A STATISZTIKAI ADATBÁZISOK ÖSSZEKAPCSOLÁSÁNAK TAPASZTALATAI ÉS LEHETŐSÉGEI

JÓNÁS ISTVÁN – DR. NOVÁK ZOLTÁN

A tanulmány a Magyar Statisztikai Társaság éves konferenciáján (Balatonfüred, 2004. október 14–15.) elhangzott előadások szerkesztett változata, melyet a szerzők *Kovács Tibornak*, a Hivatal közelmúltban elhunyt tagjának ajánlanak, akinek hathatós segítsége, értékes szakmai észrevételei és javaslatai nagyban hozzájárultak a feladat végrehajtásához, illetve ezen írás megszületéséhez. A cikk két, időben egymáshoz közel eső, teljes körű összeírásból kialakított statisztikai adatbázis kifejezetten statisztikai célból történő összekapcsolásával, az ebből származó elemzési lehetőségekkel foglalkozik. Bemutatja ennek törvényi hátterét, az elemi adatok összekapcsolásának technikai megoldását, az ennek során felmerült cím-, valamint az ezeken található személyekkel kapcsolatos azonosítási problémákat.

TÁRGYSZÓ: Statisztikai adatbázis. Cenzus. Adatbázisok összekapcsolása.

Magyarországon, az 1999. évi XLVI. törvény alapján, 2000 áprilisában, a Központi Statisztikai Hivatal lebonyolította a 6. Általános Mezőgazdasági Összeírást (ÁMÖ), az 1999. évi CVIII. törvény alapján pedig, 2001 februárjában, a 14. hivatalos népszámlálást. A két összeírásból bőséges, időben egymáshoz közeli információhalmaz állt rendelkezésre. Aligha kell bizonyítani azt, hogy a két cenzus adatainak közös felhasználása, az információk együttes kezelése mennyire gazdagítja mindkét felvétel elemzési lehetőségeit, ugyanakkor a létrehozható új állomány mennyi új vizsgálat, értékelés, elemzés elvégzésére ad lehetőséget. A népszámlálás elsősorban a társadalomstatisztika számára biztosít alapadatokat, ugyanakkor a gazdaságstatisztikai elemzések elkészítéséhez is nélkülözhetetlenek, míg az általános mezőgazdasági összeírás a gazdaságstatisztikának egyik legjelentősebb felvétele, amelynek számos információja a társadalomstatisztikusok számára is kiemelt jelentőségű, például a népesedési folyamatok mozgatórugóinak mélyebb megismeréséhez.

A közös felhasználást, a praktikus szempontok mellett, fokozottan indokolta, hogy mindkét összeírás felvételi programjából eleve hiányoztak, vagy a felvételi program véglegesítése során – részben a felvételek időigénye, a kérdőívek terjedelme, részben pedig anyagi okok miatt – nagyrészt törlésre kerültek azok az információk, amelyek elsősorban a másik felvétel vizsgálati körébe tartoznak. Emiatt született például olyan döntés, hogy a

népszámlálás adatfelvételi programjában ne szerepeljenek a háztartások mezőgazdasági termeléssel való kapcsolatát vizsgáló olyan, a korábbi népszámlálásban egyébként hagyományosnak tekinthető, átfogó jellegű mezőgazdasági kérdések, amilyeneket sokkal részletesebben vizsgált a mezőgazdasági felvétel. Az elemzések szempontjából fontos kiemelni, hogy mindkét cenzus teljes körű volt, tehát elemi szinten álltak rendelkezésre az adatok, melyek így az ország bármely területi egységére, bármely választott csoportképző ismérv szerint aggregálhatók, összerakhatók, vizsgálhatók lettek. Természetesen ehhez elengedhetetlenül fontos napjaink fejlett számítástechnikai infrastruktúrája, amely a két felvétel adatainak gyakorlatilag korlátlan együttes kezelését lehetővé teszi.

Munkánk során a két cenzusból nyert adathalmazok segítségével első ízben kínálkozott lehetőség a hagyományosnak mondható, a megyei és régiós kötetekben már megjelent településsoros, településcsoportos bontásban közölt információk felhasználásával készíthető elemzések mellett, az adatfelvételek adatainak elemi szintű, kifejezetten statisztikai célú összekapcsolására.[1] Ez nem csak a két nagy hazai összeírás történetében újdonság. A nemzetközi gyakorlatban is úttörő módszertani megoldás. A munkálatok megkezdéséhez lökést adott az a nem éppen elhanyagolható szempont is, hogy még az ország Európai Unióba való belépése előtt lehetőség kínálkozott a magyar mezőgazdaság egyéni gazdaságainak demográfiai és lakáshelyzetének részletes, területi szintű bemutatására, amely a későbbiekben jó alapot nyújt a további vizsgálatokhoz.

A két cenzus adatai elemi szintű összekapcsolásának alapvető feltétele a mindkét felvételben hasonló tartalmú csatlakozási pontoknak és a fogalmi rendszer feltárása. Lehetőség szerint igyekeztünk minél pontosabban közelíteni egymáshoz az egymással teljesen meg nem feleltethető fogalmakat, valamint igyekeztünk meghatározni kompromisszumokkal lehetséges használatukat. A két felvételből létrehozható adatbázis kialakításához azonosítható, egymással megegyező, vagy egymáshoz közel álló elemeket kellett keresni. A két felvétel elemei azonosító információinak vizsgálata során a szakértők arra a megállapításra jutottak, hogy a legkisebb egység, amelynek a szintjén a csatlakozási felület biztosítható, a gazdaság, illetve a háztartás, azaz a mezőgazdasági felvételből az egyéni gazdaságok,[2] a népszámlálásból pedig a háztartások, azon belül is elsősorban az ún. lakásháztartások.[3] Egyetértetés született abban, hogy bár a két felvétel más-más halmazokra terjed ki, különböző fogalmaik bizonyos szinten ugyanazt a sokaságot tartalmazzák, közöttük nagy biztonsággal megállapítható az azonosság. Ugyanakkor ezen elemek olyan nagy számban fordulnak elő a felvételekben, hogy ennek köszönhetően a két nagy összeírás megfigyelési egységeinek ezen a szinten meglehetősen nagy a közös metszete, és viszonylag kicsi azoknak az egységeknek a száma, amelyek csak az egyik, vagy csak a másik megfigyelési körébe tartoznak.

Az adatok közös adatbázisba szervezésének feltétele a közös metszet egyedei azonosító adatainak leválogatása, az így létrehozott két adatállomány elemeinek egyértelmű

[1] Az összekapcsolás megoldásához kiindulási alapul szolgált *Laczka Sándorné* és *Czibulka Zoltán*: Az ÁMÖ 2000 és a népszámlálás adatainak együttes kezelése c., a KSH Népszámlálás az ezredfordulón 3. (Tanulmányok) című kötetben megjelent tanulmánya.

[2] Egyéni gazdaság összefoglaló névvel szerepelnek az egy lakásban élő, mezőgazdasági termeléssel egyéni gazdaként, önfoglalkoztatóként, egyéni gazdaságban segítő családtagként dolgozó, vagy a mezőgazdasági termékeket döntően saját fogyasztásra termelő családok rokoni vagy nem rokoni közösségek.

[3] Lakásháztartáson azokat a magánháztartásokat értjük, amikor a lakásban egyetlen háztartás tagjai élnek. A következőkben a háztartás fogalmát a lakásháztartás szinonimájaként használjuk.

azonosítása. A közös adatbázis további vizsgálatainak alapját azok az elemek (egyéni gazdaságok, illetve háztartások) képezhetik, amelyekhez mindkét felvétel azonosítója kapcsolódik. A két adatállomány elemeinek megfeleltetéséhez egy természetes azonosító, az összeírás helyének címe használható.

ADATVÉDELMI FELTÉTELEK

A technikai feltételek biztosítása mellett természetesen szem előtt kellett tartani a statisztikai és az adatvédelmi törvények rendelkezéseit, az együttes kezelés kizárólag statisztikai célból történhet.[4] A két felvétel adatainak összekapcsolásához a jogi hátteret a jelenleg hatályos 1993. évi XLVI. statisztikai törvény 21. §. /5/ bekezdése biztosítja:

„ /5/ A hivatalos statisztikai szolgálathoz tartozó szerv a saját statisztikai adatállományait statisztikai célra összehasonlíthatja. Személyes adatállományok összekapcsolása csak az adatgyűjtés meghatározott célját meg nem haladó mértékben, a cél eléréséhez szükséges ideig történhet. Az eredeti célt meghaladó adatkezelés új adatkezelésnek minősül."

A személyes adat fogalmát a személyes adatok védelméről és a közérdekű adatok nyilvánosságáról szóló 1992. évi LXIII. törvény a következőkben határozza meg:

„Személyes adat: a meghatározott természetes személlyel kapcsolatba hozható adat, az adatból levonható, az érintettre vonatkozó következtetés. A személyes adat az adatkezelés során megőrzi e minőségét, amíg kapcsolata az érintettel helyreállítható."

Mindezek alapján megállapítottuk, hogy az elemzések elvégzése céljából a vonatkozó jogszabályok lehetőséget biztosítanak az összekapcsoláshoz, vagyis a két cenzus adatállománya együttes kezelésének nincs jogi akadálya, hiszen az statisztikai célból történik, az összekapcsoláshoz felhasznált információ pedig személyes adatot nem tartalmaz. Az összekapcsoláshoz a mezőgazdasági összeírás állományából a gazdaság vezetőjének családi és utónevét felhasználni azért is felesleges volt, mert a név nemcsak a népszámlálási állományban nem szerepelt, hanem az összekapcsoláskor már megsemmisített népszámlálási kérdőíveken sem.

Az előbbiek tisztázása után az összekapcsoláshoz szükséges számítástechnikai programok elkészítésével egyidejűleg azt a célt tűztük ki, hogy a létrejött egyeztetési állomány alkalmassá váljék a két teljes körű összeírás anyagából olyan adatbázis létrehozására, amely mind a mezőgazdasági összeírás, mind a népszámlálás részletes adataiból vehessen át információkat. Az adatbázis tartalmát, az együttes vizsgálatra alkalmas információkat a vizsgálat céljának megfelelően kellett meghatározni. Első lépésben a szakemberek az egyéni gazdaságok demográfiai és lakáshelyzetét két csoportképző ismérv alapján javasolták elemezni,[5] mégpedig a gazdaságok termelési típusa (növénytermesztő, állattartó, vegyes gazdálkodást folytató), valamint a termelés célja (saját fogyasztásra, saját fogyasztásra és értékesítésre, értékesítésre termelő) szerinti csoportosításban. Az így képzett csoportokba tartozó egyéni gazdaságokban vizsgáltuk az ott élő népesség népszámláláskor megfigyelt nemét és korát, a 7 évesek és idősebbek legmagasabb iskolai végzett-

[4] A munkálatok megkezdése előtt kikértük a Hivatal Jogi és igazgatási osztályának szakvéleményét, amely szerint a vonatkozó jogszabályok lehetőséget biztosítanak az összekapcsoláshoz az elemzések elvégzése céljából.

[5] Az összekapcsolt állományok segítségével további vizsgálatokat tervezünk.

ségét, gazdasági aktivitását, a foglalkoztatottakat összevont foglalkozási főcsoport és nemzetgazdasági ág szerint, a háztartás összetételét, valamint a lakott lakások felszereltségét és komfortosságát. Az elemzés a vidéki agrárgazdaságokra vonatkozóan (tehát Budapest nélkül) tervezési és statisztikai régiónként készült, a régiókon belül pedig megyei és statisztikai kistérségenkénti részletezésben.[6]

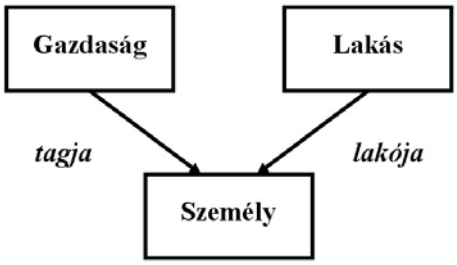
AZ ÁMÖ ÉS A NÉPSZÁMLÁLÁS ELEMI ADATAINAK ÖSSZEKAPCSOLÁSA

Egy adatfelvétel elemi adatain a megfigyelt sokaság elemeinek (az egyes előfordulásoknak) az adatait értjük.[7] Az adatok összekapcsolása az egyes előfordulások közötti elemi kapcsolatok alapján történik (a létező kapcsolatok közül kell kiválasztani az összekapcsolás céljának megfelelő kapcsolattípust). Az összekapcsolás egy olyan elrendezés, amelynek felhasználásával az elemi kapcsolatok szerint összetartozó előfordulások adatai között műveletek végezhetők.

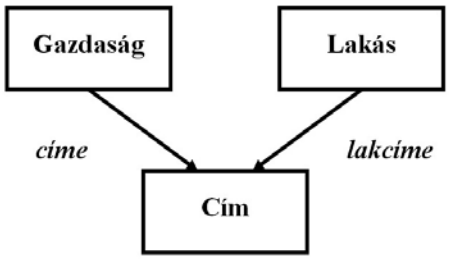
Ebben az esetben két adatbázis összekapcsolásának célja a mezőgazdasági összeírás egyéni gazdaság és a népszámlálás háztartás fogalma közötti összefüggések vizsgálata. Az összefüggések részben a személyek összetételében mutatkozó átfedés, részben a mindkét fogalomhoz kapcsolódó tulajdon azonossága alapján vizsgálhatók.

A megfigyelt sokaságok az ÁMÖ esetében az egyéni gazdaság, a népszámlálás esetében a lakás és a személy. Az ÁMÖ rögzített néhány információt a gazdaság tevékenységében nem alkalmazottként részt vevő személyekről is (például nem, életkor).[8] Ennek alapján a sokaságok elemei között a gazdaság tagja, illetve a lakás lakója nevű kapcsolat feleltethető meg. (Lásd az 1. ábrát). A kapcsolódó tulajdon meghatározásának egyetlen támpontja most a közös cím. (Lásd a 2. ábrát.). A gazdaság és a háztartás kapcsolata a lakásban lakó személyek adatai alapján vizsgálható.

1. ábra. Kapcsolat az azonos személyek szerint



2. ábra. Kapcsolat a közös cím alapján



Az adatkapcsolatok megadásához az adatbázis-elmélet az egyedi azonosító fogalmát használja. Az egyedi azonosító olyan (esetleg összetett) ismérv (tulajdonságtípus), amely

[6] A kiadványsorozat „Az egyéni agrárgazdaságok és népességük” címmel 2005 májusában jelent meg nyomtatásban és CD-n.

[7] Az elemi adatok lényegében az összeírás során keletkezett adatok. Nem elemi adatok a különféle ismérvek szerinti csoportosítások aggregátumai (például a településsoros adatok).

[8] Részletesen lásd. ÁMÖ-kérdőív 23. pont.

a sokaság egy elemét egyértelműen azonosítja. A kapcsolatot az egyes előfordulások egyedi azonosítóiból kialakított párok reprezentálják.

Mindkét adatfelvétel önállóan képzett összetett ismérveket használt az egyes előfordulások azonosítására (ezekben csak a településazonosító közös, a személyi szám adatvédelmi megfontolásokból nem szerepelt az azonosítók között). Az elemi kapcsolatok leírására egyrészt az egyéni gazdaság tagja, másrészt a lakás lakója egyedi azonosítóiból képzett párok használhatók. Ezek a következők:

település – ÁMÖ körzet – gazdaságsorszám – személy sorszáma a gazdaságban ↔
település – számlálókörzet – címsorszám – személy sorszáma a lakásban

A gazdaság tagja egyedi azonosítójának része a gazdaság egyedi azonosítója, így a *tagja* kapcsolatot az ismérvek közös része meghatározza. Hasonlóan igaz ez a lakás és a lakás lakója vonatkozásában. A személyek azonosságát leíró táblázat (a fenti módon kialakított párok) előállítása önálló logikai feladatként jelent meg, mivel a két adatfelvétel különböző egyedi azonosítókat használt. A kapcsolatok meghatározása arra a feltételezésre épült, hogy az ÁMÖ-beli egyéni gazdasághoz és a népszámlálásbeli háztartáshoz tartozó személyek a közös cím, és néhány demográfiai jellemző alapján azonosíthatók.[9]

A megvalósítás a következő lépésekben történt.

A két címállományt egy előzetes számítógépes programfutással párosítottuk. Közel egymillió gazdaság címeinek 90 százalékát sikerült így azonosítani.

A „páratlan” ÁMÖ-címeket egy interaktív program segítségével azonosítottuk. Körülbelül tíz hónap élőmunka befektetésével 97 százalékra javult a „találati arány”.

A személyek azonosítását egy utólagos programfutással végeztük. A gazdaságtagok 78 százalékát sikerült megtalálni.

A feltárt logikai kapcsolatok felhasználásával az adatállományok fizikai összekapcsolása már kidolgozott informatikai technológiával (például adatbázis-lekérdezés) valósítható meg.

A CÍMAZONOSÍTÁS PROBLÉMÁI

A cím a KSH minden statisztikai adatfelvételénél szerepet kap, ez a fogalom tehát generálisan összekapcsolja a KSH adatgyűjtéseit. A feldolgozások néhány regiszter köré csoportosíthatók, ezek közül a vizsgált két terület (ÁMÖ, népszámlálás) mellett a Gazdálkodó Szervezetek Regisztere (GSZR), illetve a KISkereskedelmi REGiszter

[9] Számolni kell a két adatfelvétel közötti változásokkal (költözés, elhalálozás, házszámváltozás), valamint azzal, hogy ez a stratégia nem biztosítja azoknak a személyeknek az azonosítását, akik a gazdaság címétől eltérő című lakásban laknak. A kapcsolatok pontos meghatározására a személyi szám ismeretében lenne lehetőség.

(KISREG) a jelentősebbek.[10] Az ÁMÖ-adatok – népszámlálási adatok kapcsolatának felderítéséhez hasonlóan egyéb adatfelvételek összekapcsolása is felvethető. Ahol a kapcsolat azonosítása a címek alapján lehetséges, feltehetőleg ugyanazon (illetve nagyon hasonló) technika alkalmazható.

A címek megfeleltetését az eltérő írásmód, a pontatlan címzés (például csak a házszám szerepel), valamint az adatfelvételek között eltelt idő alatti címmódosítások (utcák átnevezése, házszám-változások) nehezítik.

A számítógépes párosítás könnyen elvégezhető, ha a címek pontosak. Egy egyszerű program által végzett összehasonlítást úgy képzelhetünk el, mintha egy-egy átlátszó papírra írnánk a két címet, majd a papírokat egymásra téve megnéznénk, hogy fedésbe hozható-e a két szöveg. Ha tökéletes a fedés, akkor a két cím megegyezik. Egy ilyen szigorú feltételnek persze nem felelnek meg például a KOSSUTH UTCA – Kossuth utca, Petőfi Sándor utca – Petőfi utca, Fő utca 1./a – Fő utca 1/a párok. Az „intelligens", írástudó személy ezek fölött átsiklik, annyira természetesnek veszi az azonosságot.

Két címállomány vizsgálatánál tehát problémát okozhat, hogy az állományok eltérő írásmóddal tartalmazzák a címeket. Ez akár ugyanazon állományban is előfordulhat. Például a GSZR-ben a pécsi Bajcsy-Zsilinszky utca többféle változatban is előfordult.

3. ábra. Eltérő írásmód előfordulása ugyanazon utca címeinél

Bajcsy Sz. utca 4.
BAJCSY Z E UT 33 ZOLYOM U 2
BAJCSY ZS. U. 4. IX.27.
Bajcsy Zsilinszky u 33
Bajcsy Zsilinszky utca 23.1.em.2.
BAJCSY-ZS. E. UTCA 25
BAJCSY-ZSILINSZKI ENDRE U 25
Bajcsy-Zsilinszky utca 5.
Bajcs-Zs u 11. "Exclusive Gentlemans"
Bajzsy-Zs.u.14-16.

Az agyunk egy szempillantás alatt azonosítja, hogy az itt látható címek mindegyike a Bajcsy-Zsilinszky utcához tartozik, még a betűhibák esetében is. Ezért a gépi azonosítást is a „majdnem egyezik" stratégiával oldottuk meg. Fontos feltétele a sikeres azonosításnak, hogy legyen egy „normának" tekinthető címállományunk, mint például a népszámlálás címállománya.[11]

[10] Számos előny származna abból, ha a címadatok egy közös címállományban lennének, minden feldolgozás innen venné ezt az információt. Erre a szerepkörre a népszámlálás címregisztere lehet esélyes (teljes körű, az önkormányzatok által megállapított hivatalos címjegyzék). Egy ilyen címállomány karbantartása azonban a KSH-n belül aligha lehet sikeres, ezt az önkormányzatokkal (okmányirodákkal) közösen lehetne megoldani.

[11] A GSZR-utcanevek esetében egy olyan kiértékelő függvényt használtunk, amelyik az egyes címekben egyező szövegfoszlányokat keres, ezek hossza és elhelyezkedése szerint minősíti a hasonlóságot. Korábban (a 90-es évek elején) már foglalkoztunk – viszonylag kevés sikerrel – hasonló címek normalizálásával. Ez a feladat azt jelenti, hogy például a 3. ábrában szereplő címek közül meg kell találni a legjobbat, a „normát". Lényegesen egyszerűbb feladat egy meglévő „normához" (például a népszámlálási utcajegyzékhez) való hasonlóság vizsgálata. A természetes intelligencia számára a normát például az helyettesítheti, ha történelmi tanulmányai alapján ismeri Bajcsy-Zsilinszki Endre nevét.

Megnehezíti a címek azonosítását, ha egyes ingatlanok nincsenek megfelelő fizikai azonosítóval ellátva. Ha egy többlakásos épületben nincsenek sorszámozva a lakások, vagy egy házszám alatt több lakóépület is található, akkor a számlálóbiztos egy fiktív azonosítót regisztrálhatott. A 4. ábra felső részén három gazdaság azonosítóit látjuk, mindegyiket a KÖVES-FÖLDI ÚT 9. alatt találta az ÁMÖ. Hogy a valóságban van-e A és B épület, és az ajtókon van-e sorszám, azt nem tudjuk, mindenesetre a népszámlálás itt 9 különböző címet regisztrált.

A címek azonosításához ilyenkor további információ szükséges. Ilyen lehet, ha van valamilyen adatunk az ott lakó személyekről. Az összeírások során számos olyan információ keletkezik, amely nem kerülhet a túlnyomórészt eldöntendő (zárt) kérdéseket tartalmazó kérdőívekre.

4. ábra. Pontatlan címzés – a gazdaság címe csak a házszámot tartalmazza

19415 009 169	KÖVES-FÖLDI ÚT 9	61F 58N
19415 009 170	KÖVES-FÖLDI ÚT 9	57F 52N
19415 009 171	KÖVES-FÖLDI ÚT 9	45F 44N 23F 21F 22N

KÖVES-FÖLDI ÚT 9 A	
KÖVES-FÖLDI ÚT 9 B	57F 52N 27F
KÖVES-FÖLDI ÚT 9 A 2	61F 58N 39F
KÖVES-FÖLDI ÚT 9 A 3	78N
KÖVES-FÖLDI ÚT 9 B 1	48F
KÖVES-FÖLDI ÚT 9 B 2	34F 33N 11F 7N
KÖVES-FÖLDI ÚT 9 B 3	25F
KÖVES-FÖLDI ÚT 9 B 4	38F
KÖVES-FÖLDI ÚT 9 B 5	45F 44N 22F 23N 24F 4N

A két adatfelvétel között eltelt időben megváltozhatott az utcák elnevezése, esetleg a házszámozás is. A népszámlálás általában gerjesztője ennek a folyamatnak: rákényszeríti az önkormányzatokat a rendrakásra. A számozás megváltozása miatt előfordulhat olyan eset, hogy a két adatfelvétel formailag teljesen megegyező címe nem ugyanarra az ingatlanra utal.

A címváltozások kellemetlen hatását egy „örökös” egyedi azonosító segítségével lehetne kiküszöbölni. A régebbi azonosítók közül az általánosan használt, évtizedek óta változatlan helyrajzi szám a legalkalmasabb az azonosításra, de a helyrajzi szám a KSH címállományaiban csak elvétve szerepel, így általános kapcsolóelemként nem jöhet szóba.[12] A térinformatikai megoldások a jövőben tökéletes azonosítást tesznek lehetővé.

Az ÁMÖ és a népszámlálás címállománya hasonló szerkezetben tartalmazza a címeket: település, közterületnév, közterületjelleg, házszám, épület, lépcsőház, emelet, ajtó.

Szerencsére itt nyoma sincs a GSZR-t jellemző sokszínűségnek. A szigorú egyezésnek ugyan egyetlen ÁMÖ-cím sem felelne meg, de néhány egyszerű korrekció után a cí-

[12] Az agrárstatisztikai rendszer használ helyrajzi számokat (például az ültetvényösszeírásoknál).

mek jelentős része azonosítható. A kisbetű–nagybetű eltéréseket, az előnullázott házszámokat, a felesleges pontokat és szóközöket meg kellett szüntetni, valamint a hosszú és rövid ékezetes magyar betűket kellett azonos formára hozni.

5. ábra. A népszámlálás előtti „rendrakás" a házszámok megváltozásával járhat

Az előzetes párosítás során a címek 90-92 százalékát sikerült számítógéppel azonosítani.

1. tábla

A címek előzetes azonosítása

Terület	A gazdaságok száma (darab)	A cím teljesen azonosítható	Csak a házszám található meg	Nincs egyezés
		százalék		
Vidék összesen	954 110	88	4	8
Ebből:				
városok	295 593	82	8	10
községek	658 517	90	2	8

Az átlagosan magas egyezési arány településenként igen eltérő mértékű volt. A párosítást követő interaktív címazonosításnak elsősorban az volt a feladata, hogy a hiányosan azonosított területek megfeleltetését javítsa. A kezelő itt egyedi információkat is felhasználhatott (például emlékezhetett rá, hogy valójában hol történt az összeírás, átvezethette az utcanévváltozásokat stb.). Emellett azt is felhasználtuk, hogy váratlan

helyzetekben a „természetes intelligencia” hatékonyabb megoldást talál a számítógépnél. A program itt segédmunkát végez: egy gombnyomásra leválogatja az utca összes címét, megkeresi a házszámot, az utcanévváltozások fogadása után újra elvégzi a házszámok azonosítását.[13]

AZ EGYÉNI GAZDASÁGOKNAK ÉS A GAZDASÁG TAGJAINAK AZONOSÍTÁSA

Az interaktív címazonosítás végére a címek 97 százalékát sikerült megtalálni. Ezután azt vizsgáltuk, hogy a címeken azonosíthatók-e a személyek. A gazdaság azonosítását akkor tekintettük sikeresnek, ha a gazdasághoz tartozó személyek jelentős része azonosítható volt.

A megtalált személyek alapján a kapcsolatot így jellemezhetjük:

– megegyező szerkezet: a gazdaság minden tagját azonosítottuk, azok megegyeznek a népszámlálás során regisztrált felnőtt korú személyekkel,

– beépülő szerkezet: a gazdaság tagjait megtaláltuk, de további felnőtt korú személyek találhatók ezen a címen,[14]

– hiányos/külső szerkezet: a személyeket csak részben sikerült megtalálni.

Az ÁMÖ csupán néhány információt rögzített a gazdaság tevékenységében nem alkalmazottként részt vevő személyekről. Ezen információkat a népszámlálás során felvett hasonló tartalmú adatokkal összevetve lehetőségünk nyílt a gazdaság tagjainak azonosítására. A személyeket jellemző adatok közül a nem, életkor, valamint legmagasabb iskolai végzettség mutatkozott leginkább alkalmasnak az azonosításra.

Az egyes tulajdonságok azonosításra való felhasználásához néhány megjegyzést fűzünk. A személy neme alapján csak az egyezés fogadható el. Az ÁMÖ esetében a személy életkora lett regisztrálva, a népszámlálásnál a születési ideje. Az előbbi valószínűleg pontatlanabb információ, a 60 éves életkor utalhat betöltött 60 évre, de egy majdnem 60 éves is válaszolhatott így a kérdésre. Figyelembe véve azt is, hogy az ÁMÖ 10 hónappal megelőzte a népszámlálást, az életkorok néhány éves eltérése is elfogadható. Az iskolai végzettséget az ÁMÖ a nincs, alapfok, középfok, felsőfok megkülönböztetésével jellemezte, a népszámlálás részletesebb osztályozása ennek megfeleltethető. Amennyiben az egyéb jellemzők alapján többféle választási lehetőség is kínálkozott a keresett személy kiválasztására, ezt a jellemzőt is felhasználtuk az azonosításhoz, a szomszédos kategóriát is egyezőnek tekintve.

A személyek azonosítását az említett tulajdonságok egyezésére épülő kiértékelő táblázat segítségével végeztük el. A 6. ábrán bemutatott példában a gazdasághoz egy 47 éves férfi, és egy 44 éves nő tartozott. A gazdaságtagok adatait rendre összehasonlítottuk a négy népszámlálási személy adataival. Az összes lehetséges variációt kiértékelve, a legjobb eredményt adó változatot választottuk. Az ábrán látható választáshoz 20+19 = 39

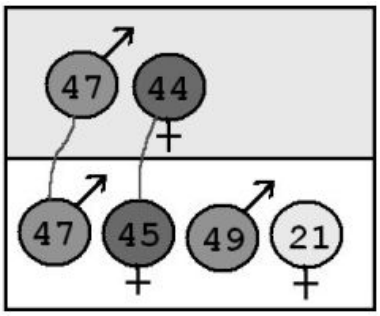
[13] A program tesztváltozatát felhasználtuk a „Nagyvárosok belső tagozódása” c. munkához. Itt a GSZR-, illetve a KISREG-címek népszámlálási címekkel való kapcsolatát használtuk a cégek városrész szerinti besorolásához. Öt megyeszékhely – Pécs, Szeged, Kaposvár, Békéscsaba és Nyíregyháza – feldolgozásának tapasztalatai szerint a problémás címek azonosításához címenként átlagosan 1 perc élőmunkára volt szükség. Ez az információ segítette az ÁMÖ – népszámlálás feldolgozás élőmunkaigényének becslését.

[14] Az ÁMÖ-ben csak a mezőgazdasági munkát végző családtagok szerepeltek.

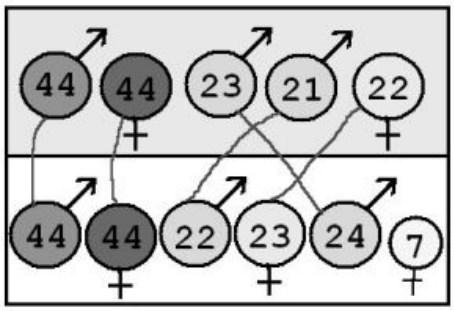
pont tartozik, a lehetséges $4 \times 3 = 12$ variáció közül ez a választás adja a legmagasabb pontszámot.

neme egyezik	10 pont
életkora egyezik	10 pont
+– 1 év eltérés	9 pont
+ 2 év eltérés	7 pont
+ 3 év eltérés	4 pont
iskolai végzettsége egyezik vagy hasonló	+1 pont

6. ábra. Beépülő szerkezetű kapcsolat

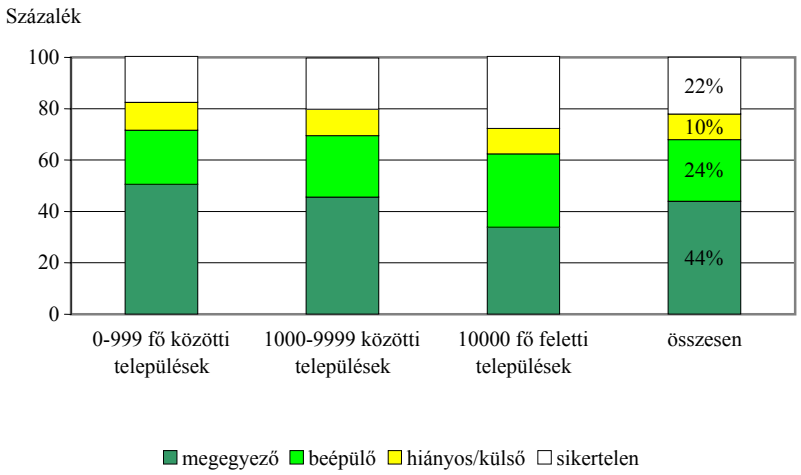


7. ábra. Megegyező szerkezetű kapcsolat



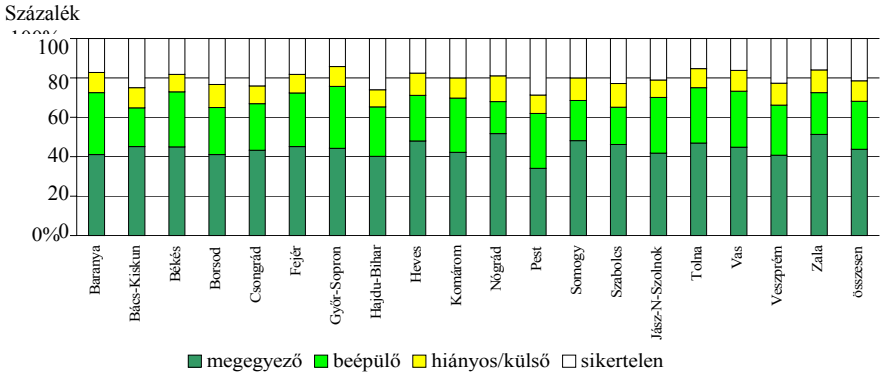
A 7. ábrán bemutatott példában öt gazdaságtagot kell azonosítani 6 népszámlálási személy közül, így a variációk száma jóval több: 6×5×4×3×2 = 720. Itt a lakásban lakó felnőttkorú személyek a gazdaságnak is tagjai, az azonosítás megegyező szerkezetet talált.

8. ábra. A kapcsolattípusok megoszlása a települések népességnagyság-csoportjai szerint



A személyek azonosítását úgy végeztük, hogy a gazdaság minden tagja megtalálja „a maga párját”. Ezzel lényegében kísérletet tettünk a gazdaság azonosítására. A gazdaságok 44 százalékában a lakásban lakó felnőttek és a gazdaságtagok ugyanazok voltak, 24 százalék esetében további felnőtteket is regisztrált a népszámlálás. Az esetek 10 százalékában nem sikerült minden gazdaságtagot azonosítani, 22 százalék esetében sikertelennek bizonyult az azonosítás. A hiány okaként a halálozás, az elköltözés, a címváltozások, az adatfelvételi és azonosítási hibák említhetők meg.

9. ábra. A kapcsolattípusok megoszlása megyénként



Az alkalmazott eljárás az 1. ábrán bemutatott kapcsolatok részleges feltárására volt alkalmas (a gazdaság címétől eltérő címen lakó személyeket akkor sem találtuk volna meg, ha a két adatfelvétel között nem történik semmiféle cím-, illetve demográfiai változás). Csak egy közös egyedi azonosító (mint például a személyi szám) tenné lehetővé a kapcsolatok teljes feltárását.

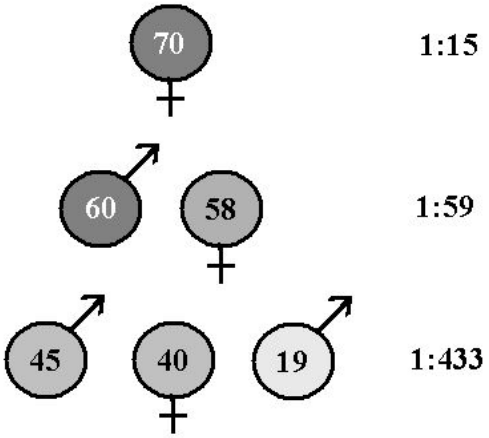
Ennek hiányában a személyek azonosításához a cím és néhány demográfiai jellemző mellett a csoporthoz tartozás tulajdonságát használtuk fel. A gazdaságok azonosítását ugyancsak a cím és a személyi összetétel szerkezete alapján végeztük el.

A címek 97 százalékos formai egyeztetése után csak 78 százalékban bizonyult sikeresnek a gazdaságok azonosítása. A különbség részben valódi változásokból adódik (költözés, halálozás), részben a címek azonosítóinak (utcanév, házszám) változásából. Ez a bizonytalanság összhangban van azzal, hogy a cím nem olyan erős tulajdonsága egy személynek, mint például a nem és az életkor.

Hasonlóan jellemző tulajdonsága egy gazdaságnak a személyi összetétel szerkezete.

Ennek igazolására Baranya megyei adatokkal végeztünk kísérletet. A gazdaságtagokat nemcsak a kijelölt címen, hanem a település összes címén is összevetettük a népszámlálás által regisztrált személyekkel. A 10. ábrán látható, hogy az egy-, két-, illetve háromszemélyes gazdaságok személyi összetétele átlagosan milyen gyakorisággal fordul elő a településen. A kijelölt címen így rendre 6,7, 1,7, illetve 0,2 százalék a tévedés valószínűsége. A 4. ábrán bemutatott pontatlan címzés a személyi összetétel ismeretében számítógéppel is megbízhatóan korrigálható.

10. ábra. Az egy-, két- és háromszemélyes gazdaság-szerkezet előfordulási gyakorisága



AZ EGYÉNI AGRÁRGAZDASÁGOKBAN ÉLŐ NÉPESSÉG NÉHÁNY FŐBB DEMOGRÁFIAI JELLEMZŐJE

A 2001. évi népszámlálás 3 millió 863 ezer háztartást írt össze, ezt megelőzően a 2000. évi ÁMÖ alkalmával pedig 2,1 millió háztartást kerestek fel a számlálóbiztosok. A budapesti 770 ezer háztartásból mindössze 4400 volt az ÁMÖ-nek is adatszolgáltatója. Emiatt, a két cenzus adatszolgáltatói körének összevetése alapján, az adatok összekapcsolása, elemzése és értékelése során, Budapest eredményeitől eltekintettünk. Az adatok értékelése így a vidéki 3 millió 93 ezer háztartásra vonatkozik, amelyeknek mintegy kétharmadát vizsgálta az ÁMÖ, ugyanis a városok belső övezeteiben élők közül csak az előre összeállított címlistán szereplőket kellett felkeresniük a számlálóbiztosoknak. A megfigyelt háztartásokból mintegy 960 ezer bizonyult gazdaságméretet elérő adatszolgáltatónak,[15] további 835 ezer pedig a meghatározott gazdaságküszöbnél kisebb méretben foglalkozott mezőgazdasággal, 300 ezer viszont egyáltalán nem végzett mezőgazdasági tevékenységet. A közel 960 ezer egyéni gazdaságból, az előbbiekben ismertetett módszerek segítségével, 784 ezer (82 százalékot) sikerült népszámlálási adatokkal összekapcsolni és így megvizsgálni. Ez az arányszám az egyes megyék, sőt a statisztikai kistérségek között is csak nagyon kis mértékben szóródott, ezáltal a részletesebb elemzésre is jó lehetőség kínálkozott. A továbbiakban ezen lehetőségek közül a tervezési, statisztikai régiókra (továbbiakban régiókra) vonatkozóan villantunk fel röviden néhányat.[16]

A korábban említett közel 960 ezer egyéni gazdaságban 2001-ben több mint 2,6 millió ember élt, ami a vidéken élő teljes népesség közel egyharmada. Különösen ma-

[15] Gazdaságnak a felvétel során azt a háztartást kellett minősíteni, melynek mezőgazdasági tevékenysége az előírt – piaci értéke alapján számottevő – alsó küszöbértéket elérte, illetve meghaladta.

[16] E témában részletesebb elemzés az áprilisban megjelent régiós kiadványok mellett a *Területi Statisztika* c. folyóirat 2005. évi júliusi számában jelenik meg.

gas volt ez az arány az Alföldön, de a Dél-Dunántúlon is meghaladta a 30 százalékot, a legalacsonyabb pedig Pest megyében (15,9%) volt. Nyugat-Dunántúlon és Észak-Magyarországon a teljes népesség egynegyede, Közép-Dunántúlon pedig egyötöde élt egyéni agrárgazdaságokban.[17] A nemenkénti megoszlást illetőn vidéken nem volt nagyobb különbség a teljes népesség és az egyéni gazdaságokban élők között, az utóbbiakban a férfiak aránya egy százalékponttal magasabb (49, illetve 48 százalék), a nőké ugyanennyivel alacsonyabb (51, illetve 52 százalék). A kormegoszlás tekintetében már jelentősebb eltérések tapasztalhatók. Az egyéni gazdaságokhoz tartozók 23 százaléka töltötte be a 60. életévét, 3 százalékponttal többen, mint ami a vidéki teljes népességre jellemző, ugyanakkor a fiatal, 40 év alatti korosztály aránya 5 százalékponttal elmaradt attól (47százalék az 52 százalékkal szemben). Ez nyilvánvalóan összefügg azzal, hogy az egyéni gazdaságokhoz tartozók többsége községi lakos, akiknek körében közismerten az átlagosnál magasabb az időskorúak aránya. A korösszetételben – régiónként vizsgálva –, ha nem is jelentős, de kisebb eltérések tapasztalhatók. A gyermekkorúak aránya Nyugat-Dunántúlon a legkisebb, Észak-Alföldön pedig a legmagasabb, a 60. életév felettieké ugyanakkor Pest megye mellett az utóbbiban a legalacsonyabb, Nyugat-Dunántúlon és Észak-Magyarországon a legnagyobb.

2. tábla

Az egyéni gazdaságokban élők korcsoportonkénti megoszlása a régiókban, 2001
(Százalék)

Régió	A 14 és fiatalabb	15–39	40–49	60 és idősebb	*Összesen*
	éves népesség aránya				
Pest megye	15	32	31	22	*100*
Közép-Dunántúl	15	32	30	23	*100*
Nyugat-Dunántúl	14	31	30	25	*100*
Dél-Dunántúl	16	31	30	24	*100*
Észak-Magyarország	15	30	30	25	*100*
Észak-Alföld	17	32	29	22	*100*
Dél-Alföld	16	30	31	24	*100*
Összesen	16	31	30	23	*100*
Vidék összesen	17	35	28	20	*100*

Megjegyzés. Itt és a következő táblákban közölt megoszlási viszonyszámok kerekítése egyedileg történt, ezért a részadatok összege nem mindig egyezik pontosan az összesített adatok kerekített értékeivel.

Az egyéni gazdaságokban és a vidéken élő teljes népesség iskolai végzettsége között is jelentős eltérés mutatkozott. A vidéken élő 7 éves és idősebb népességen belül az agrárnépességhez képest 5 százalékponttal magasabb a középiskolai végzettséggel, több

[17] Magyarországon a 2000-ben regisztrált egyéni agrárgazdaságok száma 436 ezerrel volt kevesebb, mint az ezt megelőzően, 1991-ben összeírt. A gazdaságok számának csökkenése mellett ugyanebben az időszakban ez a nagyon kívánatos koncentrációs folyamat is érzékelhető volt, amely remélhetőleg a jövőben tovább folytatódik. Egy évtized alatt az egy hektárnál kisebb területet használók száma (amely még 2000-ben is meghaladta az 580 ezret) a felére csökkent, ugyanekkor az ennél nagyobb területtel rendelkezők száma 139 ezerről 283 ezerre, ezen belül ezerről 51 ezerre gyarapodott a 10 hektár felettieké.

mint 3 százalékponttal pedig az egyetemet és főiskolát végzettek aránya. Ugyanakkor az utóbbiak közül valamivel meghaladta (2 százalékpont) a 8. általános iskolai osztálynál alacsonyabb, jelentősebben (6 százalékpont) az általános iskola 8. évfolyamát végzettek aránya a vidéki népesség hasonló arányszámát. Régiónként vizsgálva, az egyéni gazdaságokban élők iskolai végzettségét, eléggé változatos kép rajzolódott ki. Az országos átlagtól az egyes régiókban jelentősebb eltérések a 8. általános osztálynál alacsonyabb végzettséggel, valamint a középiskolai érettségivel rendelkezők aránya tekintetében jelentkeztek, az általános iskolai, továbbá az egyetemi és főiskolai végzettséggel rendelkezőknél viszont sokkal kiegyenlítettebb volt a kép. A 8. általános iskolai osztálynál kevesebbel rendelkezők aránya az Észak-Alföldön volt a legmagasabb és a Nyugat-Dunántúlon a legalacsonyabb, amely az eltérő korösszetétellel lehet összefüggésben. Az érettségizettek aránya Pest megyében volt a legnagyobb, de alig maradt el tőle Nyugat-Dunántúl és Észak-Magyarország.

3. tábla

Az egyéni gazdaságokban élők
legmagasabb iskolai végzettség szerinti megoszlása a régiókban, 2001
(Százalék)

Régió	A 7 éves és idősebb népességből				*Összesen*
	a 8. általános iskolai osztálynál kevesebbet	az általános iskola 8. osztályát	középiskolai érettségivel	egyetemi, főiskolai végzettséggel	
	végzettek		rendelkezők		
	aránya				
Pest megye	21	55	19	5	*100*
Közép-Dunántúl	21	58	17	4	*100*
Nyugat-Dunántúl	20	58	18	5	*100*
Dél-Dunántúl	24	58	15	4	*100*
Észak-Magyarország	23	55	18	5	*100*
Észak-Alföld	26	55	15	4	*100*
Dél-Alföld	23	57	15	4	*100*
Összesen	23	56	16	5	*100*
Vidék összesen	21	50	21	8	*100*

Az egyéni gazdaságokhoz tartozók gazdasági aktivitása jelentősen eltért a teljes vidéki népességétől, körükben alacsonyabb volt a foglalkoztatottak és – már az eltérő korstruktúrából adódóan is – magasabb az inaktív keresők aránya. Szembetűnő, hogy a munkanélküliek aránya tekintetében nem volt tapasztalható eltérés, az eltartottaké – valószínűleg szintén az eltérő korösszetétel következményeként – viszont vidéki átlagban magasabb, mint az agrárgazdaságokban. Az ország egyes térségei, egyéni gazdaságaiban élők közül Nyugat-Dunántúlon volt a legmagasabb a foglalkoztatottság, míg az Észak-Alföldön a legalacsonyabb. Ez utóbbiban a munkanélküliek aránya megközelítette a 6 százalékot. Az eltartottak arányának két szélső értéke szintén e két régióban volt megfigyelhető.

Figyelmet érdemel, hogy az inaktív keresők aránya az agrárgazdaságokban 4 százalékponttal meghaladta a foglalkoztatottakét, a vidéki népesség egészében viszont 2 százalékponttal elmaradt a foglalkoztatottakétól. A hét régió közül az egyéni gazdaságokban élőkön belül csupán a Közép-Dunántúlon és Nyugat-Dunántúlon volt magasabb a foglalkoztatottak aránya, Pest megyében azonos volt, a további négy régióban viszont az inaktív keresők aránya meghaladta a foglalkoztatottakét, sőt a különbség Észak-Magyarországon és Észak-Alföldön elérte a 10 százalékpontot.

4. tábla

Az egyéni gazdaságokban élők gazdasági aktivitás szerinti megoszlása a régiókban, 2001
(Százalék)

Régió	A foglalkoztatottak	A munkanélküliek	Az inaktív keresők	Az eltartottak	*Összesen*
	aránya				
Pest megye	36	3	36	25	*100*
Közép-Dunántúl	38	3	34	25	*100*
Nyugat-Dunántúl	40	2	35	23	*100*
Dél-Dunántúl	32	5	39	25	*100*
Észak-Magyarország	30	5	40	25	*100*
Észak-Alföld	28	6	38	28	*100*
Dél-Alföld	33	4	37	26	*100*
Összesen	33	4	37	26	*100*
Vidék összesen	35	4	33	28	*100*

Az egyéni gazdaságokban élő foglalkoztatottak közül 2001-ben értelemszerűen a vidéki átlagnál magasabb volt a mezőgazdasági foglalkozásúak, illetve a mezőgazdaságban, erdőgazdálkodásban dolgozók aránya, de még ebben a körben is meghatározó volt a más foglalkozási főcsoportokban, illetve nemzetgazdasági ágba tartozók aránya. A mezőgazdaság legnagyobb szerepe a foglalkoztatásban a Dél-Alföldön figyelhető meg, míg a legkisebb Észak-Magyarországon.

*

Itt megszakítjuk az elemzést. Tesszük ezt abból a meggondolásból, mert e cikk megírásával célunk alapvetően az volt, hogy a két statisztikai adatbázis összekapcsolásának körülményeit, annak módszerét, technikáját, az ennek során jelentkező problémákat ismertessük, hiszen, már csak a terjedelmi korlátok miatt is, az összekapcsolásból kapott adathalmaz részletesebb elemzése külön cikk megírását igényli. Mint ahogy arra korábban utaltunk, erre rövidesen sor is kerül a *Területi Statisztika* című folyóirat hasábjain.

SUMMARY

The study is an edited version of the presentations that were given at the annual conference of the Hungarian Statistical association (held on October 14-15, 2004, in Balatonfüred). The authors dedicate this

writing to the memory of *Tibor Kovács,* recently deceased member of the HCSO, whose guidance, valuable professional advice, comments were a big help to the accomplishment of this task. The study deals with the matching of data of two cenzuses that are close in time, and also with the analyzing possibilities deriving from it. Legal background of this matching is described, as well as the technical solution of matching of primary data, and the problems arising concerning identification of addresses and the persons living under this address.