Adathozzáférés kutatói szemmel

Hozzászólás az MTA Statisztikai Tudományos Albizottság ülésén

Kézdi Gábor

CEU & MTAKRTK

STAB, 2012. május 24.

A színvonalas, értékes és releváns empirikus kutatás záloga a megfelelő adat

- Adattartalom
  - Megfelelő információkat kell tartalmaznia (egyszerre)
- Adatminőség
  - Megfelelő minőségben
- Adatmennyiség
  - Megbízható statisztikai elemzéshez kell az elemszám

Miben látom én a fejlődési lehetőségeket?

- Meglevő (statisztikai) adatfelvételek tökéletesítése
  - Mintavételi keret minősége
  - Nemválaszolás problémája
  - Szolgáltatott adat minősége
- Új (statisztikai) adatfelvételek
  - Hiányoznak standardnak tűnő információk (pl. háztartások vagyona, pénzügyi döntései)
  - Új irányok lehetnek szükségesek (pl. mintavétel társadalmi kapcsolathálók mentén)
- Adminisztratív adatok kutatói felhasználása
  - Pl. személyi jövedelemadó rekordok, társadalombiztosítási rekordok elemzése

Adminisztratív adaok összekapcsolása egyéb információkkal

Kutatási céllal

- Talán ebben van a legtöbb lehetőség
- Az adminisztratív adatok általában jók
  - adatmennyiség
  - adatminőség: az adatok gyűjtésének technológiája miatt
    - ★ pl. munkaviszony "spell"-ek a társadalombiztosítási rekordokban szemben retrospektív kérdőíves adatfelvételekből nyerhető hasonló változókkal
- De az adminisztratív adatfelvételek általában szűkek
  - kevés változóra terjednek ki
  - a változók tartalma kutatási szempontból gyakran nem tökéletes
  - a lefedett változók halmazát és tartalmát nem a kutatás szempontjai vezérlik
- Standard vagy célzott kiegészítő adatfelvételek és adminisztratív adatok kapcsolásával hatalmas lehetőségek nyílnak a kutatás számára

Egy példa: A Health and Retirement Study összekötése TB adatokkal (USA)

- A Health and Retirement Study (HRS) egy nagyméretű panel
  - az 51+ éves népességet vizsgáló, kétévente lekérdezett panel survey
  - multidiszciplináris adatfelvétel az öregedés kérdéseinek vizsgálatára
- Az 51 éves kor előtti jövedelmeket és nyugdíj-jogosultságokat nem mér (vagy nem jól méri) az adatfelvétel
- Összekapcsolták az egyének survey adatait a TB adatokkal
  - A válaszadókkal aláirattak egy nyilatkozatot, hogy a megfelelő adatkezelési szabályokat betartva összeköthetik-e a kérdőívből származó adataikat TB adataikkal
  - Háromnegyedük aláírta
- A kapcsolt adatok kutatószobában elemezhetők
  - A University of Michigan Survey Research Center-ében
    - ★ én magam is dolgoztam az adatokkal a kutatószobában

Egyéb példák

- A norvég, svéd, finn, dán admin adatok összekapcsolhatók szinte bármivel
  - egymással (pl. a norvég csecsemők születési adatai későbbi TB adataikkal)
  - kiegészítő adatfelvételekkel (pl. a norvég csecsemők közül kik egypetéjű ikrek)
- A HRS európai testvér-felvételének (SHARE) német mintáját is összekapcsolták TB adatokkal
  - A SHARE biomarkereket is elkezdett gyűjteni, és elindult az egyéni adatok összekapcsolása egészségbiztosítási adatokkal
- A privát szférában keletkező egyéni adatok összekapcsolhatók survey adatokkal
  - pl. bankok ügyfelei: tranzakciók teljes története

Kutatószoba a KSH-ban

- A KSH saját szenzitív adatbáziait kutathatóvá teszi kutatószobai környezetben
- Kollegáimmal én is elkezdtem dolgozni ilyen adatokon
  - A 1970-2010 közötti szülések adatain
- A környezet jó
  - Up to date számítástechnikai háttér
  - Rendkívül segítőkész kollegák
- Az adathasználat szabályai sztenderdek
  - Bent bármi vizsgálható, kihozni csak szigorúan ellenőrzött outputokat lehet
- Adatkapcsolásra módvan aggregált szinten
  - pl.településszintű adatok "bevihetők"

A jövő itthon?

- A kutatószoba kialakítása rendkívül pozitív fejlemény a KSH-ban
  - A nemzetközi statisztikai élvonalhoz csatlakozik
- Az adatok egyéni szintű kapcsolása nem lehetséges
  - Nem egyszerű feladat, de nagy lehetősegeket rejt magában
- A többi hazai adminisztratív adat felhasználása még várat magára
  - Én egy-két kivételről tudok, adatkapcsolás nélkül

Köszönöm a figyelmet