Hivatalos statisztika és a Big Data

Giczi Johanna,
a Központi Statisztikai Hivatal statisztikai tanácsadója
E-mail: Johanna.Giczi@ksh.hu

Szőke Katalin,
a Központi Statisztikai Hivatal tanácsosa
E-mail: kato.szoke@gmail.com

A klasszikus statisztikai adatgyűjtés kérdőívezésen alapul. Ez a mai napig gyakran használt adatfelvételi módszer viszonylag alacsony elemszám mellett, bizonyos korlátokon belül megbízható, érvényes adatokat „produkál", és állandó, magas adatminőséget garantál – ami a hivatalos statisztikában kulcsfontosságú. Problémát jelent azonban, hogy a válaszok nem minden esetben tükrözik a valóságot, az információk feldolgozása esetenként lassú, így egy-egy adat már akkorra elveszti az aktualitását, mire a döntéshozók elé kerül. E hátrányok kiküszöbölésére kiváló megoldást nyújthat a Big Data, amelynek statisztikai célú alkalmazásakor viszont számos nehézséggel kell megküzdeni. Jelen tanulmányban a szerzők arra tesznek kísérletet, hogy bemutassák, az akadályok ellenére e hatalmas adatállományok miként integrálhatók a hivatalos statisztikába. Feltárják és rendszerezik azokat a módszertani dilemmákat, amelyek kezelése elengedhetetlen a megfelelő adatminőség biztosítása érdekében, ismertetik a Big Data lehetséges felhasználási területeit a hivatalos statisztikában, végül pedig felvázolják a Központi Statisztikai Hivatal egyik projektjét, amelyben Big Data-forrásokat használnak egy hivatalos statisztikai adat előállítása érdekében.

TÁRGYSZÓ:
Big Data.
Hivatalos statisztika.
Statisztikai módszertan.

DOI: 10.20311/stat2017.05.hu0461

Megoszlanak a vélemények arról, hogy mennyi adat létezik a világon. Egyes informatikai szakemberek szerint 2,5 exabyte (10^{18}) adat keletkezik naponta (5 exabyte-nyi adattárhelyen már az összes valaha kiejtett emberi szó elférne), míg az IBM szakértői becslésük alapján arra jutottak, hogy napjainkban kétévente megduplázódik az összes adatmennyiség, vagyis huszonnégy hónap alatt annyi adat termelődik, mint a történelemben előtte összesen. Egyértelmű tehát, hogy az információs és kommunikációs technológia fejlődésének következtében hatalmas mennyiségű adat jön létre, amelynek kihasználatlansága az adattudósok számára pazarlásnak tűnik. A Big Data nóvuma azonban nemcsak az adatok számosságában rejlik, hanem – elsősorban a közösségi média és a mobiltelefonok szolgáltatásainak széles körű terjedése miatt – azok változó természetében is. Noha napjaink technikai fejlettsége egyre inkább lehetővé teszi e hatalmas adatmennyiség összegyűjtését, feldolgozását, tárolását és rendszerezését, a hivatalos statisztika számára mégis nehézséget okoz Big Data-alapú módszertanok kimunkálása és alkalmazása. A hivatalos statisztikai szolgálat keretén belül, az Európai Unió legfőbb statisztikai szervezetében, az Eurostatban már közel öt éve indult el az a munka, amelynek célja, hogy a Big Data-forrásokat a hivatalos statisztikai rendszerbe illeszthető adatforrásokká „szelídítse”,[1] és biztosítsa az adatszolgáltatók terheinek csökkentése mellett, illetve a hatalmas adathalmazok adta lehetőségek kihasználásával a hivatalos statisztikai adatok jelenleginél gyorsabb és jobb minőségű előállítását, az adatokból készülő elemzések változatosabbá, esetenként részletesebbé tételét, valamint pontosabb következtetések és előrejelzések készítését. E célok eléréséhez a Big Data integráns részét kell, hogy képezze a hivatalos statisztikai adatgyűjtésnek.

1. Definíciók, avagy mi is az a Big Data

Míg a Big Data definíciója az Oxford-szótárak szerint: „Extrém nagy adathalmazok, amelyek számításigényes analizálása során mintázatokat, trendeket és összefüggéseket lehet feltárni különösen az emberi viselkedés és interakciók terén”,[2] a Wikipédián a következő olvasható: „A Big Data olyan nagy és komplex adathalmazok összessége, amelyek kezelése hagyományos adatbázis-kezelő eszközökkel nem

[1] A téma európai uniós kiemeltségét bizonyítja az is, hogy az Európai Statisztikai Rendszer 2020-ig tartó közös víziójának egyik alapprojektje az ún. „ESS.VIP Big Data”, ami a Big Data hivatalos statisztikai célú fejlesztésével foglalkozik.

[2] https://en.oxforddictionaries.com/definition/big_data

lehetséges."[3] *Gartner, Inc*. [2017] egy harmadik meghatározást ad: „Olyan adatforrások, amelyek általánosságban így írhatók le: nagy mennyiségű, sebességű és változatos adatok, amelyek költséghatékony módon, innovatív formában segítik a folyamatokba való jobb betekintést és a döntéshozatalt." Klasszikusan tehát a következő három fogalommal jellemezhető a Big Data (ezt az angol elnevezések kezdőbetűit alkalmazva 3V-definíciónak is szokták nevezni) (*Glasson et al*. [2013]):

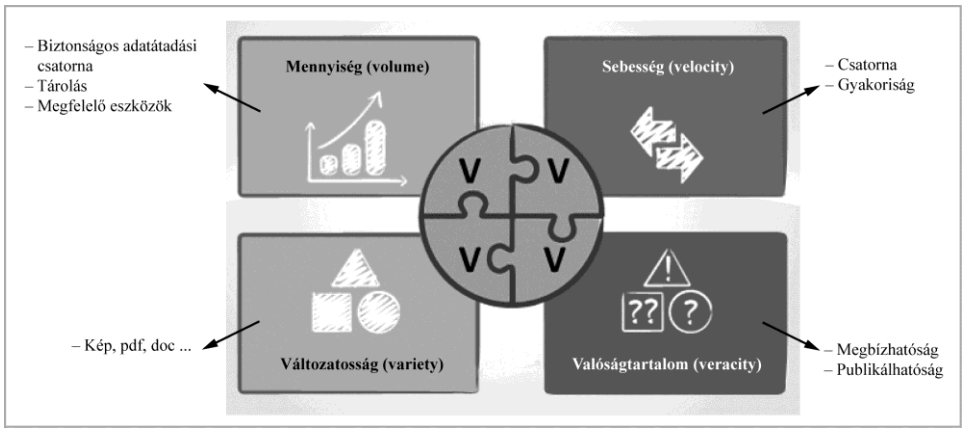
> *1*. *Mennyiség* (*volume*). Nehéz meghatározni, hogy mennyire nagy ez az adatmennyiség, abban azonban mindenki egyetért, hogy amit ma soknak tartunk, az holnapra még több lesz.
>
> *2*. *Változatosság* (*variety*). A Big Data-állományok típusukat, strukturáltságukat tekintve nagyon különbözőek (leginkább strukturálatlanok vagy félig strukturáltak, csak nagy ritkán strukturáltak), és számos forrásból származnak. A teljesség igénye nélkül, ebbe az adatkörbe tartoznak a szenzorok által érzékelt és az okos eszközök adatai, illetve a közösségi hálózatok által generált „lenyomatok", vagyis minden olyan információ, ami valamilyen emberi tevékenység vagy eszköz által nyomot hagy az interneten (számítógépeken). Ilyenek például az sms-ek, a tweetek, a hipertextek, a geolokalizációs információk, az audio- és a videofájlok, a klikkek, a log fájlok, a tranzakciók és az érzékelők adatai stb. A statisztikában több esetben paraadatokként[4] hivatkozott információk egy része is előállhat Big Data-forrásból.
>
> *3*. *Sebesség* (*velocity*). *Groves* [2011] megfogalmazásával élve a Big Data élő adat szemben a survey-típusú felvételek tervezett adataival. A hatalmas adatállományok létrejöttének sebessége elsősorban az adatok „élő" jellege miatt növekszik, hiszen a folyamatosan keletkező adatok szüntelenül áramlanak. Ezzel párhuzamosan gyorsul feldolgozásuk és értelmezésük sebessége is.

A 3V-definíción túl a szakirodalom említést teszt más (ugyancsak v betűvel kezdődő) jellemzőkről is, amelyek közül a hivatalos statisztika szempontjából az adatok *valóságtartalma* (*veracity*) kiemelt fontosságú. E kifejezés arra utal, hogy az adatok mennyire jó minőségűek, milyen mértékben tükrözik a valóságot (*DeVan* [2016]). A jó adatminőség a hagyományos statisztikai eszközökkel folyó adatgyűjtés esetében a legfontosabb szempontok egyike, ugyanakkor óriási kihívás is.

[3] https://en.wikipedia.org/wiki/Big_data

[4] A paraadatok olyan kiegészítő információk, amelyek az adatgyűjtési folyamat során állnak elő. Például (automatikusan rögzített) hívásadatok, összeírók megfigyelései (szomszédtól kapott információk, a kapcsolatfelvétel adatai, az összeírás körülményei stb.) és teljesítményadatai (ledolgozott órák száma, utazási távolság stb.), rögzítési adatok (a validáló üzenetek hatása).

1. ábra. A Big Data 3+1V tulajdonsága a hivatalos statisztika szempontjából



Forrás: *Infodiagram.com* [2014] saját kiegészítéssel.

A Big Data definiálásának szempontjából további fontos jellemzők még: a *változékonyság* (variability), a *megjelenítés*, *vizualizáció* (visualization), az *értékes, felhasználható eredmény* (value), az *érvényesség* (validity), valamint az *illékonyság*, azaz az *érvényesség hossza* (volatility) (*DeVan* [2016]).

Mielőtt rátérnénk azoknak a problémáknak a bemutatására, amelyeket ezek a jellemzők okoznak a hivatalos statisztika számára, illetve ismertetnénk kezelésük módjait, sorra vesszük a Big Data keletkezés szerinti csoportjait.

2. Big Data-taxonómia

A Big Data-forrásokat több rendszerező elv szerint csoportosíthatjuk (*Glasson et al*. [2013]). Ahogy arra már korábban utaltunk, az adatok keletkezésük szerint három nagy csoportba sorolhatók (*Vale* [2013]).

– Az *emberi eredetű adatok* kategóriája az emberi tapasztalatok szubjektív rekordjait takarja, amiket korábban könyvek, művészeti alkotások, majd fotók, videók és audioeszközök tároltak, és amelyek napjainkban csaknem mindig digitálisan (személyi számítógépeken, a közösségi hálón) keletkeznek. A hivatalos statisztika ezekhez a jellemzően gyengén strukturált, gyakran ellenőrizetlen adatokhoz csak korlátozottan fér hozzá. E típusba tartoznak a Facebook-kommentek, a

lájkok és a posztok, a tweetek, a blogok, a vlogok, a személyes dokumentumok, a közösségi képmegosztókra (Pinterestre, Instagramra, Youtube-ra) feltett képek, videók, az interneten lefuttatott keresések, a mobiltelefonon küldött üzenetek és az e-mailek is.

– A *folyamateredetű adatok* közé a különböző (elsősorban az üzleti) folyamatok során keletkező adatokat soroljuk. Ezek jól strukturált, jellemzően RDBMS- (relational database management system – relációs adatbázis-kezelő rendszer) adatok vagy metaadatok. Egy típusukat a *nyilvántartások adatai* alkotják, melyek tipikusan állami intézmények (például a közhivatalok) által fenntartott források adatai, de idetartoznak az elektronikus egészségügyi nyilvántartások, az orvosi rekordok, a kórházi látogatások nyilvántartása, a biztosítási nyilvántartások, a banki vagy részvényadatok, a vállalkozások üzleti adatai is (ha az utóbbiakról nyilvántartás vezetését jogszabály írja elő). A folyamateredetű adatok másik csoportját a *tranzakciós adatok* adják; ezek közös jellemzője, hogy két entitás közötti tranzakcióból származnak. Ilyenek például a kereskedelmi tranzakciók (például az internetes vásárlások), a bank- és hitelkártya-tranzakciók, valamint az e-kereskedelem adatai (ideértve a mobilkészülékről indított tranzakciókat is) stb.

– *A gépek által előállított adatokat* klasszikusan a hangzatos Internet of Things (A dolgok internete) néven emlegetik. Idetartoznak a fix és mozgó szenzoros adatok, valamint a log fájlok. Definícióját tekintve a szenzoros adatokból származó információköteg nem más, mint a fizikai világ eseményeit rögzítő és mérő érzékelők milliárdjainak adatai. Ahogy egyre több érzékelő kerül a világban bevezetésre és aktiválásra, úgy nő az ilyen jellegű adatok volumene is. Mindent összevetve, ennek az adattípusnak a mennyisége növekszik a leggyorsabban. Szenzoros adatoknak tekinthetjük például a háztartási eszközök érzékelőinek, az időjárás- vagy a légszennyezettség-érzékelőknek, a műholdképeknek, a forgalomfigyelőknek/webkameráknak az adatait; a nyomkövető eszközös adatok közé pedig például a mobiltelefonok útvonal-/követési és a földrajzi helyzetre vonatkozó (például GPS-) adatok sorolhatók. A log fájlok a számítógépek működése során, szöveges (text) formában létrehozott, rendszereseményekről szóló ún. naplóbejegyzések.

A három típust elkülöníthetjük olyan szempontból is, hogy mely két szereplő között történik a kommunikáció: az emberi eredetű adatok ember és ember, a folyamateredetűek ember és gép, míg a gépek által előállítottak gépek közötti kommunikáció eredményeként keletkeznek.

3. Big Data-paradigma

Már az eddigiekből is világosan látszik, hogy a Big Data egészen eltérő tulajdonságokkal bír, más logika alapján „működik", mint a hagyományos adatgyűjtési eljárások, módszerek. A problémakört alapvetően információtechnológiai kérdések és szakmai dilemmák alkotják. Az előbbiekre jelen tanulmányban nem térünk ki részletesen, csak néhány pontban vázoljuk fel őket.

3.1. Információtechnológiai kérdések

Könnyen belátható, hogy a Big Data-dömping kezeléséhez, a folyamatosan áramló adatok gyűjtéséhez, tárolásához, előkészítéséhez és feldolgozásához meg kell felelni bizonyos előfeltételeknek.

1. Egyre növekvő számítástechnikai teljesítményre van szükség, amelyet MPP- (massive parallel processing – masszív párhuzamos feldolgozás) megoldásokkal lehet kezelni.

2. Elengedhetetlen az adatredisztribúció és a párhuzamos feldolgozás lehetőségének megteremtése (a MapReduce, a Hadoop, a Hortonworks Data Platform, az R-Rstudio stb. ismerete és alkalmazhatósága).[5]

3. Nélkülözhetetlen a nem csak SQL-re épülő, adatmennyiségredukáló szoftvertechnológia használata és az abban való jártasság. A statisztikusok szempontjából ugyanakkor kérdéses, hogy a Big Data alkalmazása milyen IT-ismereteket kíván meg.

3.2. Szakmai kérdések

Azoknak a problémáknak a tisztázása érdekében, melyekkel egy statisztikusnak a Big Data hivatalos statisztikai alkalmazása során kell szembesülnie, elsőként a hagyományos adatgyűjtést tekintjük át és azt, hogy ahhoz képest a Big Data gyűjtésekor milyen változásokkal kell számolni. Ezt követően a különböző adatforrástípusokat vetjük össze, és bemutatjuk, hogy a hatalmas adathalmazok miben térnek el ezektől. Majd a hivatalos statisztika minőségi kritériumai szerint haladva tárgyaljuk a

[5] Ez utóbbira kiváló példa a SETI Intézet programja, melynek keretében önkéntesek bevonásával elemzik a világűrből érkező jeleket, értelmes élet nyomait (legalábbis mintázatokat) keresve. Az önkéntesek saját számítógépükön futtatják a programokat, megsokszorozva ezáltal az elemzések tempóját.

Big Data és a hagyományos statisztikai eljárások jellemzőit, végül a Big Data elemzési problémáival foglalkozunk.

3.2.1. Adatgyűjtés és a Big Data

Hagyományos megközelítés, avagy a top-down paradigma. A hivatalos statisztika általános gyakorlata szerint egy adatfelvétel előtt elsőként azt kell meghatározni, hogy milyen információkra van szükségünk, és ehhez hipotéziseket fogalmazunk meg. Majd a következő lépéseket hajtjuk végre: *1*. adatgyűjtés-tervezés, *2*. adatgyűjtés, *3*. adat-előkészítés, *4*. adatelemzés, *5*. információkinyerés az adatbázisból/a felállított hipotézis igazolása vagy cáfolata.

A top-down paradigma lényege, hogy az adatgyűjtés megtervezése során az elemzési cél(ok) meghatározásán van a hangsúly. A hagyományos adatfelvételeknek tehát kulcsfontosságú eleme a tervezés, melynek részei a következők: *1.* változók, definíciók kialakítása, konceptualizálás, majd operacionalizálás, *2*. a vizsgálni kívánt sokaság kiválasztása (ez lehet teljes körű, vagy alapulhat mintavételen), *3*. az alapsokaság elérésére listák, regiszterek alkalmazása, *4*. osztályozások, kérdőívek készítése.

Az elemzési célok eléréséhez specifikus információ(k)ra/hipotézis(ek)re van szükség, amely(ek) megszerzése/megfogalmazása után modellépítés következik. A folyamat zárása lehet valamilyen leíró statisztika, becslés vagy előrejelzés megadása.

Big Data-megközelítés, avagy a bottom-up paradigma. A Big Data-paradigma esetében az előzőhöz képest egészen más logikát kell követnünk. Mivel itt nincs szükség az adatgyűjtés tradicionális értelemben vett megtervezésére (hiszen az adatok már megvannak, pontosabban mindenütt ott vannak), felborul a klasszikus sorrend. A tervezés helyett ilyenkor magával az *1*. adat(be)gyűjtéssel indítunk, ezt követi az *2*. adatelőkészítés, az *3*. adatfeltárás (ami többnyire korrelációk keresését jelenti), *4*. az algoritmusok testreszabása (elsősorban skálázható algoritmusok választása aggregálás kerülésével), végül *5*. új tudás felfedezése/és az eredmények validálása (heurisztikus [mintakereső] technológiák használata az előrejelzésekhez/becslésekhez).

E megközelítés esetében a hangsúly a hozzáférhető adatok felfedezésén, vagyis olyan információértékek keresésén van, amiket ezekből mások még nem nyertek ki. Nyilvánvalóan ez a logika inkább az adattudósok[6] (data scientists) által vizsgált problémákra kínál megoldást, akiket sokkal inkább a „Mi történik?” kérdés érdekel, mint a „Miért?” és a „Hogyan?”. E speciális jellemzők miatt a Big Data integrálása a hivatalos statisztikába egyáltalán nem megy gördülékenyen.

[6] Ez egy viszonylag új foglalkozás. Művelőjének a matematikai és a statisztikai készségeken túl programozói ismeretekkel, az adott területen tényleges tapasztalattal és magabiztos szakértői tudással is rendelkeznie kell.

3.2.2. Az adatforrástípusok összevetése

Az adatgyűjtési paradigmákon túllépve, vizsgáljuk meg, hogy milyen jellemzők szerint definiálhatók az elsődleges és a másodlagos statisztikai adatforrások. Elsődleges adatforrásnak hívjuk azokat az adatfelvételeket, amelyek kérdőíves technikát alkalmaznak (ilyen például a népszámlálás is), függetlenül az adatgyűjtés mintavételes vagy teljes körű jellegétől. Másodlagos adatforrások pedig az adminisztratív forrásból származó adatfelvételek és a Big Data-jellegű adatforrások.[7] Az 1. táblázatban a különféle adatforrások jellemzésének fő szempontjait foglaljuk össze.

1. táblázat

Az adatforrások jellemzői

Jellemző	Elsődleges statisztikai adatforrás	Másodlagos statisztikai adatforrás	
		Adminisztratív adatforrás	Big Data-jellegű adatforrás
Az adatok statisztikai cél(ok)ra tervezettek	igen	nem	nem
A fogalmak, a definíciók és az osztályozási rendszerek egyértelműek és ismertek	igen	gyakran	ritkán
A célsokaság jól definiált	igen	gyakran	nem
Rendelkezésre állnak metaadatok	igen	gyakran	nem
Az adatok strukturáltak	igen	igen	ritkán
Az adatok a vizsgált alapsokaságra vonatkoznak	igen	rendszerint	nem
A statisztikai adatok „kinyeréséhez” az adatok előfeldolgozása szükséges	nem	nem	igen
A lényeges/érdeklődésre számot tartó adatok közvetlenül elérhetők	igen	gyakran	nem
A segédváltozók közvetlenül elérhetők	igen	gyakran	nem
Az adatok teljes körűen lefedik a vizsgálni kívánt sokaságot	igen (cenzus) nem (survey)	gyakran	még nem
Az adatok reprezentatívak vagy adott elemzésekre reprezentatívvá tehetők	igen	gyakran	nem

Forrás: *Istat ESTP* [2016].

Az összehasonlításból kiderül, hogy a Big Data-jellegű adatforrások jellemzői, definíciós és osztályozási rendszerek szerinti adekvátságukat és (bizonyos esetekben) strukturáltságukat tekintve, „csak közelítenek” a klasszikus adatgyűjtési módszerek-

[7] Az Istat (Olasz Statisztikai Hivatal) ezekre harmadlagos adatforrásokként hivatkozik.

kel szemben megfogalmazott követelményrendszerhez. Elsősorban az alapsokaság jó definiálhatóságán, illetve a célsokaság meghatározhatóságán kell javítani (melyen elsősorban „lefedettségi" problémák kezelésére gondolunk) ahhoz, hogy az ilyen jellegű adatok a hivatalos statisztika minőségi követelményeinek is megfeleljenek. Amennyiben ezeket sikerül jobbá tenni, azaz minőségüket mérni és megfelelő módszereket beépíteni a statisztikai adat-előállítási folyamatba, úgy közelebb kerülünk a reprezentativitás problémájának megoldásához is. A metaadatok és a segédváltozók meghatározásának kérdése ugyancsak az előbbi kérdések kezelésének a függvénye. A megoldáshoz a számítási kapacitások növelése mellett számos módszertani, adatvédelmi és etikai dilemma újratárgyalására, illetve új szabályrendszerek lefektetésére, majd nemzetközi szintű megvitatására lehet szükség.

3.2.3. Minőségi kritériumok és a Big Data

A hivatalos statisztikának mint adatnak és mint intézménynek számos minőségi kritériumnak kell megfelelnie. Ezek közül a továbbiakban azokat vesszük sorra, amelyek megfontolásra érdemesek a Big Data alkalmazásakor.[8]

A hivatalos statisztikában kulcsfontosságú a *reprezentativitás*. A hagyományos mintavételi eljárásokban alkalmazott, jól megválasztott, bizonyos szempontból reprezentatív minta jól jellemzi a sokaságot. A Big Data-jellegű adatforrások esetén ezzel szemben már rendelkezésre állnak az adatok, ám a statisztikai definíció szempontjából általában nem teljes körűen. A teljes sokaságot tekintve, lefedettségi hiány és többlet is felmerülhet, ami torzításhoz vezet. Így a Big Data-források nem reprezentatív adatbázisoknak tekinthetők, amelyekhez biztosan kellenek az érvényesség[9] vizsgálatára alkalmas referenciaadatok is. Ugyancsak fontos a szelektivitási/reprezentativitási mutató[10]. Ez azt mutatja meg, hogy a Big Data-forrásból származó adatok miben térnek el a tényleges sokaságtól. Az ún. ignorálhatósági feltételek felállításával pedig a lefedettségi, a mintavételi, a mérési és a válaszadói torzítás kezelhető (lásd bővebben *Couper* [2013]).

A hivatalos statisztikában kulcsfontosságú szempont az adatok *összehasonlíthatósága*. Az, hogy országonként eltérő fogalmakkal dolgozunk (gondoljunk itt akár a háztartás, a család vagy a munkanélküliség definíciójára), alkalmanként problémát

[8] A Függelék F1. táblázatában további minőségi dimenziók alapján hasonlítjuk össze az adatgyűjtéseket és a Big Data-jellegű adatforrásokat.

[9] Kutatás-módszertani szempontból az érvényesség annyit tesz, hogy a kutatásunk valóban a vizsgálat tárgyára irányul, vagyis a módszer arra a kérdéskörre szolgáltat információt, amit meg akarunk vizsgálni, ismerni. A nemzeti statisztikai hivatalok a megtervezett felvételektől egyre inkább a termékek és az outputok széles köre felé „mozdulnak el", ami még inkább felhívja a figyelmet az érvényességre.

[10] A szelektivitás/reprezentativitás az egyik legfőbb aggodalomra okot adó dimenzió. Egy nem reprezentatív adatbázis lehet hasznos bizonyos célokra, de nem megfelelő másokra. A kérdés, hogy vannak-e referenciaadatok, amelyekkel meg tudjuk vizsgálni az érvényességet.

jelent a különböző szakstatisztikák közötti, a területi vagy akár az időbeli összehasonlíthatóság szempontjából a hagyományos adatgyűjtésekben. Hasonló nehézségek merülnek fel a Big Data-jellegű és az adatgyűjtésekből származó adatok összevetése során is:

– *Definíciós különbségek*. A hivatalos statisztikában a mérendő fogalom a különböző hivatalos statisztikai szolgálatok közötti harmonizáció eredményeként, valamint az európai uniós és a hazai elvárások, szabályok alapján pontosan definiált. Egy Big Data-jellegű adatforrás alkalmazásakor viszont figyelembe kell venni, hogy az ez alapján kialakított változó fogalma általában nem egyezik meg a statisztikaival, így az eltérő fogalmi struktúrák összehangolása az elsődleges feladat.

– *A sokaság fogalma*. A hivatalos statisztikában a vizsgálat tárgyát képező egységek összességét, halmazát (statisztikai) sokaságnak nevezzük. Az egységek tulajdonságaik megadásával jellemezhetők. A Big Data-jellegű adatforrásokban elérhető sokaság általában eltér a jellemezni kívánt sokaságtól, ezért olyan módszertani eljárások kialakítására van szükség, amelyekkel az előbbi alapján az utóbbi előállítható. Egy mobilszolgáltató esetében például a szolgáltatásokat igénybe vevők alkotják a sokaságot, ami így biztosan magában hordoz bizonyos torzításokat, eltér a jellemezni kívánt sokaság fogalmától. (Ha az általunk jellemezni kívánt sokaság Magyarország lakónépessége, akkor lesz olyan, akit ezen a módon nem tudunk megfigyelni [például azokat <gyerekeket, időseket>, akiknek nincs mobiltelefon-előfizetésük], illetve olyan is, akinek több előfizetése is van. Így torzul a jellemezni kívánt sokaság.)

– *A statisztikai egység fogalma*. A hivatalos statisztika a gyűjtött adatok alapján eltérő egységeket vagy az egységek eltérő csoportjait figyeli meg, elemzi, illetve azokról tájékoztat. A Big Data felhasználásakor ugyanakkor végig kell gondolni, hogy minden olyan információ rendelkezésre áll-e, ami az eltérő statisztikai egységek kezeléséhez szükséges. A Big Data-forrás vonatkozási köre, statisztikai egységei ugyanis eltérnek a hivatalos statisztikáéitól, így további módszerek, modellek alkalmazására van szükség az információk előállítása céljából. Példaként hozható erre az az eset, mikor egy Big Data-jellegű adatforrásban a mobiltelefonok (előfizetések) és nem a személyek a statisztikai egységek. Mivel a statisztikusok a személyek viselkedéséről, szokásairól szeretnének megállapításokat tenni, ilyenkor problémát jelenthet, hogy egyeseknek több mobiltelefonja (előfizetése) is van, míg másoknak egy sincs.

3.2.4. Módszertani kérdések pro és kontra

Az eddig leírtakat mintegy összefoglalva, illetve kiegészítve, a 2. táblázatban foglaljuk össze azokat a tapasztalatokat, amelyek a Big Data hivatalos statisztikai alkalmazása mellett, illetve ellene szólnak.

2. táblázat

Érvek a Big Data módszertana mellett és ellen

Érv	Ellenérv/Kihívás
Nincs minta	Nincs minta – reprezentativitás
Valós idejű	Lefedettség (többlet/hiány)→torzítás
Valós viselkedés, nem önbevallás	Input-/output-adatok minőségének mérése
Válaszadói tehercsökkentés	Adatforrás felhasználásának módja, potenciális validálási adatforrás elvesztése
Társíthatók más adatbázissal	Összehasonlíthatóság (jelenlegi statisztikával)
Új ismeret feltárása	IT-felszereltség, támogatás
Költségek (hosszú távon)	Költségek (rövid távon)
	Adathozzáférés
	Adatvédelem
	Stabilitás

Forrás: Saját összeállítás.

A 2. táblázat érvei és ellenérvei között szerepel a minta hiánya. Ez egyrészről jó, hiszen nem merülnek fel mintavétel okozta hibák, másrészről viszont, ahogy arra már korábban utaltunk, gondot okoz a reprezentativitás szempontjából. Probléma, hogy csak nagyon korlátozott a tudásunk az alapsokaságról, ebből következően nem világos a mintaegységek kiléte sem. A sokaság beható ismerete nélkül azonban nem biztosítható, hogy a teljes célsokaságra vonatkozó statisztikákat állítsuk elő; vagyis „oda" a kvantitatív kutatások egyik legfőbb erénye, az általánosíthatóság.

A kihívások oldalon felmerül a stabilitás kérdése is, ami még a hagyományos adatgyűjtések esetében is gondot okoz (a magas nemválaszolási arány például instabillá tehet egy felvételt). A Big Data könnyen és gyorsan változó adatáram; ezért bármikor előfordulhat, hogy megszűnik egy honlap, letörli valaki a mobiljáról az adatokat gyűjtő applikációt, vagy letiltja a hozzáférést a telefonjához stb.

A Big Data határozott előnye ugyanakkor, hogy valós idejű. Ezek az adatok akár azonnal rendelkezésre állnak, így a hagyományos adatgyűjtések során keletkező adatokhoz képest gyorsabban lehet őket elemezni. A valós idejű adatok gyűjtése azonban több „falba is ütközhet", ami miatt elveszhet ez az előny. A hivatalos sta-

tisztikai szervezetek számára ugyanis probléma, hogy a Big Data más intézmények, szervezetek, személyek tulajdonában van, ezért az ahhoz való hozzáférés bizonyos esetekben költséges, a valós idejű, egyedi szintű, személyes adatokhoz való hozzáférés pedig etikai, illetve adatvédelmi problémák miatt nehézkes lehet. Az ilyen jellegű adatok a legtöbb esetben erősen strukturálatlanok, és szinte minden esetben vannak közöttük „céltalanok" is. Ezért a gyors felhasználhatóság korlátja lehet, ha ezeket a zajokat nem sikerül megfelelően kiszűrni az adatbázisból.

A Big Data egy másik nagy pozitívuma, hogy bizonyos fajtái, szemben az önbevallásos adatokkal, valós viselkedést mutatnak. Ezáltal kiküszöbölhetjük a hagyományos adatgyűjtés néhány nem elhanyagolható nem mintavételi hibáját (például a nemválaszolást, a válaszadói hibát, a torzítást, a kérdezőbiztos hatását).[11]

A hivatalos statisztikának – ahogy erre már utaltunk korábban – jelenleg is fontos célja a válaszadói terhek csökkentése. Ezért amennyiben a kérdőívekre adandó válaszok részben vagy egészben rendelkezésre állnak más (Big Data- vagy adminisztratív) adatforrás(ok)ból, illetve az/azok alapján kikövetkeztethetők, akkor nem terheljük az adatszolgáltatókat a megkérdezéssel, és ezáltal az adatgyűjtési költségek is csökkennek.

Ugyancsak a Big Data használata mellett szól, hogy az ilyen jellegű adathalmazok könnyen társíthatók más adatbázisokkal. A jelenlegi álláspont szerint a Big Data-források kiegészítő, validáló jelleggel, megfelelő adatfúziós eljárások alkalmazásával hasznosíthatók a hivatalos statisztikában.

3.2.5. Elemzési gondok

Noha eddig már számos, Big Data-val kapcsolatos problémát ismertettünk, szeretnénk bemutatni azt is, hogy az adatelemzés során milyen gondokkal kell szembesülnünk.

A hagyományos elemzési eljárások nem működnek a Big Data esetében. Egyrészt a hatalmas adathalmazok elemzése során a számítási kapacitás és a komplexitás határaiba ütközünk (például inverz mátrixok képzése, legkisebb négyzetek elve, maximum likelihood vs. Newton–Raphson-algoritmus). A legtöbb hagyományos algoritmust nehéz párhuzamosítani, tehát nagyon körülményes megoldani, hogy egyszerre több processzor dolgozzon a részleteiken (a Hadoop például nem tudja ezt kezelni). Márpedig az óriási adatmennyiség miatt a számítási kapacitás növelése másként nem oldható meg. Másrészt, a hagyományos statisztikai eljárások nagyon érzékenyek a hibás adatokra és a szélsőértékekre, ezért kötelező ellenőrzést és adattisztítást végrehajtani. A Big Data nagy része azonban „zajos" és strukturálatlan, ráadásul olyan óriási adatmennyiség, amelyben nem lehet „egyszerűen" editálni, imputálni,

[11] A teljesség kedvéért meg kell jegyeznünk, hogy a Big Data-ban lehetnek olyan, akár szisztematikus torzítások is, amelyeket nem ismerünk, így azokat kontrollálni sem tudjuk.

outliereket kezelni. Nehézséget jelent még a duplikátumok kezelése is. A statisztikai hivatalokban hagyományos adatkezelései eljárások, jól felépített ellenőrző rendszerek és adattisztítási algoritmusok biztosítják, hogy a survey-típusú felvételekből előállított adatbázis ne tartalmazzon duplikátumokat. A Big Data esetében erre szintén külön eljárásokat kell kidolgozni.

További problémát jelent, hogy a Big Data-n alapuló elemzések többségükben korrelációvizsgálatra épülnek. Ez a módszer azonban magában hordozza a hamis korrelációk (lásd ökológiai tévkövetkeztetések) lehetőségét; a nem egyértelmű korrelációk pedig az „okozat halálához" vezetnek (*Scannapieco–Virgillito–Zardetto* [2013]).

Amennyiben a Big Data-t statisztikai célra használjuk fel, kiváltva a statisztikai adatgyűjtésre épülő adatokat, az előbb tárgyalt problémák „árnyaltan" jelentkeznek, hiszen a begyűjtött Big Data-nak is ugyanazokon az eljárásokon (adatelőkészítésen, mikrovalidáláson, editáláson, outlier-kezelésen, aggregáláson) kell átesnie, mint a hagyományos adatfelvételekből származóknak. A számítási komplexitás és a kapacitás kérdése azonban továbbra is problematikus.

A hivatalos statisztika jelenlegi eljárásai (tervezett, modellre épülő mintavételi eljárások, regresszió, általános lineáris modellek stb.), melyek a hagyományos alapadatok specifikus tulajdonságain állnak vagy buknak, jó minőségű, de (a Big Data-hoz mérten) kevés adat kezelésére, elemzésére alkalmasak.

Az előbbiekben megfogalmazottak alapján tehát úgy tűnik, hogy a jelenlegi elemző eljárásoknak semmi közük a Big Data-hoz. Mi lehet akkor a megoldás? A szakirodalom egyetért abban, hogy a Big Data kezeléséhez radikális paradigmaváltásra van szükség a statisztikai metodológiában:

> – Robusztus eljárásokat kell használni még akkor is, ha az némileg a pontosság rovására megy. Ugyanakkor mindig ki kell kötni a pontosság és a minőség kritériumait. A pontosság csak akkor romolhat, ha párhuzamosan más minőségi összetevők megfelelő mértékben javulnak.
>
> – A Big Data elemzési metódusának közelítő és nem egzakt optimalizációs technikákon kell alapulnia, amelyek képesek megbirkózni a zajos célfüggvényekkel.[12]
>
> – Szemléletbeli változás szükséges. El kell fogadni, hogy a Big Data más típusú elemzéseket tesz lehetővé (*Scannapieco–Virgillito–Zardetto* [2013]).

[12] A kvadratikus célfüggvényt gyakran (így a Big Data esetében is) több-kevesebb mérési hibát (zajt) tartalmazó mérési adatok határozzák meg. A célfüggvény becslése történhet közvetlenül a nyers mérési adatokból (historikus becslés) vagy úgy, hogy előbb a nyers adatokra egy eloszlásfüggvényt illesztünk (parametrikus becslés).

E kompromisszumokat ugyanakkor árnyalhatja az a tény, hogy ha a hivatalos statisztikát akárcsak részben Big Data-alapon akarjuk fejleszteni, előállítani, közzétenni, akkor az utóbbinak maradéktalanul meg kell felelnie a hivatalos statisztikával szembeni elvárásoknak. Mindebből pedig az következik, hogy a hatalmas adathalmazok hivatalos statisztikává válásukkor (részben) elvesztik Big Data-jellegüket.

4. A Big Data alkalmazása a hivatalos statisztikában, nemzetközi tapasztalatok

A Big Data-val kapcsolatos módszerek fejlesztésében és alkalmazásában az Olasz és a Holland Statisztikai Hivatal jár az élen. A következőkben néhány olyan projektet mutatunk be, amelyek eredményeit már sikerrel alkalmazzák a hivatalos statisztikai szolgálatok.

4.1. Közösségi médiaelemzés a hivatalos statisztikában

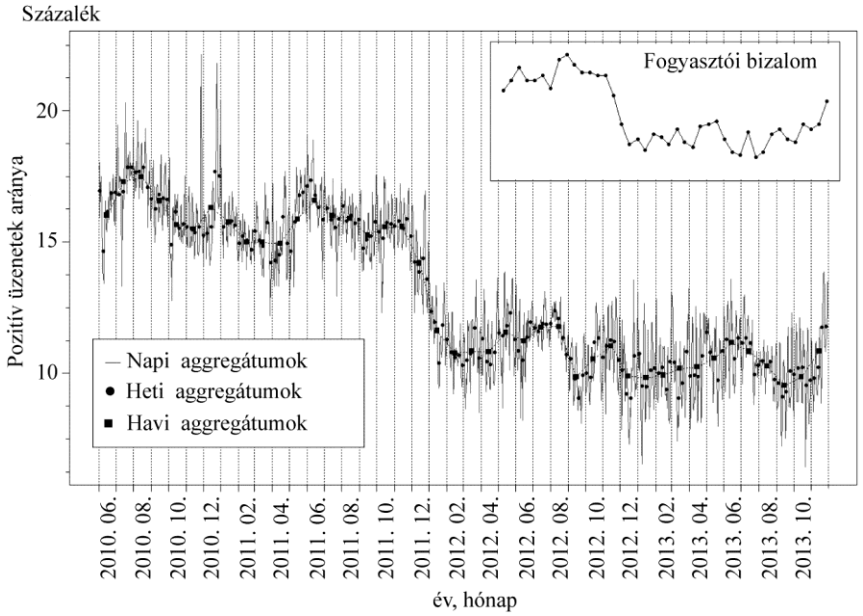
Hollandiában a lakosság körülbelül 70 százaléka használ egy vagy több közösségi oldalt (*Daas–van der Loo* [2013]), amelyek közül a Facebook és a Tweeter a legnépszerűbb. Kutatók a holland Tweeteren – ahol a legtöbb holland nyelvű, nyilvánosan elérhető tartalom található – közzétett üzeneteket elemezték, azok tartalma és „általános hangulata" közötti összefüggést vizsgálva.[13] A szövegeket tanulmányozva kiderült, hogy az általános hangulat erős korrelációt mutat a gazdasági helyzettel és a fogyasztói bizalommal.[14] Az előbbivel való összefüggése olyannyira stabilnak mutatkozott, hogy azt heti és havi gyakoriságban is vizsgálták. (Az eredmények kritikájaként megjegyezzük, hogy miként arra több tanulmány [például *Pléh–Unoka* [2016]] is rámutatott, a közösségi oldalak posztjaiban nem feltétlenül az egyén valós véleménye, sokkal inkább egyfajta elvárt normához való igazodás jelenik meg, ami pozitívabb képet mutat az adott személyről, mint a valóság.[15])

[13] Az elemzésekből az is kiderült, hogy a beszélgetések közel 50 százaléka értelmetlen „gügyögés" volt; a fennmaradó hányad többek között a szabadidős tevékenységgel (10%), a munkával (7%), a televíziós és a rádiós médiával (5%), valamint a politikával (3%) foglalkozott.

[14] A szociológiai szakirodalomban vitatott kérdés az általánosított és a partikuláris bizalom között aszszociáció. Az előbbi a gazdasági fejlődés hordozója, míg az utóbbi akadályozza azt (*Fukuyama* [1995], *Knack–Keefer* [1997], *Raiser et al*. [2001]). A hivatalos statisztikában a wellbeing-vizsgálatok foglalkoznak a bizalom szintjével.

[15] A decemberi közösségi médiaüzenetek sokkal pozitívabbak voltak, mint az előtte vagy utána levő időszakokban.

2. ábra. A fogyasztói bizalom és az általános hangulat összefüggése a holland közösségi médiaüzenetekben, 2010. június – 2013. december



Forrás: Daas–Puts [2014].

4.2. Szenzorok által generált adatok alkalmazása a hivatalos statisztikában

Hollandiában a közlekedési szenzorok több mint 6000 km hosszú úthálózatra vonatkozóan gyűjtenek adatokat, percenként közel 24 ezer adatot dolgoznak fel, majd tesznek közzé 75 másodpercen belül, lehetőséget adva ezzel a gépjárműben utazók számára a dugók elkerülésére és egyben az utak biztonságosabbá tételére. E részletes adatok elemzésével a hivatalos statisztika fontos szállításstatisztikai becsléseket hajthat végre (például a határon belépő és kilépő személyforgalmon kívül becsülhetővé válik akár járműtípusonként, akár járműhonossági bontásban a szállítási forgalom is), illetve kiegészítheti azokkal adatgyűjtéseinek adatait. Jelenleg a projektnek még nagy problémája, hogy a szokványos statisztikai eszközökkel egy napnyi szenzoros adatmennyiség feldolgozása lehetséges, háromhavi adat elemzése azonban már Big Data-eszközöket kíván.[16] Ez a probléma azonban a számítási kapacitások növelésével könnyen áthidalható lesz.

[16] Hogy nagyságrendben lássuk: 1 perc alatt ~ 460 000, 1 óra alatt ~ 27 millió, 1 nap alatt ~ 600 millió, 1 év alatt ~ 240 billió adat keletkezik.

A műholdképek ugyancsak szenzoros adatforrások. Segítségükkel gyakran és pontosan lehet tanulmányozni a földhasználati jellemzőket. Jó példa erre az Ausztrál Statisztikai Hivatal műholdképes földhasználati felmérése (*Tam–Clarke* [2015]), melynek adataira nemcsak a hivatalos statisztika egésze, de a környezetvédelemmel kapcsolatos szakstatisztika is támaszkodik. A földfelszíni területek felhasználásának meghatározására Ausztráliában ugyancsak műholdképeket használnak. A kutatási programban mezőgazdasági földhasználati jellemzők alapján elemzik a képeket, hogy megbecsüljék a növényfajok termesztésének arányát. A területi földhasználati jellemzőket képelemző algoritmus segítségével határozzák meg (*Daas–van der Loo* [2013]).

4.3. Mobileszközök által generált adatok alkalmazása a hivatalos statisztikában

Az Észt Központi Bank, a Tartui Egyetem és a Positium LBS cég együttműködése révén, konzorciumi projekt keretében jött létre egy fejlesztés a mobileszközök által generált adatok hivatalos statisztikai alkalmazására. A Positium LBS (amit kizárólag e célra alapítottak) a mobilszolgáltatóktól gyűjti össze és dolgozza fel statisztikai modellek segítségével a névtelen adatokat, amelyek megbízható képet adnak az országhatárt átlépőkről (mind a külföldre utazókról, mind az Észtországba belépőkről). Egy PDM (product data management – termékadat-kezelő) szoftver segítségével – ami részben a mobilszolgáltatók rendszerében működik és üzemeltetői ellenőrzés alatt áll, részben pedig a Positium LBS mint adatmediátor által vezérelt – biztosított az üzleti titok és a személyes adatok védelme (mivel a válaszadók egy véletlenszerűen kiválasztott álnevet/kódot kapnak, lehetetlen az adott telefonszám/-tulajdonos beazonosítása). Az adatgyűjtés aktív és passzív helymeghatározáson alapul. Az előbbi lényege, hogy MPS (mobile positioning system – mobil helymeghatározó rendszer) segítségével a mobilkészülékek helyének azonosítása, nyomon követése valós időben (okostelefonok esetében a GPS segítségével) történik. Passzív helymeghatározás esetén a (főleg belső üzleti vagy marketing célokra használt) adatok automatikusan tárolódnak a mobilszolgáltatóknál (memóriában vagy log fájlokban).

A három intézmény együttműködését nehezíti, hogy a mobilszolgáltatók az eladások számának növelésében érdekeltek, fontosak számukra a vevők, tehát a titoktartás és a szavahihetőségük megőrzése. A partnereknek így számos szakmai, módszertani és jogi kérdéssel kell megküzdeniük, hogy összegyűjtsék a helymeghatározási adatokat.

A Positium LBS az adatok begyűjtése után minőségellenőrzést végez. Mivel óriási adatmennyiségről van szó, ki kell szűrnie a karakterisztikus hibákat, és javítania kell azokat. Következő lépésként a cég az adatokat térben interpolálja egy speciális térin-

formatikai modul használatával. A mobilkészülék-használat tér- és időbeli vizsgálatával kapcsolatos statisztikák gyűjtésének számos módszertani sajátossága van. A mobilkészülékek használata jövedelemtől, életkortól és más társadalmi ismérvtől függetlenül (ám a hálózati lefedettségtől és a sűrűségtől függően) elterjedt a fejlett és a fejlődő országokban egyaránt. Ezáltal könnyen és széles körben folyhat az adatgyűjtés. A költséghatékonyság mindenképpen pozitív aspektusa a módszernek, hiszen az eredmények automatikusan rögzülnek, és itt nem jelentkezik az adatfelvételekre általában jellemző válaszadói felkeresés közvetlen költsége (*Daas–van der Loo* [2013]).

5. A Big Data lehetséges felhasználási területei a KSH-ban

Az európai statisztikai hivatalok egy része – ahogy azt korábban bemutattuk – már Big Data-alapú vagy a hagyományos adatgyűjtési technikát és a Big Data-t ötvöző módszereket is használ. A következőkben ismertetjük azokat a területeket, ahol e projektek tapasztalatai sikerrel lennének alkalmazhatók a KSH-ban akár az adatgyűjtés Big Data-forrásokkal való kiegészítésével, akár az eddig gyűjtött információk validálásával. A rendszerezést a Big Data-adatforrások típusai – a mobilkommunikáció, valamint az internetes, a szenzoros és a folyamatgenerált tranzakciók – alapján végeztük.

5.1. A mobileszközös kommunikáció során keletkező adatok alkalmazási lehetőségei a hazai hivatalos statisztikában

A belföldi és a nemzetközi vándorlásról szóló adatgyűjtés a KSH-nál éves gyakorisággal folyik, több OSAP (Országos Statisztikai Adatfelvételi Program) keretében. Az adatok forrásai a népszámlálás, a mikrocenzus és a LUSZ- (lakosság utazási szokásai) felvétel. Az első kettőre tízévente, míg az utóbbira évente kerül sor. Mobiltelefon-helymeghatározási adatokat használva azonban, ha a korábban említett lefedettségi problémák miatt nem is teljes körűen, de a jelenleginél gyakrabban kaphatunk információkat a népesség mobilitásáról.

A mobiltelefonos cellaadatok a turizmusstatisztikában is segítséget nyújthatnak az éves adatgyűjtésnél gyakoribb felvételekben, illetve egy mainál pontosabb becslési eljárás kidolgozásában. Kifejezetten jó támpontul szolgálhatnának például a határforgalom monitorozásában. A schengeni határszakasz kiterjesztésével és így a határátlépések ellenőrzésének megszűntével ugyanis a korábbinál lényegesen kevesebb információnk van a határátlépők számáról, honosságáról. Ezekre a problémákra – hasonlóan a korábban bemutatott észt projekthez – a telefonos cellainformációkhoz

való hozzáférés szolgálna megoldással. A Magyarországon hatályos törvényi szabályozás miatt azonban ezeknek az adatoknak az átvétele adatvédelmi szempontból meglehetősen aggályos és (jelenleg még) költséges is, noha a terület sok lehetőséget tartogat (mint említettük, az észt migrációs vizsgálatokban már a gyakorlatban is alkalmaznak passzív helyzeti adatokat).

A mobiltelefonok cellaadatainak felhasználására egy Eurostatos pályázat (grant) keretében a KSH-ban is folyik módszertani kísérlet, ami az időmérleg-felvételek naplózási adatainak Big Data-val való kiváltását célozza. Ennek megvalósításához egy mobil applikáció is készül, melynek segítségével képesek leszünk az okostelefonok GPS-adatait összevetni az időmérlegnaplót kitöltő személyek válaszaival.

5.2. A szenzoros adatok alkalmazási lehetőségei a hazai hivatalos statisztikában

Az utazások és a mobilitás vizsgálatában szenzoros Big Data-források is használhatók. Mint azt bemutattuk, Hollandia jól kiépített útszenzoros rendszerrel rendelkezik, ugyanakkor Magyarországon is egyre több ilyen jellegű forrás létezik (például a Nemzeti Útdíj Szolgáltató vagy az Országos Rendőr-főkapitányság kameraadatai). Ezek alapján, megfelelő kódolási, adatvédelmi technikák alkalmazásával nemcsak a határátlépések számát lehetne becsülni, de a migráció, illetve az ingázási és a turisztikai szokások is megfigyelhetők lennének.

A szállítás- és gépjármű-statisztikában már több esetben adminisztratív adatforrásokból való adatátvétellel állítjuk elő a statisztikai adatokat. Szenzorokkal a jelenleginél gyakrabban és gyorsabban lehetne ezt megtenni, illetve más szempontok (terület, honosság, típus) is vizsgálhatók lennének. A hivatalos statisztikán kívül még más területeken (például a várostervezésben és a közlekedés átalakításában) is jól hasznosíthatók lennének a szenzoros adatok.

Az ún. „okos mérők” (smart meters) képesek a környezet (hőmérsékleti, légnyomás-, szén-dioxid-szint- stb.) adatait eltárolni, az ezekből származó információk pedig az energia- és a környezetstatisztikát segíthetik. Az e téren jelentkező probléma abban gyökerezik, hogy hiába működnek és végeznek percenként mérést már most is szenzorok, az adatátadás ennél ritkábban történik (a KSH többnyire havi, negyedéves, éves adatokat kap). Más adatokkal összekapcsolva viszont e szenzorok lehetővé tehetnék, hogy valós idejű képet kapjunk egy város működéséről, így további dimenziók szerint is vizsgálhatók lennének például az energiafogyasztási vagy a közlekedési adatok.[17]

[17] A Massachusettsi Technológiai Intézet „Senseable City Lab” (Városkutató Laboratórium) elnevezésű projektje keretében szenzorok segítségével, valós időben figyelik a városban történő (például energiafogyasztási, közlekedési) eseményeket.

5.3. A web scraping módszer alkalmazási lehetőségei a hazai hivatalos statisztikában

Jelenleg az árindex kalkulálásához szükséges adatok nagy részét a KSH összeírói gyűjtik a kijelölt üzletekben, kisebb részük pedig online felületekről származik. A web scraping módszerrel azonban, mellyel webes felületekről, strukturált formába rendezve nyerhetők adatok egy speciális szoftvert használatával, letölthetők az ingatlanközvetítői oldalak adatai, és ezáltal becsülhető a lakáspiaci árak változása. Ugyanilyen módon szerezhetne a hivatal információkat a fogyasztói árindex legtöbb összetevőjéről is.

Szintén a vállalatok oldalait elemezve olyan információkat is találhatunk az információs és kommunikációs eszközök alkalmazásáról az üzleti életben, amelyeket a jelenleg használt kérdőívek tartalmaznak.[18] Az Olasz Statisztikai Hivatal tapasztalatai szerint a web scrapinggel gyűjtött adatok jól kiegészítik az adatfelvételekkel szerzetteket (*Barcaroli et al*. [2014]).

E módszerrel az álláshirdető portálokról, jellemzően a meghirdetett álláshelyekről is nyerhetők információk (milyen városban/cégnél, milyen típusú munkaerőre van szükség), amelyek felhasználhatók az üres álláshelyek számának statisztikai becsléséhez.

5.4. Folyamatgenerált adatok alkalmazási lehetőségei a hazai hivatalos statisztikában

A hivatalos statisztikai adatgyűjtésben a KSH háztartási költségvetési felvétele szolgál alapul a háztartások fogyasztási karakterisztikájának vizsgálatához. A mintába került válaszadóknak az év során fogyasztási naplót kell vezetniük, amelyben tételesen feltüntetik a megvásárolt termékeket azok mennyiségével és árával együtt. Ez a felvétel meglehetősen nagy terhet ró a válaszadókra, nélküle azonban nem rendelkeznénk adatokkal a háztartások fogyasztási jellemzőiről és kiadásairól. A folyamatgenerált adatok (például a bankkártyás fizetések eredményeként keletkező vagy az üzletek eladási adatai, melyek esetén tételesen látjuk, hogy mikor, hol, milyen áron, mennyi terméket vettek meg) jelentősen növelnék a fogyasztásstatisztikai adatok minőségét és pontosságát.

Ugyancsak a folyamatgenerált adatok kategóriájába tartoznak a Nemzeti Adó- és Vámhivatal online pénztárgépadatai is. Ezek mennyiségük és gyakoriságuk alapján már Big Data-nak számítanak; átvételük kiváló lehetőséget teremtene arra, hogy a KSH kiegészítse adatfelvételi adatait.

[18] OSAP 1840: Az információs és kommunikációs technológiák állományának minőségi és mennyiségi adatai.

6. A hazai napi repülőjegyárak figyelése és árindexszámítás web scraping módszerrel

A hazai fogyasztói árindex[19] számításának egy apró szelete a repülőjegyárak változásának követése. Ehhez az adatgyűjtés jelenleg manuális úton folyik, melylyel azonban csak korlátozott mennyiségű adat szerezhető, és az árváltozások gyors követésére sincs mód. A repülőjegyárak figyelése Big Data-alapú módszerrel éppen ezekre a problémákra kínál megoldást. Egy, a KSH-nál jelenleg is futó projekt keretében web scraping módszerrel automatikus árgyűjtés folyik a Google repülőjegy-keresőjével. Céljainktól függően az „adatleszívás" gyakorisága tetszőlegesen beállítható, napi többszöri, akár óránkénti/percenkénti árváltozás is figyelhető.

A projekttel a KSH elsősorban a jelenlegi manuális adatgyűjtést kívánja „reprodukálni". A web scraping azonban nemcsak gyorsítja, hanem egyszerűsíti is ezt a folyamatot, azaz a minőségjavításon túl az adatgyűjtés reformja önmagában is növeli a hatékonyságot (ez a technika nem igényel humán erőforrást, így a felszabaduló munkaerő más elemzési, fejlesztési feladatokra csoportosítható át). Az információkhoz ily módon gyorsan és költséghatékonyan (ingyen) lehet hozzájutni, az árváltozást akár napi szinten is össze lehet hasonlítani.

A módszer hátránya, hogy a hivatal munkatársai egyelőre nem minden repülőjegy-árusító weboldalt figyelnek.

6.1. Módszertan

A projektben budapesti indulással négy úti célt, Rómát, Berlint, Londont és Párizst tekintve vizsgáljuk a repülőjegyek árának változását. Az utazási időszak minden hónap 10-12-e (+/– 2 nap). Az árakat a tárgyhónap előtt öt hónappal (például a júliusi utazásra februártól júniusig, napi szinten) gyűjtjük, majd ezekből átlagot számolva képezzük az árindexet.

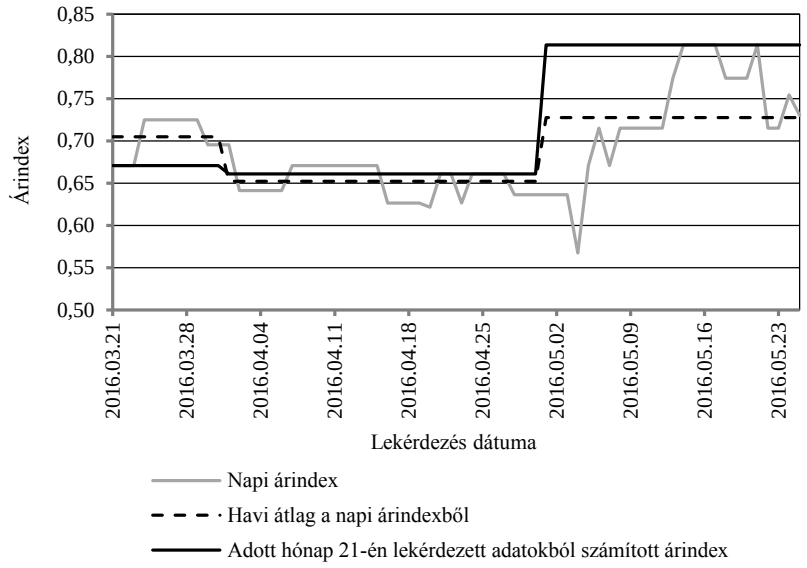
A hagyományos adatgyűjtési módszerrel a hivatal havonta egy árat gyűjt minden úti célhoz, a web scraping alkalmazásával viszont napi gyakorisággal állnak rendelkezésre adatok, így a feldolgozási lehetőségek is bővülnek. Számolható a napi minimumárakból például havi átlagár, vizsgálható az árak szóródása, vagy kalkulálható akár napi szintű árindex. Változtatható ezen kívül a bázis is: lehet az előző év azonos időszaka vagy az előző év egy tetszőleges hónapja.

[19] A fogyasztói árindex a lakosság (a háztartások) által vásárolt termékek, igénybe vett szolgáltatások fogyasztói árainak átlagos változását, röviden a fogyasztói árszínvonal változását mérő mutatószám.

6.2. Eredmények

A havi árindexek kiszámolásához a KSH a tárgyhónapot megelőző öt hónapon keresztül gyűjti az adatokat. A pilot projekt során viszont csak két hónapnyi adatot sikerült gyűjteni, így az ezekből számolt árindex nem összevethető a hivatali módszertannal, de a mintázatbeli jellemzők már vizsgálhatók. (Folyamatos, öt hónapon keresztül folyó web scrapinggel azonban már össze tudnánk vetni az eredményeket.) A 3. ábrán, ami a napi árindex, az abból számított havi átlagos árindex és a minden hónap 21-én lekérdezett adatokból kalkulált árindex alakulását mutatja be, megfigyelhetők a napi és a havi árindex közötti eltérések.

3. ábra. A Párizsba tartó repülőjáratok jegyárindexe, 2016. március 21. – május 26.
(bázisidőszak: 2015. december)



Megjegyzés. Az ábra a Párizsba tartó, 2016. júliusi repülőjáratokra szóló jegyfoglalásokra vonatkozik.
Forrás: Saját ábra.

Ebben az esetben a Big Data a minőségi, a pontossági és a gyorsasági követelmények terén kínál előrelépést. A nemrég indult projekt következő lépéseként a KSH az adatokat a jelenleginél hosszabb idősoron vizsgálja majd, illetve összehasonlítja őket a manuális adatgyűjtéssel szerzettekkel.

7. Összefoglalás

Számos hivatalos statisztikával foglalkozó fórumon elhangzanak a következő kérdések: „Miért olyan érdekes a Big Data a hivatalos statisztika számára?" „Miért nem elegendők a hagyományos adatgyűjtési technikák?" Ezekre a tanulmányban leírtak alapján a következő érvekkel válaszolhatunk:

Finanszírozási kényszerek. A 2007-ben kirobbant gazdasági válság nemcsak a piaci szférát, de a hivatalos állami szervezeteket is arra kényszerítette, hogy tevékenységük finanszírozásához a korábbinál költséghatékonyabb módszereket találjanak. A hagyományos adatfelvételek drágák lehetnek, ezért a hivatalos statisztikának más, alternatív adatforrások után kell néznie. Az adminisztratív adatokban rejlő potenciál kihasználása mellett a Big Data-ra mint alternatív adatforrásra lehet építeni, remélve azt, hogy adatfelvételi költségekkel az utóbbi esetén már nem kell számolni.

A társadalmi és a piaci változásokra való reakción túl, a Big Data-típusú adatforrások használata aktívan is alakíthatja a hivatalos statisztikai munkát. Itt elsősorban a következő területeken mutatkozhat előrelépés:

A hagyományos adatgyűjtések minőségének javítása. A hagyományos adatgyűjtések számos problémával küzdenek. A Big Data-val olyan kiegészítő információkhoz juthatunk, amelyek segíthetnek abban, hogy a jelenleginél jobb és könnyebben karbantartható mintavételi kerettel dolgozzunk, fejleszthessük a mintavételi technikákat, pontosabb kalibrációs, becslési és imputálási eljárásokat dolgozzunk ki, a más forrásokból (például a hagyományosnak számító adatgyűjtésekből vagy adatátvételből) származó adatokat validáljuk, csökkentsük a nemválaszolási arányt, vagy árnyaljuk annak jellegzetes karakterisztikáját (bizonyos társadalmi csoportok nehezen vagy szinte egyáltalán nem elérhetők kérdőíves módszerekkel), illetve gazdagítsuk az adatok elemzési tárházát.

A Big Data használatával *csökkenthetők lennének a válaszadói terhek*. Ez a szempont feltehetően nem szorul különösebb magyarázatra. Minden adatszolgáltató − vegyen részt akár lakossági, akár gazdaságstatisztikai adatgyűjtésben − üdvözli, ha rövidebb kérdőívvel, űrlappal keressük meg, így kevesebb időt kell a hivatalos statisztikai adatszolgáltatással töltenie. Amennyiben az adat más forrásból is hozzáférhető, szükségtelenné válik az adatszolgáltatók megkérdezése.

A Big Data aktív használatának az egyik legnagyobb előnye az, hogy *új ismeretek, korábbi technikákkal nem gyűjthető adattípusok szerezhetők, új összefüggések tárhatók fel*, amelyekre az óriási adat-

halmazok hiányában nem derülhetett volna fény, valamint olyan *innovatív eszközök, módszertanok hozhatók létre*, amelyek később akár mérföldkőnek bizonyulhatnak a hivatalos statisztikai eljárásokban. A rövid távú célokat tekintve úgy véljük, a Big Data használatával lehetőség nyílhat újfajta jóléti indikátorok kidolgozására, az általános gazdasági, mezőgazdasági és környezetstatisztika több szempontú összekapcsolására, a háztartási fogyasztás- és jövedelemfelvételek kiegészítéseként új mérési technikák kidolgozására, a fogyasztói bizalom mérésére és a fogyasztói magatartás megértésére.

Az itt felsorolt szempontok a Big Data alkalmazási lehetőségeinek csak töredékét képezik. Azonban továbbra is sok kérdésünk maradt, technikai és szakmai jellegűek egyaránt. Abban bizonyosak vagyunk, hogy a statisztikai adatgyűjtés olyan paradigmaváltás előtt áll, ami gyökeresen megváltoztatja a hivatalos statisztika mivoltát. Az irányok kidolgozásában a következő kérdések megválaszolása segíthet: „Mi a célunk, reprodukció vagy egy új számítási módszer kialakítása?" „Mi a teendő, ha statisztikusoknak nincsenek a Big Data kezeléséhez megfelelő IT-eszközeik és -szakértelmük?" „A statisztikus inkább az IT-tudását (például a programozási nyelvek ismeretét) fejlessze, vagy inkább váljon adattudóssá, hogy hatékonyan kezelhesse az új szemléletet?" „Beépíthető-e a jelenlegi adat-előállítási folyamatba a Big Data?" „Megbízhatóbb, pontosabb lesz-e ettől a statisztika?" „Gyorsaság vs. pontosság, avagy mi a hivatalos statisztika feladata? A kettő közül melyik a fontosabb?" Véleményünk szerint az utóbbi két tényező között ki kell alakítani az egyensúlyt, hiszen a cél az, hogy ne csak gyors eredményeket, de módszertani garanciát is tudjunk biztosítani.

Meglátásunk szerint a Big Data, hasonlóan az adatfelvételi módszereknek mára már integráns részévé vált internetes survey-ekhez, ugyancsak megtalálja majd a helyét a hivatalos statisztikában anélkül, hogy a hagyományos adatgyűjtési eljárásokat feleslegessé tenné.

Függelék

F1. táblázat

Az adatgyűjtések és a Big Data-jellegű adatforrások minőségi dimenzióinak összehasonlítása

Minőségi dimenzió	Adatgyűjtés-jellegű adatforrás	Big Data-forrás
Lefedettség	– ismert, kontrollálható – valószínűségi mintavétel – teljes körű megfigyelésre van lehetőség	– gyakran nincs róla információ – a lefedettségi hiány vagy többlet nehezen kezelhető
Adatforrásból előállított statisztikák adatvédelme	– szabályozott	– összetett kérdés
Metaadatok elérhetősége az adatok megértéséhez és felhasználásához	– metaadatok rendelkezésre állnak	– metaadatok korlátozottan állnak rendelkezésre, illetve korlátozottan hozzáférhetők/megismerhetők
Pontosság (az adatok mennyire helyesen írják le a jelenséget)	– a nem mintavételi hiba mérhető; a mintavételi hiba számszerűsíthető/jellemezhető	– jellemzően nem mérhető/jellemezhető
Használhatóság (a nemzeti statisztikai hivatalok mennyire tudnak az adatokkal dolgozni anélkül, hogy a specializált források jelentősen leterhelnék meglevő forrásaikat, illetve mennyire könnyen tudják ezeket integrálni rendszerükbe és szabványaikba)	– nem igényel specializált szaktudást és erőforrást	– specializált szaktudásra és erőforrásra lehet szükség – a meglevő rendszerbe való integrálás összetett feladat
Alapadatok elérhetősége és időszerűség	– adott hónapban/negyedévben/évben csak egy bizonyos időszakot vagy időpontot figyelnek meg – az adatok beérkezésének van átfutási ideje	– lehetőség van napi vagy annál gyakoribb megfigyelésre – az adatok elvileg azonnal rendelkezésre állnak
Reprezentativitás	– bizonyos jellemzőkre vizsgálható, értelmezhető	– nehezen vizsgálható, értelmezhető
Érvényesség	– az adatgyűjtés azt tudja mérni, amit a statisztika jellemezni szeretne	– a statisztika nem határozhatja meg, hogy a Big Data-forrás mit mérjen; lehetnek viszont olyan elemek is, amelyeket mérni szeretnénk, de azokat a forrás nem tartalmazza

(A táblázat folytatása a következő oldalon.)

(Folytatás.)

Minőségi dimenzió	Adatgyűjtés-jellegű adatforrás	Big Data-forrás
Hozzáférhetőség (mennyire könnyen tudnak hozzájutni a felhasználók az adatokhoz, metaadatokhoz) és egyértelműség (rendelkezésre állnak-e világos, egyértelműen leírt* adatok)	– adatvédelmi szabályok korlátozzák – a metaadatokhoz való hozzáférés biztosított – egyértelmű, módszertani információk érhetők el	– típusuktól függően könnyebben vagy nehezebben hozzáférhetők – a metaadatokról többnyire hiányos információk állnak rendelkezésre
Relevancia (mennyire felel meg a statisztikai termék a felhasználói igényeknek)	– módosítható adatforrás, ami a felhasználói igények, illetve a hazai és a nemzetközi elvárások alapján alakítható; a változó igényekhez alkalmazkodva át kell alakítani az adatgyűjtést	– módosuló adatforrás: a felhasználói igények nem az adatgyűjtésnél értelmezett módon érvényesülnek; technikai, képességbeli fejlődés is változást indukál

* Az egyértelmű leírásokhoz tartoznak például a megfigyelési egységek, változók definíciói, az adatkezelésre vonatkozó leírások (eljárások, technikák stb.).

Megjegyzés. Az adatforrásból előállított statisztikák adatvédelme mind az adatgazdákra, mind a nemzeti statisztikai szervezetekre vonatkozik. Ezek, tekintettel a jogi és szervezeti korlátokra, valamint a titoktartási és adatvédelmi kérdésekre, jelentősen befolyásolhatják az adatok tervezett felhasználását. Fontos az adatszolgáltatók adatvédelme, hiszen bizalmas információt nyújtanak, és az általuk közölt adatok kizárólag statisztikai célra használhatók.

A pontosságot a statisztikai becslések esetén a hiba szempontjából jellemezhetjük, ami két komponensből épül fel: a torzításból (szisztematikus hiba) és a szórásból (véletlen hiba). Hibaforrásnak tekinthető a lefedettség, a mintavétel, a nemválaszolás stb.

Az időszerűség és a gyakoriság a Big Data két legfontosabb minőségi aspektusa. Bár sok esetben ezek biztosítják a legnagyobb hozzáadott értéket, más minőségi faktorok rovására is mehetnek.

Forrás: Saját összeállítás az *UNECE Big Data Quality Task Team* [2014] alapján.

F2. táblázat

Lehetséges felhasználási területek a KSH-ban adatforrások szerint

Adattípus	Statisztika	KSH-adatfelvétel/publikált adat
	Mobilkommunikáció	
Mobiladatok	Turisztika, népességstatisztika, migráció	Népszámlálás (mobilitási adatok tízévenkénti publikálása) OSAP 2290 – Állandó népesség településenkénti adatai (nemenkénti és koréves bontásban) OSAP 2228 – Nemzetközi vándorlásban részt vevő állampolgárok adatai OSAP 1943 – A külföldiek magyarországi turisztikai és egyéb kiadásai OSAP 1114 – Alapadatok a fizikai és szellemi foglalkozásúak munkaidőmérlegéhez

(A táblázat folytatása a következő oldalon.)

(Folytatás.)

Adattípus	Statisztika	KSH-adatfelvétel/publikált adat
	Internet	
Internetes keresések	Munkaerő-statisztika, migrációs statisztika	OSAP 2238 – Havi munkaügyi jelentés OSAP 2009 – Jelentés a betöltött és az üres álláshelyek számáról OSAP 1114 – Alapadatok a fizikai és szellemi foglalkozásúak munkaidőmérlegéhez
E-kereskedelmek oldalai	Árstatisztika	OSAP 1009 – Fogyasztói árösszeírás (fogyasztói árindex) OSAP 1712 – Jelentés a lakás- és lakótelek-forgalom alakulásáról (lakáspiaci árak) OSAP 1007 – Az ipari termékek és szolgáltatások árjelentése OSAP 1831 – Építőipari tevékenységek ára OSAP 2193 – Adatszolgáltatás az Európai Unión kívüli külkereskedelmi termékforgalomról OSAP 2130 – Az üzleti szolgáltatások kibocsátási árjelentése
Vállalatok oldalai	Információs társadalom statisztikája	OSAP 1840 – Az információs és kommunikációs technológiák állományának minőségi és mennyiségi adatai
Vállalatok oldalai	Gazdasági Szervezetek Regisztere	Gazdasági Szervezetek Regiszterének pontosítása
Álláshirdetők oldalai	Üres álláshelyek statisztikája	OSAP 2009 – Jelentés a betöltött és az üres álláshelyek számáról
Ingatlanhirdetések oldalai	Árstatisztika (ingatlanpiac)	OSAP 1712 – Jelentés a lakás- és lakótelek-forgalom alakulásáról OSAP 2418 – Építési költségbecslés standard lakástípusokra
Közösségi média	Fogyasztói elégedettség, GDP, információs társadalom	GDP

(A táblázat folytatása a következő oldalon.)

(Folytatás.)

Adattípus	Statisztika	KSH-adatfelvétel/publikált adat
	Szenzoros adatforrás	
Közlekedési szenzorok	Közlekedés-/szállításstatisztika, turisztika	OSAP 1390/03 – A helyi közutak és hidak adatai OSAP 2297 – Közúti járműállomány OSAP 1183 – A közúti személyszállítás adatai OSAP 1189 – Közúti és kötöttpályás személyszállítási teljesítmények OSAP 1654 – A közúti teherszállítás belföldi és nemzetközi teljesítményei Népszámlálás (mobilitási adatok tízévenkénti publikálása) OSAP 2290 – Állandó népesség településenkénti adatai (nemenkénti és koréves bontásban) OSAP 2228 – Nemzetközi vándorlásban részt vevő állampolgárok adatai OSAP 1943 – A külföldiek magyarországi turisztikai és egyéb kiadásai OSAP 1114 – Alapadatok a fizikai és szellemi foglalkozásúak munkaidőmérlegéhez
„Okos” mérőeszközök	Energiastatisztika	Energia és környezet: OSAP 1321 – Energiamérleg, ipari szektor OSAP 2221 – Energiamérleg, energiaszektor, energiahordozók OSAP 1324 – Jelentés a távhőtermelők és távhőszolgáltatók adatairól OSAP 1329 – Energiagazdálkodási operatív jelentés OSAP 1335 – Energiafelhasználási beszámoló
Műholdképek	Mezőgazdasági, földhasználati, környezetvédelmi statisztika	Földhasználat: OSAP 1082 – Földterület és vetésterület, május 31. OSAP 2218 – Egyéni gazdaságok júniusi összeírása OSAP 1709 – Tájékoztató az országos jelentőségű védett természeti területekről és a Natura 2000 területekről

(A táblázat folytatása a következő oldalon.)

(Folytatás.)

Adattípus	Statisztika	KSH-adatfelvétel/publikált adat
Repülőgép-mozgások	Szállítási és levegőszennyezési statisztika	Légi szállítás: OSAP 1725 – Repülőterek forgalmi adatai OSAP 1966 – Jelentés a repülőterek forgalmáról OSAP 2160 – Vízi, légi és csővezetékes szállítási teljesítmények OSAP 1066 – Levegőtisztaság-védelmi adatok
	Folyamatgenerált tranzakció	
Szupermarketek szkennerei és eladási adatai	Árstatisztika, háztartásifogyasztás-statisztika	OSAP 1006 – Fogyasztói árösszeírás (fogyasztói árindex) OSAP 2153 – Háztartási költségvetési és életkörülmény adatfelvétel, naplóvezetés OSAP 2154 – Háztartási költségvetési és életkörülmény adatfelvétel, éves kikérdezés OSAP 1045 – Jelentés a kiskereskedelem és vendéglátás eladási forgalmáról OSAP 1646 – Jelentés a kiskereskedelem és vendéglátás eladási forgalmáról árucsoportonként OSAP 2130 – Az üzleti szolgáltatások kibocsátási árjelentése
Pénzügyi tranzakciós adatok	Háztartásifogyasztás-statisztika	OSAP 2153 – Háztartási költségvetési és életkörülmény adatfelvétel, naplóvezetés OSAP 2154 – Háztartási költségvetési és életkörülmény adatfelvétel, éves kikérdezés
Önként szerkesztett földrajzi információk (VGI, weboldalak [OpenStreetMap, Wikimapia, Geowiki])	Földhasználat	OSAP 1082 – Földterület és vetésterület, május 31. OSAP2218 – Egyéni gazdaságok júniusi összeírása

Megjegyzés. VGI (volunteered geographic information): önkéntes térinformáció.
Forrás: Saját összeállítás.

Irodalom

BARCAROLI, G. – NURRA, A. – SCARNÒ, M. – SUMMA, D. [2014]: *Use of Web Scraping and Text Mining Techniques in the Istat Survey on* "*Information and Communication Technology in Enterprises*". Istat. http://www.q2014.at/fileadmin/user_upload/Iad_in_ICT_survey_PAPER.pdf

COUPER, M. P. [2013]: Is the sky falling? New technology, changing media and the future of surveys. *Survey Research Methods*. Vol. 7. No. 3. pp. 145–156. http://dx.doi.org/10.18148/srm/2013.v7i3.5751

DAAS, P. J. H. – PUTS, M. J. H. [2014]: *Social Media Sentiment and Consumer Confidence*. Statistics Paper Series No. 5. European Central Bank. Frankfurt.

DAAS, P. J. H. – VAN DER LOO, M. [2013]: *Big Data and Official Statistics*. United Nations Economic Commission for Europe, Eurostat, Organisation for Economic Cooperation and Development, United Nations Economic and Social Commission for Asia and the Pacific. Discussion paper. https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.50/2013/Topic_4_Daas.pdf

DEVAN, A. [2016]: *The 7 V's of Big Data*. Impact Radius blog. https://www.impactradius.com/blog/7-vs-big-data/

FUKUYAMA, F. [1995]: *Trust*. The Free Press. New York.

GARTNER, INC. [2017]: *Big Data*. IT Glossary. http://www.gartner.com/it-glossary/big-data/

GLASSON, M. – TREPANIER, J. – PATRUNO, V. – DAAS, P. – SKALIOTIS, M. – KHAN, A. [2013]: *What Does "Big Data" Mean for Official Statistics?* Paper prepared for the High-Level Group for the Modernization of Statistical Production and Services. 10 March. http://www1.unece.org/stat/platform/pages/viewpage.action?pageId=77170622

GROVES, R. M. [2011]: Three eras of survey research. *Public Opinion Quarterly*. Vol. 75. No. 5. pp. 861–871. https://doi.org/10.1093/poq/nfr057

INFODIAGRAM.COM [2014]: *Visualizing Big Data concepts – Strong and Loose Relation Diagram*. https://blog.infodiagram.com/2014/04/visualizing-big-data-concepts-strong.html

ISTAT ESTP (EUROPEAN STATISTICAL TRAINING PROGRAM) [2016]: "*Introduction to Big Data and its tools*" Course. 29 February – 2 March. Rome.

KNACKS, S. – KEEFER, P. [1997]: Does Social Capital have an economic payoff? *Quarterly Journal of Economics*. Vol. 112. No. 4. pp. 1251–1288. https://doi.org/10.1162/003355300555475

PLÉH CS. – UNOKA ZS. [2016]: *Hány barátod is van?* Oriold és társai Kft. Budapest.

RAISER, M. – HAERPFER, C. – NOWOTNY, T. – WALLACE, C. [2001]: *Social Capital in Transition: A First Look at the Evidence.* EBRD Working paper. No. 61. EBRD. London.

SCANNAPIECO, M. – VIRGILLITO, A. – ZARDETTO, D. [2013]: *Placing Big Data in Official Statistics*: *A Big Challenge*? Eurostat, Collaboration in Research and Methodology for Official Statistics. https://ec.europa.eu/eurostat/cros/system/files/NTTS2013fullPaper_214.pdf

TAM, S.-M. – CLARKE, F. [2015]: *Big Data, Statistical Inference and Official Statistics*. Research Paper No. 1351.0.55.054. Australian Bureau of Statistics. Canberra. http://www.ausstats.abs.gov.au/Ausstats/subscriber.nsf/0/015937BADB90186BCA257E0B000E428A/$File/1351055054_mar%202015.pdf

UNECE BIG DATA QUALITY TASK TEAM [2014]: *A Suggested Framework for the Quality of Big Data*. December. http://www1.unece.org/stat/platform/download/attachments/108102944/Big%20Data%20Quality%20Framework%20-%20final-%20Jan08-2015.pdf?version=1&modificationDate=1420725063663&api=v2

VALE, S. [2013]: *Classification of Types of Big Data*. UNECE Statistics Wikis. http://www1.unece.org/stat/platform/display/bigdata/Classification+of+Types+of+Big+Data

VUKOVICH G. [2015]: Adatforradalom és hivatalos statisztika. *Statisztikai Szemle*. 93. évf. 8–9. sz. 745–758. old.

Summary

The paper presents the main methodological dilemmas as well as the advantages and disadvantages of using Big Data sources in official statistics. It tries to find answers for the following questions: "Can Big Data be integrated into the existing data collection (and production) processes?" "Is it the right solution?" or "Is it better to use such massive volumes of data only for validation?"

The authors identify those fields in official statistics where Big Data can be used and give insight into one of the projects of the Hungarian Central Statistical Office that uses Big Data sources for data production.