A mintavétel és a súlyozás

a Magyarországon élő harmadik országbeliek felvételében

Tartalom

Bevezetés

Az Európai Integrációs Alap támogatásával megvalósuló EIA/2013/2.6.1. számú, a „Migránsokra vonatkozó társadalomstatisztikai adatgyűjtések megalapozása” című projekt részeként a Központi Statisztikai Hivatal 2014. második negyedévében mintavételes lakossági felvételt hajtott végre a Magyarországon élő harmadik országbeliek vizsgálatára; az egyszerűség kedvéért a következőkben migrációs felvételként és migrációs mintaként hivatkozunk rá.

A migrációs minta egy komplex minta, három mintavételi keretből választott minta uniója. A három keret forrása:

- a 2011-es népszámlálás adatállomány,
- a KEKKH (Közigazgatási és Elektronikus Közszolgáltatások Központi Hivatala) személyiadat- és lakcímnyilvántartása és
- a BÁH (Bevándorlási és Állampolgársági Hivatal) nyilvántartása.

A migrációs felvétel többcélú. Célja elsősorban a Magyarországon élő harmadik országbeliek bizonyos munkaerő-piaci és migrációs jellemzőinek becslése, a célsokaság számosságának becslése, valamint megtudni, hogy mennyire alkalmasak a fenti adatforrások és a használt módszerek (idegen-nyelvű kérdőívek, internetes kikérdezés) a Magyarországon élő harmadik országbeliek elérésére.

Jelen dokumentum célja, hogy a migrációs felvétel mintavételi tervét és a megvalósult minta súlyozását leírja.

Egy hagyományos lakossági felvétel súlyozása alapvetően három lépésre bontható.

(1) A mintába választott elemek ún. design súlyát definiálja a mintavételi és kiválasztási terv.

(2) A design súlyt az egység szintű meghiúsulás kompenzálása miatt korrigálni kell.

(3) A súlyozás utolsó lépésében a mintát sarokszámokhoz igazítjuk: úgy módosítjuk a súlyokat, hogy a súlyozott mintában bizonyos eloszlások megegyezzenek a célsokaság egyéb forrásokból ismert eloszlásaival.

A migrációs felvétel súlyozása eltér ettől a sablontól. Látni fogjuk, hogy egyrészt újra kellett definiálni a design súlyokat, mert a címek tételes ellenőrzése során kiderült, hogy jelentős arányú ismétlődés van a mintavételi keretben. A (2) pontban említett korrekció természetesen itt is létezik, ez ráadásul több lépésben történt. A (3) pontban említett sarokszámokhoz igazítás a migrációs felvétel súlyozásából hiányzik, esetünkben nem létezik megbízható forrás a Magyarországon élő harmadik országbeliek számáról. Végül, de nem utolsósorban: a migrációs felvétel súlyozása alapvetően a három keretnek megfelelő három részmintán külön-külön történt; külön súlyt készítettünk a cenzus, KEKKH és BÁH címekre, amiket a súlyozás utolsó lépésében egyesítettünk.

Az első fejezetben bemutatjuk a három különböző forrásból kapott állományokat (lényegében a korábbi dokumentum 2. fejezete). A második (rövid) fejezet definiálja a migrációs felvétel célsokaságát és keretsokaságát. A harmadik fejezet a migrációs minta tervének, a negyedik fejezet az adatgyűjtés meghiúsulási kérdőívének rövid ismertetése (a korábbi dokumentum azonos fejezetei). Az ötödik fejezetben utalunk a keretek megbízhatósági vizsgálatának

tanulságaira és hiányosságaira, ami miatt újra kellett gondolni a minta elsődleges súlyozását. A hatodik fejezet az új elsődleges súly korrekciójának, a megvalósult minta súlyozásának leírása.

1. Az adatforrások, a mintavételi keret

A migrációs felvétel három forrásból merítette a mintavételi keretet. Meg kell jegyezni, hogy bár a KEKKH és BÁH forrású listák személyi szintű azonosításra is alkalmasak, mintavételi keretként minket elsősorban a listákban szereplő címek érdekeltek, hogy azokon a címeken elérhetők-e a célsokaság tagjai. A migrációs felvétel mintavételében a végső kiválasztási egység a cím.

A **2011-es népszámlálás** adatállománya alkalmas arra, hogy leszűrjük belőle a harmadik országbeli személyeket. A 18861 címet tartalmazó lista az egyik forrás. A 2014-es felvétel számára ennek a listának a legnagyobb hátránya, hogy nem időszerű adatokat tartalmaz, a cenzus óta eltelt két és fél év kifejezetten hosszú időnek számít.

A **KEKKH** személyiadat- és lakcímnyilvántartásából 43429 harmadik országbeli személy listáját kaptuk (az eredeti állományból töröltünk mintegy 1000, ismeretlen állampolgárságú személyt). A listából elhagytunk mintegy 600 rekordot, a közterület címmező üres értékei miatt, további ~600-at azért, mert kijelentett vagy érvénytelenített lakóhely, így 42783 lakóhellyel rendelkező célszemélyünk lett. Mivel minél pontosabban kívántuk elérni a célsokaságot, a listába belevettük azokat is, akiknek tartózkodási helyük van. A fentihez hasonló szűrések után ez plusz 3815 rekordot jelentett. Az összesen 45998 személy 28227 címet adott a mintavételi keretnek (az előzetes egyszerű vizsgálatok alapján 2-2 olyan címet találtunk, amik az adatforrás szerint különbözők, de vélhetően ugyanarra mutatnak). Nagy előnye a népszámlálási forráshoz képest, hogy 2013. novemberi állapotot tükröz, az adatok jóval frissebbek.

A **BÁH**-tól két állomány érkezett. Jelentős különbség a KEKKH-hoz képest, hogy ezek eset-alapú adatbázisok, vagyis egy személy annyiszor szerepel benne, ahány 'ügye' volt a hivatallal. Ez rendkívül megnehezítette az állomány előkészítését, a mintavételi keret kialakítását. Az IDTV állomány a BÁH „Tartózkodási és letelepedési engedélyek rendszeréből" leválogatott állomány, mely a 2007/II. törvény alapján Magyarországon tartózkodó harmadik országbeli (nem EGT-s) állampolgárokat tartalmazza. Az EGT-EU állomány a BÁH „Szabad mozgás és tartózkodás jogával rendelkező állampolgárok tartózkodási engedélyeinek nyilvántartásából" leválogatott állomány, mely a 2007/I. törvény alapján Magyarországon a szabad mozgás és tartózkodás jogával tartózkodó személyeket tartalmazza (EGT állampolgárok + magyar vagy más EGT ország állampolgárainak harmadik országbeli családtagjai).

Elvileg létezik személyt azonosító 'id' változó az IDTV állományon, de a tesztelések alapján kiderült, hogy ezzel nem lehet kiszűrni az ismétlődéseket. Ezért az 'id' változón túl ügyindítási dátum, születési idő, útlevélszám, név változók segítségével történt a személyek azonosítása. Az EGT-EU állományon egyszerűbb volt a személyek azonosítása, a duplikációk megszüntetése. A két állomány összefűzését és némi további szűréseket követően az együttes állományban 187920 személy lett. Meg kel jegyezni, hogy szakértőink szerint ebből mindössze 50460 személy az, aki valóban Magyarországon él. A többi, mintegy 137 ezer fő

esetében nem egyértelmű, hogy velük mi történt, pl. lejárt az engedélyük, de nem tudjuk, hogy az országban vannak-e még.

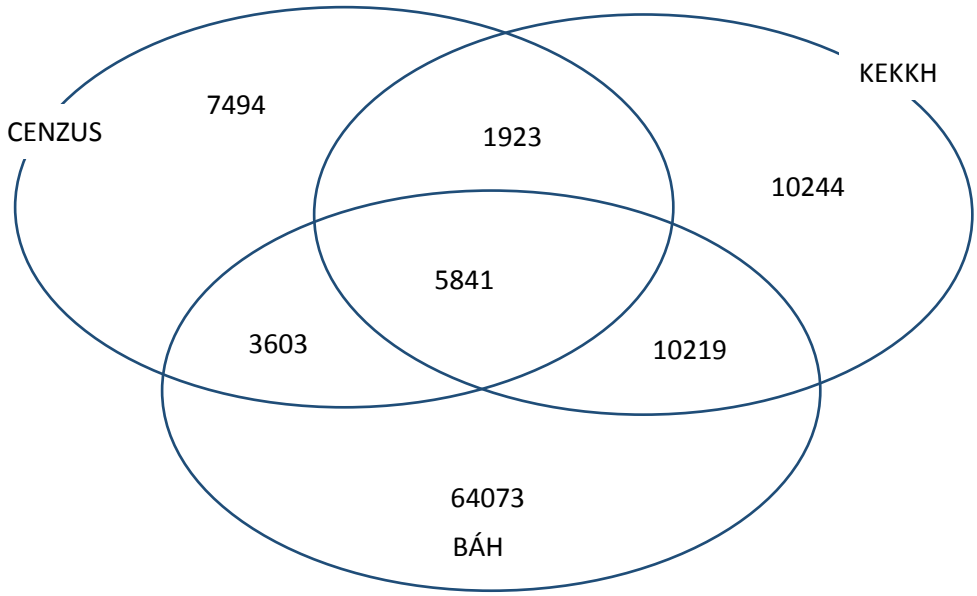
A 187920 fő az állományokon szereplő címazonosító mezők szerint 90311 címen található. Az gyorsan kiderült, hogy a címek kezelése ebben a forrásban némileg kifogásolható, még az állományon belül sem egységes, és a legkevésbé sem pontos. Egy viszonylag primitív eljárással új címazonosítót hoztunk létre, ezzel már csak 83736 cím maradt, ez szolgáltatja keretet. Meg kell jegyezni, hogy sajnos valószínűleg számos címismétlődés maradhatott a keretben, ami ronthat a majdani becslések megbízhatóságán.

A KEKKH-s forráshoz hasonlóan a BÁH adatforrás 2013. novemberi állapotot tükröz.

A három forrásból tehát rendre 18861, 28227 és 83736 címünk van. A három keret összefésülésével összesen **103397** címet tartalmazó mintavételi keretet kaptunk, ebből választottunk ki egy 3995 címet tartalmazó mintát.

A keret címeinek megoszlása a források szerint az 1. ábrán látható.

1. ábra



2. A célsokaság és a keretsokaság

Ideális esetben egy, a harmadik országbelieket vizsgáló felvétel **célsokasága** a Magyarországon élő összes harmadik országbeli[1], a felvételből származó becslések erre a célsokaságra vonatkoznának. A gyakorlatban azonban ez a felvétel nem érheti el az összes migránst.

[1] A továbbiakban az egyszerűség kedvéért a hazánkban élő harmadik országbeli kifejezés helyett a migráns szót használjuk.

- Mivel címmintával dolgoztunk, ahol a mintavételi keret nem az ország összes címének listája, értelemszerűen a mintával nem tudjuk elérni azon migránsokat, akik olyan címen élnek, ami egyik keretforrásban sincs benne.
- A KSH önkéntes lakossági felvételeinek jellemzője, hogy a célsokaság a nem intézeti népességre szorítkozik. Ennek több oka van: nemzetközi előírás a célsokaságra; összeírási nehézség intézeti címek esetén; az intézeti 'lakások' pontos keretének hiánya.

A fentiek miatt a migrációs felvétel **keretsokasága**, vagyis az alkalmazott mintavételi keret és összeírási gyakorlat által elérhető sokaság tehát a népszámlálás, a KEKKH és a BÁH forrásokban fellelhető címeken magánháztartásban élő harmadik országbeli személy. A megvalósult mintából származó becslések erre a sokaságra vonatkoznak.

3. A mintavételi terv

A feladat egy ~4000 címet tartalmazó minta kiválasztása az egyesített keretből. Összeírás-szervezési okokból a keretnek külön kezeltük azon címeit, amik olyan településen vannak, ahol a munkaerő-felmérés (MEF) nincs jelen. Ezért a migrációs felvétel mintavételi terve kétféle, attól függően, hogy MEF-es vagy nem MEF-es település címeiről van szó.

A nem MEF-es településeken a keretben mindössze 8524 cím van, összesen 1611 településen. Ebből csak azokat a településeket tekintettük, ahol van legalább 8 cím a keretben: a 303 ilyen településen összesen 5065 cím van a keretben. A 303 településből 50-et választottunk ki nagysággal arányos valószínűséggel (megye és településnagyság-kategória szerinti implicit rétegezéssel), és minden kiválasztott településen 8 címet. Összességében tehát ezen a részen 400 kiválasztott címünk van.

A migrációs felvétel mintavételi keretének túlnyomó része a MEF településmintáján található, összesen, közel 95 ezer cím. Ezen a sokaságon egylépcsős rétegzett kiválasztással jutottunk a mintához, összesen 3595 címet választva. A rétegzés szempontjai:

- a cím milyen forrás(ok)ból származik
- KEKKH-s cím esetén lakóhely-e a cím (1) vagy csak tartózkodási (0)
- BÁH-os cím esetén lakik-e ott szakértőink szerint célsokasághoz tartozó személy (erv) vagy sem (old)

A fenti tényezők értékei 17 réteget definiálnak, ahogyan az 1. táblázatban látható. A táblázatban feltüntettük a mintavételi keret címeinek számát és a tervezett mintaelemszámot is. Az allokáció a rétegek között arányos, közel minden 15. cím kerül a mintába. Ez alól kivétel azon BÁH-os címek rétege, amit csak a BÁH forrásban találtunk meg és ahol csak olyan személyek vannak a forrásban, akik szakértőink szerint már nem tartoznak a célsokaságba: ebben a rétegben a kiválasztási arány jóval kisebb, összesen 200 címet figyelünk meg a 45 ezerből.

Az 1. táblázat tehát egy rétegzett mintát definiál, mintaallokációval együtt. Egy rétegen belül a címek kiválasztása véletlen szisztematikus módon történt, ahol a címeket megye, település, közterület, házszám szerint rendeztük sorba.

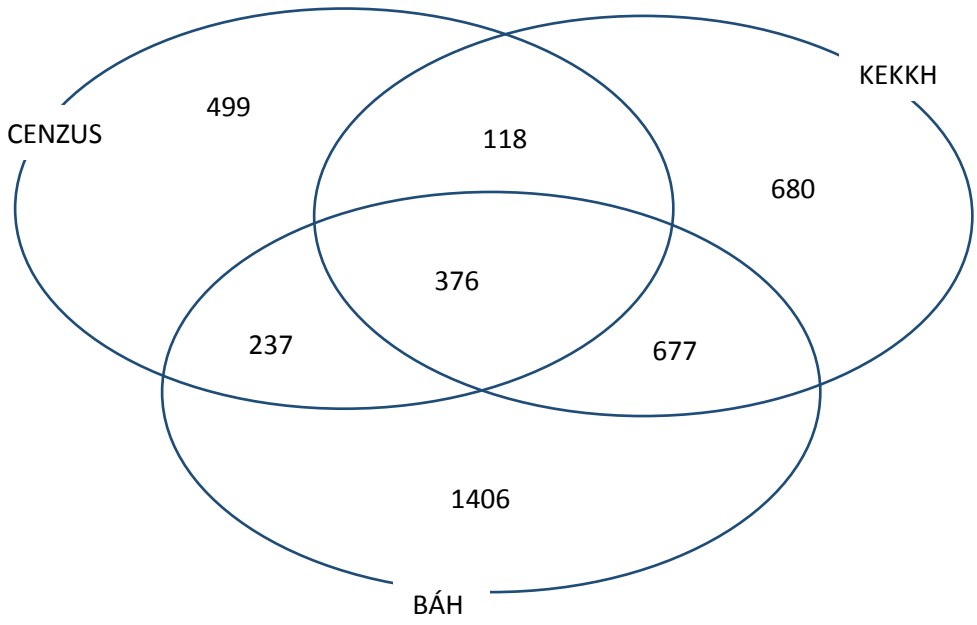
A mintakiválasztást követően, a terepmunkára készülés közben derült ki, hogy a 3995 elemű kiválasztott mintában két esetben duplán szerepelnek a címek, ennek megfelelően a tényleges

végső minta elemszám 3993. Az elemszámok források szerinti megoszlását mutatja az alábbi ábra. A cenzus, KEKKH és BÁH részmintákban rendre 1230, 1851 és 2696 cím van.

1. táblázat

Forrás			BÁH cím réteg	KEKKH cím réteg	Címek száma	Minta elemszám
Cenzus	BÁH	KEKKH				
0	1	0	erv		14604	1007
0	1	1	erv	0	97	6
0	1	1	erv	1	6525	450
1	1	0	erv		1726	119
1	1	1	erv	0	49	3
1	1	1	erv	1	4266	294
0	**1**	**0**	**old**		**45558**	**200**
0	1	1	old	0	146	10
0	1	1	old	1	2268	156
1	1	0	old		1481	102
1	1	1	old	0	25	1
1	1	1	old	1	727	50
0	0	1		0	1507	103
0	0	1		1	7592	523
1	0	0			6720	463
1	0	1		0	138	9
1	0	1		1	1444	99

2. ábra



A mintavételi terv a kiválasztott minta elemein definiálja az ún. **design súlyt**, ami a mintába kerülés valószínűségének reciproka. A design súllyal a kiválasztott minta reprezentálja a keretsokaságot. A fenti leírásnak megfelelően a design súly a minta legnagyobb részén ~15, a

'régi' BÁH-os címek sokaságán ~228 (1. táblázat vastagon szedett sora). Ha az összeírás során nem történt volna meghiúsulás, illetve a keretek minősége elfogadható, akkor a design súly, becslésre alkalmas súly. A gyakorlatban ezt a design súlyt módosítani kell ahhoz, hogy a megvalósult minta reprezentálja a keretsokaságot. A súlyozás leírása olvasható az ötödik és hatodik fejezetekben.

4. A meghiúsulási kérdőív

A migrációs felvétel szempontjából a legfontosabb természetesen az, hogy a kiválasztott minta hány megvalósult kérdőívet eredményez. A megvalósult címek esetén a keret pontosnak bizonyul, hiszen a címen elérhető harmadik országbeli személy. A minta súlyozása szempontjából azonban az igazán érdekes az, hogy mit tudunk a felvételben meghiúsuló címekről, mert ezen információk alapján tudjuk módosítani a megvalósult minta címeinek design súlyát.

Az összeíróknak a felvétel céljainak megfelelően ezért nem csupán egy meghiúsulási kóddal kellett ellátniuk a kérdőívet, hanem ki kellett tölteniük egy meghiúsulási kérdőívet. Ezt úgy alakítottuk ki, hogy minél pontosabban kiderüljön számunkra, hogy

- egyrészt a lakott (ám meghiúsult) címeken él-e harmadik országbeli személy,
- másrészt az üres lakásoknál, illetve ahol nem jelenleg nem él célszemély, ott lakott-e az elmúlt egy évben harmadik országbeli?

A meghiúsulási kérdőív és segédlete a mellékletben látható.

A **nem azonosítható** (meghi=21) és **nem létező** címek (meghi=22) kerethibás címek, miként a **nem lakáscím** is (meghi=24). Ezeknél az eseteknél nem kérünk további információt az összeírótól. Itt meg kell jegyezni, hogy a nem lakáscím kategóriába tartoznak az intézetek is (kollégium, munkásszállás, kereskedelmi szálláshely). Annak ellenére, hogy ezeken a címeken élhetnek (és élnek is) harmadik országbeliek, ezek mégsem nem összeírandó címek. Információként rendelkezésre fog állni, hogy a mintában hány címet találtak intézetnek, de a végső becslésben ezek és az intézeti lakók nem jelennek meg.

Az **elérhetetlen háztartás** (meghi=31), **válaszmegtagadás** (meghi=41), **válaszképtelenség** (meghi=42) és **nyelvi nehézség** (meghi=43) mind olyan kategóriák, amik lakott lakásokra vonatkoznak. Ekkor az összeírónak a meghiúsulás ellenére információt kellett szereznie arról, hogy lakik-e célcsoportbeli személy a címen (meghiúsulási kérdőív 2. kérdés, **celcsop1** néven hivatkozunk rá).

Az **üres lakás** (meghi=23) és a **nincs célszemély** (meghi=50) kategóriákban pedig arról kellett információt szereznie, hogy lakott-e célcsoportbeli személy a címen az elmúlt egy évben (meghiúsulási kérdőív 3. kérdés, **celcsop2** néven hivatkozunk rá).

Az alábbi, 2. táblázat mutatja a *meghi*, *celcsop1* és *celcsop2* mezők értékeinek lehetséges kombinációit.

A lehetséges kombinációk közül a *celcsop1=5* az egyik kategória, ahol nincs információnk arról, hogy a lakott lakásban él-e jelenleg harmadik országbeli. A *celcsop2=6* kategóriáról csak azt tudjuk, hogy ott nem él jelenleg célszemély, de nem ismert, hogy korábban lakott-e?

Bár a *celcsop1=1,3* és *celcsop2=1,4* kategóriák vélelmezésen alapuló információk, a keretek jellemzésekor ezeket önálló kategóriának tekintjük[2].

2. táblázat

MEGHI	CELCSOP1	CELCSOP2
12		
21,22,24		
31,41,42,43	1	
	2	
	3	
	4	
	5	
23,50		1
		2
		3
		4
		5
		6

A kiválasztott minta minden címéhez kötődik egy úgynevezett design súly, ami a cím mintába kerülési valószínűségének reciproka. Ezt a súlyt használva a 2. táblázat feltölthető mindhárom részmintából. Ahhoz azonban, hogy az információhiányos címeket kiküszöböljük, a design súlyt módosítani kell a többi címen, ahol van információ a keretről. Ennek a súlyozásnak a leírását mutatjuk be a következő fejezetekben.

Megjegyezzük, hogy minden kiválasztott cím esetén (a megvalósultaknál is) az összeíró lejegyezte a cím környezeti jellegét, az épület típusát és állagát.

5. A részminták új design és elsődleges súlyozása

A keretek megbízhatóságát elemző tanulmányban leírtuk azt a hagyományosnak nevezhető elsődleges súlyozást, ami alapján minősítettük a kereteket, és ami súly alapja lehetett volna a további súlyozási lépéseknek. Az eredeti elsődleges súlyozás leírása a mellékletben olvasható.

A keretminőség vizsgálatának egyik, mostani feldolgozásunkat is érintő, kellemetlen következménye volt, hogy mindhárom forrás esetében nem kevés olyan címet találtunk, ami nem része a másik kettő forrásnak. Tanulságként megfogalmaztuk, hogy *a megvalósult minta súlyozásakor különös figyelmet kell fordítani a lehetséges címismétlődések kezelésére*.

Címismétlődések a keretben

A fentieknek megfelelően először a közel 4000 elemű kiválasztott minta címeit vizsgáltuk meg tételesen. Azt kerestük, található-e benne számottevő ismétlődés. Ennek eredménye mintegy három tucat ismétlődés. Ez önmagában nem tűnik túlságosan nagyarányúnak a teljes minta elemszámhoz képest, de ha figyelembe vesszük, hogy a címeket véletlen szisztematikus

[2] A megvalósult minta végső súlyozásakor, a végső becslések elkészítésekor dönteni kell, hogy elfogadjuk-e az összeírók feltételezését vagy a vélelmezés esetén modellezzük a bizonytalanság mérték.

módon választottuk ki, ahol a sorba rendezés éppen a címek szerint történt, akkor ez a szám már gyanakodásra ad okot.

Ezek után vette kezdetét a teljes keret tételes átvizsgálása: ennek során a 103397 elemű teljes keretben kerestük azokat a címeket, amik megegyezhetnek a 3995 elemű kiválasztott minta címeivel, amik a nem egységes és pontatlan címhasználat miatt a keretben több soron is megjelenhettek. A tételes átvizsgálás jelentős mennyiségű, előre nem tervezett munkaórát jelentett. Ideális esetben nem csupán a kiválasztott minta, hanem a keret összes címének ismétlődését kerestük volna, azonban ez kivitelezhetetlenül sok időt vett volna igénybe, ezen kívül egy korrekt súlyozáshoz elegendő ez a szűkített vizsgálat is.

Ezen a ponton új fogalmakat kell bevezetni a pontosság kedvvért. Mivel kiderült, hogy a mintavételi keret számos sora ugyanarra a címre mutat, ezért a keret soraiban megtalálható címekre **keretcímként** hivatkozunk. Ennek megfelelően tehát a felvételben több keretcím mutathat ugyanarra a címre. A vizsgálat célja, hogy azonosítsa a keretben az összes olyan keretcímet, ami ugyanarra a címre mutat, amire a kiválasztott minta elemei. A vizsgálat eredményeként tudjuk, hogy a kiválasztott minta által elért minden egyes címet, a keretből hány keretcímmel választhattuk volna ki. A mellékletben bemutatunk néhány példát: ezek közül az első három olyan, ahol két keretcím mutat ugyanarra a címre, a negyedik példában már három. Ezeknek a csoportoknak a mérete (a cím kapcsolódásainak száma a kerethez) lesz az alapja az új súlyozásnak. Meg kell jegyezni, hogy a kapcsolatok feltérképezése nem lehetett tökéletesen pontos:

- egyrészt helyismeret nélkül sok esetben nem eldönthető a kapcsolódás (bár itt az internet több esetbe segítségünkre volt, ahol kideríthettük, hogy adott címen a tetőtér és a 3. emelet ugyanaz);
- másrészt nem tudtunk figyelembe venni olyan pontatlan címhasználatot, ami pl. egy számjegy elírásnak köszönhető;
- végül bizonyos esetekben feltételezéssel éltünk.

A vizsgálat számszerű eredménye: a 3995 elemű mintán belül **1137** esetben találtunk a keretben ismétlődéseket. Ez rendkívül nagy szám, alapjaiban átírja a súlyozás lépéseit, illetve az egyes címek hovatartozását: sok esetben pl. a mintába kiválasztott pusztán BÁH-os keretcímről kiderül, hogy ugyanarra a címre mutat, mint egy cenzus/KEKKH-beli keretcím.

Mivel a vizsgálatot csak a minta elemeire szűkítettük, így azt sem tudjuk megmondani ezen a szinten, hogy a keretben végül hány darab cím van. Ez majd becslésként fog előállni a feldolgozás során.

Az intézeti címek kiszűrése

A keretvizsgálat jó alkalmat jelentett az intézeti (keret)címek kiszűrésére is. Ehhez rendelkezésünkre állt a népszámlálásban megfigyelt intézetek listája. A fentiekhez hasonló módon azonosítottuk a keretben lévő intézeti keretcímeket.

Érdekes tanulságként megjegyezzük, hogy az intézeti címek egészen másképp jelennek meg a különböző forrásokban. A népszámlálás listáján jellemzően az intézeti épületek szerepelnek, míg pl. a BÁH forrásban akár szobaszám pontossággal rögzített keretcímnek vannak. Ennek köszönhetően egy kollégium címe a népszámlálásban egy rekordon jelenik meg, míg a BÁH forrásban akár több száz soron is.

A népszámlálási intézeti címek keretcímeit töröltük a keretből, amiben így **100463** keretcím maradt (vagyis közel 3000 intézeti keretcímet töröltünk, zömmel BÁH-os forrásból valót). Ebben még mindig vannak intézeti címek, olyanok, amik a népszámlálás listájából hiányoztak. A kiválasztott minta meghiúsulási kérdőíveinek köszönhetően ezek azonosíthatók. A mintában maradó intézeti címek a súlyozás megfelelő lépése után kerültek törlésre.

Új design súly

A keretben fellelt számos ismétlődés átírta a súlyozást. A 3995 elemű kiválasztott minta minden egyes elemére a design súlyt a mintavételi és kiválasztási terv definiálta. Csakhogy ezek a design súlyok nem a címekre, hanem a keretcímekre vonatkoznak. Egy címre pedig több keretcím is mutathat, ez pedig az ún. indirekt mintavétel tipikus esete. Indirekt mintavételnél pedig a címekhez köthető súly nem egyezik meg a keretcímek súlyával.

A Generalized Weight Share Method (GWSM)[3] alkalmazásával azonban tudunk címekhez súlyt rendelni. Itt nem ismertetjük az eljárást matematikai részletezettséggel, főbb lépései:

(1) minden mintabeli cím esetén meg kell számlálni, hogy összesen hány keretbeli keretcím mutat arra a címre, ez a kapcsolódások száma (ez történt a címismétlődések vizsgálatakor);

(2) minden mintabeli cím esetén összegezzük a rá mutató mintabeli keretcímek design súlyát;

(3) végül a súlyösszeget osztjuk a kapcsolódások számával.

Itt fontos kiemelni, hogy a GWSM súly már nem keretcímhez, hanem címhez kötődik, az állományok kezelésében erre különös figyelmet kell fordítani.

Végső soron a GWSM eljárással a minta címeire előállítottuk az új, címekre vonatkozó design súlyokat.

A következő lépéseket már részmintánként végezzük el, hogy a végén újra egyesítsük azokat. A súlyozás fő lépései:

- a címekhez új design súly meghatározása (GWSM);
- az egyes részmintákon elsődleges súlyozás (az információhiányos címek kiküszöbölése);
- az egyes részmintákon a megvalósult címek súlyozása;
- végül a három részminta egyesítése, a részminták metszetébe eső címeken az egyes részmintákra kiszámolt súlyok átlagolása.

Új elsődleges súly

Az új elsődleges súly képzése két alapvető ponton tér el az eredetitől (az eredetit lásd a mellékletben).

[3] Deville J. and Levallee P (2006): Indirect Sampling: The Foundations of the Generalized Weight Share Method. Survey Methodology, December 2006, Vol.32. No.2, pp.165-176,

Egyrészt a kiinduló súly nem a keretcímek design súlya, hanem a címekre kiszámított új design súly (GWSM).

Másrészt a súlyozás cím szinten történik, nem keretcím szinten. A modellezésnél használt információkat címszinten kellett felhasználni.

Utóbbi technikailag meglehetősen bonyolulttá tette a feldolgozást, de lényegében ugyanazok a lépések történtek, mint az eredeti változatban, ezért itt nem is részletezzük.

Az új elsődleges súlyok részmintánként minden olyan címhez rendelkezésre állnak, ahol van információ (a kiválasztott minta, kivéve a *celcsop1=5* és *celcsop2=6* címek). Ezekkel a súlyokkal a részminták reprezentálják a megfelelő részkeretsokaságot, immár cím szinten.

6. A megvalósult minta súlyozása

A megvalósult minta súlyozása az elsődleges súlyok módosítása. Több lépésben történt.

(1) Részmintánként a nem azonosítható (*meghi=21*) és nem létező (*meghi=22*) címek miatti korrekció. E mögött az a feltételezés van, hogy valójában ezek a címek is léteznek, csak a keretben szerepelnek annyira pontatlanul, hogy ilyen kódokkal hiúsult meg a felvétel. Azoknak a címeknek az elsődleges súlyát korrigáltuk, amiknél *meghi nem 21 vagy 22*. A korrekció az elsődleges súlyozásnál is alkalmazott logisztikus regresszió segítségével becsült valószínűségek szerint történt. A korrigált súllyal a maradék részminta ugyanazt a sokaságot reprezentálja, mint a korrekció előtti minta az elsődleges súllyal.

(2) Részmintánként az intézeti címek elhagyása. Ez elsősorban a BÁH címeket érintette, hiszen a cenzus címek közül a feldolgozás elején kigyomláltuk az intézeti címeket. A mintában ezek után találtak az összeírók még 45 intézeti címet (amik nem voltak benne a cenzus listájában). Ezzel a maradék részmintákban a magánháztartási címek maradtak. Ezek közül a *celcsop1=1,2* és a *meghi=12* típusú címek a migránsok által lakott (vagy feltehetően lakott) cím. Ezzel a módosított súllyal a *celcsop1=1,2* és a *meghi=12* típusú címek a magánháztartásban lakó migránsok címeit reprezentálják.

(3) Az utolsó korrekció a migránsok által lakott címek körében történt, részmintánként. Ennek során a megvalósulás (meghi=12) valószínűségét modelleztük, és ezzel a valószínűséggel korrigáltuk az előző súlyt. Ez a súly az adott részminta megvalósult elemeinek végleges súlya. Ezzel a súllyal a megvalósult minta részmintánként reprezentálja a keretsokaságot.

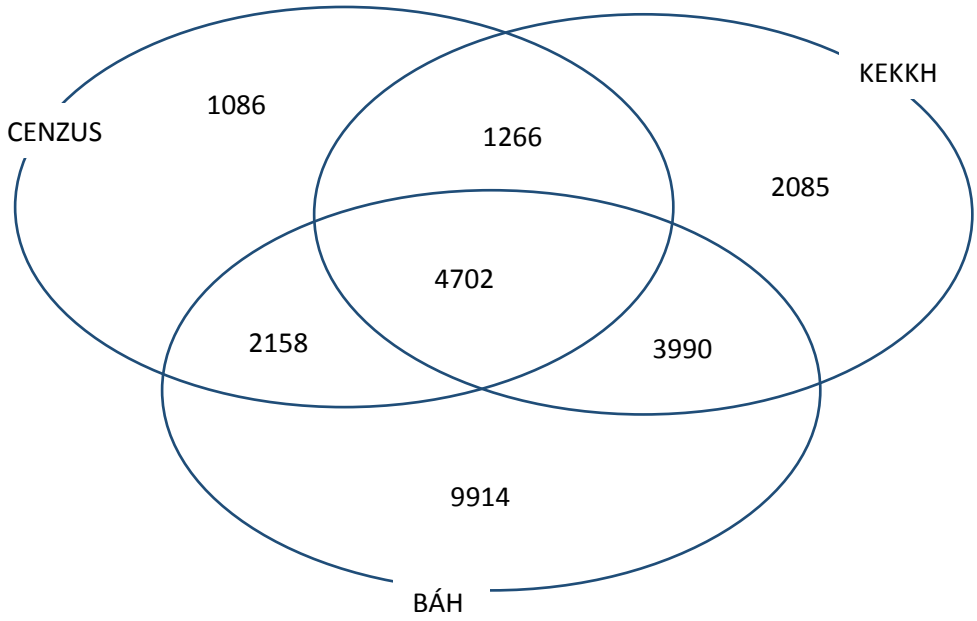
(4) A súlyozás utolsó lépése a három részminta egyesítése és a súlyok átlagolása. Az átlagolás nyilván csak a részminták metszetein értelmezhető. A súlyozás tervezése kezdetén a legkézenfekvőbb megoldást terveztük, az egyszerű számtani átlagolást. A feldolgozás során azonban azzal szembesültünk, hogy a BÁH-keret a legkevésbé megbízható ebből a szempontból, ennek köszönhetően a BÁH részminta becsléseit tartjuk önmagukban a legkevésbé megbízhatónak. A részminták metszetiben a súlyokat az alábbi módon egyesítettük:

- a cenzus-KEKKH metszetben a két részminta súlyainak egyszerű számtani átlagát vettük;

- a cenzus-BÁH metszetben a két részminta súlyainak súlyozott számtani átlagát vettük, a cenzusbeli és BÁH-beli címek súlyai rendre 0.7 és 0.3;
- a KEKKH-BÁH metszetben a két részminta súlyainak súlyozott számtani átlagát vettük, a KEKKH-beli és BÁH-beli címek súlyai rendre 0.7 és 0.3;
- a cenzus-KEKKH-BÁH hármas metszetben a három részminta súlyainak súlyozott számtani átlagát vettük, a cenzusbeli, KEKKH-beli és BÁH-beli címek súlyai rendre 0.4, 0.4 és 0.2.

A végső egyesített súlyok címekhez kötődnek. Ezen a szinten becslést lehet adni a migránsok által lakott lakások számára a keret forrásai szerint. Ennek eredményei a 3. ábrán láthatók. Hasonló ábrát készítettünk a korábbi, keretminőséget vizsgáló dokumentumban is, az ott alkalmazott klasszikus elsődleges súlyozás felhasználásával[4]. Érdemes összevetni a mostanit az akkori számokkal, amiket itt a 4. ábrában tüntetünk fel.

3. ábra

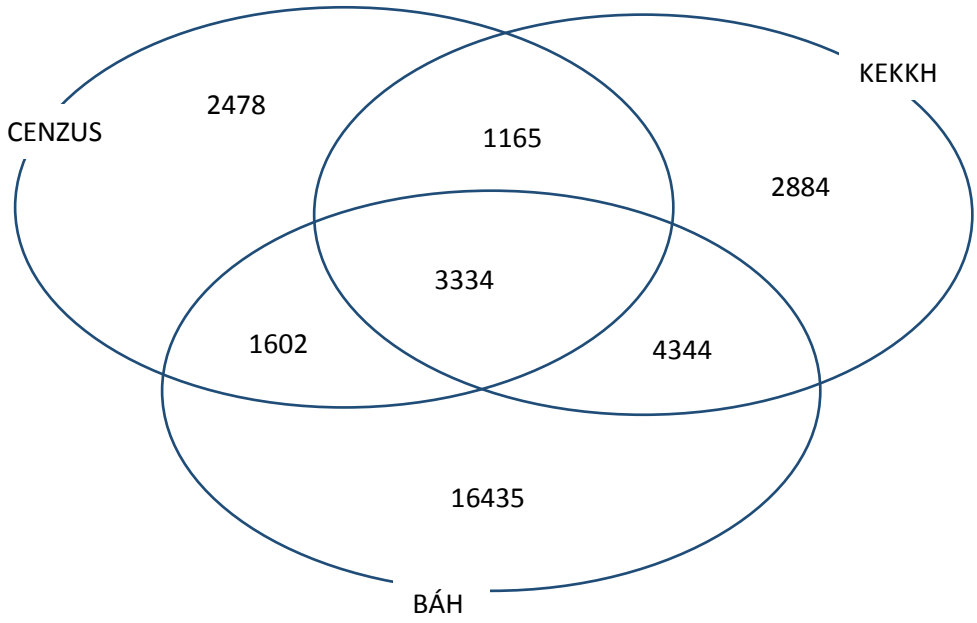


A különbség szembetűnő, ami a keretcím ismétlődések számlájára írható. Nem kérdéses, hogy a most alkalmazott súlyozás ad pontosabb becsléseket, a GWSM eljárással reményeink szerint az eredeti, keretcímekhez köthető design súlyok közvetlen használatával adódó torzítások jelentős része kiküszöbölt.

Igazán számottevő mértékben a BÁH-címek száma változott, jelentősen csökkent, a cenzus és KEKKH címek száma kis mértékben növekedett. Azon túl, hogy csökkent a BÁH címek száma, az igazán jelentős változás az a metszetek felé való eltolódás, vagyis csökkent azoknak a címeknek a száma, amik csak az egyik forrásban elérhetők, bár megjegyezzük, hogy számuk még így is meglehetősen nagy.

[4] Amiről időközben kiderül, hogy erősen torzított eredményt ad, köszönhetően a rendkívül nagyarányú keretcím ismétlődésnek.

4. ábra



Összefoglaló

A migrációs felvétel három forrásból származó, címeket tartalmazó mintavételi kerettel kívánta elérni a Magyarországon élő harmadik országbelieket: a 2011-es népszámlálás, a KEKKH személyiadat- és lakcímnyilvántartása, valamint a BÁH adatbázisai álltak rendelkezésünkre. A feldolgozás jelen szakaszában rendelkezésünkre áll egy végső becslő súly, amellyel a megvalósult minta reprezentálja a keretsokaságot, vagyis a három forrásban elérhető címeken a magánháztartásban lakó migránsokat.

A feldolgozás során számos nehézségbe ütköztünk, többször kiemeltük a címkezelés hiányosságait. Ezen a pontos ismét hangsúlyozzuk, hogy az itt feltárt komoly hiányosságokból nem feltétlenül lehet következtetni az egyes (adminisztratív) adatforrásokban megjelenő személyek nem lakcímet érintő helytelen nyilvántartására.

A legfontosabb megállapítások.

(1) Mint minden felvétel, természetesen a migrációs felvétel is hibákkal terhelt. A megvalósult minta súlyozása nem egy determinisztikus folyamat. Több olyan pontja van, ami bizonytalanságot hordoz, illetve több lehetséges megoldás közül választ.

   a. Az egyik ilyen pont a keretcím ismétlődések azonosítása. Értelemszerűen ezt nem lehet tökéletesen pontosan kivitelezni. Bizonyosan kreáltunk hibás kapcsolatokat, illetve bizonyosan elsiklottunk létező kapcsolatok felépítése fölött.

   b. Bizonytalanságot növelő további elem, hogy a kapcsolatokat csak a minta címeire szűkítettük. Itt megjegyezzük, hogy ez önmagában torzítást nem okoz, csupán a becslések szórását növelheti.

   c. A meghiúsulások modellezése, a megvalósult minta súlyozása nem csupán az itt bemutatott megoldásokkal kivitelezhető.

Elsősorban a cenzus és KEKKH részminták súlyozásánál lehetett volna más technikát alkalmazni: mivel ezek a források a címkezelésben pontosabbnak bizonyultak, itt alkalmazhattunk volna a lakossági felvételnél bevált ún. kalibrálást a meghiúsulások kompenzálására. Hogy ez mennyire eltérő eredményt adna, egy későbbi módszertani kutatás témája lehet. Mivel a BÁH részmintán a keretcímek nem megbízhatóak, ezért azon a részmintán ennek a gondolatát is elvetettük, és úgy döntöttünk, hogy mindhárom részmintán hasonló technikával történjen a súlyozás.

Ehhez a ponthoz tartozik az a döntés is, miszerint elfogadtuk az összeírók által vélelmezett címen lakást is (*celcsop1=1*).

d. A három részminta egyesítésekor a metszetekben lévő címek végső súlyának megállapítása sem egyértelmű: érzésünk alapján indokolt a BÁH-súlyoknak az átlagolásnál kisebb súllyal történő figyelembe vétele, de az alkalmazott súlyok meghatározása (6. fejezet (4) pont) mögött nincs tudományos magyarázat.

(2) Az önmagában megnyugtató, hogy az új GWSM súlyokkal jelentősen átrendeződött a kép, már ami a lakott címek számának eloszlását illet (3. és 4. ábra). Továbbra is magas azonban azon címek száma, amik csak egyetlen forrásból elérhetők. Ennek a kérdésnek a vizsgálata a későbbi elemzések tárgya lehet.

Tanulságok.

(3) Itt is ki kell emelnünk, hogy a közigazgatásban komoly szüksége lenne az egységes címszabvány használatára adminisztratív nyilvántartásokban, regiszterekben.

(4) Komoly tanulságként szolgált az előkészítés és feldolgozás során tapasztalt számos nehézség. Ennek lényege, hogy nagyon óvatosan, körültekintően kell eljárni egy ismeretlen, korábban nem használt adatforrás mintavételi keretként való alkalmazásánál. Már az adatfelvétel folyamatának kezdetén azonosítani kell a keretekben rejlő problémákat (címazonosítás, ismétlődés), szükség esetén betervezni a kerettisztítást, szélsőséges esetben a keret használatának elvetése is lehet az eredmény.

Mellékletek

Meghiúsulási kérdőív

MEGHIÚSULÁSI KÉRDŐÍV

A felkeresett cím azonosítója: L[ ][ ][ ][ ][ ][ ][ ][ ][ ][ ][ ] **MEGHI** [ ][ ]

Ha elérhetetlen háztartás (31), vagy válaszmegtagadás történt (41)

1. **Volt-e sikeresnek induló felkeresése ezen a címen?**
   (1) Igen, volt olyan felkeresése, amikor kérdőívet hagyott a címen
   (2) Igen, volt olyan felkeresése, amikor internetes elérést hagyott a címen
   (3) Igen, volt olyan felkeresése, amikor kérdőívet és internetes elérést is hagyott a címen
   (4) Nem volt sikeresnek induló felkeresése ezen a címen (megtagadták a válaszadást vagy nem érte el őket)

Ha elérhetetlen háztartás (31; *szomszéd, névtábla alapján*) vagy válaszmegtagadás (41; *a kérdőívet nem tölti ki, de lehet, hogy ebben a kérdésben együttműködő*) vagy válaszképtelenség (42) vagy nyelvi nehézség (43)

2. **Lakik-e a címen harmadik országbeli állampolgár?**
   (1) Vélhetően IGEN
   (2) Biztosan IGEN
   → **Tovább a 4. kérdésre!**
   (3) Vélhetően NEM
   (4) Biztosan NEM
   (5) Nem tudja
   → **Tovább a 8. kérdésre!**

Ha üres lakás (23) vagy nem célcsoportbeli a címen élő (50)

3. **Lakott-e a címen korábban, az elmúlt egy évben célszemély?**
   (1) Vélhetően IGEN
   (2) Biztosan IGEN
   (3) Biztosan IGEN, de már Európai Uniós állampolgárságot (is) kapott
   (4) Vélhetően NEM
   (5) Biztosan NEM
   (6) Nem tudja
   → **Tovább a 8. kérdésre!**

4. **Tudja-e milyen nyelven valósulhatott volna meg a kikérdezés? (Akár plusz információk alapján.)**
   (1) Biztosan tudja (ha például a nyelvválasztó alapján tudja, vagy a szomszéd biztos benne)
   (2) Vélelmezi
   (3) Nem tudja
   → **Tovább a 6. kérdésre!**

5. **Véleménye szerint, milyen nyelven valósulhatott volna meg az adatfelvétel?**
   (1) angol (2) orosz (3) kínai (4) vietnami (5) arab (6) magyar
   (7) egyéb, éspedig.....

6. **Hány fő él(t) a címen?**
   [ ][ ] → **Ha nincs információja a címen élők számáról folytassa a 8. kérdéssel!**

7. **Vélhetően hány harmadik országbeli él(t) a címen ? (Próbálja kitölteni akkor is ha nem tudja pontosan, ha egyáltalán nincs róla információja, hagyja üresen!)**
   [ ][ ]

8. **MEGJEGYZÉS: Kérjük, adja meg honnan származik a plusz információ ha van, a „célcsoporthoz tartozást"/nyelvet, mi alapján állapította meg; ha megtagadás történt mi lehetett az oka stb.!**

.....
.....
.....

8. MEGJEGYZÉS:	
Címismétlés (22) *(már korábban ebben a felvételben felkeresett cím)*	"Dupla"
MEF mintában is szereplő cím (22)	"MEF cím"
Intézményt talált (24)	"Intézmény"

SEGÉDLET

KÓD	KATEGÓRIA	JELENTÉS	
12	*Sikeresen megvalósult*	**laptop, papír, web**	**Nem kell kitöltenie ezt a kérdőívet!**
21	*nem azonosítható cím*	**mint a MEF**	**Nem kell kitöltenie ezt a kérdőívet!**
22	*nem létező cím*	**mint a MEF +címismétlés, vagy a MEF mintában is szerepel**	**Kezdje a 8. kérdéssel!**
23	*nem lakott (üres) lakás*	**mint a MEF**	**Kezdje a 3. kérdéssel!**
24	*nem lakáscím*	**mint a MEF+intézmény**	**Kezdje a 8. kérdéssel!**
31	*elérhetetlen háztartás*	**mint a MEF (utolsó kontaktkísérlet alapján)**	**Kezdje a 1. kérdéssel!**
41	*válaszmegtagadás*	**mint a MEF (utolsó kontaktkísérlet alapján)**	**Kezdje a 1. kérdéssel!**
42	*válaszképtelenség*	**mint a MEF**	**Kezdje a 2. kérdéssel!**
43	*nyelvi nehézség*	**mint a MEF**	**Kezdje a 2. kérdéssel!**
50	*nincs célszemély*	**biztos információ az ott lakóktól – volt kapcsolatfelvétel)**	**Kezdje a 3. kérdéssel!**

A részminták elsődleges súlyozása – eredeti

Az előző fejezetben ismertetett okok miatt tehát azoknak a címeknek a design súlyát módosítottuk (növeltük), ahol van valamilyen információ a cím és a harmadik országbeliek kapcsolatáról. A súlymódosítás hatására a

- *celcsop1=1,2,3,4* típusú címek a módosított (elsődleges) súllyal reprezentálják a *celcsop1=1,2,3,4,5* típusú címeket, illetve a
- *celcsop2=1,2,3,4,5* típusú címek a módosított (elsődleges) súllyal reprezentálják a *celcsop2=1,2,3,4,5,6* típusú címeket.

A *meghi=12, 21, 22 és 24* típusú címek elsődleges súlya megegyezik a design súllyal.

Az elsődleges súlyozás, miként a keretek megbízhatósági vizsgálata a három részmintán (a három adatforrásra) külön történik, külön eredményünk lesz a három keretre[5].

A design súly módosítása a logisztikus regresszió segítségével történt. Ez az eljárás modellezi annak a valószínűségét, hogy adott címről van-e információnk vagy nincs (benne van-e *celcsop1=1,2,3,4* vagy *celcsop2=1,2,3,4,5* kategóriákban). Ezt követően a címek design súlyát osztjuk a kapott valószínűséggel, ez adja az elsődleges súlyt. A modellezéshez olyan címszintű információkat használhattunk, amik rendelkezése állnak a részminta egészén.

Az alábbi magyarázó változók mindhárom részminta súlyozásánál szerepeltek:

- településtípus és régió
- a nyelv (lásd meghiúsulási kérdőív 5. kérdés)
- a cím környezete, az épület típusa és állaga (ez is része a kérdőívnek)
- annak indikátora, hogy a cím melyik forrásokból származik (lásd 1. ábra)
- adott forrás szerint milyen földrészről származók köthetők a címhez, és hányan vannak

Ezen túl az egyes adatforrások sajátosságainak megfelelően további magyarázó változókat építettünk be.

A cenzus részminta súlyozásánál a címen lakók aktivitását, illetve a lakáshasználat jogcímét.

A KEKKH részminta súlyozásánál a címen lakók családi állapotát, nemét és korcsoportját.

A BÁH részminta súlyozásánál a címen lakók korcsoportját, illetve azt az információt, hogy a cím melyik input adatbázisból került a keretbe.

A bemutatott eljárást követően a cenzus, KEKKH és BÁH részminták mindegyikére előállt az az elsődleges súly, ami alapja a további vizsgálatoknak.

[5] A megvalósult minta végső súlyozásánál a három részmintát már nem külön kezeljük.

Címismétlődés, néhány példa

cenzus	KEKKH	BÁH	település	közterület neve	közterület jellege	házszám	épület	lépcső	emelet	ajtó
		1	Budapest 17. ker.	MINÁR GYULA	UTCA	17	B			
1			Budapest 17. ker.	MINÁR GYULA	UTCA	17/B				
		1	Budapest 02. ker.	BUDAKESZI	ÚT	55/D	C	1/4		
		1	Budapest 02. ker.	BUDAKESZI	ÚT	55/D	C	I/4		
1			Budapest 16. ker.	ÚJSZÁSZ	UTCA	45/B	G		2	207
	1	1	Budapest 16. ker.	ÚJSZÁSZ	UTCA	45/B	G		2	7
	1	1	Budapest 08. ker.	BAROSS	UTCA	103	A		9	32
	1		Budapest 08. ker.	BAROSS	UTCA	103-A			9	32
1			Budapest 08. ker.	BAROSS	UTCA	103/A			9	32