KÖZGAZDASÁGI SZEMLE, LXVI. ÉVF., 2019. NOVEMBER (1230–1236. o.)

SEBŐK ANNA

A KRTK Adatbank Kapcsolt Államigazgatási Paneladatbázisa

A Közgazdaság- és Regionális Tudományi Kutatóközpont Adatbankjában létrejött a legújabb Kapcsolt Államigazgatási Paneladatbázis, az Admin3. A különböző államigazgatási nyilvántartások személyi szintű adatösszekötése – a korábbi hullámokhoz hasonlóan (Admin1 és Admin2) – lehetővé teszi a magyar lakosság 50 százalékos mintáján a népesség munkaerőpiaci, munkanélküliségi, oktatási és egészségügyi jellemzőinek tudományos vizsgálatát 2003 és 2017 között. Az egyéni és vállalati szintű, hosszú idősoros, ugyanakkor természetes azonosítókat nem tartalmazó paneladatbázis egyedülállóan szerteágazó tartalmú. Az Admin3 forrásregiszterei között szerepelnek a Nemzeti Egészségbiztosítási Alapkezelő, a Magyar Államkincstár, az Oktatási Hivatal, a Pénzügyminisztérium és a Nemzeti Adó- és Vámhivatal adatbázisai.*

Journal of Economic Literature (JEL) kód: C8, C80, C81, C82, C89.

2019 nyarán a Közgazdaság- és Regionális Tudományi Kutatóközpont (KRTK) Adatbankja[1] harmadik alkalommal hozta létre az Kapcsolt Államigazgatási Paneladatbázist. A kapcsolt államigazgatási paneladatbázisok – így az Admin3 (2003–2017) is – adatintegrációs eljárással készültek, anonimizált módon, ugyanakkor egyéni szinten tartalmazzák a magyar lakosság felének egészségügyi, oktatási, munkaerőpiaci és munkanélküliségi adatait, valamint a magyar vállalatok Bértarifa-felvételben szereplő tulajdonságait. A létrejött adatbázisok nem tartalmaznak természetes azonosítókat és háztartástáblát, ugyanakkor kutatási szempontból egyedülállóan részletesek.

Az adatbázis korábbi hullámai széles körben ismertek és elismertek a hazai és a nemzetközi tudományos élet különböző területein egyaránt. A korszerű nemzetközi

* Köszönettel tartozom az KRTK Adatbank munkatársainak: *Bálint Mónikának, Czethoffer Évának*, *Köllő Jánosnak*, *Sinka-Grósz Zsuzsannának, Szabó Endrének* és *Tir Melindának*.

[1] Korábban: Magyar Tudományos Akadémia Közgazdaság- és Regionális Tudományi Kutatóközpont (MTA KRTK) Adatbankja.

Sebők Anna, a KRTK KTI fiatal kutatója és a KRTK Adatbank munkatársa.

A kézirat első változata 2019. október 28-án érkezett szerkesztőségünkbe.
DOI: http://dx.doi.org/10.18414/KSZ.2019.11.1230

tudományos célú adatkezelő rendszerek kritériumainak megfelelően, az adatok biztonságos szerverkapcsolattal, kontrollált körülmények között, kizárólag kutatási céllal kerülnek felhasználásra.

Az adatösszekötés

Az államigazgatásban felgyűlő információk kutathatósága nagy múltra tekint vissza Magyarországon. A Központi Statisztikai Hivatal adminisztratív adatkezelési gyakorlatát, valamint az államigazgatási tevékenység során összegyűlt adatok kutatási felhasználásának lehetőségét a vonatkozó nemzetközi és magyar jogszabályok – lásd például az általános adatvédelmi rendeletet *(General Data Protection Regulation, GDPR)* – mind rögzítik. Ezek alapján forrásregisztert érintő, illetve több nyilvántartás együttes vizsgálatát célzó kutatási adatbázisok egyaránt létrehozhatók. Ez utóbbit a döntés-előkészítéshez szükséges adatok hozzáférhetőségének biztosításáról szóló 2007. évi CI. törvény teszi lehetővé. Az egyesített adatbázisok előállításához szükséges adatintegrációt a fenti törvényben rögzített költségvetési szervek vezetői indíthatnak, amelyet technikailag kizárólag a Nemzeti Infokommunikációs Szolgáltató Zártkörűen Működő Részvénytársaság (NISZ Zrt.) hajthat végre.

A Kapcsolt Államigazgatási Paneladatbázis összes hullámának alapelve, hogy az adott összekötés időpontjában fellelhető és összekapcsolható összes kutatási célra felhasználható regisztert egyesíti. Így a legutóbb összekapcsolt Admin3-ban a Nemzeti Egészségbiztosítási Alapkezelő (NEAK), a Magyar Államkincstár (MÁK), az Oktatási Hivatal (OH), a Pénzügyminisztérium (PM) és a Nemzeti Adó- és Vámhivatal (NAV) egyéni és vállalati szintű adatai lettek összekötve. Az összekapcsolásnak köszönhetően összefüggéseikben is kutatható adatkörök – hasonlóan a korábbi kapcsolt államigazgatási paneladatbázisokhoz – a következők.

Egészségügyi terület: a társadalombiztosítási azonosító jelből (TAJ) képzett anonimizált azonosító és a TAJ-regiszter, lakcímadatok, társadalombiztosítási jogviszonyra vonatkozó, közgyógyellátási, háziorvosi, járóbeteg- és fekvőbetegellátási, halálozási, vénykiváltási, társadalombiztosítási és pénzbeli ellátási adatok, egyéni szinten.

Munkaerőpiaci terület: munkavállalói, munkaerőpiaci, közfoglalkoztatási és munkaerő-kiközvetítési adatkörök, egyéni szinten.

Szociális transzferek területe: nyugdíjfolyósítási, pénzbeli ellátási, munkanélküliségi, munkaerőpiaci programokhoz kapcsolódó adatok, egyéni szinten.

Oktatási terület: felsőoktatási képzés, felsőoktatási jogviszony, köznevelési jogviszony, érettségi, országos kompetenciamérés adatai, egyéni szinten.

Vállalati terület: társasági adó (tao) bevallásból és a Bértarifa-felvételből származó adatok, vállalati szinten, ugyanakkor egyénekhez köthetően.

Az adatintegrációs eljárás alapja a különböző államigazgatási regiszterek egyéni szintű, esetleg valamely egyéb egyedi azonosítóval rendelkező egységek szerinti összekötése. Így annak eredményeként a különféle adminisztratív regiszterek

tartalma együttesen válik elemezhetővé, megfigyelési egységenként, ugyanakkor anonim módon. Az adatgazdáknál fellelhető adatokat kapcsolati kódok alapján kötik össze, majd törlik a természetes azonosítókat. Kapcsolati kód lehet minden olyan egyedi azonosító, amely több adatgazdánál is fellelhető, így az eredmény-adatbázisban a különböző regiszterekből származó adatokat módunkban áll együttesen vizsgálni.

Mivel a Kapcsolt Államigazgatási Paneladatbázis teljes népességet érintő, univerzális kutatási alapanyag, így a vizsgált mintában szereplők száma nem haladhatja meg a teljes sokaság 50 százalékát [335/2007. (XII. 13.) kormányrendelet]. A minta leválogatását ez alkalommal is a Nemzeti Egészségbiztosítási Alapkezelő mint a teljes magyar népességet kvázilefedő nyilvántartó végezte, a 2003-ban TAJ-számmal rendelkezők állományából. Az alapsokaság leválogatása a többi adatgazda számára is ismeretes kapcsolati kódok (TAJ-szám) listájának létrehozásával kezdődik. Az anonim adatkapcsolásért felelős NISZ Zrt. által kifejezetten az adott összekötéshez generált hash-algoritmus[2] segítségével, az alapsokaságot leválogató regiszter kezelője az eredeti kódokhoz egyedi, technikai azonosítókat rendel. A következő lépésben a többi adatközlő is leválogatja az alapsokasághoz kapcsolódó adatait, és *hashelt* állapotban továbbadja azt a NISZ Zrt. számára, amely végül egyesíti és anonimizálja az adatbázist. Így a létrejött adatbázis nem tartalmaz természetes azonosítókat. A kapcsolt államigazgatási paneladatbázisok esetében az egyéni adatok összekapcsolásán (TAJ-szám-sokaságon) kívül a munkáltatói adatok is hasheltek: a foglalkoztatói adószám segítségével a munkáltatók pénzügyminisztériumi bértarifa-, NAV- és MÁK-adatokat is egyesíti az adatbázis. Az egyesítés után az eredeti azonosítókat (például a TAJ-számot és a foglalkoztatói adószámot) elvesztve, ugyanakkor a személyek adatait összekötve kapta meg az KRTK Adatbank nyers, kutatásra még alkalmatlan formában. Ezt követően kezdődik az adatok tisztítása, kutatási kérdésekhez simítása.

Az adminisztratív adatok sajátosságai

Az eljárásból következően az adatintegráció kényszerűen magában hordozza és továbbörökíti az adatszolgáltatók tartalmi és adatrögzítési hibáit. Az elemzéshez szükséges előzetes tisztítás és adatértelmezés során figyelembe kell venni, hogy az adminisztratív adatoknak célhoz kötött tartalma, terminológiája és struktúrája van, azaz elsősorban az államigazgatási nyilvántartások szempontrendszere által orientált logika alapján jönnek létre. Ezek ismerete a kutatási kérdés szempontjából releváns adatkörök tekintetében feltétlenül szükséges az adatok kutatási célú feldolgozásához és elemzéséhez.

[2] A Hash-algoritmus egyirányú kódolási gyakorlat, amely a bemeneti adatból a következő feltételek teljesülése mellett képez kimeneti adatot: adott bemeneti információból mindig ugyanazt a kimenetet adja, valamint a kimeneti adat egyértelműen utal a bemeneti adatra, de a kimeneti adatból nem állítható elő a bemeneti adat. Az eljárásban a bemeneti adat legkisebb változása is teljesen más kimenetet eredményez. Hash-módszereket használnak tömörítési, jelszótárolási, keresési eljárásokhoz is. Esetünkben az eljárás az anonim technikai azonosítók létrehozását szolgálja.

A következőkben az adminisztratív alapú adatbázisok tulajdonságait a hagyományos kutatási terminológiába illesztve mutatjuk be. A tisztítást követően előálló kutatási adatbázis érvényessége azt jelenti, hogy az összeállított változók milyen mértékben fedik le a kutatási koncepció alapján azonosított kérdéseket, vagyis a mérés tárgya mennyire ragadható meg a rendelkezésre álló adatokkal. Az érvényesség a megalapozó vizsgálatokat követő tisztítási folyamattal, valamint az értelmezési keretek pontos rögzítésével és az ezek szerinti adatkezeléssel fokozódhat, de a kutatási kérdésekhez teljes mértékben illeszkedő adattartalom csak közelítőleg érhető el. Az adattartalom egyes kutatási kérdésekre vonatkozó megbízhatósági szintje jellemzően magas, vagyis ismételt mintavétel hatására csak kismértékben változik, de az adatbázisok belső dinamizmusa, fluktuációja miatt itt is tapasztalhatók minimális eltérések. A konceptualizálás (a mérés tárgyának definiálása), illetve az operacionalizálás (a mérés módjának meghatározása) – szemben a hagyományos, jellemzően tudományos kérdésfeltevésből kiinduló vizsgálatokkal – adminisztratív adatforrások felhasználása esetén iteratív módon, párhuzamosan, több lépésben történik. A végső realizáció, azaz a létrejött változók tényleges számítási menete tehát utólagosan, általában hosszas kísérletezés után rögzíthető.

Az adatösszekötést követően hosszú idősoros paneladatbázist készít a KRTK Adatbank, amely a legfrissebb összekötés során több száz nyers mező harmonizálásával jön létre. A szakszerű – több mint tízéves időszakot átfogó – adattisztítás hosszú időt vesz igénybe. Éppen ezért a jelenleg összekapcsolt adatbázis tudományos célú elemzésére leghamarabb 2020–2021 folyamán kerülhet sor.

Az adattisztítás

Az adattisztítás és harmonizálás sajátos módon, az adatok adminisztratív tulajdonságaihoz alkalmazkodik. Az egyesített paneladatbázisok adatai részben strukturáltak, az időbeliség dimenzióját tekintve statikusak, ugyanakkor időben visszafele longitudinális kutatást tesznek lehetővé. A megfigyelhető egység az adminisztratív státusok változása, tehát vélemények helyett konkrét viselkedések vizsgálhatók.

Az adminisztratív adatok itt részletezett tulajdonságai miatt a longitudinális és a keresztmetszeti konzisztenciavizsgálatok egyaránt részét képezik az adattisztítás folyamatának. A keresztmetszeti konzisztenciavizsgálat során a különböző mezők tartalmának adott időpontra vonatkozó összevetését végezzük. Az eljárás során felmerülő anomáliák megismerésével képet kaphatunk az adatok korlátairól is. Az adminisztratív jelleg miatt a tág értelemben vett adatkörnyezet vizsgálata is elkerülhetetlen a teljes vizsgálati időtartamra vonatkozóan – amennyiben a folyamatok, időbeli változások követése kutatási cél –, retrospektív módon (*Veroszta* [2015]). Ezzel párhuzamosan, az adatok longitudinális ellenőrzésével az évről évre változó adatkörök tartalmát, változásait figyeljük meg. Emellett az adatok értelmezéséhez elengedhetetlen a kódszótárak frissítése is a teljes megfigyelési időszakra. A tisztítás során a fent részletezett megalapozó vizsgálatokat követően, elemzői döntések nyomán nyerhető ki a kutatási kérdéseknek megfelelő adattartalom az

önmagukban csupán adminisztratív (az adott adatgyűjtés célja és kontextusa által körülhatárolt) jelentésű cellákból.

A szűk értelemben vett adatbanki tisztítás a beérkezett adatok, tehát az adatszolgáltatók és a NISZ Zrt. munkájának áttekintésével kezdődik. Az adatátadás ellenőrzése is a korábban részletezett adminisztratívadat-sajátosságok figyelembevételével történik. Az átadott adatok ellenőrzése után, a nyers adatmezők iteratív átalakítása során különféle kutatási kérdéseknek megfelelő változók létrehozása következik, először csupán az egyes adatforrások regisztertartalmán belül.

Miután adatforrásonként megtörtént az adatok tisztítása, megismerése és harmonizációja, kezdetét veszi a szűkített, ugyanakkor legfontosabb változókat összekapcsolva tartalmazó óriás adatbázis felépítése. Az így létrejött adatbázis az Admin3 esetében az egyének 2003 és 2017 közötti, havi szintű státusait tartalmazza. A longitudinális (időbeli) és keresztmetszeti (különböző adatközlők összevetésén alapuló) harmonizációs, ellenőrzési és tisztítási hullám is ebben a fázisban zajlik, valamint ekkor jönnek létre a több adatforrásból származó változók. Az összekapcsolást követően megmutatkozó adatbázis-inkozisztenciákat (folyamatában) kezeljük.

Ezáltal létrejön egy nagy részben konzisztens óriás adatbázis. Ezen a bonyolultabb változókat felhasználva összetett vizsgálatok, minikutatások indulnak, annak érdekében, hogy a későbbi adatfelhasználáskor felmerülő potenciális adathibákra fény derüljön. Az ilyen módon felmerülő kérdések és megoldások szintén beépülnek a javított adatbázisba.

Ezt követi a harmonizáció kollektív fázisa, amelynek során a tudományos közösség különböző területeiről érkező szakértők lehetőséget kapnak a Kapcsolt Államigazgatási Paneladatbázis használatára, ezzel együtt a korábbi tapasztalataik tisztítási eljárásba építésére, annak tökéletesítésére. A kollektív harmonizáció fázisában gyűjtött információk (akár megírt programkódok) beépülnek a következő Admin-hullám első körös adatbanki tisztításaiba, organikus módon folyamatosan továbbfejlesztve azt. Hasonló módon, minden későbbi adathasználat során felmerülő kérdés, probléma és visszajelzés, valamint az azokra megírt programkódok is beépülnek az Admin-fáljok tisztító programjába.

Hozzáférés

Az adatintegráció során a NISZ Zrt. anonimizálja az adatokat, így azok elvesztik az utólagos azonosításra alkalmas vonásaikat. Az ilyen módon létrehozott adatbázisok nyers adattartalmának közzétételét szintén a NISZ végzi. Az adatösszekötések nyers tartalmát leíró változókat és a partnerek listáját a kapcsolódó szerződés melléklete tartalmazza, amihez a NISZ Zrt. kérésre hozzáférést biztosít a már említett 335/2007. (XII. 13.) kormányrendelet alapján.

A fenti szempontok szerint tisztított, jellemzően havi bontású egyéni státusokat tartalmazó óriás adatbázist az Adatbank a tisztítás kollektív szakaszában csupán a KRTK kutatói számára teszi közzé, akik ekkor részt vesznek a szakirányú adattisztítási munkában.

A szakirányú tisztítást követően a kapcsolt államigazgatási paneladatbázisokhoz doktori disszertációkhoz és szakdolgozatokhoz minden esetben, valamint megfelelő affiliáció és kutatási cél esetében adhatunk hozzáférést. Jelenleg az Admin1 (2002–2008) és az Admin2 (2003–2011) adatbázis érhetők el. Az adatbázisokon folyó munka biztonságos szerver- és STATA-alapú szoftverkörnyezetben történik, amely használatára az adatkérő lapon szereplő, előre rögzített határidőig van mód.

Kutatási relevancia, jelentőség

A Kelet- és Közép-Európában egyedülállóan gazdag Admin-adatbázisokat széles körben használják hazai és nemzetközi kutatói körökben. Kifejezetten alkalmasak színvonalas nemzetközi folyóiratokban megjelenő tudományos eredmények, publikációk előállítására, hiszen szerteágazó tudományterületeken lehetőséget nyújtanak összetett longitudinális és keresztmetszeti elemzésekre. A Kapcsolt Államigazgatási Paneladatbázis jelenleg kutatási céllal elérhető Admin1 és Admin2 hullámait eddig több mint 80 kutatás során használta közel félszáz kutató. Az egészségtudomány, az egészségpolitika, a regionális tudományok, a munkagazdaságtan, a vállalatkutatás, a migrációkutatás, az agronómia és szociálpolitika területén egyaránt találunk Admin-alapú, éppen folyó kutatásokat, illetve már megjelent nemzetközi tudományos publikációkat.

A létrejött tudományos produktumok számos hazai szakmai folyóirat mellett nemzetközi tudományos lapokban is megjelennek, ilyenek többek közt az American Economic Journal (*Lindner–Reizer* [2019]), a Quarterly Journal of Economics (*DellaVigna és szerzőtársai* [2017]), a Health Economics (*Bíró–Elek* [2018]), az IZA Journal of European Labor Studies (*Czafit–Köllő* [2015]) és a Scandinavian Journal of Public Health *(Adamecz-Völgyi Anna és szerzőtársai* [2018]) vagy a Research in Labor Economics (*Csillag* [2019]) könyvsorozat.

A kapcsolt államigazgatási paneladatbázisok legutolsó, Admin3 hullámának megvalósulása az államigazgatási regiszterek adattartalmának folyamatos javulása miatt hosszabb idősorokat és pontosabb adatokat tartalmaz. Ennek mértéke elsősorban az oktatás területén, különösen az egyelőre még rövid múltra visszatekintő Országos kompetenciamérés szempontjából jelentős.

Kapcsolat

A KRTK Adatbankja Magyarországon egyedülállóan szerteágazó témákban gyűjt *survey*- és adminisztratív alapú kutatási adatbázisokat, alakít ki és üzemeltet teljes körű mikroadatok vizsgálatára alkalmas kutatószobát, hoz létre regiszteralapú, adatintegrációs eljárással készült adatbázisokat, szervez STATA-képzéseket, valamint tart fenn kísérleti labort és ahhoz tartozó gépparkot. Általános tájékoztatást az adatbank@krtk.mta.hu email-címen nyújtunk, míg adatkéréssel kapcsolatosan az adatkeres@krtk.mta.hu email-címen lehet érdeklődni.

Hivatkozások

Adamecz-Völgyi Anna–Bördős Katalin–Lévay–Scharle Ágota [2018]: Impact of a personalised active labour market programme for persons with disabilities. Scandinavian Journal of Public Health, Vol. 46. Suppl. 19. 32–48. o. https://doi.org/10.1177/1403494817738421.

Bíró Anikó–Elek Péter [2018]: How does retirement affect healthcare expenditures? Evidence from a change in the retirement age. Health Economics, Vol. 27. No. 5. 803–818. o. https://doi.org/10.1002/hec.3639.

Czafit Bence–Köllő János [2015]: Employment and wages before and after incarceration. Evidence from Hungary. IZA Journal of European Labor Studies, No. 4. 1–21. o. https://doi.org/10.1186/s40174-015-0044-z.

Csillag Márton [2019]: The Incentive Effects of Sickness Absence Compensation – Analysis of a Natural Experiment in Eastern Europe. Research in Labor Economics, (Health and Labor Markets), Vol. 47. 195–225. o. https://doi.org/10.1108/S0147-912120190000047007.

DellaVigna, S.–Lindner Attila–Reizer Balázs– Schmieder, J. F. [2017]: Reference-dependent job search: evidence from Hungary. Quarterly Journal of Economics, Vol. 132. No. 4. 1969–2018. o.

Lindner Attila–Reizer Balázs [2019]: Frontloading the unemployment benefit: an empirical assessment. Megjelenés alatt, American Economic Journal: Applied Economics.

Veroszta Zsuzsanna (2015): Adminisztratív adatok társadalomkutatási kezelése. Educatio, 23. évf. 3. sz. 3–14. o. http://ofi.hu/sites/default/files/attachments/educatio_2015-3_web_0.pdf.

335/2007. (XII. 13.) Korm. rendelet a döntéselőkészítéshez szükséges adatok hozzáférhetőségének biztosításáról szóló 2007. évi CI. törvény végrehajtásáról. https://net.jogtar.hu/jogszabaly?docid=A0700335.KOR.

2007. évi CI. törvény a döntéselőkészítéshez szükséges adatok hozzáférhetőségének biztosításáról.https://net.jogtar.hu/jogszabaly?docid=a0700101.tv.