VEROSZTA ZSUZSANNA

Adminisztratív adatok társadalomkutatási kezelése

Tanulmányunk célja számba venni az adminisztratív adatok társadalomkutatási felhasználásának elméleti és gyakorlati feltételeit és következményeit. Ennek során vizsgálódásunk elvi kiindulópontja az az állítás, miszerint az adminisztratív adat avagy adatbázis – mégha kutatási célra használják is - nem kutatási adat avagy adatbázis. Kutatási célú felhasználásához tehát számos validálási folyamaton szükséges végigmennie, hogy teljesítse a kutatási adatokkal kapcsolatos általános kritériumokat. Itt nyilvánvalóan nem hierarchikus minőségi különbségről van szó, hanem a társadalomkutatási módszertan által az adatok irányában megkövetelt szigorú feltételekről. A nehézséget ez esetben az jelenti, hogy míg a kutatási célú adatgyűjtések esetében ezek a validálási eljárások kidolgozottak, módszertanuknak története és kultúrája van, addig az adminisztratív adatok esetében – a megnövekedett mennyiség és a növekvő igény együttállása miatt – épp most válik egyre sürgetőbbé a kutatási befogadás útjának kiépítése. A tanulmány ehhez a jelenleg is zajló folyamathoz kíván néhány szempont átgondolásával hozzájárulni. Enélkül ugyanis az adatok és adatbázisok önmagukban valóságán alapuló kutatások egész sora indulhat útjára és ez minden bizonnyal nem kedvez a társadalomkutatás nívójának, hosszú távon.

E szemlélettel és céllal foglaljuk össze az alábbiakban elsőként azokat a tényezőket, amelyek az adminisztratív adatok iránti megnövekedett figyelmet táplálják, majd megvizsgáljuk ezen adattípus természetét és viszonyát a kutatási céllal gyűjtött adatokhoz. Ezt követően az adminisztratív adatokon nyugvó társadalomtudományi kutatás szükséges validálási kritériumait és eljárásait tekintjük át, illusztratív példákat is beemelve. Gondolatmenetünket az adatbázisok összekapcsolásában rejlő lehetőségek áttekintéséig és néhány gyakorlat bemutatásig visszük el.

Adminisztratív adatbázisok

Az adminisztratív adat igazgatási céllal gyűjtött adattípus. Tekintettel arra, hogy az ilyen jellegű adatbázisba kerüléshez jog avagy kötelezettség társul, az adminisztratív adatbázisokat egyedi beazonosíthatóság és a célcsoport teljes lefedése jellemzi.[1] Előbbi tulajdonságuk adatvédelmi és etikai szempontból számos kérdést nyit meg a kutatási felhasz-

[1] Az adattípusok leírását lásd Gárdos Éva és Salomvári György e számban megjelent tanulmányaiban.

➤ Educatio 2015/3. **Veroszta Zsuzsanna:** *Adminisztratív adatok társadalomkutatási kezelése*, 3–14. pp.

nálás felé, ám ugyanennyi potenciált rejt az elemi szintű adatkapcsolás számára. Utóbbi jellemzőjük pedig sok módszertani lehetőséget és kezelendő sajátosságot nyújt a kutatás számára.

Vizsgálódásunk számára az adminisztratív adatok lényeges sajátossága az, hogy ezt az adattípust nem kutatási – hanem igazgatási, szabályozási, regisztrációs, szolgáltatási stb. – céllal gyűjtik. Mindez főként az olyan nagy állami rendszerek feladata, mint például az adózásért, társadalombiztosításért, oktatásért fenntartott szervezetek. Az adatok egy része személyekre, más része szervezetekre vagy egyéb elemzési egységekre vonatkozik. Az adatgyűjtésben meghatározott célon túli felhasználást összességben másodlagos felhasználásnak tekintjük. Ezen belül az adatok statisztikai jellegű feldolgozása már korábban megindult és sokszor hivatalos formát kapott, kutatási hozzáférésükre azonban csak a törvényi szabályozási keretek változása után nyílt lehetőség.[2] Hasznosulási – de még nem kutatási – szempontból nézve az adminisztratív nyilvántartások másodlagos felhasználása mellett szóló legfőbb érv gazdaságossági. Ezen adatok közjószágnak számítanak, amennyiben gyűjtésüket, rögzítésüket az adatkezelő közpénzből finanszírozza még akkor is, ha az adatátadás (előbbihez képest elhanyagolható) költsége az adatkérőt terheli. A másodlagos felhasználás szakpolitikai döntésmegalapozás formájában is hasznos lehet. Az adminisztratív adatok alapján végzett hatásvizsgálatok mind a szándékolt intézkedés hatásainak előrejelzésében, mind a beavatkozás utáni állapot-összevetésben fontosak, ami a célpopuláción végzett előzetes és utólagos mérésekben ölt testet. Közvetett haszonként mindezt egy használat által megvalósuló minőségellenőrzési funkció is kiegészíti, javítva a rendszer által végzett kötelező ellenőrzések hatékonyságát (*Cseres-Gergely – Scharle*, 2008). Mindezeket felismerve számos országban állami szervezetek, ügynökségek jöttek létre kimondottan az adminisztratív adatbázisok összegyűjtése és másodlagos felhasználásának menedzselése céljából. E szempontból erős gyakorlatot és tradíciót találunk Dániában, Hollandiában, Svédországban vagy Finnországban, de az Egyesült Királyságban is kiemelt figyelemmel foglalkoznak a problémával (*Dibben et al*. 2009).[3] E folyamatot számos nemzetközi szervezet (mint az Európai Bizottság, ENSZ, OECD, Világbank) is támogatja és szorgalmazza.[4]

Az adminisztratív adatok szerepének felértékelődése

E helyt nem foglalkoznánk részletesen az evidence based szemlélet nemzeti és nemzetközi szintű szakpolitikai előretörésével. Az adatbőség és adatéhség általános tendenciáját immár vizsgálatunk stabil háttérközegének tekinthetjük (*Halász*, 2009; 2010). Egy olyan alaphelyzetnek, melyben az adat főszereplővé vált a szakpolitikai döntéshozás legitimációjában, kitermelve mindeközben ennek kritikai megközelítéseit is (lásd erről *Hammersley*, 2001; *Veroszta*, 2011).

Az adatfelhasználás jelentőségét ezúttal a társadalomkutatás felől megközelítve azokra a folyamatokra utalnánk, amelyek a kutatók államilag gyűjtött mikroadatokhoz bizto-

[2] Az adathozzáférésről és annak szabályozásáról részletesen szól Székely Iván e számban megjelent tanulmánya.

[3] Az adminisztratív adatok kezelésének nemzetközi gyakorlatáról ír Széll Krisztián e számban megjelent tanulmányában.

[4] Ennek egyik alapdokumentuma a 2007-es Istanbul Declaration (OECD 2007).

sított hozzáférését dominálják[5] (*National Research Council*, 2005). Ilyen folyamat a közvéleményben növekvő aggodalom a magánélet védelme és a titoktartás iránt, amely az önkéntes részvételen alapuló kutatások iránti bizalmatlansággal jár együtt, magas válaszmegtagadást és ezáltal romló minőséget eredményezve. Mindeközben a társadalomról való tudás igénye – az adatéhség – egyre növekvő mértékben és szinteken jelentkezik. A társadalomban zajló differenciálódással speciális, kislétszámú vagy nehezen elérhető csoportok sokaságára fókuszálunk, például célzott beavatkozások, támogatási rendszerek fejlesztése kapcsán. A gazdasági és társadalmi helyzet pontos feltérképezéséhez a részletező adatok mellett azok időbeni alakulása is fokozott figyelmet kap, melyet hatékonyan csak az adminisztratív rendszerek biztosítanak. Ezt a tendenciát az informatikai közeg gyors fejlődése hálózatkutatási, adatbányászati eszközökkel, adattárházakkal, fejlett elemzési és disszeminációs eszközökkel kellőképpen támogatja és rendkívül hatékonnyá teheti. Ez persze csak egy megfelelő adatvédelmi-adathozzáférési szabályozási közegben tud jól működni, amely ennélfogva az elmúlt időszakban szintén jelentős változásokon ment keresztül. Az állami adatbázisok mikroadataihoz biztosított kutatói hozzáférést mindemellett az anonimizálási technikákban lezajlott fejlődés és a korlátozott hozzáférés kezelésére kidolgozott speciális módszertani és szervezeti fejlesztések (pl. adatközpontok, ellenőrzött távoli hozzáférési pontok, licenszálás) is támogatják (*National Research Council*, 2005). Az adathasznosítás témájában született programok és szakértői elemzések alapján jól látszik az is, ahogyan az adminisztratív adatok kutatási célú felhasználásáról való gondolkodás a szükségesség és a lehetségesség lépésein keresztül tartalmi szempontból rendszerint az adatkapcsolásban rejlő fejlesztési potenciálig jut el (*Hotz et al*., 2008; *The UK Administrative Data Research Network*, 2012).

Szervezeti szempontból szintén egységesnek tűnnek a nemzetközi szakirodalomban és szakmai háttéranyagokban fellelhető ajánlások a kihívások kezelésére (*OECD* 2013; The UK *Administrative Data Research Network*, 2012; *Mosley*, 2012; *Card et al*., 2011). Ezek rendszerint olyan lépéseket jelentenek, mint a nemzeti avagy nemzetközi szintű adatkezelési szervezetek (ügynökségek) létrehozása, a speciális kutatóintézetek kiépítése, az adatfelhasználás és adatkapcsolás jogi kereteinek megteremtése, a működésben az etikai szempontok rögzítése és érvényesítése.

Adminisztratív adat versus kutatási adat

Az adminisztratív adatok kutatási felhasználásának módszertanai áttekintése előtt azokat a főbb jellemzőket határozzuk meg, amelyek ezt az adat- és adatbázistípust a kutatási céllal gyűjtött adattól/adatbázistól megkülönböztetik. Az összegzés alapjául több, a témával foglalkozó munka főbb megállapításait foglaljuk össze (*Dixon*, 2000; *Roos* ,2008; *Smith et al*., 2004; *Hotz et al*., 2008; *McNabb et al*., 2009; *Garai-Veroszta*, 2013; *Elias*, 2015)

[5] Jóllehet vizsgálatunkban alapvetően az oktatási szférára és annak releváns adatbázisaira fókuszálunk, a kutatási felhasználás folyamata itt is – akárcsak az evidence based szemlélet elterjedésében (Halász 2009) – az egészségügyből indult el. Az átfogó egészségügyi adminisztrációs adatbázisok – beleértve, sőt kiemelten kezelve a finanszírozási adatokat – lehetővé tették a szolgáltatások, ráfordítások időbeni alakulásának vizsgálatát, az egyéni életutak elemzését, az ellátás regionális különbségeinek feltérképezését, melyeket aztán survey adatokkal kiegészítve, vagy éppen összekapcsolva részletesebb vizsgálatokkal tudtak kiegészíteni (Mor 2009).

Adatmennyiség

Az adminisztratív adatbázisok nagymennyiségű adatot tartalmaznak mind elemszám, mind változószám tekintetében. A fő különbség a kutatási adatbázisokkal szemben ez esetben az, hogy a kutatónak nem szükségképpen kell korlátoznia az elemszám avagy változószám alakulását, mérlegelve anyagi szempontokat, illetve a válaszadók toleranciaküszöbét. A terjedelmi korlátok jóval gyengébb volta miatt az adminisztratív adatbázis esetében a „szükség esetén elhagyható" vagy „biztonsági" változók beemelése sem igényel olyan komoly kutatói mérlegelést, mint egy kérdőív esetében. Emellett a nagy elemszám lehetővé teszi relatíve kislétszámú alcsoportok vizsgálatát is.

Rugalmasság

A kutatás elindulása alapvető kutatói döntések előzetes meghozatalát feltételezi. A sorrendiség ennélfogva lényegesen jobban determinálja a survey kutatásokat, hiszen az elemzés céljára létrehozott, vagy szűrt adminisztratív adatbázisba bevont változók köre ismételt adatkéréssel legtöbbször akár utólag is módosítható (ez az integrált adatbázisok esetében nyilvánvalóan nem feltétlenül van így, erről a későbbiekben lesz szó). A rugalmasság időben is értelmezhető. A survey kutatások esetében a longitudinális vizsgálatok tudják ugyan biztosítani az időbeliséget, de például a vizsgálati időhöz képest múltra vonatkozó adatok már csak áttételesen ragadhatók meg. Ehhez képest az adminisztratív adatbázisok egy-egy személy adatait (az adatbázis létrejöttétől kezdődően és utána is kiegészülve) azonos struktúrában, azonos minőségben (megbízhatósággal és érvényességgel) tudják folyamatosan, esetenként nagyon gyakori időbontásban is biztosítani.

Lefedettség

Tekintettel arra, hogy az adminisztratív adatbázisok célja igazgatási, adatai elvben a célcsoport minden tagára kiterjednek. Ez a tulajdonság survey kutatáshoz képest lényegében eliminálja a mintavételi, illeszkedési kérdéseket, megszünteti a válaszadási torzítás mérésének és kezelésének problémáit. Ez az elvi teljeskörűség azonban gyakorlati szinten nem feltétlenül igaz. Semmiképpen nem jelenti például azt, hogy minden egyénről minden változó tartalmaz információt. Az adminisztratív adatbázisoknak is megvannak azok a „vakfoltjai" amelyek nem tudnak megfelelő mennyiségű adatot adni (a minőségről a későbbiekben esik majd szó). Ennek hátterében gyakran szervezési, szabályozási okok állnak (pl. hogy az adat rögzítése nem kötelező) vagy strukturális okok (pl. az adminisztrációs rendszeren, vagy input rendszereken, kategorizáción zajlott időbeni módosítás).

Költségek

A survey kutatások lebonyolításának magas költségigényéhez képest az adminisztratív adatok esetében egy már meglévő adatvagyonról van szó, ami kiiktatja az adatgyűjtés költségeit. Ugyanakkor az adatok kutatási felhasználásához szükséges szervezeti, informatikai és humán feltételek biztosítása – e célra létrehozott professzionális közvetítő szervezetben gondolkodva – tetemes beruházási és fenntartási költség. Utóbbi ráadásul

folyamatos, szemben a survey kutatási programok ad hoc finanszírozási jellegével (ez a longitudinális avagy ciklikus vizsgálatok esetében nyilván nem áll fenn.)

Érvényesség

A validitás kérdése különös jelentőséget kap az adminisztratív adatbázisok esetében. Amíg a kutatási adatgyűjtés esetében a konceptualizálás és operacionalizálás folyamatát előzetesen a kutató végzi el, addig az adminisztratív adatok operacionalizálása egy más – nyilvántartási, szabályozási – célszerűséget szem előtt tartva korábban már lezajlott. Ebből következően az adminisztratív adatok kutatási adatként történő felhasználása véleményünk szerint csak egy ismételt – kutatási célú – validálási szakaszt követően lehetséges, amelynek során feltárjuk az adott változó kontextusát, jelentését, forrását, és mérési problémáit az adott kutatás szempontjából. Ez az újra-operacionalizálás teszi lehetővé, hogy a bevont adatra kutatási adatként építhessünk a maga kutatási érvényességi korlátain belül (avagy vessük el kutatási felhasználását).

Hiányok

Az adminisztratív és kutatási adatbázisokat összevető elemzések mindegyike hangsúlyozza, hogy míg bizonyos tartalmak adminisztratív adatokkal jól lefedhetőek, addig e források alapján a társadalomkutatás számos, gyakran vizsgált releváns területe bizonyosan homályban marad. Hogy mást ne említsünk, a motivációk, percepciók, értékek, tervek, vélemények és vágyak továbbra sem rögzíthetők „kartoték-adat"-ként. Az e kontextusban kevésbé evidensen előkerülő – nem szubjektív – területeken is sok a tartalmi hiány. Az adminisztratív adatbázisok mindegyike rendelkezik egyrészt a lefedettség során már említett vakfoltokkal, azaz nem teljes körűen rögzített adatsorokkal.[6] Másrészt nyilvánvaló, hogy a társadalomkutatás számára elengedhetetlenül fontos objektív adatok jó része sem tartozik az adminisztratív nyilvántartások célkörébe – gondolhatunk itt elsősorban a származási adatokra. Más esetben a hiányt az okozza, hogy változószinten elérhető ugyan adminisztratív adatbázisból a kutatás számára szükséges adat, ám tartalmában nem illeszkedik a vizsgálat céljaihoz. Jellemzően ilyen eset a regionális vizsgálathoz elengedhetetlen lakóhelyi adat, amely adminisztratív megközelítésben a hivatalosan bejelentett, állandó lakcímet jelenti, míg a kutatás számára a tényleges, életvitelszerű tartózkodás helye lehet informatívabb.[7] Összességében a fő problémát nyilván az jelenti, hogy adminisztratív adatok felhasználása esetén a kutatói érdeklődésnek utólag kell hozzáidomulnia a rendelkezésre álló adatok által kínált információtartalomhoz és nem előre tervezetten és strukturáltan gyűjteni be azt, primer kutatás során.

Kontextus

Ahogy az előbbiekben is rendre az adatok mögötti nem kutatási motiváció okozta a felhasználási nehézségeket, az adatok kontextusát tekintve is erre hivatkozhatunk. Az ad-

[6] A diplomás pályakövetési célú adatbázis-felhasználás esetében ilyen jelentős hiány például a külföldi munkavállalás hivatalos rögzítetlensége.

[7] Különösen így van ez olyan a fiatalabb, vagy átmeneti időszakban lévő célcsoportok esetében, mint például a frissdiplomások.

minisztratív adatbázisok kutatási felhasználása során ugyanis nem rendelkezünk az adatbázis létrejöttére vonatkozó kontextuális adatokkal, háttér-információkkal. Pusztán a már rögzített adatokat, adatbázisokat látva azok keletkezésének módja, története és sokszor oka sem ismert, holott alapvetően határozhatja meg az adat vagy adatbázis természetét. Sokszor a kutatási célú felhasználásnak az operacionalizálás részeként az adatok keletkezési közegének feltárása is evidens része. Adatbázisok esetében ez jelentheti a létrehozás és alkalmazás céljának és az adatgyűjtés menetének megismerését. Egyes adatkörök esetében szintén felmerülhet ilyen kontextualizálási igény. Jó példa erre a foglalkozásokat jelző FEOR kódok változója, amely egyfelől az alkalmazott statisztikai besorolás révén objektív mutatóként kínálja magát, az adat keletkezésének vizsgálata azonban számos esetlegességet, további feltárás iránti igényt mutat.[8]

Tervezettség

Az adminisztratív adatokat és adatbázisokat kutatási szempontból érintő érvényességi és kontextus-problémák összességében a tervezhetőség, sorrendiség sajátosságában azonosíthatóak. Primer kutatás esetén az adatgyűjtés tartalma, menete és módja az előzetes elméleti keretrendszerbe ágyazódik be. Az adminisztratív adatok felhasználása esetén a kutató nem tudja kontrollálni az adatok tartalmát és létrejöttük folyamatát sem, ami óhatatlanul gyengíti a módszertani megalapozás minőségét. Hozzátehetnénk, hogy a már meglévő adattartalom mindemellett olykor gyengébb elméleti megalapozásra is csábíthatja a kutatót, de ezt nyilvánvalóan nem tekinthetjük törvényszerűnek. A sorrendiség felborulása egyébként az összevethetőségre is negatívan hat. A nemzetközi survey kutatási projekteket megelőző hosszas adatgyűjtési standardizálási folyamatok épp az egyes adatkörök összevethetősége felé törekszenek. Adminisztratív adatok esetében ez az összevethetőséget lehetővé tévő azonosság visszamenőleg kisebb mértékben ellenőrizhető, biztosítható. Az esetlegesség – annak kis esélye, hogy az adott probléma az egyes országokban egymással összevethető adatok alapján kutatható – nem csak adat, hanem adatbázis szinten is fennáll.

Kontroll

Az adminisztratív és kutatási adatok közti fontos eltérés emellett az ellenőrizhetőség is. Egy kontrollált kutatási helyzetben az adatok keletkezése folyamatosan kézben tartott, ezzel szemben az adminisztratív adatbázisok készen kapott, utólag nem ellenőrizhető adattömeget jelentenek. Nyilván ez esetben az adattisztítási folyamat is egészen más – ám nem elhagyható – eljárásokat követel meg.

Etika

Összevetésünkben nem feledkezhetünk meg az adatok felhasználásához és hozzáféréséhez kapcsolódó etikai vonatkozásokról, még ha a téma részletes kifejtésére ezúttal nem is nyílik tér. A kutatásetika alapelvei nyilván az adminisztratív adatbázisok felhasználása

[8] Például annak tisztázását, hogy a munkaadói adminisztráció során helyi szinten kik és milyen megfontolások, eseti döntések alapján rögzítik e foglalkozási kódokat.

esetében is működnek, ám sok szempontból más a kritikus kérdések köre. Máshogy, más technikákkal szükséges kezelni és biztosítani például az anonimitást. Ugyanakkor az önkéntesség, a nyilvánosság, a kutatási felhasználás kapcsán lényegében a kutatási adatok másodelemzésekor alkalmazott általános elvek tűnnek alkalmazhatónak.

Hozzáférés

Tulajdonképpen szintén etikai kérdéseket vet fel az adatokhoz való kutatói hozzáférés biztosítása is. Ennek jogi szabályozása adott lehet ugyan, de fontos annak átgondolása, hogy az elvileg szabad hozzáférés ténylegesen mennyire függ az adatgazdától, információs egyenlőtlenségektől, kapcsolati hálótól. Sőt, a szabad hozzáférés nem feltétlenül jelenti az eredmények szabad publikálását.

Adminisztratív adat alapú kutatás

Az adminisztratív és kutatási adatok és adatbázisok közti fenti különbségek figyelembe vételével kell eljárnia a társadalomkutatónak akkor, amikor adminisztratív mikroadatokra alapozott kutatást tervez és végez. A (köz)igazgatás igényeihez igazodó adatbázisok esetében a kutatási újraértelmezés elengedhetetlennek tűnik, kiindulva abból a megfontolásból, hogy a kutatás számára ezen adatokat önmagukban véve üresnek kell tekintenünk és operacionalizálási igénnyel szükséges feléjük fordulnunk (*Veroszta*, 2011).[9]

A kutatói munka számára mindez tehát új feladatok ellátása és új kompetenciák fejlesztésének irányába mutat. Az adminisztratív adatbázisokban rejlő kutatási potenciál kiaknázásához megfelelő kapacitás szükségeltetik mind a tudás és szemléletmód, mind a készségek, mind pedig az eszközök szintjén (*Elias*, 2015). Tartalmi szempontból felértékelődik az adattudomány szerepe. Kompetencia-szempontból fontossá válik az adatgondozás és újrafeldolgozás (content curation) képessége. Eszköz-szempontból pedig a nagy adattömegeket kezelni képes megfelelő informatikai háttér lesz elengedhetetlen.

Az adminisztratív adat alapú kutatások főbb lépéseit a brit Administrative Data Liaison Service (ADLS)[10] és az amerikai National Research Council (2005)[11] iránymutatása alapján áttekintve számos olyan szükséges eljárást és megfontolást emelhetünk ki, amelyek a klasszikus empirikus társadalomkutatásnak nem részei. Részben ilyen a korábbi azonos tartalmú adatkérések előzetes feltáró vizsgálata (csak részben, mivel nyilván ezzel párhuzamba állítható a téma kutatási előzményeinek feltárása, amely minden kutatás része). Ezekből a korábbi eljárásokból sok előzetes információ gyűjthető a hatékony adatkéréshez és feldolgozáshoz. Az adminisztratív adatbázisok hozzáférésének kiterjedése sok szempontból új helyzetet teremthet a kutatásban. Az adatmegosztás általában is a kutatás és a kutatói közeg nyitottságát növelheti, bizonyos értelemben az átláthatóság irányába mozdítva a társadalomtudományi kutatást. Azonos empirikus bázison az eredmények megismételhetővé, ellenőrizhetővé válnak. Ebben az adminisztratív adatokhoz való szabad hozzáférés előmozdító folyamat.

---

[9] Az adatok társadalomtudományi újrafelhasználásáról ír e számban megjelent tanulmányában Nagy Péter Tibor.

[10] Az Egyesült Királyságban alapított, az adminisztratív adatok kutatási felhasználást támogató hivatal. Internetes elérhetősége: http://www.adls.ac.uk/

[11] http://www.nationalacademies.org/

Szintén szükséges előzetes döntést hozni arról, hogy a kutatás céljaihoz mely aggregáltsági szintű adatok szükségesek. A legtöbb esetben elengedhetetlen az egyedi szintű (mikro)adatok lekérése, ám kétségtelenül ez a leginkább nehezített eljárás. Ezért több ponton érdemes lehet mérlegelni a jóval könnyebben elérhető aggregált adatok, makrováltozók bevonását is, ahol ez nem sérti a várt kutatási eredményeket. A döntés e tekintetben igen nagy jelentőségű, hiszen a nagyobb részletezettségű adatok jobb használhatósága egyértelműnek tűnik, kifinomult többváltozós modellépítéshez a mikroadatokra feltétlenül szükség van. A kérdés ezért inkább az, hogy az egyéni szintű adathiányok milyen külső makrováltozók bevonásával pótolhatók. Az adatokra vonatkozó kutatói döntésben külön is tekintetbe kell venni a longitudinalitás lehetőségét, amely az adminisztratív adatbázisok esetében gyakran jól hozzáférhető opció.

Az adatokra vonatkozó kutatói döntés utáni fontos szakasz az adatbázist birtokló szervezetek megkeresése és adatgyűjtési és adatkérési dokumentációjuk tanulmányozása. Előbbi rendszerint a kutatás ismertetését, az érzékeny adatok kérésének indokoltságát, a vállalt biztonsági követelményeket tartalmazza. Utóbbi pedig annak vizsgálatát, hogy egészen pontosan milyen változók állnak rendelkezésre. Ennek természetesen elengedhetetlen feltétele a megfelelő transzparencia az adatgazda részéről. A kutatás előkészítése során azzal is számolni kell, hogy az adatgyűjtési – ez esetben adatkérési – szakasz a primer társadalomkutatási adatfelvételhez képest jóval kevésbé kiszámítható, tervezhető. Nagyban múlik ugyanis az érintett szervezet(ek) adatközlési feltételeitől és gyakorlatától (sőt gyakorlottságától).

Adatkapcsolások

Az adminisztratív adatok és adatbázisok kutatási felhasználásának áttekintése után a mikroszintű adatok összekapcsolásáról érdemes szót ejtenünk, hiszen ezen eljárások maguktól értetődően követik az adatbázisok egyre kiterjedtebb rendelkezésre állását. Az adatkapcsolás „data linkage" ezen áttekintésünkben mindazon eljárásokat jelenti, amelyben egy adminisztratív adatbázis valamely megfigyelési egységének adataihoz, illetve adatsorához más adatokat kapcsolunk. A technikai jellemzőkön túl tartalmilag az adatkapcsolásra olyan eljárásként tekintünk, amely több adatbázis információtartalmának integrálását célozza meg. Ez a lehetőség a készen kapott, limitált és rigid adattartalommal, ám teljes lefedettséggel rendelkező adminisztratív adatbázisok esetében különösen fontos szerephez jut és új módszertani eljárások felé nyit utat. Az összekapcsolás kulcsa – mint azt a továbbiakban jól látjuk majd – egy (vagy több) egyedi azonosítást lehetővé tévő változó avagy azok kombinációja.

Az adminisztratív adatbázis alapú adatkapcsolásoknak több típusa is lehet, főbb módjait az alábbiakban kategorizáljuk.

Adminisztratív adatbázison belüli összekapcsolás

Ezzel az eljárással ugyanazon adatbázis megfigyelési egységének egyedi adatsoraihoz kapcsolunk időbeliségében vagy tartalmában kiegészítő változókat. Ilyen lehetőség például a longitudinális, időben visszamenőleges vizsgálat egyazon adatbázison belül. Azonos adatbázison belül az egyedi azonosítók rendelkezésre állása nem okoz problémát (hacsak a feltöltöttség időben nem különbözik).

Adminisztratív adatbázisok közti összekapcsolás

A több adminisztratív – azaz teljeskörű – adatbázis közti egyénsoros összekapcsolás feltétele az ezt lehetővé tévő, mindkét adatbázisban megtalálható egyedi azonosító kód (match-merge eljárás), avagy a személyes adatok olyan kombinációja, ami egyéni megkülönböztető adatképpé áll össze (deterministic linkage). Mindkét módon eljártak már a hazai társadalomtudományi gyakorlatban is. A hazai diplomás pályakövetési célú adatkapcsolás során például különböző időszakokban az összekapcsolás mindkét módját alkalmazták. A cél itt a felsőfokú tanulmányi adatok és a munkaerő-piaci kimeneti jellemzők összekapcsolása volt. Előbbi esetben a Felsőoktatási Információs Rendszerben a 2009-ben abszolutóriumot szerzettek teljes évfolyamára vonatkozóan rendelkezésre álló képzési és szociodemográfiai adatsorokat egészítették ki több állami adminisztratív rendszer adataival,[12] a kapcsolati kód kialakításához felhasználva – majd később törölve – a végzettek személyes azonosításra alkalmas adatait (*Fodor – Veroszta*, 2013). Ilyen személyes adatok jellemzően a név, születési dátum, lakcím stb. A pályakövetési célú adatkapcsolás későbbi – 2013[13]-as illetve 2014[14]-es – fázisaiban már a másik ismertetett metódus, a több adatbázisban megtalálható egyedi azonosítók alapján történt az összekapcsolás (mely ez esetben a személyek TAJ számából, a kapott hashkód-képző eljárással előállított anonim kód volt). Ez az eljárás lényegesen jobb lefedettséget eredményez (*Nyüsti-Veroszta*, 2014a; *Nyüsti-Veroszta*, 2015).

Adminisztratív adatok összekapcsolása makrováltozókkal

Ezekben az adatkapcsolásokban az adminisztratív adatbázisban lévő egyénsoros (ill. a megfigyelési egységre vonatkozó) mikroadatokhoz nem egyéni szintű adatok kapcsolódnak más adatbázisokból, hanem az egyedi adatok magasabb aggregáltsági szintjéhez kapcsolunk külső makrováltozókat. Ilyen kontextuális információt nyújthatnak például a hivatalos statisztikai adatgyűjtés mutatói, indexei. Ez az eljárás adminisztratív és survey adatbázisokon egyaránt gyakran alkalmazott a társadalomtudományi kutatásban, nagy szolgálatot tesz hiányzó kutatási dimenziók pótlásában.

Adminisztratív és survey adatbázisok közti összekapcsolás

Ezekben a keresztmetszeti avagy longitudinális adatkapcsolásokban a teljeskörű adminisztratív adatok megfigyelési egységre vonatkozó egyedi adatsorai egészülnek ki survey

[12] Az adatkapcsolásba bevont szervezetek: Felsőoktatási Információs Rendszer (FIR), Adó- es Pénzügyi Ellenőrzési Hivatal (APEH), Országos Egészségbiztosítási Pénztár (OEP), Foglalkoztatási es Szociális Hivatal (FSZH). A vizsgálat alapsokasága a 2008/2009-ben végzettek teljes köre, az adatgyűjtés a 2010-es státuszra vonatkozott.

[13] Az adatkapcsolásba bevont szervezetek: Felsőoktatási Információs Rendszer (FIR), Nemzeti Adó- és Vámhivatal (NAV), Országos Egészségbiztosítási Pénztár (OEP). A vizsgálat alapsokasága a 2009/2010-ben végzettek teljes köre, az adatgyűjtés a 2012-es státuszra vonatkozott.

[14] Az adatkapcsolásba bevont szervezetek: Felsőoktatási Információs Rendszer (FIR), Diákhitel Központ Zrt., Magyar Államkincstár (MÁK), Nemzeti Adó- és Vámhivatal (NAV), Országos Egészségbiztosítási Pénztár (OEP), Országos Nyugdíjbiztosító Főigazgatóság (ONYF), Nemzeti Munkaügyi Hivatal (NMH). A vizsgálat alapsokasága a 2009/2010-ben illetve 2011/12-ben végzettek teljes köre, az adatgyűjtés a 2013-as státuszra illetve visszamenőlegesen három évre vonatkozott.

kutatásból származó szintén egyedi adatsorokkal. A regiszter alapú összekapcsolás nagy értéke a fentiekben részletezett két típusú (kutatási, ill. adminisztratív) adattartalom erényeinek kombinálása. A megvalósítás történhet a teljes populációra vonatkozó adminisztratív adatbázison végzett, ahhoz kötött utólagos survey kutatással, amelynek adatai aztán az adminisztratív adatbázis egyedi azonosítói alapján visszavezethetőek (*Sakshaug et al*., 2012; *McNabb*, 2009).

Más esetben már meglévő különböző, ám azonos alappopulációt lefedő adminisztratív, illetve survey adatbázisok egyedi szintű összekötéséről van szó (*Groen*, 2012). Az ehhez alkalmazott eljárásként rendelkezésre álló valószínűségi adatkapcsolás (probabilistic record linkage) statisztikai eljárással azonosítja a két, azonos alappopulációt lefedő adatbázis tagjai közti kapcsolatot. Az eljárás matematikai alapjainak kidolgozása Fellegi-Sunter (1969) nevéhez fűződik. A módszer (akárcsak a deterministic linkage) a mindkét adatbázisban rendelkezésre álló egyéni szintű, személyes adatok kombinációjára épül, de nem követeli meg a megfigyelési egység szintű teljes egyezést. Az eljárás megkülönböztet egyező, nem egyező és bizonytalanul egyező kapcsolódást a különböző forrású adatsorok között. Utóbbi esetben az egyezés statisztikai valószínűségével számol, és ezt rendeli hozzá az egyedi adatsorhoz. A probabilistic record linkage alkalmazása a társadalomtudományokban az erős statisztikai háttér mellett a számítástechnikai lehetőségek és ismeretek fejlődését is előfeltételezi (*Winkler*, 1999). Amellett, hogy az eljárás a különböző adatbázisok adattartalmának egyesítésével azok információtartalmát tudja megsokszorozni, fontos szereplő az adatminőség javítására – a lefedettség, mintavétel, válaszadás problémáinak kezelésére – alkalmazott statisztikai eszközök között (*Kreuter et al*., 2010; *Schnell*, 2013). Ilyen jellegű és célú adatkapcsolást kísérleti jelleggel a Diplomás Pályakövetési Rendszer keretein belül is végezteka survey jellegű online adatfelvételekben feltételezett szisztematikus hiba azonosítása céljából. Ennek során ugyanazon végzett évfolyamra vonatkozóan vetették össze a survey kutatási adatok eltérésének mintázatait az adminisztratív adatbázis (FIR) alapján lefedett teljes célpopulációtól (*Nyüsti-Veroszta*, 2014b).

***

Az adminisztratív adatok kutatási felhasználásának elméleti és gyakorlati feltételeit és következményeit vizsgálva tanulmányunkban az adminisztratív és survey kutatási adatok eltéréseit és felhasználásuk módszertani lehetőségeit vetettük össze. Az adminisztratív adatok kutatási felhasználásához kapcsolódó szempontokat ezt célzó gyakorlati eljárásokkal egészítettük ki. Jóllehet munkánk során kevésbé elemző, sokkal inkább áttekintő megközelítést alkalmaztunk, elméleti kiindulópontunkhoz ennek során mindvégig igazodtunk. Eszerint a társadalomkutatásban egyre nagyobb figyelmet és súlyt kapó adminisztratív adatok és adatbázisok csak a megfelelő kutatási validálási eljárások után tekinthetők kutatási felhasználásra alkalmas adatállománynak. A validálási eljárás során pedig – az eltérő adatközeg ellenére – a társadalomkutatás hagyományos szemléletmódjához szükséges és lehetséges igazodnunk. Tanulmányunkban ennek a fontosnak ítélt adatmunkának a főbb megfontolásait és lépéseit vettük számba.

IRODALOM

CARD, D., CHETTEY, R., FELDSTEIN, M.& SAEZ, E. (2011): *Expanding Access to Administrative Data for Research in the United States*. National Science Foundation. pp. 6.

CSERES-GERGELY Zs. & SCHARLE Á. (2008): *Az államigazgatásban keletkező adatok nyilvánosságáról*. Kézirat. MTA-KRTK-KTI. pp.12.

DIBBEN, C., GOWANS, H., ELLIOT, M., ANTTILA, C.,BOYLE, P., KAYE, J., MCcLENNAN, D., NOBLE, M., SMITH, G. & WILKINSON, K. (2009): *Encouraging the wider and more creative use of administrative data in the UK* - the 'Administrative Data Liaison Service. pp.10.

DIXON, S. (2000): Using administrative data sources in labour market Research. *Labour Market Bulletin* 2000/02 Special Issue, pp. 26-30.

ELIAS, P. (2015): *New forms of data – new opportunities for research.* Trans-Atlantic Platform Social Sciences and Humanities, 11th February 2015.

FELLEGI, I. A.& SUNTER, A. B. (1969): A theory for record linkage. *Journal of the American Statistical Association*. 64. 328. pp. 1183–1210.

FODOR SZ.& VEROSZTA ZS. (2013): Államigazgatási adatok pályakövetési célú integrációja a hazai gyakorlatban. In: GARAI O. & VEROSZTA ZS. (Eds.) *Államigazgatási adatbázisok a diplomás pályakövetésben*. Educatio Társadalmi Szolgáltató Nonprofit Kft. pp. 83-128.

GARAI O. & VEROSZTA ZS. (Eds.) (2013): *Államigazgatási adatbázisok a diplomás pályakövetésben*. Educatio Társadalmi Szolgáltató Nonprofit Kft. pp. 128 .

GROEN, J. A. (2012): Sources of Error in Survey and Administrative Data: The Importance of Reporting Procedures. *Journal of Official Statistics*. 28. 2. pp. 173–198.

GROGGER, J. (2013): Bounding the Effects of Social Experiments: Accounting for Attrition in Administrative Data. *NBER Working Paper* No. 18838.

HALÁSZ G. (2009): Tényekre alapozott oktatáspolitika és oktatásfejlesztés. In: Pusztai G.– Rébay M. (szerk.) *Kié az oktatáskutatás?* Csokonai Könyvkiadó, Debrecen. pp.187-191.

HALÁSZ G. (2010): *Az oktatáskutatás globális trendjei*. Vitaanyag. A Magyar Tudományos Akadémia Pedagógiai Bizottsága. Kézirat. Budapest. pp. 67.

HAMMERSLY, M. (2001): *Some Questions about Evidence-based Practice in Education*. Paper presented at the symposium on „Evidence-based practice in education" at the Annual Conference of the British Educational Research Association, University of Leeds, England, September 13-15.

HOTZ, V.J., GOERGE, J., BALZEKAS, J. & MARGOLIN, F. (2008): *Administrative Data for Policy-Relevant Research: Assessment of Current Utility and Recommendations for Development*. A Report of The Advisory Panel on Research Uses of Administrative Data of the Northwestern University/ University of Chicago Joint Center for Poverty Research. pp. 116 .

KREUTER, F., MÜLLER, G. & TRAPPMANN, M. (2010): *Nonresponse and measurement error in employment research: Making use of administrative data. Public Opinion Quarterly*, Vol. 74, No. 5. pp. 880–906.

MCNABB, J., TIMMONS, D., SONG, J.& PUCKETT, C. (2009): Uses of Administrative Data at the Social Security Administration. *Social Security Bulletin*, Vol. 69. No. 1.

MOR, V. (2009): Administrative Data Systems in Social Science Research on Health and Aging. In: McKinlay, J.B. – Marceau, L.D. (Eds.): Behavioral and Social Science Research Interactive Textbook.

MOSLEY, H. (2012): *Pes Data Products and Services for Labour Market Policy Evaluation in Germany*. Peer Review on "Evaluation of Labour Market Policies and Programmes: the use of data-driven analyses, Brussels. 19-20 November, pp. 11.

NATIONAL RESEARCH COUNCIL (2005): *Expanding access to research data: Reconciling risks and opportunities*. In Panel on Data Access for Research Purposes, Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington, DC: National Academies Press. pp. 120.

NYÜSTI SZ. & VEROSZTA ZS. (2014a): *Diplomás pályakövetési adatok 2013.* Adminisztratív adatbázisok integrációja. Educatio Társadalmi Szolgáltató Nonprofit Kft. pp. 114.

NYÜSTI SZ. & VEROSZTA ZS. (2014b): *Linking Administrative And Survey Data In Educational Research. Possibilities And Limitations Of Using Probabilistic Record Linkage In Graduate Career Tracking.* Poster presented at ECER 2014 „The Past, the Present and Future of Educational Research in Europe" Porto, 1 - 5 September.

NYÜSTI SZ. & VEROSZTA ZS. (2015): *Adminisztratív adatbázisok integrációja 2014* - Gyorsjelentés. Educatio Társadalmi Szolgáltató Nonprofit Kft. pp. 20.

OECD (2007) *Istanbul Declaration*

OECD (2013) *New Data for Understanding the Human Condition: International Perspectives*. OECD Global Science Forum Report on Data and Research Infrastructure for the Social Sciences pp. 52 .

ROOS, L.L., BROWNELL, M., LIX, L., ROOS, N.P. WALLD, R. & MACWILLIAM L. (2008): From Health Research to Social Research: Privacy, Methods, Approaches. Social Science & Medicine 66(1) pp.117-129.

SAKSHAUG, J.W., COUPER, M.W., OFSTEDAL, M.B. & WEIR, D.R. (2012): Linking Survey and Administrative Records. Mechanisms of Consent. *Sociological Methods Research, 2012* vol. 4,1 no. 4. pp. 535-569.

SCHNELL, R. (2013): *Linking Surveys and Administrative Data*. German Record Linkage Center, Working Paper Series.

SMITH, G., NOBLE, M., ANTTILLA, C., GILL, L., ZAIDI, A., WRIGHT, G., DIBBEN, C. & BARNES, H. (2004): *The Value of Linked Administrative Records for Longitudinal Analysis*. Report to the ESRC National Longitudinal Strategy Committee. pp. 48.

The UK Administrative Data Research Network (2012): *Improving Access for Research and Policy*. Report from the Administrative Data Taskforce. pp. 59. .

VEROSZTA ZS. (2011): Adatok az oktatáspolitika és a kutatás közti térben. *Educatio* 2011/4 pp. 521–534.

WINKLER, W.E. (1999): *The state of record linkage and current research problems*. Statistical Research Division, US Census Bureau, pp. 15.