhatnak az újfajta adatokból, míg másokat akár érintetlenül is hagyhat. Mind a kettőre hozok egy-egy példát.

A kultúra-, a fogyasztás- és az ízlésszociológia lehet az egyik nagy nyertes az új paradigmán belül. A mit és hogyan fogyasztunk gyakori témái a közösségi oldalaknak, utazásokról és ételekről teszünk fel leírásokat vagy fotókat a Facebookra, és az Instagrammra. A Vivinóra feltöltjük, hogy milyen bort ittunk, a Foursquare-re hogy hol jártunk, a TripAdvisorra, hogy miként értékeljük az éttermet, ahol ettünk, az Amazonra (vagy itthon a Molyra), hogy mennyire tetszett az olvasott könyv. Ez a hatalmas digitalizált adatfolyam teljesen újraírhatja azt, hogy mit tudunk kultúráról, ízlésről, fogyasztásról, olyan komplex mintákat mutathat meg nekünk, amelyek a standard surveyekből sosem bukkanhattak elő. Persze itt ismét nyithatunk egy zárójelet és visszautalhatunk a korábban írtakra a mintaszerkezet kapcsán. Az, hogy kik töltenek fel ételfotókat, kik értékelnek könyveket, kik használnak közösségi alkalmazásokat, egyáltalán nem véletlenszerű, a magasabb státusz mellett (több alkalmazás is implicit módon megköveteli az okostelefon használatát), a digitális tudás vagy akár valamilyen önreprezentációs kényszer is megjelenhet a megosztási minták között, ami megnehezíti azt, hogy az eredményeinket általánosíthassuk.

A szegénység, társadalmi kirekesztettség, kizáródás vizsgálata viszont vélhetően azon szociológiai szegmensek közé tartozik, amit csak kismértékben fog elérni a Big Data-paradigma. Ezeknek a kérdéseknek, problémáknak lenyomatai elsősorban nem az *online* térben keletkeznek (bár akár ott is megjelenhetnek), tehát nem is vizsgálhatók jól digitális adatforrások segítésével. A szegénységgel foglalkozó kvantitatív kutatók tehát többségében továbbra is hagyományos surveyekre fognak támaszkodni, legalábbis a közeljövőben nem várható ide a Big Data-paradigma betörése.

A két kiragadott példa csak szemléltetni hivatott azt, hogy az egyes szakszociológiák nagyon eltérő módon fognak találkozni a Big Data jelentette lehetőséggel, kihívással, ami akár a szakma további differenciálódásához, fragmentálódásához is vezethet.

ZÁRÓ GONDOLATOK

Rövid dolgozatomat azzal a felkiáltással indítottam, hogy a társadalomtudományoknak és ezen belül a szociológiának nem szabad félnie a Big Datától, lehetőségként kell tekintenie rá. Ha a Big Data segítségével szeretne valaki beszélni a társadalomról, nem lehet majd kikerülni a szociológusokat. Nem elég az adathalmaz, tudni kell releváns kérdéseket és válaszokat is megfogalmazni, láttatni kell

az adatok érvényességének korlátait, a minta torzulásait, és úgy összességében a strukturálatlanság mögött megbúvó struktúrát. Ezek mind-mind olyan kérdések és feladatok, amelyekre véleményem szerint a szociológusok tudnak legjobban megfelelni. A másik irányból is hasonló konklúzióra juthatunk. A kvantitatív szociológiának is szüksége lesz a Big Datára, az egyre nehezebbé váló adatgyűjtési környezetben fel fog értékelődni minden más típusú adatforrás. Ez bizonyos szakszociológiákat előtérbe helyezhet majd, mások esetleg háttérbe szorulnak, ez még nem látható most pontosan, csak a körvonalak rajzolódtak ki. Az összességében biztos, hogy a szociológiát sem fogja érintetlenül hagyni a Big Data, abban viszont csak bízhatunk, hogy jól tudja majd szakmánk ezt az „akadályt" venni.

IRODALOM

Barabási A.-L. (2010): *Villanások: a jövő kiszámítható.* (ford. Kepes J.) Budapest: Nyitott Könyvműhely

Csepeli Gy. (2015): A szociológia és a Big Data. *Replika*, 92–93, 169–174. http://www.replika.hu/system/files/archivum/92-93_12_csepeli.pdf

Dessewffy T.– Láng L. (2015): Big Data és a társadalomtudományok véletlen találkozása a műtőasztalon. *Replika*, 92–93, 155–168. http://www.replika.hu/system/files/archivum/92-93_11_dessewffy_lang.pdf

Diesner, J. (2015): Small Decisions with Big Impact on Data Analytics. *Big Data & Society*, 2, 2. DOI: 10.1177/2053951715617185, http://journals.sagepub.com/doi/pdf/10.1177/2053951715617185

Fraley, C. – Raftery, A. E. (1998): How Many Clusters? Which Clustering Method? Answers via Model-based Cluster Analysis. *The Computer Journal*, 41, 8, 578–588. DOI: 10.1093/comjnl/41.8.578

Groves, R. M. (2011): Three Eras of Survey Research. *Public Opinion Quarterly,* 75, 5, 861–871. http://www.uvm.edu/~dguber/POLS234/articles/groves.pdf

Grusky, D. B. – Weeden, K. M: (2005): The Case for a New Class Map. *American Journal of Sociology*, 111, 1, 141–212. https://inequality.stanford.edu/sites/default/files/media/_media/pdf/key_issues/social%20class_research.pdf

Lee, M. – Martin, J. L. (2015): Surfeit and Surface. *Big Data & Society*, 2, 2, DOI: 2053951715604334, http://journals.sagepub.com/doi/full/10.1177/2053951715604334

Lewis, K. (2015): Three Fallacies of Digital Footprints. *Big Data & Society*, 2, 2, DOI: 2053951715602496 DOI: 10.1177/205395171560249, http://journals.sagepub.com/doi/pdf/10.1177/2053951715602496

Nooy, W. (2015): Structure from Interaction Events. *Big Data & Society*, 2, 2, 1–4. DOI: 10.1177/2053951715603732, http://journals.sagepub.com/doi/pdf/10.1177/2053951715603732

Shaw, R. (2015): Big Data and Reality. *Big Data & Society*, 2, 2, DOI: 2053951715608877, http://journals.sagepub.com/doi/pdf/10.1177/2053951715608877

Vermunt, J. K. – Magidson, J. (2002): Latent Class Cluster Analysis. In: Hagenaars, J. – McCutcheon, A. (eds.): *Applied Latent Class Analysis.* Cambridge: Cambridge University Press, 89–106. https://pure.uvt.nl/ws/files/487979/hagenaars2002b.pdf