

MTA Közgazdaságtudományi Intézet – Adatbank

Olyan érdeklődőt keresünk, aki közreműködne egy különösen finom felbontású elérhetőségi adatbázis fejlesztésében és/vagy azon önálló kutatást végezne.

A GEO névre hallgató adatbázis tartalmazza (a) 45 500 népszámlálási számlálókörzet (a továbbiakban szlok) adatait: népesség, iskolázottság, foglalkoztatás, munkanélküliség, lakások, közintézmények és a többi (b) az ott telephellyel rendelkező vállalatok és költségvetési intézmények pénzügyi és létszámadatait (c) a szlokok közötti utazás távolságát, időigényét és költségét tömegközlekedés és autós ingázás esetén.

Természetesen léteznek ennél nagyobb és pontosabb elérhetőségi adatbázisok, Google Maps, stb., amik naprakész információval szolgálnak arról, hogyan jutunk el A-ból, B-be. Olyan azonban nincs, ahol azt is tudjuk, hogy mi van A-ban és B-ben. Ezért a GEO számos közgazdasági, munkaerőpiaci, oktatási, népegészségügyi probléma kutatásához teremt alapot, különösen azért, mert az MTA-KSH kutatószobában szlok szinten összekapcsolható a 2011-es népszámlálás teljeskörű, egyéni szintű adatállományával.

(a) Ami a folyamatban lévő fejlesztést illeti, a cél az, hogy megbecsüljük az egyes szlokok meghatározott iskolázottságú népessége számára elérhető munkaerőpiacok minőségét, amit az elérhető üres munkahelyekkel (V), és azoknak a másutt élő munkanélkülieknek a számával (U) mérünk, aki számára ugyanezek a munkahelyek szintén elérhetők. Ehhez értékelni kell a lehetséges útvonalakat a bérek, a segélyekből megszerezhető jövedelmek és az utazási költség alapján. A releváns útvonalak leválogatásához a szlok \times iskolázottság interakciókra, a vállalat \times foglalkozás interakciókra és az útvonalakra vonatkozó információkra egyidejűleg van szükség.

Technikai nehézséget jelent a gráf mérete, amit még a rendelkezésünkre álló szerveren (24 mag, 780 Gbyte RAM) sem könnyű kezelni: szigorúan véve 180 ezer kiinduló pontról (szlok \times iskola interakcióról), kb. 4 millió végpontról (munkáltató \times foglalkozás interakcióról) és sokmilliárd potenciális kapcsolatról van szó. Mivel bizonyos ismérvek csúcs-specifikusak, mások pedig él-specifikusak és függetlenek az iskolázottságtól, a foglalkozástól és a munkáltató kilététől, az útvonalak értékelése megoldható úgy is, hogy csak 45 500 kiinduló pontunk, kb. 500 ezer végpontunk és 800 millió élünk van.

Az általunk használt Stata programcsomag képes befogadni ekkora adattömeget, de ezt kevésbé intelligens módon teszi, a szlokokra, a munkáltatókra és az élekre vonatkozó fix formátumú mátrixok rengeteg redundáns információt tartalmaznak, bizonyos kódértékek csoportonként ismétlődnek. Az eljárás Statában nagyon időigényes: vagy nagyon nagy mátrixokat kell beolvasni, vagy több tízezer lépéses ciklusokat kell futtatni. Nagy segítséget jelentene, ha valaki, megfelelő gráfelméleti ismeretekkel felvértezve, segítene a futásidő csökkentésében.

(b) Az adatbázis-fejlesztést és az elemzést is megkönnyítené, ha rendelkeznénk a számlálókörzetek közlekedési szempontból vett összekapcsoltságára (connectedness) vonatkozó eredményekkel.

(c) Az adatbázist a fenti munkálatokban részt vevő gyakornok rendelkezésére bocsátjuk saját kutatás végzéséhez.

A szerverhez távoli hozzáférést biztosítunk. A KTI jelenleg a Budaörsi úton, 2017-től kezdve a Vaskapu utcában (a lágymányosi kampusszal szemben, a túlparton) működik.

Köllő János

az Adatbank vezetője, kollo.janos@krtk.mta.hu, 30/816 4273