A harmadik országbeli állampolgárok adatait tartalmazó adatbázisok integrálásának módszertanát elemző tanulmány

Budapest, 2015. június

A tanulmány „*Migránsokra vonatkozó társadalomstatisztikai adatgyűjtések megalapozása*” c. projekt (EIA/2013/2.6.1.) keretében Kővári Zsolt tanulmányának felhasználásával készült.

Bevezetés

A harmadik országbeliekre vonatkozó migrációs statisztika fejlesztésének kiemelt területe a rendelkezésre álló adatforrások minél hatékonyabb felhasználása. Ebbe a körbe tartoznak az adminisztratív adatforrások és az ország teljes lakónépességére vonatkozó népszámlálás. Az adminisztratív adatforrások tekintetében a KSH legfontosabb adatátvételei, adatgazdák szerint megnevezve a következőek: a Bevándorlási és Állampolgársági Hivatal (továbbiakban BÁH) idegenrendészeti állománya, a Közigazgatási és Elektronikus Közszolgáltatások Központi Hivatala (továbbiakban KEKKH) lakcímnyilvántartása, a Nemzeti Adó- és Vámhivatal (továbbiakban NAV) nyilvántartása a személyi jövedelemadót fizető külföldiekről, Országos Egészségügyi Pénztár (továbbiakban OEP) nyilvántartása a TAJ számmal rendelkező külföldiekről.

A felsorolt szervektől átvett adatállományok tisztítása, olyan eljárások és módszerek kidolgozása, melyek segítségével az említett adatbázisok statisztikai célra alkalmasabbá válnak, további adatforrások felkutatása, az adatbázisok integrálása egyaránt fontos lépések a harmadik országbeliekre vonatkozó statisztikák fejlesztése irányában.

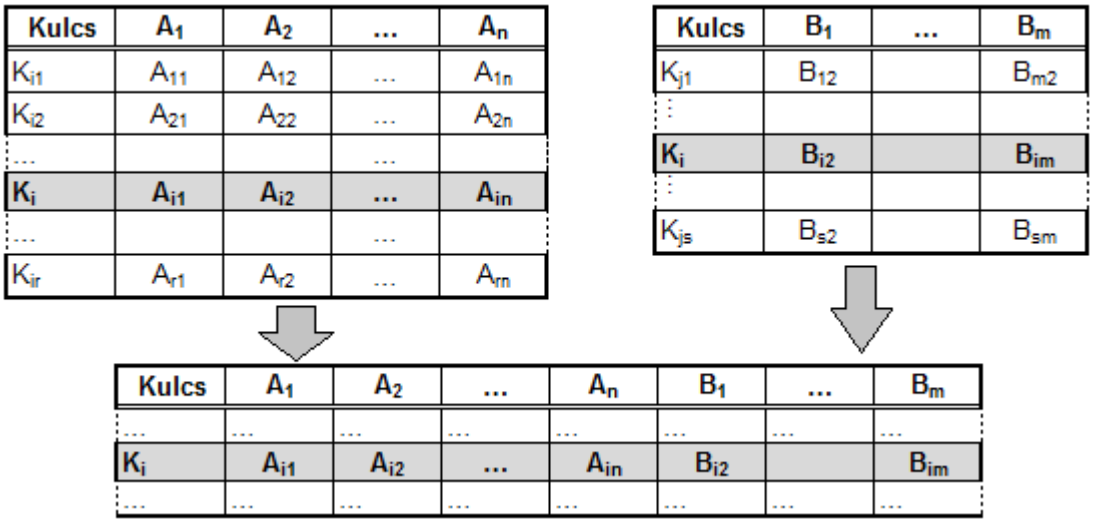
Az EIA/2013/2.6.1 számú, „Társadalomstatisztikai adatgyűjtések megalapozása” című KSH projekt keretei között egyrészt adattisztítási módszereket, technikákat dolgoztunk ki a KSH rendelkezésére álló, harmadik országbeliek adatait tartalmazó adatforrások minőségének javítására, másrészt új szoftver, a RELAIS alkalmazása segítségével kísérletet tettünk egy integrált migrációs adatbázis kialakítására.

1. Adatbázisok összekapcsolásának elméleti áttekintése

Informatikai értelemben adatbázis alatt az adatok és a közöttük levő kapcsolatok valamilyen adatmodell szerint kialakított tárolását értjük. A ma használatos adatbázis-kezelők nagy részében alkalmazott relációs adatmodellt az 1970-es években kezdték kidolgozni. Ennek lényege, hogy a leírni kívánt egységeket (amelyek lehetnek személyek, tárgyak, vállalatok, országok, stb. egy szóval egyedek) több elemi adat azonosítja. Ezek az elemi adatok logikailag kétdimenziós táblázatba szervezhetők, ezeket nevezzük relációknak. A táblázat első sora a reláció fejléce, amely az oszlopok (táblázaton belül egyértelmű) azonosítóit – a reláció attribútumait – tartalmazza. A többi sorban egy adott egyedre vonatkozó adatok szerepelnek. Tekintettel arra, hogy a reláció matematikai értelemben egy halmaz, amelynek elemeit csak egyszer adjuk meg, így nem szerepelhet a táblázatban két azonos sor. Ha ez így van, akkor biztosan található oszlopok olyan összessége, amelyekben előforduló adatok együttese minden sorban különböző. Az ilyen oszlop együttest a reláció kulcsának mondják.

A kapcsolatok leírása a relációs adatmodellben kulcsokon keresztül valósul meg (determinisztikus vagy empirikus adatösszekapcsolás). A legegyszerűbb esetben a kulcs egyetlen oszlop, amely garantáltan minden sorban más értéket vesz fel. Ez lehet egy sorszámozás, vagy valamilyen mesterséges azonosító, mint pl. a TAJ szám, az adószám vagy a személyi szám. Ha léteznek két adattáblában ilyen kulcsok, akkor a mindkettőben előforduló személyek adatai egy új táblázatban előállíthatók. Nem kell ugyanis egyebet tenni, mint az egyik relációban szereplő egyedhez tartozó kulcs értékét kikeresni a másik relációból, és annak többi attribútumát az első mellé másolni, ahogy ezt az 1. ábra mutatja.

1. ábra: Empirikus adatösszekapcsolás



A fent leírt eljárás nyilván nem használható, ha a kulcsok különböző tartalmú azonosítók, azaz értékük egy adott táblában egyedi, de két különböző adatbázisból származó relációban nincsenek közös értékeik. Ilyen eset alakul ki, ha az egyik adatbázisban például a személyi igazolvány számát, a másikban az adószámot használták azonosításra. Előfordulhat az is, hogy az adatlistát nem valamilyen adatbázis-kezelő rendszerben vezetik, így nincs, ami betartassa a kulcsokra vonatkozó kényszereket, megszorításokat, vagy esetleg a kulcsok sérülnek, módosulnak az adatátadás során használt exportálási műveletek alatt. Ilyenkor lehetséges, hogy nincs használható kulcs, ezért az összekapcsolásra más utat kell keresni.

A valószíűségi alapon történő adatösszekapcsolás (probabilistic matching) célja, hogy azonosítsa az azonos való világbeli entitásokat, amelyek különböző módon vannak reprezentálva az adatforrásokban, még akkor is, ha egyedi azonosítók nem állnak rendelkezésre, vagy hibákkal terheltek. A statisztikában egyedi szintű adat-összekapcsolásra (record linkage-re) van szükség számos alkalmazásban, beleértve a különböző adatbázisokban tárolt információk gazdagítását, az adatbázisok duplikáció mentesítését, egy forrás adatminőségének javítását, egy populáció nagyságának jelölés-visszafogás módszerével történő mérésénél, nyilvánosan használt mikroadatok bizalmasságának ellenőrzését.

Ezekben az esetekben az összekapcsolás végrehajtásához szükség van egy döntési szabályra, amely segítségével egy rekordpárról megmondható, hogy ugyanazt az egyedet írja-e le, vagy sem. 1969 decemberében jelent meg Fellegi Iván és Allan B. Santer cikke, (A theory of record linkage", Journal of the American Statistical Association 64) amelyben ismertetnek egy döntési szabály megkonstruálására használható eljárást. Az Olasz Statisztikai Hivatal koordinálásával a közelmúltban kifejlesztettek egy ingyenesen használható szoftvert, amely alkalmas a Fellegi-Santer modell megvalósítására, a neve RELAIS (Record Linkage At ISTAT). Ennek a szoftvernek az alkalmazását, alkalmazhatóságát, korlátait mutatjuk be a következőkben.

2. Record linkage folymatok a RELAIS-ben

A RELAIS projekt célja, hogy a record linkage technikákat a nem szakértő felhasználók számára is könnyen hozzáférhetővé tegye. Valójában a programnak van egy grafikus felhasználói felülete, amely lehetővé teszi, hogy a record linkage munkafolyamatot megfelelő rugalmassággal építsük fel. Másrészt ellenőrzi a különböző felajánlott technikák között a végrehajtás sorrendjét, mivel a precedencia szabályokat muszáj kontrollálni.

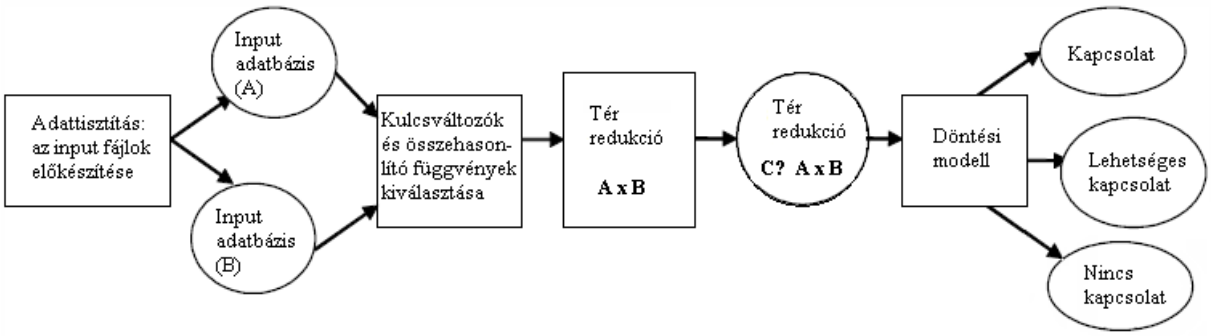
A teljes összekapcsolási folyamat összetettsége számos különböző természetű problémától függ. Ha elérhetők egyedi azonosítók az érintett adatforrásokban, a probléma elég egyszerűen kezelhető, mert a bonyolultság pusztán számítástechnikai kényszerekre redukálódik. Általában azonban nincsenek felhasználható egyedi azonosítók, így komplikáltabb statisztikai folyamatokra van szükség, amelyek a kulcsváltozók megválasztásától függenek.

A RELAIS célja, hogy összekapcsolja a record linkage problémakör lényeges matematikai és számítástechnikai eszközrendszereit.

A record linkage folyamat fázisai

A record linkage folyamat fázisokra történő bontása jelenti a RELAIS eszközkészlet magját. A teljes munkafolyamatot könnyen vezérelhetővé teszi, hogy mindegyik fázis saját ablakkal rendelkezik. Jelen tanulmányban általánosan tekintjük át a főbb fázisokat.

2.ábra: A record linkage folyamat fázisai



Általánosan szólva az input fájlok előkészítése az első fázis, amely Gill (2001) szerint a record linkage folyamat megvalósítására tett erőfeszítések 75%-át elviszi, mivel az adatok különböző formátumokban fordulhatnak elő, lehetnek közöttük hiányzók, következetlenek vagy hibásak. Ennek a fázisnak a kulcstevékenysége, hogy az input adatokat egy megadott formába konvertáljuk, feloldva a következetlenségeket, hogy csökkentsük az inkorrekt módon jelentett adatokból eredő hibákat. Ebben a fázisban töröljük a nullahosszúságú szövegeket, a rövidítéseket, írásjeleket, kis- és nagybetűket egységes alakra hozzuk, és ha szükséges, megfelelő transzformációkat hajtunk végre a változók standardizálása érdekében. A gyakori szavak különböző helyesírással írt változatait egységes helyesírásúra cseréljük le.

Az előzetes fázis végrehajtása után a következő fontos lépés a kulcsváltozók kiválasztása, amelyek annyira alkalmasak a tekintett összekapcsolás számára, amennyire csak lehetséges. Az összekapcsoló tulajdonságokat általában a terület szakértője választja, a RELAIS metaadatok[1] biztosításával támogatja a felhasználókat a kulcstulajdonságok kiválasztásában. Ha vannak egyedi azonosítók az összekapcsolni kívánt adatbázisokban, a legegyszerűbb és leghatékonyabb út ezek használata, de szigorú ellenőrzés szükséges, ha egyedül numerikus azonosítót használunk. Az olyan változók, mint a név, vezetéknév, cím és születési dátum használhatók

[1] Az alábbi metaadatok használhatók fel jelenleg: Teljesség (Completeness); Pontosság (Accuracy); Következetesség (Consistency); Entrópia (Entropy) – a változó szerinti heterongenitást mutatja be; Korreláció (Correlation); Gyakorisági eloszlás (Frequency distribution)

együttesen, ahelyett, hogy külön-külön dolgoznánk velük. Ilyen módon csökkenthetőek az olyan problémák, mint a nevek leírásának széleskörű változatossága, vagy amelyek amiatt lépnek fel, hogy a családi név megváltozik például a családi állapot módosulása okán. Az nyilvánvaló, hogy minél heterogénebbek egy változó értékei, annál inkább alkalmasak azonosításra, továbbá, ha a hiányzó esetek száma egy mezőben magas, akkor az nem használható kulcsváltozóként.

A harmadik fázis az összehasonlító függvények kiválasztása. Az összehasonlító függvények segítségével a program kiszámítja a rekordok közötti távolságot a kiválasztott kulcsváltozók alapján. A RELAIS jelenlegi verziójában számos összehasonlító függvény érhető el, mégpedig a következők:

- Egyenlőség (Equality)
- Numerikus összehasonlítás (Numeric Comparison)
- 3Grams
- Dice
- Jaro
- JaroWinkler
- Levenshtein
- Soundex

Az 1. Számú mellékletben részletesen kitérünk ezekre a függvényekre.

A negyedik fázis a kapcsolásra jelölt párok keresési terének létrehozása és redukálása. Az A és B adathalmazok összekapcsolási folyamatában a párokat osztályoznunk kell a következőképpen: kapcsolat, nem kapcsolat és lehetséges kapcsolat. Ezek a párok az $A \times B$ halmaz elemei. Amikor nagy adathalmazokkal foglalkozunk, összehasonlítván minden egyes (a;b) párt, ahol a az A, b pedig a B adathalmazhoz tartozik, a keresztszorzás majdnem hogy megvalósíthatatlan. Valójában mialatt a lehetséges párok száma lineárisan nő, a számolási feladatok négyzetesen növekszenek, a bonyolultság $O(n^2)$. A komplexitás csökkentése érdekében az összehasonlítások számának csökkentése szükséges. Sok különböző technika létezik amelyek a keresési tér csökkentésére alkalmazhatók, a két legfontosabb ezek közül a blocking neighbourhood és a sorting neighbourhood. A blokkolás azt jelenti, hogy a két adathalmazt particionáljuk, és csak azokat a rekordokat tekintjük összehasonlíthatónak, amelyek ugyanazon blokkon belül vannak. A partíciók blokkoló kulcsokon keresztül készülnek, két rekord ugyanahhoz a blokkhoz tartozik, ha minden blokkoló kulcs értéke azonos, vagy ha a két rekord blokkoló kulcsához alkalmazott hash függvény ugyanazt az eredményt adja. A sorting neighbourhood rendezi a két input állományt a blokkoló kulcs alapján és csak egy fix méretű ablakon belül keresi a lehetséges kapcsolatokat, amely végigszalad a két rendezett rekordhalmazon.

A csökkentett keresési tértől kezdve alkalmazhatunk különböző döntési modelleket, amelyek definiálják azokat a szabályokat, amelyek meghatározzák, hogy egy (a;b) rekordpár kapcsolat, nem kapcsolat vagy lehetséges kapcsolat.

Az ötödik fázis, a record linkage folyamat magja egy döntési modell kiválasztása, amely lehetővé teszi a párok besorolását egy M (kapcsolat) és egy U (nem kapcsolat) halmazba. A RELAIS jelenlegi változatában kétféle döntési modell érhető el, mégpedig a determinisztikus (empirikus) és a valószínűségi. A determinisztikus megközelítés során egy pár kapcsolat, ha teljesen egyezik minden választott kulcsváltozó (exact matching)[2], vagy kielégít egy megadott

[2] Ilyenkor nem lehet az egyenlőségen kívül más összehasonlító szabályt alkalmazni.

szabályrendszert (rule-based, szabály-alapú összekapcsolás)[3]. Ez utóbbi azt jelenti, hogy az alkalmazott összehasonlító függvény eredménye túl van egy megadott határértéken.

A valószínűségi megközelítés a Fellegi-Sunter modellen alapul. Ez megköveteli a modell paraméterek egy becslését, amely elvégezhető EM algoritmussal, Bayesi módszerrel, stb. Az eljárás azon alapul, hogy számszerűsíthető a rekordok „hasonlósága", így előírhatók numerikus határértékek, amelyeknél „jobban hasonlító" rekordokat össze kell, és egy másik, amelynél „kevésbé jobban hasonlítókat" nem szabad összekapcsolni. Az eljárás jelentősége az, hogy e két adott határérték mellett minimalizálja azoknak a rekordoknak a számát, amelyek összekapcsolásáról nem lehet egyértelmű döntést hozni. A kétféle döntési modell közötti választás módszertani szempontjaira külön kitérünk a következő fejezetben.

A hatodik fázis az egyedi azonosítók kiválasztása. Ebben a fázisban történik az M:N kapcsolás redukálása 1:1 kapcsolássá. A kapcsoló folyamat a következő módon osztályozható:

1:1 probléma: amikor az A-ban egy rekord a B-ben pontosan egy rekordhoz kapcsolható.

1:N probléma: ha egy rekord egy adathalmazban többhöz kapcsolható az összehasonlított állományban.

N:M probléma: ha több mint egy rekord kapcsolható minden egyes fájlban több mint egy rekorddal a másik állományban.

Az utóbbi két eset magában foglalhatja a duplikált rekordok létezését az összekapcsolni kívánt adathalmazokban.

Végül, mivel a record linkage folyamatban nem minden összekapcsolt rekord hivatkozik ugyanolyan azonosítóra, a „record linkage folyamat kiértékelésében" fontos megállapítani, hogy vajon egy kapcsolat helyes-e, vagy sem. Más szavakkal a kapcsolás projekt alatt osztályozni kell a rekordokat, mint igazi kapcsolat, vagy igazi nem kapcsolat, hogy minimálisra csökkentsük a két lehetséges hibatípust: a hibás kapcsolatot vagy a hibás nem kapcsolatot. Az első típusú hiba olyan összekapcsolt rekordokra utal, amelyek nem reprezentálnak azonos entitást, míg a második olyan – valójában összetartozó entitásokat – jelent, amelyek nem kapcsolódtak össze.

3. Determinisztikus és sztochasztikus adatösszekapcsolás

Ahogy az előző fejezetben jeleztük, a RELAIS célja, hogy különböző megközelítésekről és technikákról gondoskodjon, amelyek segítségével változatos record linkage problémákkal lehet foglalkozni. A RELAIS implementál egy módszert a valószínűségi alapon történő record linkage megvalósítására, amely a Fellegi-Sunter elméleten alapszik, és két módszert a determinisztikus record linkage-re, amelyek a kulcsváltozók összehasonlításán alapulnak. A következőkben megadunk néhány általános szempontot, amely az egyes módszerek közötti választást segítik, bemutatva azok előnyeit, hátrányait.

A szakirodalomban a determinisztikus és a valószínűségi megközelítés között gyakori a megkülönböztetés, az előbbi a formális döntési szabályokat, míg az utóbbi a valószínűségek explicit megadását használja arra, hogy eldöntse, egy rekordpár valójában mikor jelent egyezést. Valójában nehéz tisztán megkülönböztetni a két megközelítést. Néhány szerző (Statistics Canada) a determinisztikus record linkaget csak úgy definiálja, mint azt a módszert, mely az egyedeket akkor és csak akkor kapcsolja össze, ha az egyedi azonosító, vagy a közös azonosítók halmaza,

[3] A Rule based módszer esetén komplex szabályok definiálhatók, amelyek alszabályokba vannak szervezve. Minden egyes alszabály állhat feltételekből, amelyeket AND operátor választ el. Az egyes alszabályokat az OR operátor választja el, és egyszerű logikai feltételek is megadhatók. Ha egy rekordpár előírt változóira teljesülnek a szabályok, akkor összekapcsolódnak, egyébként nem.

a kulcsváltozók teljesen megegyeznek. Más szerzők azt mondják, hogy a determinisztikus record linkageben egy pár kapcsolat, ha kielégít néhány speciális kritériumot, amelyeket a priori definiálnak, valójában nem csak a kulcsváltozókat kell kiválasztani és kombinálni, de az összehasonlító függvény egy küszöbértékét is rögzíteni kell, hogy megállapítsuk, vajon egy párt tekintetbe kell-e venni, mint párt, vagy nem. , és hogy ez a fajta linkage majdnem egzakt, vagy szigorúan véve nem is egzakt. A determinisztikus megközelítésben mind az egzakt, mind a majdnem egzakt esetben a két különböző adatbázis közötti kapcsolat bizonytalansága minimális, de a kapcsolási arány nagyon alacsony lehet.

A determinisztikus record linkaget alkalmazhatjuk a valószínűségi módszer helyett, ha az egyedi azonosítók hibamentesek, (mint a TB szám vagy az adószám) vagy amikor a kulcsváltozóknak jó a minősége és nagy megkülönböztető erő érhető el, és kombinálhatók, hogy a kapcsolati státuszt megállapítsuk. Ilyenkor a determinisztikus megközelítés nagyon gyors, hatékony, és alkalmazása megfelelő. Más oldalról a szabályok definíciói szigorúan függenek az adatoktól és a gyakorlati alkalmazást végző személy ismereteitől. Továbbá a kulcsváltozók minőségétől való szigorú függés miatt a determinisztikus eljárásban néhány kapcsolat hiányozhat a kulcsváltozók hiánya vagy a bennük levő hibák miatt, így a determinisztikus és a valószínűségi módszer közötti választás során figyelembe kell venni a fájlokban levő változók stabilitását, elérhetőségét és egyediségét. **A valószínűségi megközelítés összetettebb és formálisan illeszkedik a rossz minőségű adatok okozta problémákhoz.** Különösen akkor segíthet, amikor különféle helyesírási hibák, felcserélt vagy rosszul jelentett változókat tárolnak a két adatfájlban. A valószínűségi eljárás során lehetséges a kapcsolási hibák kiértékelése, mivel a korrekt kapcsolatok valószínűsége meghatározható.

Általánosan szólva a determinisztikus és a valószínűségi eljárások kétlépéses folyamatban kombinálhatók. Először e determinisztikus kapcsolást kell végrehajtani, majd a valószínűségi módszert a maradék állományokon alkalmazni.

4. Adatösszekapcsolás a migrációs statisztikában

Az elméleti áttekintés, a record linkage és a RELAIS munkafolyamatának megismerése után, ebben a fejezetben a migrációs statisztika területén rendelkezésre álló adatbázisokra alkalmazzuk a módszertant, és a harmadik országbeliek adatait tartalmazó adatbázisok összekapcsolását mutatjuk be lépésenként.

4.1.Az adatbázisok közös attribútumai

Az előző fejezet rövid elméleti ismertetőjéből levonható egyik legfontosabb következtetés, hogy az összekapcsoláshoz az esetleges azonosítókon és közös attribútumokon keresztül vezet az út. Ennek érdekében először át kell tekinteni, hogy mely mezők használhatók fel. Az 1. táblázat tartalmazza azokat az attribútumokat, amelyek legalább két adatbázisban előfordulnak.

1. táblázat: A harmadik országbeli állampolgárok adatait tartalmazó adatbázisok közös változói

Attribútumok	BÁH	KEKKH	OEP	NÉPSZ	NAV
Személy leírása					
Családnév	X	X			
Utónév	X	X			
Anyja neve	X	X			X
Születési idő	X	X			X
Nem	X	X	X	X	
Családi állapot	X	X	X	X	
Születési országkód	X	X		X	
Születési ország		X	X		
Születési település	X	X			
Állampolg. Kód	X	X	X	X	X
Lakóhely					
Irányítószám		X	X		X
Település		X	X	X	X
Tartózkodási hely					
Irányítószám	X	X	X		
Megye	X				
Település	X	X	X	X	

A fenti táblázat jól mutatja, hogy a BÁH, KEKKH adatbázis összekapcsolására érdemes a legnagyobb hangsúlyt fektetni, hiszen ezek olyan közös attribútumokat tartalmaznak, melyek alapján a megfelelő döntési szabály alkalmazása esetén megbízhatóan azonosíthatják a közös rekordokat. Emellett a NAV adatbázis tartalmaz még azonosításra jól használható anyja neve, születési dátum változót. Az OEP és Népszámlálás esetében jelenleg kevés az összekapcsolásra felhasználható változók száma, és azok egyedisége is kicsi.

A fenti táblázatba szereplő közös attribútumok egy részét elő kellett állítani az egyes adatbázisokban rendelkezésre álló változókból, hiszen az összekapcsoláshoz szükséges feltétel, hogy legyenek (kellő számban) közös oszlopok a relációkban, amelyek tartalma és kitöltési módja azonos. Nem működhet az összekapcsolás, ha például az egyik adattáblában egy Név oszlopban szerepel egy ember teljes neve, míg a másikban van külön Vezetéknév és Utónév oszlop, vagy mondjuk a születési dátumot az egyik helyen folytonos 8 karakteres szövegként (19620427), a másikban dátumként (1962.04.27) tárolják. Az ilyen jellegű problémák megszüntetése az esetek többségében megfelelő előkészítő és adattisztító műveleteket igényel.

Az adattisztítás lépései két csoportba oszthatók. Egyrészt, ahogy már említettük, szükséges az 1.táblázatban bemutatott közös attribútumok létrehozása érdekében olyan műveleteket végrehajtani, hogy ezeknek a változóknak a tartalma azonos legyen. A kódjellegű változók esetében például gyakran szükség van az állományon belüli átkódolásra. Az adattisztítási lépések másik fontos részét képezik azok a műveletek, ahol a változók mélyebb tartalmát vizsgáljuk és az összekapcsolás jobb megvalósítása érdekében végzünk. Például szabad szöveges mezők esetében a nevek minél egységesebb leírása érdekekében tehetünk lépéseket (doktori fokozat törlése a névből, hiszen azt nem biztos, hogy minden adatbázis tartalmaz, ráadásul időben változhat miatta a névhasználat), a hibásan írt településnevek javítása, az országkódok tekintetében történt egységesítés, mint a Jugoszláviai utódállamok nem következetes kitöltése miatt egységes YUG kód használása, nem létező országkódokhoz (például az Afrika egyéb megnevezéshez) településnév alapján történő országkód rendelés. Fontos, hogy ezekben az esetekben nem történik adatvesztés, hiszen minden input adatállomány tartalmaz azonosításra alkalmas változót, mely segítségével az összekapcsoláshoz szükséges új változó mellett az eredeti változó visszakereshető marad. A projekt során feldolgozott adatállományokkal kapcsolatos adattisztításokat, a közös attribútumok kialakítására és az adatminőség javítására tett lépések részletes leírását a 2. számú melléklet tartalmazza.

Az összekapcsolás során felhasznált változók és értékeik tekintetében az is fontos, hogy a RELAIS az adatbázisokat szövegfájlból olvassa be. Az első rekord kötelezően az adatbázis-sémát kell, hogy definiálja. A két adatbázis-séma lehet különböző, de az összekapcsoláshoz szükséges oszlopokat mindkettőnek tartalmaznia kell, mégpedig azonos néven. A mezőnevek megadásánál figyelni kell arra, hogy a RELAIS érzékeny a kis- és nagybetűk közötti különbségre. A szövegfájlban az új sor karakter zárja le a rekordot, a listaelválasztó pedig egyértelműen elválasztja az egyes mezőket a rekordon belül. Így az nem fordulhat elő egyik mezőben sem, és egyedi kell, hogy legyen az egész adatbázisban.

Az adattisztítás utolsó mozzanataként törölni kell a duplázásokat és az egyéb nem megfelelő rekordokat. Két rekordot egy adatbázisban akkor tekintettünk egyenlőnek, ha valamennyi mező tartalma – eltekintve az egyedi azonosítóktól – pontosan megegyezik.
A BÁH esetén – mivel azt az állományt két adatlistából fűzzük össze[4] – külön-külön töröltük az azonos rekordokat, elértük, hogy az eredeti azonosító újra egyedi legyen, majd az összefűzött állományból újra töröltük az azonos rekordokat. Ezek elvégzése után a tisztított változat 557 756 rekordot tartalmaz.
A KEKKH-adatállományban 213 102 rekord található, azok a külföldiek szerepelnek benne, akik bejelentett lakóhellyel rendelkeznek.
A NÉPSZ-ben 143 197 külföldi állampolgár rekordja szerepel.
Az OEP adattáblában 320 036 rekordja tartalmazza az érvényes és érvénytelen TAJ számmal rendelkező külföldiek adatait is. Az OEP adatbázisba bekerült személyek adatait ugyanis fizikailag sohasem törlik, érvénytelenítés révén kerülnek ki a rendszerből. Az érvénytelenítés oka lehet, például elhalálozás, külföldre település, külföldön létrejött kereső tevékenység bejelentése.
A NAV állomány 80 158 rekordja az adott évben személyi jövedelemadót fizető külföldiek adatait tartalmazza.

4.2 A BÁH és a KEKKH adatbázisok összekapcsolása

A hivatalos migrációs statisztika elsősorban a BÁH idegenrendészeti állományán alapul, hiszen ez tartalmazza a legszélesebb körben és leginkább naprakészen a Magyarországon huzamos ideig tartózkodó külföldieket. Mivel a RELAIS egyszerre két adatbázissal tud dolgozni, ezért jelen kutatás során elsősorban a BÁH állománnyal való összekapcsolást vizsgáltuk.

A BÁH adatbázis előkészítése során sok esetben derültek ki az adatforrás hiányosságai, és az adatbázisok vizsgálata rávilágított az adatminőség javításának, az adattisztításnak a fontosságára. Azért, hogy a RELAIS tudjon dolgozni, létrehoztunk összekapcsolás előtti állományokat, amelyeket a művelet során felhasználunk. Ezek csak a tisztított, összekapcsoláshoz felhasznált oszlopokat tartalmazzák, hogy ne zavarjanak olyan karakterek, (vessző, pontosvessző vagy kettőspont), amelyek esetleg listaelválasztónak vannak megadva. A 2. táblázatban látható, hogy a BÁH adatbázis tisztított változata 557 756 rekordot tartalmaz. A KEKKH-ban 213 102 rekord van. Ezek kívül a táblázat felsorolja a BÁH és KEKKH adatbázisok azon attribútumait, amelyet felhasználhatók az összekapcsolás során. Tekintettel arra, hogy az eredmények annál pontosabbak, minél kisebbek a megengedett hibák, ahol lehetett a kapcsolómezők pontos egyezéséhez ragaszkodtunk. A közös attribútumok puszta léte mellett az eredményesség szempontjából azok kitöltöttsége sem lényegtelen. A következő pontokban szereplő táblázatok tartalmaznak erre vonatkozó információkat is. Ezekben a **Kitöltött** oszlop a nem üres cellák százalékos arányát mutatja.

[4] A szabad mozgás és tartózkodás jogával rendelkezőket tartalmazó egt állomány, a többi harmadik országbelit tartalmazó idtv állomány.

Minden adatbázishoz bevezettünk továbbá egy **myid** nevű saját azonosítót, amely alapján könnyebben látható, hogy az adathalmaz egy adott része pontosan mely rekordokból áll.

2. táblázat: A BÁH és KEKKH adatbázisok összekapcsolásához felhasznált változók és kitöltöttségük

Mező	Jelentés	BÁH adatbázisban		KEKKH adatbázisban	
		Nem üres	Kitöltött (%)	Nem üres	Kitöltött (%)
myid	Egy általunk bevezetett egyedi azonosító, amely egyszerű numerikus sorszámozás.	552 465	100.0	212 244	100.0
szcsnev1_k	Születéskori családi név, az eredeti adatbázisban levő nevek tisztított változata.	552 465	100.0	212 244	100.0
szunev1_k	Születéskori utónév, az eredeti adatbázisban levő nevek tisztított változata.	552 423	100.0	212 244	100.0
acsnev1_k	A személy anyjának családi neve, az eredeti adatbázisban levő nevek tisztított változata.	552 391	100.0	212 244	100.0
aunev1_k	A személy anyjának utóneve, az eredeti adatbázisban levő nevek tisztított változata.	549 846	99.5	212 244	100.0
sz_szulido	A személy születési ideje 8 karakteres szöveges adat ÉÉÉÉHHNN formátumban.	549 522	99.5	212 244	100.0
szulev	A személy születési éve négykarakteres numerikus formátumban.	552 465	100.0	212 244	100.0
sz_nem_k	A személy neme karakteres típusú adatként, az 1 jelenti a férfit, a 2 a nőt.	552 465	100.0	212 244	100.0
szulorsz_k	A születési ország ISO3166 szabvány szerinti kódja, kiegészítve néhány olyan kóddal, ami a többi adatbázis miatt kell. Ezek: YUG, (Jugoszlávia), USR, (Szovjetunió) és CSE (Csehszlovákia). A KEKKH állományban néhány esetben szükség volt a településnév alapján újrakódolni (lsd. 2. sz. melléklet)	552 465	100.0	212 010	99.9
szultel_k	A születési település neve, az eredeti adatbázisban levő nevek tisztított változata.	552 021	99.1	212 003	99.9
lt_helyseg_k	A tartózkodási hely (KEKKH-ban lakóhely) településének neve, az eredeti adatbázisban szereplő nevek tisztított változata.	551 502	99.8	213 243	100.0
lt_megye_k	A tartózkodási hely településének megyéje, amelyet a település alapján állítottunk elő.	550 209	99.6	211 684	99.7

A szultel_k változó kitöltöttsége például láthatóan megfelelő, de az adatbázis vizsgálata azt mutatta, hogy a minősége nem kellően jó.

Tekintettel arra, hogy a KEKKH adatbázisban a lakóhely kitöltöttsége jóval magasabb, mint a tartózkodási helyé, és a tartózkodási hely, illetve a lakóhely fogalma a különböző nyilvántartó helyek használatában nem pontosan fedi egymást, így az összekapcsolást a lakóhely települése alapján végezzük el.

Először megpróbáltunk **determinisztikus módon** összekapcsolni az adatbázisokat, hiszen az mindig megbízhatóbb, mint bármelyik sztochasztikus módszer. Ennek érdekében mindkét input adatbázist pontosvesszővel határolt txt kiterjesztésű szövegfájllá alakítottuk, majd a következő eljárást hajtottuk végre.

Input fájlok beolvasása:

Beolvastuk az adatállományokat és beállítottuk határoló karakternek a pontosvesszőt. Ekkor a RELAIS automatikusan a **Dataset A** (DSA) és a **Dataset B** (DSB) neveket rendeli az adatforrásokhoz. A munka során a BAH volt a DSA és a KEKKH-nak jutott a DSB. Ez egy kicsit kényelmetlen, de bármikor lekérdezhető, és utólag is visszakereshető, hogy melyik adathalmaz melyik nevet kapta. A RELAIS kiírja a beolvasott rekordok számát, ami a BÁH esetén 557 756, a KEKKH esetén 213 102

Lehetséges, hogy az általunk kiválasztott változókról különféle tájékoztató adatokat kérjünk, ezt a **Data Profiling/Matching Variables** menüpontban lehet kezdeményezni. Elkészíttettük például a változók gyakoriság eloszlását, hogy lássuk, melyek a gyakran előforduló adatok, illetve a 2. táblázatban közölt kitöltöttségi arányokat.

Kapcsolómezők megadása, összehasonlító függvények beállítása:

Ekkor állítjuk be azokat az attribútumokat, amelyek hasonlósága alapján döntünk a rekordok egyezőségéről, és hozzájuk rendeljük a RELAIS eszköztárában meghatározott függvények közül azokat, amelyek vizsgálni és számszerűsíteni tudják a numerikus és szöveges adatok hasonlóságát. A küszöbértékkel megadjuk, hogy mikor tekinthetők közel azonos értékűeknek ezek a változók.

A **Decision Model/Deterministic/Equality Match** parancs kiadása után lehetett megadni, hogy mely változókat használjuk az összekapcsoláshoz. A 3. táblázatban leírtakat választottuk:

3. táblázat: A kapcsoló változók, a döntéshez szükséges függvények és a megkövetelt küszöbértékek.

Változó neve	Metrika	Küszöb
SZCSNEV1_K	Equality	1
SZUNEV1_K	Equality	1
ACSNEV1_K	Equality	1
AUNEV1_K	Equality	1
SZ_SZULIDO	Equality	1
SZ_NEM_K	Equality	1
SZULORSZ_K	Equality	1

A táblázatot a RELAIS generálta, az első oszlop tartalmazza a kapcsoló változók nevét, a második az összehasonlításukhoz használt relációt, ami most az egyenlőség, a harmadik oszlopban található küszöbérték a pontos egyezőség esetén egy.

A változók beállítása után lefut az eljárás, és előállnak az összekapcsolt táblázatok, valamint a reziduális adatbázisok, amelyek az eredeti listák azon rekordjait tartalmazzák, amelyeket nem sikerült összekapcsolni. Ezeket txt formátumban lehet menteni.

Az összekapcsolt tábla automatikusan a Match.txt nevet kapja, szerkezetét pedig a 4. táblázat mutatja. (A neveket valamilyen karaktersorozattal helyettesítettük, de az azonos sorozatok azonos neveket takarnak. Helytakarékosságból csak az elő néhány oszlopot mutatjuk.)

4. táblázat: Az összekapcsolt rekordokat tartalmazó adatállomány szerkezete

DS	KEY_DS	MYID	SZCSNEV1_K	SZUNEV1_K	ACSNEV1_K
A	2	2	XXX	YYY	AAA
B	140173	140178	XXX	YYY	AAA
A	3	3	ZZZ	WWW	BBB
B	140404	140409	ZZZ	WWW	BBB

Az első oszlopba írja a RELAIS az adatbázis azonosítóját az A vagy B betűt. A második oszlopban a RELAIS által létrehozott, az adott rekordhoz tartozó egyedi azonosító értékei látszanak. Ezek egyszerű sorszámok. A harmadik oszloptól jönnek a közös mezők értékei.

A RELAIS egyébként nem tud minden ékezetes betűvel megbirkózni, az Ő helyére például kérdőjel került. Valójában az eredményhalmazból csak a DS és a MYID oszlop érdekes, mert ezek segítségével az eredeti adatbázisokból le tudjuk kérdezni a teljes rekordot, és létre tudjuk hozni az outputot. Ezt a munkát már Access segítségével végeztük el. Az összekapcsolás kb. egy perc alatt megtörtént, és 145 760 rekordpárt sikerült egymásnak megfeleltetni. Ezek azonban csak olyan értelemben képeznek párokat, hogy a kapcsolómezők értékei egymással egyenlők. Valójában nem biztos, hogy egymás megfelelői, mert a többi argumentum értékében eltérhetnek egymástól. Csak 132 793 olyan rekordpárt sikerült elkülöníteni, amely egy-egy kapcsolatban áll egymással. Amely rekordokra ez nem igaz, azt külön táblába gyűjtöttük.

A RELAIS automatikusan a ResidualDSA.txt és a ResidualDSB.txt nevet adja mentéskor azoknak a fájloknak, amelyek az A és B adatbázisok nem összekapcsolható rekordjait tartalmazzák. Ezek szerkezete megegyezik az eredeti adattáblák szerkezetével. Esetünkben a reziduális adattáblák a BAH esetén 407 623, a KEKKH esetén pedig 80 633 rekordot tartalmaznak. Amiatt, hogy egyes rekordok több másikkal is párba álltak a kapcsolt és a reziduális táblák sorainak összege nem adja ki az eredeti tábla sorainak összegét. A reziduális táblákban nincsenek ismétlődő sorok.

Ezt követően logikus lépésnek látszik a maradék rekordok **sztochasztikus összekapcsolása**, de az ehhez szükséges Descartes-szorzat elkészítését reménytelen feladat megkísérelni. Ezért szükség volt a *keresési tér csökkentésére*.

Blokkoló változók megadása[5]:

Először szétválasztottuk mindkét reziduálist **a román, jugoszláv és szovjet, illetve az egyéb országokban született személyeket tartalmazó részekre.** Ezeken belül az első csoportban a **születési évet, és a nemet, a másodikban ezeken felül a születési ország kódját is használtuk blokkoló változónak**, mert ezeknek majdnem minden rekordra létezik értékük, és különbözőségük kizárja az összekapcsolhatóságot.

Ezek után a születési ország szerint bontott táblákat újra betöltöttük (a BAH lett a DSA, a KEKKH a DSB), majd az alábbi lépéseket hajtottuk végre:

Megvizsgáltuk a kiválasztott blokkosító változót. (**Data Profiling/Blocking Variables**) Lehetőség van arra, hogy ellenőrizzük a kitöltöttségét, gyakoriság eloszlását, és egy adott blokk dimenzióra (ez a blokkba eső párok száma, amely a RELAIS alapértelmezése szerint 1 000 000,) kiszámoljunk egy Blocking Adequacy mutatót, amely a blokkdimenziónál kisebb számú rekordpárt tartalmazó blokkok és az összes blokk számának hányadosa.

Létrehoztuk azt a keresési teret, amelyben a Fellegi-Sunter módszer szerint vizsgáljuk majd a rekordpárokat. (**Search Space Creation/Search Space Reduction/Blocking**) A keresési tér viszonylag gyorsan elkészült, 213 blokk kicsit több mint 35 millió rekordpárt tartalmaz, miközben a Descartes-szorzat 5.3 milliárd párból állna. A párosítás kb. 4 óra alatt futott le.

Kapcsolómezők megadása, összehasonlító függvények beállítása:

---

[5] Ez a lehetőség a keresési tér redukálását célozza. Könnyű végiggondolni, hogy nagyobb adatbázisok összekapcsolása esetén az összehasonlítások száma hatalmas nagy lehet, mert, ha az egyik reláció n, a másik m sort tartalmaz, akkor (n·m)/2 összehasonlítást kell elvégezni.

Megadtuk az összekapcsoláshoz használt változókat, nevüket, az összehasonlításuk vizsgálatához használt függvényt és a hozzájuk rendelt küszöbértékeket az 5. táblázat tartalmazza. (**Decision Model/Probabilistic/Matching Variables/Variables Selection**)

Beállítottuk az összehasonlításhoz használt metrikát (Similarity metric) és a hibaszintet (**Decision Model/Probabilistic/Matching Variables/Metrics and Threshold Setting**)

5. táblázat: A valószínűségi összekapcsoláshoz használt kapcsoló változók, függvények, hibaszintek

Választott modell (3G)		
Változó	**Hasonlító metrika**	**Hibaszint**
SZCSNEV1_K	3Grams	0.9
SZUNEV1_K	3Grams	0.9
ACSNEV1_K	3Grams	0.9
AUNEV1_K	3Grams	0.9
SZ_SZULIDO	Equality	1.0
SZULTEL	3Grams	0.8

Valamennyi változó kitöltöttsége nagyobb, mint 90%, de a teljesen azonos tartalmú rekordok már hiányoznak. Az összekapcsolásra három modell készült, amelyeket az összehasonlító függvények különböztettek meg (például a születési név tekintetében próbát tettünk a Dice, Levenstein függvények alkalmazására is), az 5. táblázat a választott modellt mutatja be.

Ezt követően elő kell állítani az összehasonlító vektor komponenseit, és az előfordulások gyakoriságait tartalmazó kontingencia táblát. Ezt megtehetjük akár egyetlen blokkra is, amit előtte meg kell adnunk, (**Decision Model/Probabilistic/Fellegi-Sunter/One Block/Block Selection**, majd **Contingency table)** vagy mindre (**Decision Model/Probabilistic/Fellegi-Sunter/Contingency table)**. Ennek előállítása hosszabb időt vehet igénybe, a futási idő nagy része erre megy el.

Ezt követően optimalizáltuk a megoldást 1:1 kapcsolatot feltételezve optimalizáltuk a megold Lefuttattuk a megoldást az összekapcsolhatóságot és a nem összekapcsolhatóságot jelző küszöbértékek beállításával (**Linkage result/Threshold**).

Létrehoztuk az output táblákat (**Linkage result/1:1 Result/** és a szükséges táblák kiválasztása).

Beállítottuk az output mappát, és mentettük az eredményeket. (**Save**) Ezek szöveges (txt kiterjesztésű) állományokban állnak elő, amelyek további szoftverekkel könnyen szerkeszthetők.

A jó minőségű országkód helyett, most a rosszabb minőségű születési települést használtuk kapcsolómezőnek. Természetesen amiatt, hogy az összehasonlításkor nem követeltük meg a teljes egyezést, vannak eltérések, az egyes mezők tartalmában, de ezek valóban nem jelentősek. Pl. összekapcsolódott két rekord, amelyben a születési település az egyikben PISLOLT, a másikban PISCOLT, vagy egymásra talált a LOYOS és a LAJOS, és megfelelt a KSZENIIA a KSZENYIJA-nak. Összesen mintegy 1000 biztosan összekapcsolható rekordot eredményezett a szochasztikus összekapcsolás a Romániában, a volt Jugoszláviában vagy a volt Szovjetunióban születettek esetén, és arányaiban hasonló eredményt adott a többi országban születettek összekapcsolása is. A látszólag kicsi számnak az az oka, hogy a determinisztikus összekapcsolás más a pontosan illeszkedő rekordokat kivette a folyamat elején, és nagyrészt ugyanazokkal a kapcsolómezőkkel tudtuk a párosítást folytatni. Ez persze nem baj, mert a determinisztikus összekapcsolás mindig jobb minőségű, mint a sztochasztikus.

Outputok:

BAH_KEKKH	Az összekapcsolható rekordok, jelöltük, hogy melyek keletkeztek determinisztikusan és melyek sztochasztikusan.
R_BAH_KEKKH_B:	A BAH azon rekordjai, amelyeket nem lehetett bevonni az összekapcsolásba.

R_BAH_KEKKH_K: A KEKKH azon rekordjai, amelyeket nem lehetett bevonni az összekapcsolásba.

4.3.A BÁH összekapcsolása az OEP és NÉPSZ állományokkal

Amikor az összekapcsolt adatbázisok egyike a NÉPSZ vagy az OEP, akkor abba a sajátos problémába ütközünk, hogy nagyon kicsi a közös attribútumok száma. A probléma az, hogy ezek az adatok a legtöbb esetben nem azonosítják egyértelműen a személyeket (nem kulcsok), így amikor kiválasztunk egy sort a BÁH (vagy szintén neveket is tartalmazó KEKKH) adatbázisból, akkor ahhoz nagyon sok rekord rendelhető a NÉPSZ vagy az OEP adatbázisból.

A BÁH és OEP adatbázis viszonylatában a közös mezők (lásd 1. táblázat) a **születési év**, a **nem** és a **születési ország kódja.** Az összekapcsolás során pontos egyezést vártunk el, mivel a kapcsolásban kódok vannak, így nem indokolt semmilyen hibahatár felvétele.

Az összekapcsolás után közel pár ezer rekord áll elő, mivel pl. a 319 075 myid értékű OEP rekordot 57 BÁH-beli sorhoz lehet kapcsolni, azaz abból az 57 emberből bárki lehet pár. De ez fordítva is igaz, például a 8 187 myid értékű BÁH rekordnak négy lehetséges folytatása van az OEP adatbázisban, melyek közül nem lehet megmondani, hogy melyik a helyes.

Valójában ezekben a speciális esetekben a reziduális táblák informatívak, amelyek azokat a rekordokat tartalmazzák, amelyek nem kapcsolhatók össze. Emellett hasznos lehet azoknak a rekordoknak az ismerete is, amelyekhez csak két-három további rekord kapcsolható.

Megoldást jelentene az OEP és a NÉPSZ esetleges olyan további rekordjainak ismerete, amelyek pontosítják, hogy mely személyekről van szó, és felhasználhatók az összekapcsolás során.

Például a BÁH és OEP adatbázisok összekapcsolásakor a születési év, nem, születési ország kódja mellé felvettük kapcsolómezőnek a tartózkodási hely települését. Egyrészt megkerestük azokat a rekordokat, amelyek csak egyszer fordulnak elő a BAH táblában és csak egy OEP-beli sor csatlakoztatható hozzájuk. Másrészt látható volt, hogy a BAH minden születési év-nem kombinációja előfordul az OEP-ben is, viszont vannak olyan OEP rekordok, amely nem fordulnak elő a BÁH-ban, ezeket szintén előállítottuk.

Ezzel az információval a jövőben gazdagítani lehet a BÁH adatbázisban megtalálható személyekről rendelkezésre álló adatokat. Akit az OEP adatbázisában is megtaláltunk, azokról plusz információkkal fogunk rendelkezni (az OEP adatbázisban jobban nyomon követik az elhalálozást, külföldre távozást). Ez fontos például a lejárat nélküli engedéllyel Magyarországon tartózkodók esetében, ahol mindig kérdés, hogy az országban tartózkodnak-e még, az Ő esetükben érdemes figyelni, hogy státuszuk az OEP adatbázisban aktív-e.

Outputok

BAH_OEP**:** Azok a rekordok, amelyekben egyezik a születési év, a nem, a születési ország és a tartózkodási hely, és bármelyiket választjuk az egyik adatlistában, a másikban pontosan egy felel meg neki.

OEP_NINCS_BAH**:** Azok a rekordok, amelyek benne vannak az OEP táblában, de biztosan nincsenek benne a BÁH adatbázisban.

4.4. A BÁH és a NAV adatbázisok összekapcsolása

A következőkben a BÁH és a NAV adatállományok összekapcsolásának egy lehetséges változatát mutatjuk be. Ahogy korábban láthattuk a BÁH tisztított állománya 557 756 rekordot tartalmaz, a NAV állomány ennél jóval kisebb, hiszen az adott évben személyi jövedelemadót fizető külföldiek adatait tartalmazza, 80 158 rekordból áll.

6. táblázat: A NAV adatállomány összekapcsolásra alkalmas attribútumai és azok kitöltöttsége.

Mező	Jelentés	NAV adatbázisban	
		Nem üres	Kitöl-tött (%)
myid	Egy általunk bevezetett egyedi azonosító, amely egyszerű numerikus sorszámozás.	80 158	100,00
a_neve	Anyja neve, a BÁH állományban a családi és utó-név konkatenálása révén létrejött változó	80 158	100,00
sz_szulido	A személy születési ideje 8 karakteres szöveges adat ÉÉÉÉHHNN formátumban.	80 158	100,00
szulev	A személy születési éve négykarakteres numerikus formátumban.	80 158	100,00
helyseg_k	A tartózkodási hely településének neve, az eredeti adatbázisban szereplő nevek tisztított változata.	65 371	81,55
allampolgarsag	Állampolgárság országa.	78 167	97,52

A két adatbázisnak az összekapcsolása a születési idő és anya neve alapján történt meg, ezek a változók közel 100 %-os szinten kitöltöttek és nagy valószínűséggel egyértelműen meghatározzák a személyt. A NAV, BÁH adatbázisokban összesen 55 623 közös rekordot találtunk.
Ebben az esetben olyan összekapcsolást végeztünk el, ahol a születési idő és az anya nevének első négy karaktere alapján kerestük a teljesen egyező rekordokat.

Összegzés

A BÁH, KEKKH, OEP, NÉPSZ, NAV állományok összekapcsolása sikeres volt, közülük is a BÁH, KEKKH, NAV adatbázisok összekapcsolása volt a legeredményesebb, hiszen ezek az állományok tartalmazták a legtöbb egyedi azonosítókat. A determinisztikus összekapcsolási módszer mellett alkalmazott sztohasztikus integrálás tovább javította az összekapcsolás sikerességét. A RELAIS program sikeresen alkalmazható volt az adott állományok összekapcsolására.

Az összekapcsolások fontos előfeltétele volt az adatbázisok előkészítése, mely során az elvégzett adattisztítások és közös formátum meghatározások nagyon sokban járultak hozzá az összekapcsolhatóság hatékonyságához.

Az alkalmazott eljárás lehetőséget teremtett olyan állományok összekapcsolására is, melyek nem tartalmaztak egyedi azonosítókat (pl.: OEP). Ezekben az esetekben azonban nem bizonyult elégségesnek a rendelkezésre álló változók köre, ezért az állomány viszonylag kis részére sikerült egyértelműen azonosítható kapcsolatot kimutatni.

Fontos tanulsága a projektnek, hogy az adatbázisok összekapcsolása során nemcsak az összekapcsolt közös részből nyertünk plusz információt, láthatóan sokszor az adatbázisok nem összekapcsolható részei (reziduális adatállományok) a leginformatívabbak.

A felhasznált nyilvántartások folyamatosan frissülnek, és a KSH részére minden évben átadásra kerülnek. A kialakított módszertani dokumentum pedig segítséget nyújt ezen feldolgozása megismétléséhez. Az elvégzett adattisztítások és formátum-transzformációk alkalmasak a jövőbeni frissített nyilvántartások előkészítésére is, de számítani kell rá, hogy új rekordok bevonásával bővíteni szükséges majd az adatelőkészítés iteráló folyamat lépéseit.

Az adatállományok jövőbeli frissítése mellett érdemes további változókkal is kiegészíteni az integrálandó állományokat, különösen az OEP esetében, ahol az egyedi azonosításra alkalmas változók hiányában még fontosabb lenne további mellékváltozók bevonása az egyértelmű összekapcsolások arányának javítása érdekében.

A jövőben érdemes megvizsgálni, milyen további nyilvántartások építhetőek be a modellbe és a fellelhető adatforrásokkal kiegészíteni a folyamatot.

1. számú melléklet: A Relais-ben elérhető összehasonlító függvények tartalma

Az összehasonlító függvények a két mező közötti hasonlóságot mérik. Többet közülük javasolnak az irodalomban, és elérhetőek a RELAIS programban, ahogy alább leírjuk. Általánosan az összehasonlító függvények eredménye formálhat kategóriákat vagy folytonos értékeket. A RELAIS-ben az egyes összehasonlító függvények normalizáltak és az eredményeik a [0;1] intervallumba esnek. Elvárják továbbá, hogy a felhasználó válasszon egy küszöbértéket 0 és 1 között, következésképpen a RELAIS bináris elemekké alakítja az eredményeket kezelvén az összes eredményt a küszöbérték fölött, mint 1-et és a küszöbérték alattiakat, mint 0-t. A távolság két szöveg között annál nagyobb, minél hasonlóbbak.

Az alábbiakban listázzuk a RELAIS-ben elérhető összehasonlító függvényeket egy rövid leírással.

Numerikus összehasonlítás

Ez a metrika a két szöveget a numerikus értékük alapján hasonlít össze. Így, ha N_x és N_y jelöli a két numerikus értékét az S_x és S_y sztringeknek, a numerikus összehasonlítás eredménye a következő:

$$NC(S_x;S_y)=\frac{\mathbf{min}\{|N_x|;|N_y|\}}{\mathbf{max}\{|N_x|;|N_y|\}}$$

Ha a szövegek nem értelmezhetők numerikusan, az NC eredménye nulla.

Levenshtein távolság

Ez egy alapvető szerkesztési távolság függvény, amellyel a távolság egyszerűen a minimális szerkesztési műveletek száma, amelyek átviszik a Szöveg1-et a Szöveg2-be. A szerkesztési műveletek a következők:

- másolni egy karaktert a Szöveg1-ből a Szöveg2-be
- törölni egy karaktert a Szöveg1-ből
- beilleszteni egy karaktert a Szöveg2-be
- helyettesíteni egy karaktert egy másikkal

Néhány más összehasonlító függvény az alább felsoroltak között a Levenshtein távolság kiterjesztése és tipikusan megváltoztatják a szerkesztési műveletek költségeit, míg a Levenshtein függvény összes művelete a legalacsonyabb költséggel bír.

Dice összehasonlítás

A Dice összehasonlító függvény egy kifejezés alapú hasonlósági mérték, és úgy definiálják, mint kétszer az összehasonlított entitásokban szereplő közös kifejezések száma, osztva a kifejezések teljes számával a két vizsgált adatelemben. Ha C jelöli a közös kifejezések számát, N_a az első, N_b pedig a második kifejezésben levő szavak számát, akkor

$$DICE=\frac{2C}{N_A N_B}$$

Jaro összehasonlítás

A Jaro összehasonlító függvény figyelembe veszi a tipikus helyesírási hibákat. Röviden, két szövegben, (s és t) jelölje s' azokat a karaktereket, amelyek s-ben vannak és közösek a t-beliekkel, illetve t' azokat, amelyek t-ben vannak és közösek az s-beliekkel. Durván szólva, egy „a" karakter az s-ben közös a t-vel, ha ugyanez az „a" karakter feltűnik körülbelül ugyanazon a helyen a t-ben is.

Legyen $T_{s,t}$ az s'-beli karakterek transzpozícióinak száma a t'-re vonatkoztatva! Ekkor a Jaro hasonlósági metrika s-re és t-re a következő:

$$J(s,t) = \frac{1}{3}\left(\frac{|s'|}{|s|} + \frac{|t'|}{|t|} + \frac{|s'| - T_{s',t'}}{2|s'|}\right)$$

Jaro-Winkler összehasonlítás

Ez a metrika a Jaro függvény kiterjesztése, amely módosítja az s, t párok súlyát egy közös prefixszel, a Jaro-Winkler távolság különösen alkalmas olyan rövid szövegek esetén, mint a személynevek.

A Jaro-Winkler függvény a következő:

$$JW(s,t) = J(s,t) + (lp \cdot (1 - j(s,t)))$$

ahol l a közös prefix hossza a szöveg kezdetétől, p egy konstans skálázó faktor arra vonatkozóan, hogy mennyire lett felfelé igazítva az eredmény, hogy közös legyen az előtag. A RELAIS a p értékét Winkler munkája alapján mindig 0,1-nek veszi, és az l értéke nem lépheti túl a hatot. Ez a beállítás kedvezőbb értékelést ad egy sztringhez, amely kapcsolódik egy beállított prefix hossz elejétől.

3-Grams

A Q-Grams eljárást általában szövegek közelítő összekapcsolására használják csúsztatván egy q hosszúságú ablakot az s szöveg karakterei fölött és létrehoznak számos q hosszúságú diagramot az összekapcsoláshoz. A RELAIS esetében a q értéke 3. A kapcsolás akkor történik meg, ha számos Q-gram kapcsolat a második, t szövegen belül túl van egy elfogadható szinten. Pl.:

Legyen az egyik szó a NELSON, a másik a NEILSEN! Ha q = 3, akkor a Q-gramok az egyes szavakban a következők:

NEL | ELS | LSO | SON, illetve NEI | EIL | ILS | LSE | SEN

A szövegeket nem kapcsolnánk össze, mert nincsenek közös Q-gramok. Ugyanakkor, ha q értéke 2, akkor

NE | EL | LS | SO | ON illetve NE | EI | IL | LS | SE | EN

így kettő közös Q-gram is lenne, a NE és az LS.

Soundex összehasonlítás

A Soundex függvény egy durva fonetikus indexelő séma, amely általában egyének nevére összpontosít, ezzel a metrikával a kiejtési hibák könnyen felismerhetők, azaz a John, a Jon és a Johne ugyanarra a személyre utal. Ez egy kifejezés alapú kiértékelés, amelyben minden egyes szót egy Soundex kóddal adnak meg, minden Soundex kód betűkből áll, a szöveg első betűjéből, öt számjegyből, amelyek 0 és 6 közé esnek. Ezek a számok az alábbi táblázatban felsorolt mássalhangzókon alapulnak:

1	B,P,F,V
2	C,S,K,G,J,Q,X,Z
3	D,T
4	L
5	M,N
6	R

Magánhangzókat nem használunk, nem kódoljuk azokat, hacsak nem bukkannak fel a szó elején. Ha két szomszédos betű – amelyeket nem választanak el magánhangzók – ugyanazt a numerikus értéket kapják, csak egyet használunk fel. Ezt akkor is így van, ha az első és a második betűnek ugyanaz az értéke, ekkor a másodikat nem használjuk számjegy képzésére. Ha kevesebb, mint 6 mássalhangzó van a szövegben, a kódot nullával töltjük fel. Ez a megközelítés nagyon ígéretes az átírt, vagy rossz helyesírással írt nevekből adódó kétértelműség kezelésére.

2. számú melléklet: Adatminőség javítás, adattisztítás

Közös attribútumok létrehozása:

Az adatösszekapcsolás előkészítéseként, az attribútumok 1. táblázatban leírt rendszerének kialakításához néhány transzformáció elvégzésére volt szükség. Ezek egészen pontosan az alábbiak:

- A KEKKH adatbázisában szerepel egy **utónév2** mező. Ebbe írták az összes utónevet, eltekintve az elsőtől. Mivel a többi adatbázisban nem szerepel ez a megoldás, a KEKKH esetén konkatenáltuk az utóneveket, így állt elő az az utónév mező, amely a BÁH adatbázisban is megtalálható.
- Ugyanezt a megoldást alkalmazták a KEKKH adatbázisban az anya utónevével kapcsolatban is.
- A NAV adatbázisban az anya neve nincs külön családi és utónévre bontva, ezért ezzel a NAV állománnyal való összekapcsoláshoz a többi állományban is konkatenáltuk a családi és utóneveket teljes névvé.
- A NÉPSZ adatbázisban eredetileg nem volt születési idő, mert az évet, a hónapot és a napot külön oszlopokban tárolták. A BÁH és a KEKKH ezt az adatot egy 8 karakter hosszúságú szövegként tartalmazza. Így a NÉPSZ három dátumrészét is eggyé konkatenáltuk, majd töröltük a nem létező dátumokat. (Ilyen, ahol valamelyik dátumrész hiányzott).
- Az OEP adatbázis eredetileg csak a születési ország nevét tartalmazta. Ennek alapján azonban elő lehetett állítani az országkódot.
- A tartózkodási hely megyéje csak a BÁH adatbázisban szerepel. A tartózkodási hely települése viszont mindegyikben, és ha azt megbízhatóbbnak fogadjuk el, akkor javítjuk a minőséget azzal, ha a megyéket a helységnévtár alapján rendeljük a településekhez. Ez viszont mindegyik adatbázisban megtehető.

A fenti eljárások rejtenek magukban kockázatokat is. A nevek összeállítása során nem biztos, hogy az egyes név részeket az egyes adatbázisok azonos sorrendben tartalmazzák. Ez a probléma az adatbázisok eredeti állapotában is fennáll, a konkatenálás során csak az fordulhat elő, hogy az utónév2 mezőben szereplő több utónév más sorrendben szerepel, mint a BÁH adatbázisban, és így az egységes utóneveket elrontjuk. Ha viszont a BÁH-ban szereplő utónevet vágnánk két részre, ugyanilyen hiba fordulhatna elő, amennyiben első helyen nem a KEKKH utónév1 mezőjének tartalma szerepel.

Adattisztítás az összekapcsolás megvalósításához:

Az összekapcsolás sikerességéhez nem elegendő az, legyenek közös attribútumok, hanem az is fontos, hogy ezek tartalma azonos módon legyen megadva. Ezért szükség van bizonyos előkészítésre, amit adattisztításnak mondunk. Ezekben az adatbázisokban szereplő mezők két csoportba sorolhatók.

- Vannak kódjellegű attribútumok, amelyek csak jól meghatározott tartalmú értékeket vehetnek fel. Ezekben a hibás adattartalom felismerhető, esetleg javítható.
- Vannak szabad szövegeket tartalmazó attribútumok, amelyek értékeit rendszerint adottnak kell tekinteni, bizonyos esetekben nagy munkával lehet csak az írásukat egységessé tenni.

A kódjellegű attribútumok kezelése

Ezeknél az oszlopoknál a legfontosabb feladat, hogy azonos kódolást határozzunk meg, és minden adatbázisban azt használjuk. A következő mezőket kezeltük.

Születési idő

Négy adatbázis tartalmazza a BÁH, a KEKKH, NAV és a NÉPSZ. Mindháromban 8 karakter hosszú szöveg tartalmazza a születési dátumot ÉÉÉÉHHNN formátumban. Az OEP csak a születési évszámot tartalmazza, ami adott esetben ellenőrzés céljára hasznos lehet. Két rekordban találtunk háromjegyű évszámot, ezt töröltük. Az adattisztítás során azt vizsgáltuk, hogy a megadott karaktersorozatokat lehet-e dátummá konvertálni, de nem találtunk hibát.

Nem

A nemre utaló kód a NAV adatbázistól eltekintve valamennyi adatbázisban előfordul egy karakter hosszúságú szöveg típusú adatként. Nyilván csak kétféle értéket vehet fel, amit könnyű ellenőrizni, és ez valóban igaz minden adatbázisban.

Családi állapot

A NAV adatbázistól eltekintve valamennyi adatbázisban előfordul a családi állapotra utaló mező. Ugyanakkor ezt a mezőt nem érdemes az összekapcsoláshoz felhasználni, mert

- nem ellenőrizhető az adat helyessége, hiszen más mező tartalma alapján nem következtethetünk a kitöltés helyességére,
- könnyen megváltozhat a családi állapot, és mert az adatbázisokba nem egyszerre kerülnek be még azok sem, akik mindegyikben előfordulnak.

Ennek következében nem módosítottunk egyetlen bejegyzésen sem, pusztán annyit ellenőriztünk, hogy egy adott oszlopban tényleg csak a megengedett értékek jelennek-e meg.

Irányítószám

Különböző adatbázisokban a lakóhely és a tartózkodási hely irányítószáma szerepel. A lakóhelyhez csak települések tartoznak, így az adat nem is ellenőrizhető, hiszen egy településhez több irányítószám használatos. Elméletileg vizsgálható, hogy egy rekordban rögzített irányítószám szerepel-e az adott településhez tartozók között, ha nem, akkor viszont nehéz eldönteni, hogy a kettő közül melyik a helyes. Ezért itt csak azt ellenőriztük, hogy van-e olyan bejegyzés, amely nem lehet irányítószám.

A tartózkodási hely fontosabb, mert a nyilvántartásokban azokra helyezik a nagyobb hangsúlyt. A BÁH, és a KEKKH az irányítószám és a település neve mellett tartalmazza a közterület nevét, jellegét, a házszámot, az épületet, a lépcsőházat és az ajtót. Egy jó címlista alapján az már ellenőrizhető lenne, hogy ezek között az adatok között van-e ellentmondás, de ha van, akkor sem lehet mindig megmondani, hogy melyik a helyes.

Tekintettel arra, hogy a címet ilyen mélységben már nem feltétlenül érdemes kapcsolómezőként használni, itt is csak arra szorítkoztunk, hogy töröljük azokat a bejegyzéseket, amelyek nem lehetnek irányítószámok.

Közterület jellege

Az ilyen mezőkben általában kódokkal adnak meg olyan adatokat, mint utca, út, lépcső, stb. A BÁH adatbázisban 69-féle kód szerepel a **kozterulettipus** mezőben, de nem állt rendelkezésünkre ezek jelentését tartalmazó leírás. A KEKKH adatbázisban a **KozterJelleg** mezőben 82-féle bejegyzés szerepel, de ezek nincsenek kódolva, így a mezőt, ha akarnánk, sem tudnánk felhasználni az összekapcsolás során. Az értékeket egyik helyen sem módosítottuk.

Házszám

A BÁH és a KEKKH adatbázisokban található mezők. A BÁH sokszor tartalmaz hibás karaktereket, amelyek nem értelmezhetők házszámként, valamint érthetetlen karaktereket. Valószínűleg a helyrajzi számok is ebbe a mezőbe kerültek. A KEKKH adatbázisban jobb minőségű a házszám mező, de összekapcsolásra nem használható.

Töröltük a házszámok közül a csillagot, a kötőjelet, a kettős kötőjelet, a vesszőt, a \N bejegyzést stb.

Egyéb címrészek

Ide tartoznak az épületet, lépcsőházat, emeletet és ajtót rögzítő adatok, amelyek a BÁH és a KEKKH adatbázisban vannak. Közös jellemzőjük az alacsony kitöltöttség, hiszen nem is feltétlenül részei egy címnek. Éppen ezért nem is használjuk összekapcsolásra, így csak a nyilvánvaló hibákat töröltük az adatbázisokból.

A szabad szöveget tartalmazó mezők kezelése

Ezeknél a mezőknél a legfontosabb, hogy a különböző adatok (nevek, települések, ország nevek stb.) leírása egységes legyen. Az adattisztításban a legnagyobb nehézséget éppen az okozza, hogy sokszor nem egyértelmű, milyen írásmódot kellene követni, bármelyiket is választjuk, nem zárható ki teljes biztonsággal, hogy egy másik adatbázisban nem másikat használtak.

Családnevek

Családnevek a BÁH és a KEKKH adatbázisokban fordulnak elő. Közös gond, hogy a hosszabb családi nevet is egy mezőbe írták, így nem ritka a kettő, három vagy többtagú név. Vannak szóközhalmozódások, és nem nevekbe illő karakterek. Sok esetben tudományos fokozatot, vagy egyéb rövidítést is beleírtak a családi névbe. Máskor rövidítéseket tartalmaznak, amelyekről sok esetben nem is állapítható meg, hogy mit jelentenek: GEB, HTL, OEC, RER, ING, stb. Az adattisztítást excelben hajtottuk végre a következő módon:

- Megszüntetjük a szóközök duplikálódását, mert az alább leírt tisztító program csak egyszeres szóközökre van felkészítve.

Ez követően lefuttatunk egy tisztító programot, amely

- Törli a tudományos fokozatokat és az egyéb rövidítéseket az azokat határoló egyéb karakterekkel – rendszerint pontokkal – együtt a név elejéről, végéről és belsejéből, helyére szóközt ír.Törli a következő betűcsoportokat a név elejéről, végéről és belsejéből, majd szóközzel helyettesíti: E/V, E/W, M/E, P/V, V/D, W/V
- A nevek elején, végén és belsejében levő (FH) karakternégyest szóközre cseréli.
- Kicseréli a ' karaktersorozatot aposztrófra, a szóköz és aposztróf, illetve az aposztróf és szóköz kombinációkat aposztrófra, a pontot, a vesszőt, a pontosvesszőt, az alsó aláhúzás jelet és a kötőjelet szóközre.
- A bejegyzések elején és végén levő pontot, vesszőt, pontosvesszőt, alsó aláhúzás jelet, kötőjelet, / jelet és \ jelet szóközre cseréli.

Ezt követően meg kell szüntetni a dupla szóközöket és újra futtatni a tisztító makrót. Ezt addig kell csinálni, amíg 0 nem lesz a változtatások száma. Ha ez megtörténik, akkor célszerű manuálisan még a következőket megtenni:

- Ellenőrizni, hogy a ' valamilyen töredéke maradt-e a névben, és ha igen, vagy törölni, vagy aposztrófra cserélni.
- Megnézni, hogy maradt-e a névben csillag vagy kérdőjel karakter. A csillag rendszerint egyedül fordul elő, (ritkán két vagy három csillag szerepel névként) ez nyilván törölhető. A

kérdőjel a név belsejében egy vagy több karakter helyett szerepel, ilyenkor a hasonló nevek alapján javítunk. Ha nem voltak hasonló nevek, inkább töröltük a bejegyzést.
- Ellenőrizni, hogy maradt-e zárójel a nevekben, és ezt megfelelő módon kezelni. Előfordult, hogy egy név mellett zárójelben egy másik név olt, akkor azt – zárójelek nélkül – új oszlopba vittük. Pl.: YAMAMOTO (HORVATH) névből a HORVATH új oszlopba került, de az összekapcsolás során ezt nem használtuk fel.
- Átalakítani a bejegyzéseket nagybetűssé az egységes megjelenés érdekében, (mivel a nevek döntő többsége eleve az volt), törölni a szóközöket a nevek elejéről és végéről, megszüntetni a szóközök duplikálódását a nevek belsejében.
- Törölni a nullahosszúságú bejegyzéseket.

Az eredeti vezetékneveket eredeti tartalmukkal meghagytuk az adatbázisban, de az összekapcsoláshoz a fenti eljárás után kapott nevet használtuk.

Utónevek

Az utóneveken is végrehajtottuk a családneveknél leírt adattisztító eljárásokat. A sokféle leírás számát úgy tudtuk csökkenteni, hogy az ékezetes betűket, valamint az Ä betűt ékezet nélkülire cseréltük, továbbá a Zsuzsát Zsuzsannára, a Beát Beátára, a Katát és a Katit Katalinra változtattuk. Az eredeti utónevek megmaradtak az adatbázisokban, de az összekapcsolás során a tisztítottakat használtuk fel. A nevek cseréje során vigyázni kell arra, hogy csak akkor szabad a bejegyzést módosítani, ha a módosítani kívánt szöveg kiadja a mező teljes tartalmát. (Nem szabad a Katalin elején levő Kata szöveget is Katalinra cserélni, mert akkor Katalinlin lesz az eredmény).

Születési helyet meghatározó országkódok, ország nevek és településnevek

Ennek a három mezőnek az együttes tárgyalását az indokolja, hogy egymással szoros összefüggésben állnak és a tisztításuk is csak együtt lehetséges. Itt valójában a vizsgálatok az ellentmondás mentességet és az országok, illetve települések nevének egységes leírását célozzák. Figyelembe véve, hogy a települések pontosabban azonosítanak, mint az országok, a BÁH és a KEKKH esetén érdemes lenne a településeket használni az összekapcsoláshoz. Ekkor elsődleges fontosságú a települések azonos módon történő leírása. Ez azonban sajnos alig valósítható meg, mert sokféle települést kellene ellenőrizni. A BÁH adatbázisban eredetileg 34695-féle településnév volt, amelyek közül 240-et sikerült egységesen leírni. Ezek kb. 24 000 rekordban szerepelnek, de ez még mindig csak az adatbázis 10%-a. Ezért a születési települések valójában nem alkalmasak jelenleg arra, hogy kapcsolók legyenek, ennek ellenére megpróbáltuk javítani a minőséget, amiért is a következő lépéseket tettük:

- Töröltük a fölösleges szóközöket és a bejegyzéseket nagybetűsre konvertáltuk.
- A \ és a / előtt és utáni szóközöket megfelelő cserék segítségével töröljük, továbbá a nyitó zárójel, majd szóköz, valamint a szóköz majd csukó zárójel karakterpárokat nyitó és csukó zárójelre cseréltük. Azért, hogy minden zárójel előtt legyen szóköz, a nyitózárójelet szóköz plusz nyitózárójel, a csukó zárójelet csukó zárójel plusz szóköz karaktersorozatra cseréltük.
- Töröltük a nem megfelelő típusú bejegyzéseket, pl. dátumokat, csillagokat, kérdőjeleket. Ezek javított formáját tabelláljuk, és később használjuk fel.
- A pontot, vesszőt, kettőspontot úgy cseréltük le, hogy utána egy szóköz is következzen.
- Megszüntettük a szóközök duplikálódását.

Ez követően lefuttatunk egy tisztító programot, amely

- Törli a ' karaktereket, a vesszőt, a pontot, a kettőspontot és a kötőjelet a nevek elejéről.
- Szóközzel helyettesít néhány gyakori rövidítést (JUD, OBL, COM, SAT, CUB, BAC, CON, MUN, ORS, CD).

Végül manuálisan elvégeztük a következő tevékenységeket.

- Újra megszüntettük a szóközök, zárójelek duplikálódását, mert a tisztítás futása során keletkezhettek ilyenek.
- Manuálisan ellenőriztük a zárójeleket, / jeleket és \ jeleket tartalmazó neveket, megállapítottuk, hogy melyik a település neve, és ezt is táblázatba foglaltuk. Ezt használtuk fel a településnevek leírásának további egységesítésére, de az esetek nagy száma miatt még nem teljes.
- A táblázat alapján módosítottuk a település neveket, és ha szükséges volt, az országkódokat is.

Sajnos az országok kódolása adatbázisonként eltérő, és nem felel meg egyetlen nemzetközi szabványnak sem. Fontos tehát közös kódolásra áttérni, amihez az ISO3166 szabványban található három karakter hosszú szöveges kódot választottuk.

A **születési települések** nevét és azok módosításait tartalmazó táblázatban megadtuk az azokhoz **tartozó országot** és országkódot is, ha szükséges volt, **ennek alapján módosítottuk az eredeti országkódot.** Ezt nem lehetett mindenhol megtenni, mert mind a BÁH, mind a KEKKH tartalmazott olyan országkódokat, melynek nem felel meg az ISO3166 szabványból semmi. Ezek javítása esetén az őket tartalmazó, **nagyobb terület kódját használtuk**, ha szükséges volt, bevezettünk ilyet. Az adatbázisokban például a szerbiai és koszovói területek nincsenek egyértelműen elkülönítve, így a Jugoszlávia számára bevezetett YUG kódot használtuk ilyenkor. (példa: KEKKH-ban a 009 kód a települések neve alapján Szerbia, így az összekapcsoláshoz a YUG kódot kapja)

Végül töröltük a /N bejegyzést, amely 170 rekordban szerepelt. A későbbiekben a települések neve alapján esetleg meg lehet kísérelni az országkód megállapítását. Azért, hogy megfeleljünk a szabványnak, a ROM kódot a Romániának megfelelő ROU-ra cseréltük, továbbá a REU-t is ROU-ra változtattuk, mert az előbbi a Réunion kódja, de az ott szereplő települések Romániához tartoznak. Az MNE valószínűleg Koszovóra utal, ezért a YUG kódot alkalmaztuk helyette.

A KEKKH adatbázisban az országkódokat a következő módon állítottuk elő:

- Egy erre a célra készített táblázat alapján hozzárendeltük az országok hivatalos nevét az adatbázisban megadottakhoz.
- Az így kapott ország nevek mellé egy új oszlopba beírtuk a hivatalos országkódokat.
- Töröltük a 003-as kódot, aminek a jelentése az, hogy az ország neve nem ismert.

A KEKKH tartalmaz ugyanis olyan kódokat, amelyekhez tartozó országok már nem léteznek, vagy soha nem is léteztek, de az OEP adatbázisban szintén előfordulnak. (KEKKH, OEP adatbázisok összekapcsolásánál érdemes a közös kódokat továbbra is használni)

Ugyanakkor a KEKKH kódokhoz tartozó települések alapján több országot lehet azonosítani, ami növeli a BÁH adatbázissal való összekapcsolhatóságot, mert ott nem szerepelnek fiktív országkódok (mint Afrika egyéb, Osztrák-Magyar Monarciha). Ezért két országkód listát hoztunk létre, az egyik a fent leírt, az OEP adatbázissal való összekapcsolás miatt, a másik csak szabványos országkódokat tartalmaz, amelyeket az azonosítható településekhez tartozó országok alapján oda is beírtunk, ahol egyébként a fenti fiktív országkódok szerepelnének.

Az OEP adatbázisban szereplő országok nevét egységesítettük, majd hozzájuk rendeltük az országkódot. Itt is vannak olyan országok, amelyekhez nem rendelhető kód, ezek előfordulása nem volt jelentős. Ezek közül az országok közül egyedül a Hontalan fordul elő a NÉPSZ adatbázisban, így annak a HNT kódot adtuk.

A többi esetben kicsi az előfordulások száma, és a további mezők tartalma alapján nem lehet azonosítani, hogy pontosan melyik létező ország lehet a születési ország. Ezért ezekhez a területekhez nem tartozik kód.

Tartózkodási hely, lakóhely település és megye

A tartózkodási hely mezőben a település javítását lehetett megkísérelni. Problémát okozott, hogy sok esetben településrészek nevét adták meg, azokat egy erre a célra kialakított szótár segítségével átalakítottuk szabványos településnevekké. Az eredeti oszlopok természetesen megmaradtak. A településnevekhez a helységnév-tár alapján rendeltük a megyéket.

Az OEP adatbázisban a tartózkodási hely nagyon alacsony szinten kitöltött, és sok külföldi város is szerepel benne, így használhatósága kétséges. Itt nem is állítottunk elő megyéket.

A NÉPSZ tartalmaz a lakóterületre és a tartózkodási helyre utaló településkódokat, amelyek alapján a települések nevét könnyen elő lehetett állítani. Egyedül a 39999 kódot nem sikerült településnévnek megfeleltetni, de ez a tartózkodási hely esetén csak két rekordban szerepel, a lakóhely esetén viszont egyetlen egyben sem. A lakóterület településkódja egyébként minden rekordban megegyezik az **Terul** mezőben szereplő településkóddal.

Egyéb adatok

A többi adat zömmel csak egy adatbázisban fordul elő, így azok vizsgálatára, tisztítására nem fordítottunk figyelmet.