EURÓPAI INTEGRÁCIÓS ALAP

A projekt az Európai Unió Európai Integrációs Alapjának támogatásával valósul meg.

A mintavételi keretek megbízhatósága

a Magyarországon élő harmadik országbeliek felvétele alapján

Tartalom

Bevezetés

Az Európai Integrációs Alap támogatásával megvalósuló EIA/2013/2.6.1. számú, a „Migránsokra vonatkozó társadalomstatisztikai adatgyűjtések megalapozása” című projekt részként a Központi Statisztikai Hivatal 2014. második negyedévében mintavételes lakossági felvételt hajtott végre a Magyarországon élő harmadik országbeliek vizsgálatára; az egyszerűség kedvvért 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 alkalmas a fenti adatforrások és a használt módszerek (idegen-nyelvű kérdőívek, internetes kikérdezés) a Magyarországon élő harmadik országbeliek elérésére.

Jelen dokumentum célja, hogy a három adatforrás, mint felvételi keret megbízhatóságát bemutassa az összeírás eredményeinek tükrében. 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.

Az első fejezetben általánosságban és röviden írunk a mintavételi keretek hibáiról. A második fejezetben bemutatjuk a három különböző forrásból kapott állományokat. A harmadik fejezet a migrációs minta tervének, a negyedik fejezet az adatgyűjtés meghiúsulási kérdőívének rövid ismertetése. A hatodik fejezet a minta elsődleges súlyozásának általános leírása. A hetedik fejezet pedig a három adatforrás, mint mintavételi keret megbízhatóságának számszerű bemutatása.

1. A mintavételi keretek hibái

A klasszikus mintavételes felvételek alapja, hogy a célsokaság elemeiről (címek, személyek) létezik egy lehetőség szerint pontos és időszerű lista, jellemzően regiszterből vagy egyéb adminisztratív nyilvántartásból. Ez a lista szolgál a mintavétel keretéül: végső soron ezen lista elemiből választunk valószínűségi mintát. A keret minősége alapvetően befolyásolja a mintából számított becslések megbízhatóságát, a keret hibái a becslések torzítottságához vezethetnek.

- **Lefedettségi hiány:**

  Mintavételes felvételeknél a legveszélyesebb kerethiba. Lefedettségi hiányról akkor beszélünk, ha a keretben nincs jelen a teljes célsokaság, vagyis a kerethiba miatt a célsokaság egy része elérhetetlen a felvétel számára. Következménye a torzított becslés.

Általában a hiányt és annak hatását nehéz mérni. Mivel a migrációs felvétel egyszerre három keretet is használt, és az egyes keretek között nem teljes az átfedés, így képet tudunk alkotni arról, hogy a célsokaság mekkora része nem érhető el az egyes keretek segítségével.

- **Lefedettségi többlet:**

  Többletről akkor beszélünk, ha a keretben vannak nem célsokaságba tartozó elemek. A többlet nem okoz torzítást, ellenben költségnövelő. Általában természetes velejárója a felvételeknek, már csak azért is, mert az adatfelvétel időpontja és a mintavételhez használt keret utolsó frissítésének időpontja között mindig van egy rés.

- **Duplikátum:**

  Ha a célsokaság egyes elemei többször is megjelennek a keretben. Torzítást akkor okoz, ha az ismétlődéseket nem kezelik, akár a mintakiválasztás előtt, akár a becslés fázisában.

- **Pontatlan besorolás:**

  A mintavételi kereteken az azonosítókon kívül általában egyéb információk is elérhetők, amik felhasználhatók a mintavételi terv készítésénél vagy a becslés során segédinformációként.

A migrációs felvétel esetén mi is teszünk majd megállapításokat a keretek megbízhatóságára. Fontos azonban kiemelni, hogy számunkra az adatforrásokban szereplő címek a kiválasztási egységek, amiről olyan információink lesznek, hogy adott címen él-e, elérhető-e harmadik országbeli személy. Ennek megfelelően, ha valamely keret vizsgálatánál olyan következtetést vonunk le, hogy a keret címeinek adott százalékán nem él harmadik országbeli, akkor ez nem feltétlenül ezt jelenti, hogy az azokon a címeken nyilvántartott harmadik országbeliek már nem élnek Magyarországon, mert lehet, hogy itt élnek, csak más címen. Vagyis az adatforrásokat alapvetően mint címforrásokat fogjuk jellemezni.

Elvileg lehetséges olyan technika, ami a migrációs felvételben az egyik adatforráson kívül (csak a másik két adatforrás címein) megtalált célszemélyeket megpróbálja megkeresni a nyilvántartásban, ezzel pontosabb képet adva a személyi szintű lefedettségi hiányról, de ez a fajta összekapcsolási elemzés túlmutat ennek a vizsgálatnak a keretein.

2. Az adatforrások, a mintavételi keret

A migrációs felvétel három forrásból merítette a mintavételi keretet.

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

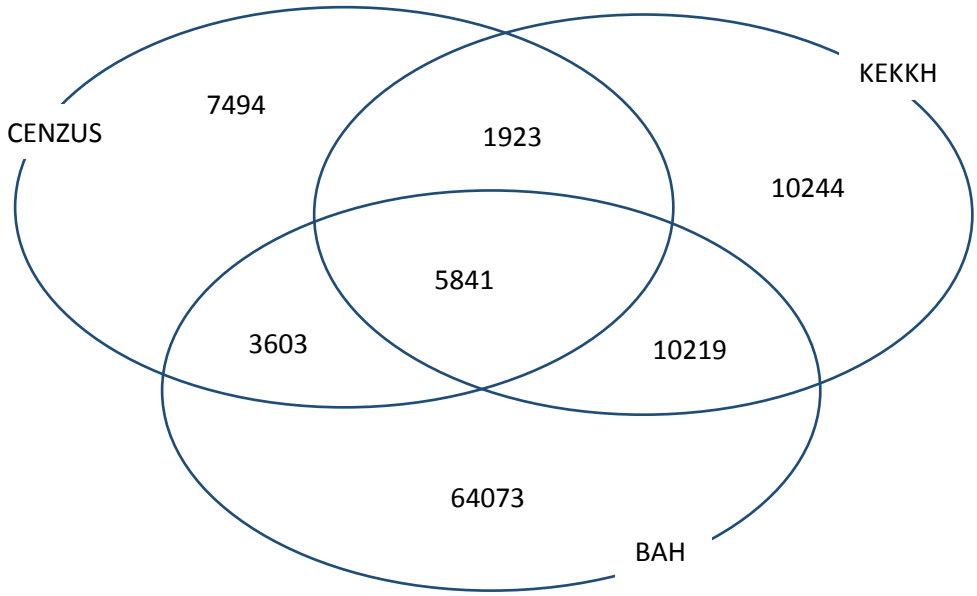
hasonló szűrések után ez plusz 3815 rekordot jelentett. Az összesen 45998 személy 28227 címet adott a mintavételi keretnek (az előzetes egyszerű vizsgálatok alapján 2-2 olyan címet találtunk, amik az adatforrás szerint különbözők, de vélhetően ugyanarra mutatnak). Nagy előnye a népszámlálási forráshoz 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. évi II. törvény alapján Magyarországon tartózkodó harmadik országbeli (nem EGT-s) állampolgárokat tartalmazza. Az EGT állomány a BÁH „szabad mozgás és tartózkodás jogával rendelkező" állampolgárok tartózkodási engedélyeinek nyilvántartásából leválogatott állomány, mely a 2007. évi 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 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.

1. ábra



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.

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)

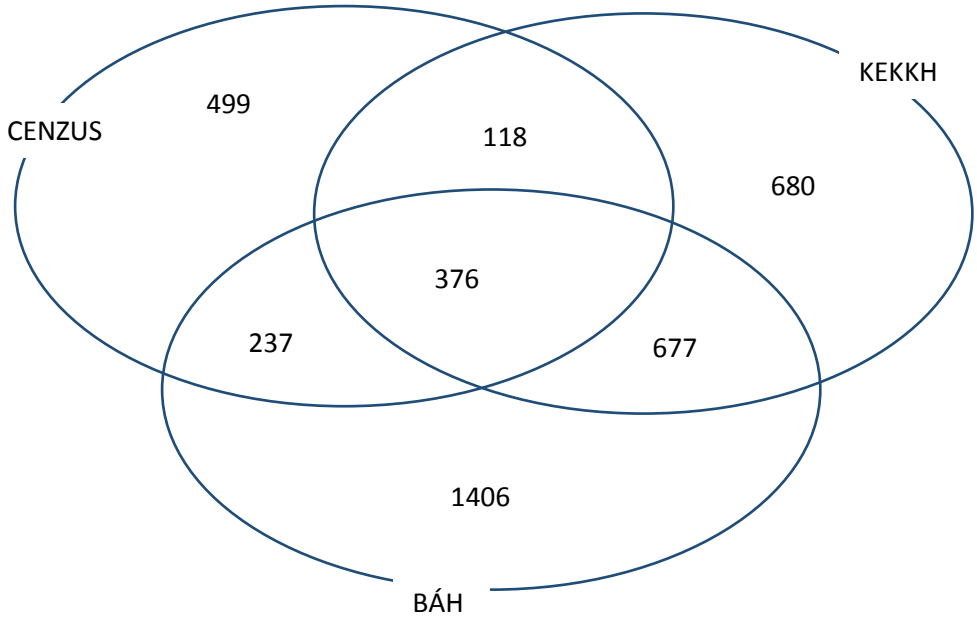
A fenti tényezők értékei 17 réteget definiálnak, ahogyan az 1. táblázatban látható. A táblázatban feltüntettük a mintavételi keret címeinek számát és a tervezett mintaelemszámot is. Az allokáció a rétegek között arányos, közel minden 15. cím kerül a mintába. Ez alól kivétel azon BÁH-os címek rétege, 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. A 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.

1. táblázat

forrás			BÁH cím réteg	KEKKH cím réteg	címek száma	mintaelemszám
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

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

2. ábra



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 mintavételi keret megbízhatósága szempontjából azonban az igazán érdekes az, hogy mit tudunk a felvételben meghiúsuló címekről. Nem minden meghiúsulás jelent kerethibát.

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 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[1].

[1] A megvalósult minta végső súlyozásakor, a végső becslések elkészítésekor ezeket a kategóriákat már másképp kell kezelni, valamilyen szempont szerint megszüntetni a bizonytalanságot.

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 2. táblázat formája lesz az alapja a három keret megbízhatósági elemzésének, kiküszöbölve az információhiányos kombinációkat. 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 fejezetben.

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

5. A részminták elsődleges súlyozása

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[2].

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

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

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.

6. A keretek jellemzése a részminták alapján

A cenzus részminta

A teljes mintavételi keretből a cenzus 18861 címet adott, ebből 18440 nem intézeti[3]. Az elsődleges súllyal ellátott 1038 cím a cenzus részmintán (a teljes részmintából kivéve a *celcsop1=5* és *celcsop2=6* típusokat) ezt a 18440 címet reprezentálja.

A keret megbízhatóságának számszerűsítéséhez elkészítünk egy, a 2. táblázathoz hasonló táblát (3. táblázat). A táblázat első sorában a cenzus részminta elemszámait tüntetjük fel.

Mivel a cenzus részminta több részre osztható aszerint, hogy a címek mely forrásokban találhatók (lásd 1. vagy 2. ábra), ezért a táblázat a mindösszesen oszlopon túl a források szerinti részeredményeket is feltüntetjük: eszerint C azokra a címekre utal, amiket csak a cenzus forrásból szedtünk, a CB és CK rendre a BÁH és KEKKH forrással közös címekre utal, míg a CKB azon címekre vonatkozik, amik mindhárom keretben benne vannak. Az elemszám alatti sorokban az elsődleges súlyokkal becsült %-os arányok láthatók *meghi, celcsop1* és *celcsop2* szerint (lásd 4. fejezet).

A táblázatban a színes kitöltéssel jelölt sorok utalnak azokra a címekre, amiken él vagy vélhetően él harmadik országbeli személy, gyakorlatilag ezek a mintavételi keret tartalmilag és formailag is pontos címei. Az ezekhez tartozó arányokat összegeztük, amik a táblázat alsó, a ’célcsoportbeli címek aránya’ sorban látható.

A táblázat számai a kerethibák közül a lefedettségi többletre és a pontatlan információra adnak becslést.

[3] A cenzus forrás esetén volt arra lehetőségünk, hogy az intézeti címeket a feldolgozás előtt kiszűrjük.

3. táblázat a cenzus részminta elemszáma és becsült %-os megoszlás

MEGHI	CELCSOP1	CELCSOP2	FORRÁS				
			mind	C	CB	CK	CKB
elemszám:			**1038**	416	187	105	330
12			**20.1**	13.3	13.9	24.4	31.1
21			**1.2**	1.8	0.0	0.8	1.4
22			**1.0**	1.0	1.6	0.0	1.0
23,50		1	**4.6**	6.1	7.0	1.7	2.2
23,50		2	**5.7**	5.5	7.5	4.7	5.0
23,50		3	**3.2**	4.5	4.2	1.6	1.4
23,50		4	**8.0**	10.0	7.6	9.9	5.2
23,50		5	**25.0**	28.8	25.9	19.8	21.5
24			**0.5**	0.7	0.0	0.0	0.8
31,41,42,43	1		**13.4**	11.7	12.0	18.2	14.7
31,41,42,43	2		**11.6**	9.6	15.1	13.7	11.3
31,41,42,43	3		**2.9**	4.1	0.4	4.2	2.6
31,41,42,43	4		**2.7**	2.8	4.8	1.0	1.8
célcsoportbeli címek aránya:			**45.1**	**34.6**	**40.9**	**56.3**	**57.1**

4. táblázat a KEKKH részminta elemszáma és becsült %-os megoszlás

MEGHI	CELCSOP1	CELCSOP2	FORRÁS				
			mind	K	KB	CK	CKB
elemszám:			**1626**	599	586	106	335
12			**18.4**	14.0	15.0	21.8	30.4
21			**2.9**	4.1	3.2	0.7	1.4
22			**1.1**	1.7	0.7	0.0	1.0
23,50		1	**3.6**	3.6	4.9	1.4	1.9
23,50		2	**4.2**	3.7	4.4	4.3	4.9
23,50		3	**2.1**	1.9	2.8	1.5	1.4
23,50		4	**6.6**	7.8	6.0	7.7	5.3
23,50		5	**29.6**	39.0	27.9	17.0	21.0
24			**2.9**	4.2	2.8	0.7	1.5
31,41,42,43	1		**11.7**	6.5	12.8	22.9	14.6
31,41,42,43	2		**11.6**	8.2	13.7	15.5	12.3
31,41,42,43	3		**2.3**	1.4	2.4	5.3	2.7
31,41,42,43	4		**3.0**	3.9	3.3	1.3	1.7
célcsoportbeli címek aránya:			**41.6**	**28.7**	**41.5**	**60.2**	**57.2**

A KEKKH részminta

A teljes mintavételi keretből a KEKKH 28227 címet adott[4]. Az elsődleges súllyal ellátott 1626 cím a KEKKH részmintán (a teljes részmintából kivéve a *celcsop1=5* és *celcsop2=6* típusokat) ezt a 28227 címet reprezentálja.

A cenzuséhoz hasonlóan a 4. táblázatban közöljük a becsléseket és elemszámokat.

A BÁH részminta

A teljes mintavételi keretből a BAH 83736 címet adott[5]. Az elsődleges súllyal ellátott 2350 cím a BÁH részmintán (a teljes részmintából kivéve a *celcsop1=5* és *celcsop2=6* típusokat) ezt a 83736 címet reprezentálja.

Az előzőekhez hasonlóan az 5. táblázatban közöljük a becsléseket és elemszámokat.

5. táblázat a BÁH részminta elemszáma és becsült %-os megoszlás

MEGHI	CELCSOP1	CELCSOP2	FORRÁS				
			mind	B	KB	CB	CKB
elemszám:			**2350**	1230	586	199	335
12			**11.6**	9.3	14.5	12.2	29.9
21			**6.4**	7.8	3.1	0.4	1.3
22			**4.5**	5.6	0.6	1.4	1.0
23,50		1	**4.0**	4.0	4.9	5.3	2.1
23,50		2	**3.4**	2.9	4.2	6.5	4.9
23,50		3	**1.7**	1.5	2.7	3.8	1.4
23,50		4	**6.9**	7.2	5.9	6.7	5.0
23,50		5	**31.4**	33.8	27.1	22.0	20.9
24			**5.1**	5.9	2.7	2.8	1.5
31,41,42,43	1		**9.7**	8.2	13.9	14.4	14.8
31,41,42,43	2		**9.8**	8.4	14.5	16.3	12.8
31,41,42,43	3		**1.8**	1.8	2.3	0.4	2.7
31,41,42,43	4		**3.6**	3.6	3.5	7.9	1.7
célcsoportbeli címek aránya:			**31.1**	**25.9**	**42.9**	**42.9**	**57.5**

A BÁH keretet már az előkészítésnél és a mintavétel tervezésénél is két részre osztottuk aszerint, hogy adott címen lakik-e vagy sem olyan harmadik országbeli, akit a szakértőink célsokaságbelinek tartanak (lásd 2. és 3. fejezet). Eszerint a BÁH-os címeket elláttuk 'erv', illetve 'old' jelzettel Utóbbi utal arra, hogy ott már 'nem kellene' célcsoportbeli személyt találni. Mivel itt nem számítottunk túlságosan nagy találati arányra, ezt a réteget a többinél kisebb arányban figyeltük meg (lásd 3. fejezet). Emiatt érdemes a BÁH részmintára vonatkozó eredményeket eszerint is vizsgálni. A becsléseket a 6. táblázatban közöljük.

[4] A KEKKH forrás esetén nem szűrtük az intézeti címeket a feldolgozás előtt, a végső becslésnél veszsük majd figyelembe.

[5] A BAH forrás esetén nem szűrtük az intézeti címeket a feldolgozás előtt, a végső becslésnél veszsük majd figyelembe.

Meglepő, hogy az ’old’ címek közel egynegyedén számíthatunk célsokaságbeli személyre.

6. táblázat a BÁH részminta elemszáma és becsült %-os megoszlás a cím jellege szerint

MEGHI	CELCSOP1	CELCSOP2	BÁH cím réteg	
			erv	old
		elemszám:	**1770**	**580**
12			**16.5**	**8.7**
21			**3.5**	**8.2**
22			**2.1**	**5.9**
23,50		1	**4.1**	**4.0**
23,50		2	**4.5**	**2.7**
23,50		3	**2.5**	**1.3**
23,50		4	**6.1**	**7.3**
23,50		5	**24.9**	**35.3**
24			**3.7**	**5.9**
31,41,42,43	1		**12.5**	**8.0**
31,41,42,43	2		**14.0**	**7.4**
31,41,42,43	3		**2.8**	**1.3**
31,41,42,43	4		**2.8**	**4.1**
célcsoportbeli címek aránya:			**43.0**	**24.1**

Az eredmények összehasonlítása

A becslések külön-külön is értelmezhetők, de érdemes egymás mellé rakni őket (lásd melléklet, 7. táblázat). A táblázatban az eredményeket a szerint rendeztük, hogy a teljes mintavételi keret mely részére vonatkoznak a becslések. A táblázatban a ’mind’ oszlopokban látható a teljes cenzus, KEKKH és BÁH részmintából kapott eredmény. A C, K és B oszlopok a keret azon részére vonatkozó becslések oszlopai, amik csak a cenzus, KEKKH és BÁH forrásból származó címeket tartalmazzák. A CK a cenzus-KEKKH közös címek keretére vonatkozó becslések: erre ezen a szinten kétféle eredményünk is van, hiszen a feldolgozás ezen szakaszában a három részmintán külön-külön végeztünk súlyozást. Az eredmények ennek megfelelően nem egyeznek meg, de meg kell jegyezni, hogy jelentős különbségek nem mutatkoznak. Hasonlóan értelmezhető a többi oszlop is.

- A számok alapján kijelenthető, hogy a legkisebb lefedettségi többlet és egyéb kerethiba a cenzus keretét terheli, itt a címek 45.1%-a célsokaságbeli. Viszonylag közel van hozzá a KEKKH keret a maga 41.6%-ával, míg a BÁH-keret a leggyengébb ebből a szempontból, a keret címeinek kevesebb mint harmadán volt találat.
- Hasonló sorrend állapítható meg, ha azokat a címeket nézzük, amik csak egyetlen forrásból származnak (C, K, B oszlopok).
- A másik két forráshoz képest különösen nagy a BÁH-keretben a nem létező és nem azonosítható címek aránya.
- Némileg meglepő, hogy annak ellenére, hogy a cenzus forrás a másik kettőnél jóval régebbi, azon lakások aránya, ahol jelenleg nincs vagy vélhetően nincs harmadik

országbeli (meghi=23,50 és celcsop=3,4 sorok), mindhárom forrásban közel azonos, ~52%.

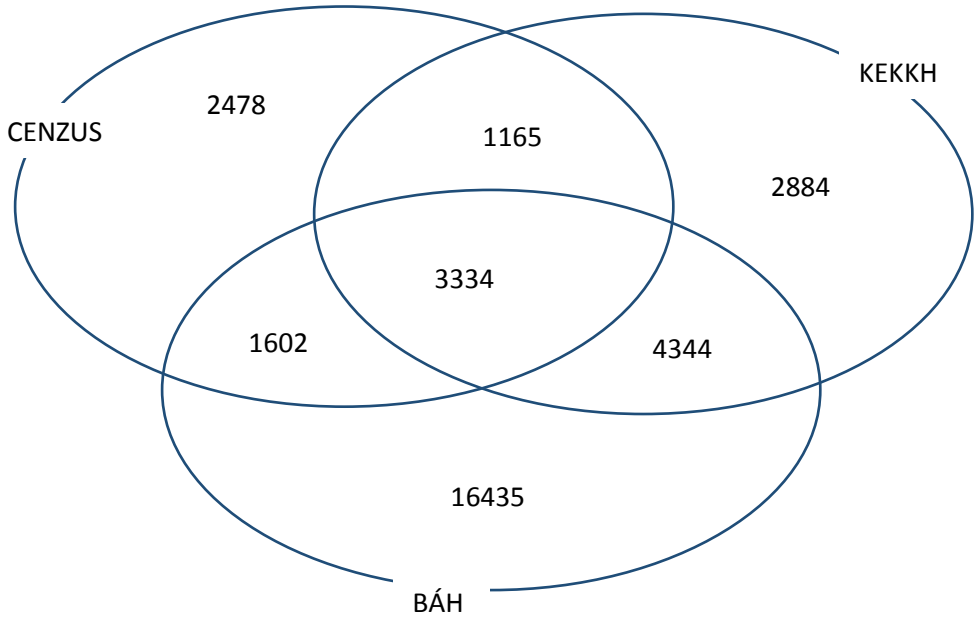
Összességében kijelenthető, hogy valójában egyik adatforrás sem megbízható abban a tekintetben, hogy a bennük lévő címek kevesebb, a BÁH esetében jóval kevesebb, mint fele juttat el a harmadik országbeliekhez. Ez persze csak akkor okoz torzítást a majdani becslésekben, ha a kerethibás címek mögött nincs Magyarországon élő harmadik országbeli személy.

Ami viszont ugyancsak torzítást okozhat egy migrációs felvételben, az a lefedettségi hiány, vagyis azoknak a címeknek a száma (aránya), amik hiányoznak az egyes keretekből. Mivel ez a felvétel egyszerre három keret használt, képet alkothatunk arról, hogy hány címen találtunk célcsoportbeli személyt az egyes kereteken kívül. Az elsődleges súlyokat felhasználva becsültük azon címek számát, amiken célcsoportbeli személyt találtunk: a keret forrása szerinti számokat a 3. ábrába írtuk bele.

Az eredmények értelmezése előtt mindenképpen meg kell jegyezni, hogy

- egyrészt a BÁH részmintából kapott becslések alapján valószínűnek tűnik, hogy a migrációs felvétel által használt mintavételi keret BÁH-os részében akár jelentősebb mértékű címismétlődés is maradhatott;
- másrészt, mivel a különböző források különböző módokon tárolták a címeket, ezek összekapcsolása nem lehetett maradéktalanul pontos, vagyis a végső keretben maradhattak címismétlődések, ami a lefedettségi hiány felülbecsléséhez vezethez.

3. ábra



A fentiek figyelembevételével, de azok ellenére is állítható, hogy mind a három keretet egy nem elhanyagolható mértékű lefedettségi hiány jellemez. A hiány nem azonos mértékű a három forrásnál, itt éppen fordított a sorrend, mint a lefedettségi többletnél: a cenzus forráson kívül találtuk a legtöbb célcsoportbeli címet, ezt követi a KEKKH, míg a BÁH keret

bizonyult ebből a szempontból a legmegbízhatóbbnak. Az elsődleges súlyok alapján tapasztalt hiány jelentős mértéke rávilágít

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 a mintán kialakított elsődleges súlyokkal jellemezni lehet az egyes források megbízhatóságát. Fontos kiemelni, hogy ez a jellemzés a címek keretére vonatkozik, az egyes (adminisztratív) adatforrásokban megjelenő személyek listájára, annak helyességére nem lehet következtetni.

A legfontosabb megállapítások.

(1) Már a tervezés, előkészítés szakaszában kiderült, hogy problémát okozhat a három adatforrásban eltérő módon és pontossággal tárol címek azonosítása. Ezen a téren a BÁH állományai tűntek a legkevésbé pontosnak. A mintából az elsődleges súlyokkal kapott bizonyos becslések arra utalnak, hogy a végső mintavételi keretben is maradhatott nem elhanyagolható mértékben címismétlődés.

(2) A meghiúsulási kérdőív válaszai alapján megállapítható, hogy mindhárom adatforrásból származó címkeret komoly többletet tartalmaz abban az értelemben, hogy a címek jelentős részén nem található harmadik országbeli személy. Ebből a felvételből nem derül ki, hogy egyes esetekben ennek mi az alapvető oka: lehet a címek pontatlansága, illetve lehet, hogy valódi lefedettségi többletről van szó, azaz az adatforrás fölöslegesen tartalmazza már a célszemélyt. Abból a szempontból, hogy a kiválasztott címek hány %-án talált a felvétel célcsoportbeli személyt, a cenzus forrás bizonyult a legmegbízhatóbbnak 45.1%-kal, őt a KEKKH forrása követte 41.6%-kal, a legkevésbé megbízhatónak pedig a BÁH forrás számított, 31.1%-kal.

(3) Mindhárom adatforrásra igaz, hogy a felvétel jelentős mértékben talált célcsoportbeli címeket az adtaforráson kívüli részen. Ebből a szempontból a sorrend fordított az előzőhöz képest: a cenzus szenvedi el a legnagyobb lefedettségi hiányt, míg a legkevésbé a BÁH érintett ebben. Fontos megjegyezni, hogy ezek a becslések meglehetősen érzékenyek a keretben található címismétlődésekre, ami a több forrásból származó címek összekapcsolásának bizonytalanságából adódik, ezért ezen a szinten itt nem is adunk számszerű jellemzést.

Tanulságok.

(4) A feldolgozás további szakaszában, a megvalósult minta súlyozásakor különös figyelmet kell fordítani a keretben maradót lehetséges címismétlődésekre, illetve torzító hatásának kiküszöbölésére.

(5) A legfontosabb tanulság talán annak az igénynek a megerősítése, hogy a közigazgatásban komoly szükség lenne az egységes címszabvány használatára adminisztratív nyilvántartásokban, regiszterekben.

Melléklet

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!**

7. táblázat *a 4-5-6. táblázatok együttese*

MEGHI	CELCSOP1	CELCSOP2	mind			C	K	B	CK		CB		KB		CKB		
			CENZUS	KEKKH	BÁH	CENZUS	KEKKH	BÁH	CENZUS	KEKKH	CENZUS	BÁH	KEKKH	BÁH	CENZUS	KEKKH	BÁH
elemszám:			1038	1626	2350	416	599	1230	105	106	187	199	586	586	330	335	335
12			20.1	18.4	11.6	13.3	14.0	9.3	24.4	21.8	13.9	12.2	15.0	14.5	31.1	30.4	29.9
21			1.2	2.9	6.4	1.8	4.1	7.8	0.8	0.7	0.0	0.4	3.2	3.1	1.4	1.4	1.3
22			1.0	1.1	4.5	1.0	1.7	5.6	0.0	0.0	1.6	1.4	0.7	0.6	1.0	1.0	1.0
23,50		1	4.6	3.6	4.0	6.1	3.6	4.0	1.7	1.4	7.0	5.3	4.9	4.9	2.2	1.9	2.1
23,50		2	5.7	4.2	3.4	5.5	3.7	2.9	4.7	4.3	7.5	6.5	4.4	4.2	5.0	4.9	4.9
23,50		3	3.2	2.1	1.7	4.5	1.9	1.5	1.6	1.5	4.2	3.8	2.8	2.7	1.4	1.4	1.4
23,50		4	8.0	6.6	6.9	10.0	7.8	7.2	9.9	7.7	7.6	6.7	6.0	5.9	5.2	5.3	5.0
23,50		5	25.0	29.6	31.4	28.8	39.0	33.8	19.8	17.0	25.9	22.0	27.9	27.1	21.5	21.0	20.9
24			0.5	2.9	5.1	0.7	4.2	5.9	0.0	0.7	0.0	2.8	2.8	2.7	0.8	1.5	1.5
31,41,42,43	1		13.4	11.7	9.7	11.7	6.5	8.2	18.2	22.9	12.0	14.4	12.8	13.9	14.7	14.6	14.8
31,41,42,43	2		11.6	11.6	9.8	9.6	8.2	8.4	13.7	15.5	15.1	16.3	13.7	14.5	11.3	12.3	12.8
31,41,42,43	3		2.9	2.3	1.8	4.1	1.4	1.8	4.2	5.3	0.4	0.4	2.4	2.3	2.6	2.7	2.7
31,41,42,43	4		2.7	3.0	3.6	2.8	3.9	3.6	1.0	1.3	4.8	7.9	3.3	3.5	1.8	1.7	1.7
célcsoportbeli címek aránya:			**45.1**	**41.6**	**31.1**	**34.6**	**28.7**	**25.9**	**56.3**	**60.2**	**40.9**	**42.9**	**41.5**	**42.9**	**57.1**	**57.2**	**57.5**