BOGÁROMI ESZTER – MÁTH ANDRÁS

Sokforrású adatbázis-építés – buktatók, nehézségek, megoldási kísérletek

Esettanulmány

Elemezni kell, információkra valamennyi döntéshozónak szüksége van ahhoz, hogy a saját területén érdemi, megalapozott döntéseket hozzon. A kutató hivatása éppen ezen döntések megalapozása, hátterének kialakítása és támogatása. A döntéselőkészítéshez szükséges anyagok, információk gyűjtése és csoportosítása egyre inkább eltolódik a megkérdezéses adatgyűjtés felől a megfigyeléses információszerzés, valamint a sokforrású adatbázisok összekapcsolása felé. Ennek oka nem elsősorban a gyorsabb elemzésben kereshető (gyakran a felmerülő nehézségek akár még több időt is igényelhetnek egy egyszerűbb megkérdezéses vizsgálatnál), hanem az elemzési szint mélysége (túl sok helyen és túl sok embert kellene kérdezni), valamint a felhasznált információk nonkonverzatív jellege miatt. Az adatbázisok forrása, struktúrája, időbeli érvényessége és a lehetséges kapcsolódási pontok egymáshoz rendelése minden esetben eltérő megfontolásokat követelnek meg – éppen ezért ezek ismerete, aktív tudása és érvényesítése a kulcsa egy-egy jól lezárt és értékes kutatásnak.

A kutatás szó használatánál talán érdemes is megállni egy pillanatra. Ennek a kifejezésnek az értelme a társadalom-kutatásban és a piackutatásban egyre inkább megváltozik. Amíg korábban teljesen természetes volt, hogy kutatáson valamilyen adatgyűjtést (értsd: kérdezés, aktív részvétel) és annak feldolgozását, elemzését értettünk, addig 2015-re eljutottunk oda, hogy a gyakorlat egyre kisebb szeletét teszi ki az ilyen jellegű vizsgálati módszerek sora – a megfigyelések, passzív adatgyűjtések segítségével létrejövő adatbázisok és értékelések javára.[1] Ez a folyamat a *„kutatás"* kifejezés fokozatos megváltozását is magával hozza, hiszen nagyon sokan még manapság is kutatáson az adatfelvé-

[1] Ugyanakkor azt is fontos leszögezni, hogy az elemzési szempontokat nem kezelő adatgyűjtések (és jelenleg még inkább ezek vannak túlsúlyban) értéktelen – bár sokszor irdatlan méretű – adatbázisokat eredményeznek).

➤ Educatio 2015/3. **Bogáromi Eszter – Máth András:** *Sokforrású adatbázis-építés – buktatók, nehézségek, megoldási kísérletek*, 86–97. pp.

telt értik; holott – főként a fiatalabb nemzedék számára – már inkább az elemzés, értékelés jelentést fedi le.

Az alábbiakban összegzett esettanulmány arra szolgál példaként, hogy egy komplex elemzési feladatban milyen akadályok merültek fel, sok, egymástól számottevően eltérő céllal, tartalommal készült adatbázis összeépítésénél, és milyen megoldások vezettek el egy sikeres elemzés kivitelezéséhez – immár egy komplex értékelési környezet felállítását követően. A sikert elsősorban azon mérhettük le, hogy a döntéshozók milyen gyakorisággal térnek vissza az elemzésre, milyen későbbi döntésekben köszönnek vissza az ott elkészített szegmensek, települési értékelések.

Az esettanulmányból bizonyos konkrét információkat kénytelenek voltunk mellőzni vagy megváltoztatni a kutatás egyedisége miatt – de ez a folyamat vagy a felmerült problémák megértése szempontjából, reméljük, nem okoz nehézséget.

A feladat

Egy magyar nagyvállalat termékportfóliója átrendeződésekor lényegesen nagyobb figyelemmel fordult egyes megyék települései felé. Ez az üzleti változás generálta azt az igényt a cég vezetői részéről, hogy szükségük volt a magyar településszerkezet mélyebb ismeretére (ehhez még nem kellett volna adatbázisokat összeépíteni, mert erről vannak kiváló tanulmányok), de amikor a beszélgetésekben az is kiderült, hogy az egyes üzleti aktivitások tervezéséhez településtipológiákat kell majd alkotni, akkorra vált egyértelművé, hogy a lehetőségek szerint legtöbb elérhető adatbázist kell felhasználni a munka során.

Tegyünk itt egy rövid kitérőt az üzleti világba, hogy jobban megérthessük, mire és hogyan volt szükséges a megrendelőnek, vagyis mire akarta használni a végeredményeket.

Az elsődleges cél és elképzelés az volt, hogy a szűkös üzletszerzési (marketing) erőforrásokat olyan területekre érdemes koncentrálni, ahol azok nagyobb valószínűséggel hoznak növekvő eladásokat – vagyis a kutatásnak célpontokat kellett kijelölnie a megrendelő számára. Ehhez természetesen azonosítottunk értékelési paramétereket – mint jövedelmi viszonyszámok, népességarányok és továbbiak. Ez a feladat arra is szolgált, hogy a meglévő értékesítési eredményeket értékelni lehessen – egyfajta elvárt/megvalósult mutatószámokat is fel lehessen vázolni – mintegy megmutatva a területen dolgozó munkatársak hatékonyságát.

A másik megfogalmazott igény a hasonlóságok meghatározása volt – vagyis hogy az egyes tesztkörnyezetben végzett aktivitások és az azokra adott reakciók után meg lehessen találni, hogy milyen más településeken várható hasonló reakció a fogyasztók részéről. Könnyű belátni ezen kérés hasznosságát: számottevően lehet csökkenteni a rosszul vagy feleslegesen elköltött marketingkiadások mértékét egy-egy próba kivitelezésével. Ha valami sikeres egy településen, akkor azt meg lehet kísérelni más helyen is – ha tudjuk, hogy mely települések tekinthetőek a szociokulturális, gazdasági és más egyéb szempontok szerint hasonlónak.

Ezekből a megfontolásokból kiindulva kettős feladatot fogalmaztunk meg a kutatás során: (A) azonosítsuk és „gyúrjuk össze" lehetőleg egy mutatóba azokat a mérőszámokat, amelyek képesek jellemezni egy-egy települést az üzleti potenciál szempontjából; (B) alkossunk olyan összetett adatbázist, amely képes arra, hogy egymástól akár földrajzilag jelentősen távol eső településeket is azonos szegmensbe soroljon, a lehető legtöbb ér-

vényes gazdasági, szociokulturális, illetve fogyasztási adat alapján. A továbbiakban lépésenként mutatjuk be ennek a feladatnak a megvalósítását .

Az adatbázisok kiválasztása

Amikor sorra vettük a lehetséges dimenziókat, melyek mentén szerettük volna felépíteni az általunk használt modellt, fontosnak tartottuk, hogy ne gondoljunk a hozzáférhetőségre. Tudtuk, hogy vannak adatbázisok, melyek könnyen érhetőek el, másokat létre kell hozni, és megint másokat nem is lehet előállítani. A különböző adatbázisok összekapcsolásának és elemzésének fontos pillanata, amikor még úgy gondoljuk, minden adat elérhető. Ezt a logikát felborítva, jelen tanulmányban nem a valóságban megtörtént adatgyűjtést mutatjuk be, hanem csoportosítva elemezzük a felhasznált forrásokat. Tanulmányunkban a bemutatott adatbázisokat három csoportba soroljuk: ingyenesen hozzáférhető adatbázisok, saját vagy fizetett adatbázisok és az elemzés céljára rögzített adatbázisok.

Az ingyenesen hozzáférhető adatbázisok köre

Az ingyenesen hozzáférhető adatbázisokba azok az adatbázisok tartoznak, melyek már adatbázisba szerkesztve érhetőek el és mindenki számára ingyenesek. Az adatok felhasználásának céljától is függ a termékek árazása. A tudományos céllal felhasznált adatok, egyedi adatszolgáltatások kutatók és egyetemi hallgatók részére gyakran ingyenesen érhetőek el. Elemzésünket piaci felhasználásra készítettük, így ebbe a kategóriába csak azok az adatbázisok tartoznak, melyek mindenki számára ingyenesek. A adatbázisba szerkesztettnek tekintjük azokat az adatforrásokat, melyeket táblázatos formában lehet menteni. Értelemszerűen nem tartoznak ide az interaktív grafikonok és térképek. Utóbbiak az eredmények bemutatásánál lehetnek hasznosak, könnyen értelmezhetővé és látványossá tehetik a prezentálást.

A legszélesebb skálán a *Központi Statisztikai Hivatal* (KSH) mutatói mozognak, melyek a Tájékoztatási Adatbázis Területi statisztikái között találhatóak.[2] A területi statisztikák között négy forrás mutatói közül válogathatunk. Az *Éves településstatisztikai adatok a 2013-as településszerkezetben* adatbázis mindig az adott időszakban legfrissebb mutatókat tartalmazza Magyarország minden településére. A nagyobb települések irányítószám szerint nincsenek bontva, aggregált eredmények találhatóak a kimutatásban. Kutatásunk idején a 2012-es adatok voltak elérhetőek. Külön adatbázisban érhetőek el *Budapest kerületeinek adatai*, az egyéb települések adataihoz igazodó frissülésben. Végül a területi statisztikák között találhatóak a *Népszámlálási adatok járások és Budapest kerületei* szerint. Mivel feladatunk a Budapesten kívüli települések jellemzése volt, a budapesti kerületekre és a járásokra vonatkozó adatbázisokat nem használtuk. A KSH területi statisztikáit tartalmazó online felületen könnyen lehívhatóak az adatok; magunk válogathatjuk össze, hogy milyen dimenziókat szánunk a sorokba, oszlopokba és munkalapokra. Az eredményeket végül menthetjük és exportálhatjuk több formátumban is (grafikon, pdf és excel).

[2] Forrás: http://statinfo.ksh.hu/Statinfo/themeSelector.jsp?page=2&szst=T

A településstatisztika tizenhat témakörben kínál több száz mutatót. A témakörök többek között kiterjednek az alap demográfiai kimutatásokra, lakásállományra, oktatásra, turizmusra, közművekre, munkanélküliségre, önkormányzati segélyezésre is. A mutatók éves aggregált adatokat tartalmaznak régiókra, megyékre és településekre.

A *Népszámlálás 2011* adatai a KSH által hozzáférhetőek település szinten is a Népszámlálás saját felületén.[3] Megyék szerint külön-külön letölthetőek a népességre vonatkozó, településsoros adatok. A megyéken belül külön kimutatás található az egyes településtípusokra, járásokra, kistérségekre, települések mérete szerinti csoportokra és az egyes településekre.

A mindenki számára hozzáférhető adatsorok közé tartoznak az *Országgyűlési választások* és az *Önkormányzati választások eredményei*. Elemzésünk idején az 2010-es országgyűlési választások eredményei számítottak a legfrissebb eredménynek, ezért a listás szavazati arányokat használtuk fel a települések jellemzésénél.[4] Az Országgyűlési választások internetes oldalán számos szempont szerinti bontásban érhetőek el az eredmények, így az egyéni választókerületek eredményei, a területi választókerületek eredményei mind listás szavazatok, mind elnyert mandátumok szerint, valamint a pártok eredményei területi listákon. Továbbá egészen a szavazóköri eredményekig bontott eredményeket is megismerhetjük.[5] Az oldal vélt funkciója egy-egy választókerület vagy település eredményeinek megismerése és nem országos adatok exportálása, így aki arra vállalkozik, hogy minden településre megismerje az eredményeket, annak a kutatás időtervének kialakításánál tervezni kell a hosszabb adatbeszerzési idővel.

A nyilvánosan elérhető adatoknak egyik körét alkotják azok a kimutatások, melyekhez csak igénylés útján lehet hozzájutni. Az igénylésnél a kutató által megadott szempontok jelennek meg elsősorban, de előfordulhat, hogy a kért kimutatások nem állíthatóak elő, esetleg nem is rögzítettek. Elemzésünkhöz a bűnelkövetők statisztikáit kívántuk beszerezni, és azon belül is arra voltunk kíváncsiak, hogy a nők ellen elkövetett bűncselekmények száma mennyire jellemző az adott településre. A *bűncselekményekre vonatkozó adatokat* a KSH Népesedési és szociális védelmi statisztikai főosztályáról igényeltük. A főosztálytól megyei bontásban kaptunk adatokat, majd a Belügyminisztériumhoz irányítottak minket, ahonnan a lehető legrészletesebb adatokat lehet igényelni a bűnelkövetés témakörében. Településszinten bontott adatokhoz így a *Belügyminisztériumon* keresztül jutottunk, de további nehézségekbe ütköztünk, melyekről a következő fejezetben írunk. Az adatok igénylésénél a kutatónak pontosan meg kell határoznia, hogy milyen kimutatásban, milyen mutatókra van szüksége, más esetben az igénylési folyamat elhúzódhat. A bűncselekményeken túl más témakörben is igényelhetők adatok az egyes minisztériumoktól és más állami intézményektől.

Nem nyilvános adatsorok

A nyilvánosan nem elérhető adatsorok fontos eleme volt a megbízó (Kiadó) *értékesítési adatai*, melyekkel a települések anyagi helyzetét, kulturális nyitottságát és fogyasztási szokásait jellemeztük. A vállalatok napról napra bővülő, frissülő adatsorok birtokában

[3] Forrás: http://www.ksh.hu/nepszamlalas/tablak_teruleti_04

[4] Forrás: http://www.valasztas.hu/hu/parval2010/354/354_0_index.html

[5] Forrás: http://www.valasztas.hu/hu/parval2010/354/354_0_index.html

vannak. A különböző szervezeti egységek más és más bontásban, tematikában állítják elő és kérik be az egyes mutatókat, melyek belső elemzéseknél felhasználhatóak. Ezeknek az adatoknak a felhasználása nehezebb, mint a beszerzésük. El kellett döntenünk, hogy milyen időszakra és milyen mutatókra vonatkozó adatsorokat illesztünk az adatbázisunkba; az egyes mutatók átlagait vagy kumulált eredményeit számoljuk az egyes településekre. A választás sokszínűségét mutatja, hogy vannak vállalatok, ahol napi, vagy akár órás bontásban állnak rendelkezésre árushely szintű adatok, minden termékre külön lebontva, illetve vannak vállalatok, ahol csak havi bontásban, nagyobb földrajzi egységekre nézve elemezhetőek az eredmények. Minél részletesebb és kisebb időegységekre vonatkozó adatokkal dolgozunk, annál tudatosabban és módszertanilag alátámasztottan kell megválasztani az elemzési egységeket, mutatókat. Saját kutatásunkban indokolt volt a termékek külön mutatóban való megjelenítése az eltérő olvasóközönség és a termékekhez kapcsolódó eltérő fogyasztási szokások miatt. Az eltérő termékekhez eltérő periodicitás is tartozik, így külön kellett döntést hozni a napilapok bemutatására és a heti- vagy havilapokéra. Átlagos példányszámokkal dolgoztunk, melyeket ezer főre vetítettünk ki. Az árushelyi eladások mellett külön változóként az előfizetéseket is beépítettük az adatbázisunkba.

A nem ingyenesen hozzáférhető adatsorok közé tartozik a *GfK Vásárlóerő kutatása* is, mely számunkra két értékes adatbázissal szolgált, egyfelől az irányítószámokra, másrészt a településekre számolt vásárlóerővel. A két adatbázis két külön adatsorként szolgál. A vásárlóerő-kutatás eredményei évente jelennek meg. A kutatásban öt dimenzió jelenik meg, a népességszám (főben és ezrelékben), a háztartásszám (darabban és ezrelékben), az összes vásárlóerő (millió euróban és ezrelékben), az egy főre jutó vásárlóerő és az egy háztartásra jutó vásárlóerő. A vásárlóerő-dimenziók három mutatót foglalnak magukba, az euró értéket, a magyarországi indexet és az európai indexet. Elemzésünkben az egy háztartásra jutó vásárlóerővel dolgoztunk. Ezzel a mutatóval a települések gazdasági erejét és a lakók gazdasági helyzetét kívántuk ábrázolni.

A települések jellemzésénél a televíziós nézettségi adatok beemelését is fontosnak tartottuk. A *Nielsen Közönségmérés* 1120 háztartásból álló panelen végzi a hivatalos nézettségi mérést. A panel nagysága és reprezentativitása csak a régiós szintű adatok használatát teszi lehetővé, ezért elemzésünkben ezeket a kimutatásokat nem tudtuk használni. Olyan kutatásoknál, melyek nagyobb földrajzi egységekre vagy nagyobb létszámú célcsoportokra vonatkoznak, a Nielsen Közönségmérés megfelelő hátteret biztosíthat. A felismerés után sem szerettük volna kihagyni az elemzésünkből a televíziózással töltött időt mint mutatót, ezért más adatforrás felkutatása mellett döntöttünk. Hazánkban egyedül a Nielsen Közönségmérés méri szoftver segítségével a televíziózást, de más elemzésekben – önbevallásos módon – megtalálható a keresett kérdés.

A *Nemzeti Olvasottság Kutatásban* (későbbiekben NOK) is kitérnek arra, hogy egy átlagos hétköznapon és egy átlagos hétvégi napon átlagosan mennyi időt töltenek tévénézéssel. A Nemzeti Olvasottság Kutatást – a *TGI (Target Group Index)* termékkel összekapcsolva – a TNS Hoffmann végzi folyamatos adatfelvétel mellett. Évente négy alkalommal adják ki a legfrissebb olvasottsági adatokat, évente két alkalommal a TGI termékhez tartozó mutatókat. A két kutatásban összesen több száz mutató érhető el, melyek kiterjednek a média területére, az FMCG termékek használatára, vásárlására, különböző attitűd-állításokra, érdeklődési körökre, pénzügyi termékekre és bizonyos életvitelt érintő kérdésekre. Mind a NOK, mind a TGI kutatásban is elérhetőek bizonyos szeg-

mentált változók, így az ESOMAR státusz változó, a TGI fogyasztói státusz, TGI Early Adopter szegmentáció, az életszakaszokat megkülönböztető változó és az informatikai ellátottságra vonatkozó mutató is. Évente a NOK-kutatásban részt vevő válaszolók száma eléri a 25.000 főt, így kisebb célcsoportok is biztosan vizsgálhatók a kutatásban. A TGI adatfelvétele személyes, papír alapú kérdőívvel történik a megkérdezett otthonában, míg a NOK kérdőívet kérdezőbiztos kérdezi le (laptop használatával) a megkérdezett otthonában. A személyek kiválasztásánál valószínűségi véletlen mintát alkalmaznak. Településszintű kimutatások nem érhetőek el a TGI-NOK kutatásban, viszont településméret, megye és régió változó használható az elemzések alkalmával. A TGI és NOK termékek változói közül – a televíziónézésen túl – az elemzésünk során használtuk a különböző lapkategóriák olvasási adatait, az attitűd-állításokat is.

Elemzés céljára rögzített adatbázisok

A hozzáférhető adatbázisok nem feltétlenül tartalmaznak minden változót és adatsort, amit a kutatás során a modellbe be kíván építeni a kutató. A hozzáférhetetlen adatsorokat akár elő is lehet állítani az elemzés céljára. Kutatásunkban két adatsort állítottunk elő, az egyik a közlekedést tartalmazta, a másik a településeken elérhető áruházak listáját.

A megfelelő *közlekedés hiánya* elszeparálhatja a kisebb települések lakóit, illetve a könnyű és dinamikus közlekedés elősegítheti az ingázást a munkahelyekre és serkentheti a szabadidős tevékenységeket is. Az első kérdés, ami felmerült ennél a dimenziónál, a települések beosztása két csoportba. Az egyik csoportba soroltuk azokat a településeket, ahová el kívánnak jutni más településekről a lakók, a másik csoportban találhatóak azok a települések, ahonnan elutaznak – akár nap nap után – más településekre. Közigazgatási szempontból a kisebb települések a kistérségi központjukhoz tartoznak, itt találhatók az okmányirodák, ahol a hivatalos iratokat kell kiállítani, az életvitelt tekintve viszont megeshet, hogy a kistérségi központtól eltérő, közelebbi városba járnak be szórakozni, vásárolni, dolgozni a település lakói. Magyarország összes településénél nem lehetett egyesével mérlegelni a lehetséges ingázási helyeket, így a közigazgatást követve, a kistérségi központokat jelöltük meg a 'cél-városoknak', a kistérségben található többi települést pedig 'kiinduló városoknak'. A városok közötti közlekedést több mutató rögzítésével terveztük megragadni, így mind a vonatközlekedést, mind a távolsági busz közlekedést rögzítettük. A közlekedésnél külön változót alakítottunk ki az induló járatok számának és a menetidőnek is, illetve külön változóban rögzítettük a két település közötti távolságot. A képzett változónk, mely a közlekedés könnyedségét mutatta, azt a közlekedési módot tartalmazta, mely a két település között a domináns lehet. Dominánsnak tekintjük az a közlekedési módot, ahol gyakrabban jár a jármű és rövidebb a menetidő. Végül a menetidőt, a járatok számát kombináltuk a két település távolságával, így azt is megtudhattuk, hogy mennyivel nehezebb tömegközlekedéssel eljutni a kistérségi központba, mint saját autóval. A képzett mutató értékei széles skálán mozogtak.

Az *áruházak számára és fajtájára* azért volt szükségünk, mert így a termékekhez való hozzájutást ismerhetjük meg, mely hatással van mind a fogyasztási szokásokra, mind a közlekedésre. Számos áruházláncot vizsgáltunk meg és rögzítettük telephelyeiket, így minden településnél tudtuk, hogy hány áruház található benne és az áruházak mekkora alapterületűek. Az áruházak minden típusa külön változóba került, majd egy képzett mutatót készítettünk, mely a település ellátottságát mutatta.

Buktatók és megoldáskeresés

Az adatbázisok begyűjtése után kezdődött a munka második fázisa, melyben egy adatbázisba kellett importálni a különböző adatsorokat. Az elemzés célja az egyes települések jellemzése volt, így az esetek értelemszerűen csakis a települések lehettek. Az adatsorok összkapcsolása számos nehézség elé állította a kutatókat. A későbbiekben felsorolt problémák közül sok azért nehezíti meg az elemzést, mert megoldásukra nagyon hosszú időt kell fordítani és egyszerű programozással nem oldhatók meg, a személyes ellenőrzés elkerülhetetlen, ami tovább növeli az elemzésre fordított időkeretet. Fontosnak tartjuk kiemelni, hogy az idő a kutatásnál egy kiemelt fontosságú dimenzió volt, mert az egyes adatbázisok eltérő időpontokra (is) vonatkoztathatnak.

Az első probléma a Nemzeti Olvasottság Kutatás felhasználásánál merült fel, ugyanis ebben az adatbázisban nincsenek a településekre vonatkozó becsült változók. Viszont a minta kellően magas elemszámú, az adatfelvétel hátterében lévő módszertan megbízható és a kutatásban felvonultatott változók egyedülállóak. A hiányosságot a megyék és a településnagyság alkalmazásával hidaltuk át. Egy képzett változót alakítottunk ki, mely a két változó kereszteződéséből született. A képzett változó értékeit az adatbázisunkban is kialakítottuk minden településre, így kapcsoltuk össze a két adatbázist. Több száz változót kapcsoltunk a NOK-kutatásból az adatbázisunkhoz, és az elemzés szakaszára hagytuk a döntést arról, hogy mely változó használható fel és melyik nem.

A második felmerülő kérdés a különböző adatsorok előállításának *időbeli eltérése* volt. Míg a népszámlálást tízévente végzik, addig a terjesztési adatok napról napra, az internetes adatsorok pedig percről percre frissülnek. A területi statisztikák adatfelvételének időpontja az elemzést két évvel előzte meg. Eltérő periodicitás és eltérő adatfelvételi időpont tartozott az egyes adatbázisokhoz. Amennyiben szigorú módszertanhoz tartottuk volna magunkat, nem lett volna két adatbázis, amelyet össze tudtunk volna kapcsolni. Külső adatforrások egyesítésénél a kutatónak meg kell hoznia azt a döntést, hogy milyen időintervallumban fogadja el a különböző adatsorokat. Ez a döntés függ a mutatók érzékenységétől, attól, hogy mennyire változhat az értékük hónapok, évek alatt. Legrégebbi adatok a népszámláláshoz tartoztak. A népszámlálás adatbázisából a vallási megoszlás-mutatókat használtuk. Az egyének felekezetváltása nem annyira jellemző, inkább a településekről való elvándorlás és az oda való bevándorlás lehet jelentősebb hatással a vallásos lakók arányára. Évente lehet nyomon követni a települések halálozási, születési és vándorlási számait, így tesztelhető, hogy melyik településen volt jelentősebb lakosságcsere. Elemzésünk idején a 2010-es népszámlálás volt a legfrissebb adatfelvétel, a KSH adatsorok adatfelvétele 2012-ben történt. A településeknél jelentősebb népességváltozás nem volt megfigyelhető, ami a vallásra vonatkozó mutatókat befolyásolta volna, ezért elfogadtuk, hogy a Népszámlálás 2010 adatsorait használjuk.

Az Országgyűlési választások eredményei is négyévente érhetőek el, ennél az adatbázisnál is a 2010-es adatsorral dolgoztunk.[6] A Központi Statisztikai Hivatal jelentései 2012-es adatokra vonatkoztak, így ezeknél az adatsoroknál a 2012-es év eredményeit illesztettük az adatbázisunkba. A saját rögzítésű adatsorainknál nem állt rendelkezésünkre régebbi állapotra vonatkozó adatsor (Volán menetrend, MÁV menetrend, áruhá-

[6] A kutatás tervezése és előkészítése 2014 tavaszán történt, így nem volt mód a 2014-es eredményeket beépíteni a rendszerbe.

zak listája), így ezeknél az adatsoroknál csak a legfrissebb állapotokat tudtuk megjeleníteni az adatbázisban.

Amennyiben minden adatsort elemzünk, jól látható, hogy az adatfelvétel időpontját tekintve akár három év eltérés is lehet két mutató között. De nem csupán az eltérést kell figyelembe venni, azt is kell mérlegelni, hogy az adott mutató értékeiben milyen mértékű változás állhat be az eltelt idő alatt. Azokat a mutatókat tartottuk meg az adatbázisban, melyeknél az eltérés feltételezett becslése még az elfogadható intervallumon belül volt. Nyilvánvaló, hogy az elfogadottság mértéke minden egyes kutatásnál az elemzési céloktól függ, annak fényében kell egyedi döntéseket hozni.

A következő probléma mellyel szembesültünk, a *települések listája* volt. Az egyes adatbázisokban nem egységes a településlista, eltérő írásmóddal szerepelnek a nevek. A kisebb települések neve változik, több település kötőjellel tartalmaz kiegészítő neveket. Ráadásul vannak adatsorok, melyekben a településnévhez számokat, kódokat is csatolnak, vagy éppen a nemzetközi szoftverek használata miatt ékezetek nélkül írják a neveket.

Az eltérő írásmód mellett a települések listája sem azonos. Alapvetően két megközelítésből lehet listázni a településeket, az egyik az *irányítószám alapú* megközelítés, a második a név szerinti. A két lista nem minden esetben fedi egymást. Az eltérés a legkisebb és legnagyobb települések esetén figyelhető meg. A nagyobb településeknél egy településen belül több irányítószám is megtalálható, míg kisebb településeknél egy irányítószám alá több település is tartozhat. A szállításnál, árusításnál ez nem probléma, viszont egy adatbázis-összevonásnál hibaüzenetként jelenik meg a település. Ezeket a '„hibákat" egyesével kell ellenőrizni, besorolni. A besorolás viszont nem minden esetben egyszerű feladat. Mutatónként kell mérlegelni, hogy a két település értékeit hogyan redukáljuk egy esetre – átlagolással, összeadással, ezer főre számolással stb. Nem minden esetben egyértelmű a döntés és nem minden esetben áll rendelkezésre minden adat. Amennyiben ezer főre számoljuk a mutató értékét, akkor lehet, hogy nem áll rendelkezésünkre az egyik település népességének száma, ebben az esetben viszont esetleg olyan adatokat kell használnunk, melyek forrása és adatfelvételi ideje eltérő. Esetünkben nem kellett ilyen kompromisszumokat kötni, de ez csak a saját adatbázis felépítésének volt köszönhető.

A kisebb települések összevonása nemcsak az irányítószámoknál jelenik meg, hanem az élet számos más területén is. Kisebb településeknél előfordul, hogy csak az egyikben van *orvosi rendelés*, és az három település lakóit szolgálja ki. Az egy év alatt összesen ellátott járóbetegek száma ebben az esetben a három település összes járóbetegére vonatkozik és nemcsak az adott településére. Természetesen lakóhely szerinti bontásban nem áll rendelkezésünkre az előbb említett mutató; ekkor felmerül a kérdés, hogy felhasználható-e a mutató. Arra biztosan nem használható, hogy a lakosság lélekszámához arányosítva tudjuk meg a betegek számát. El kell tehát döntenünk, hogy mire használjuk a mutatót. Amennyiben a település fejlettségét, érzékelt nagyságát kívánjuk megfogni, akkor informatív lehet ez a mutató, melynek értéke nulla, ahol nincs rendelés. Illetve ha a járulékos fogyasztási szokásokat, forgalmat kívánjuk ábrázolni, elemezni – mint például a gyógyszertári forgalom, közlekedés – akkor is informatív lehet a mutató. Az egyes mutatók felhasználásáról minden esetben a kutatónak kell tehát döntenie. Az egyes mutatók értelmezése és felhasználása szerteágazó lehet. A felhasználást ilyen esetekben tudományos értekezésekben pontosan kell definiálni és alátámasztani. Mivel mi az egyes betegségekre voltunk kíváncsiak, így az egyes települések egészségügyi állapotát kívántuk kimutatni, ezt a változót ki kellett emelnünk az adatbázisunkból. A kórházak, kórházi

ágyak számát, szakrendelések számát mutató változókból képzett új változót alkottunk, mely a település fejlettségi szintjét volt hivatott árnyalni.

Az adatbázis rögzítése után, az elemzésnél találkoztunk a következő szembetűnő problémával, mely a *települések lélekszámából* adódott. Magyarországon pár tíz fős, vagy éppen száz fős települések is vannak, és minél kisebb a település, annál szélsőségesebb értékei lehetnek az egyes mutatóknak. Így találkozhatunk olyan településsel, ahol az átlagos 50 százalékos férfi arányt messze elhagyva 75 százalékos a férfiak aránya. Ezek a jelentős eltérések a klaszteranalízis, főkomponens-analízis és más statisztikai eljárások alkalmával borítják fel a modellt. Gyakorlatilag lehetetlenné teszik az érdemi eltérések kimutatását. Ebben az esetben el kell döntenünk, hogy véglegesen kiemeljük a kisebb településeket az adatbázisból, vagy egy külön adatbázisban kezeljük azokat. Bármelyik megoldást is választjuk, mindkét esetben meg kell határozni az lélekszámban mért határt, mely alatt kiemeljük a településeket.

A települések lélekszámához kapcsolódik a másik véglet vizsgálata is. A legnagyobb városok városrészei számos adatsorban megjelennek külön értékekkel (ahol irányítószámot is használnak az azonosításnál, vagy kerületeket, városrészeket). Ezt a megkülönböztetést azért is alkalmazzák a hivataloknál, vállalatoknál, mert viselkedésükben, fogyasztási szokásaikban, életvitelükben, szociodemográfiai összetételükben jelentősen eltérőek lehetnek az egyes városrészek. A kutatónak kell döntenie, hogy külön egységeknek tekinti-e az egyes városrészeket, vagy a nagy városokat is egy településnek tekinti. Amennyiben a kerületeket külön eseteknek tekintjük, akkor előállhat az a probléma, hogy bizonyos adatsorokban nincsenek külön vezetve a kerületek adatai. Ekkor vagy nem használjuk ezeket az adatsorokat, vagy az ezer főre vetített mutatókat, illetve a százalékos arányokat mutató változókat megtartjuk és a város értékeit fűzzük minden kerület értékéhez. Ez nem egy az egyben megfeleltetés, mégis lehet olyan hozzáadott értéke az elemzésnél, ami miatt hasznosabb elfogadni ezt a kompromisszumot, mint elvetni az egész adatsort. Illetve egy harmadik választásunk is lehet, ha a nagyobb városokban külön adatbázisban elemezzük és a csak városokra vonatkozó adatsorokat a többi településnél használjuk fel. Jól látható, hogy a választás nem könnyű, és a mérlegelésnél figyelembe kell vennünk mind a módszertani elvárásokat, mind a mutatók tulajdonságait, a településeken belül megfigyelhető heterogenitást. Tanulmányunkban végül csak nyolc megye adatait elemeztük, adatbázisunkban csak 5-7 nagyobb város volt, melyeknél nem tartottuk indokoltnak a kerületek használatát.

Az adatbázisok felhasználása

Az adatbázisokat végül településnév alapján kötöttük össze. Voltak olyan adatbázisok, melyek egységei az irányítószámok voltak. A településneveknél az azonos irányítószám alá tartozó, eltérő településeknél – kisebb települések esetén figyelhető ez meg – minden településnél az adott irányítószám értékeit adtuk meg.

A végleges adatbázisban több, mint ötszáz változó szerepelt, és annyi eset volt, ahány település van Magyarországon.

Az elemzésnél két irányban indultunk el; egyfelől a különböző megyéket kellett jellemeznünk és minél részletesebb leírást kellett adnunk róluk, a lehető legváltozatosabb nézőpontokból. A másik irány egy dimenziócsökkentő eljárás alkalmazása volt, melynek végeredményeként egy mutatóba kellett sűrítenünk a települések változatosságát.

Megyék jellemzése

A megyék és kistérségek jellemzésénél azért volt hasznos az adatbázis, mert gyorsan és hatékonyan tudtuk egy adatbázisból bemutatni az egyes mutatókat, akár a mutatók kombinálásával. A bemutató anyag 25 témakörre terjedt ki, mint:

– Gazdasági helyzet
– Infrastruktúra
– Népesség és települések megyénként
– Nők helyzete
– Fiatalok (15–29 évesek)
– Idősek jellemzése
– Gyerekes családok
– Életmód, életvitel attitűdök
– Oktatás
– Egészségügyi helyzet
– Vallás és pártok
– Településszerkezet, utazás
– Lapolvasási szokások
– Megyei napilap-olvasók
– Sajtóhasználat-attitűd
– Tv lefedettség, napi hatókör
– Kereskedelem
– Vásárlási attitűdök
– Turizmus, nevezetességek
– Munkaügyek
– Online aktivitás
– Saját oldalak átlagos havi látogatószáma
– Átlagos havi affinitás-index
– Megyei napilap online

Az elemzés különlegességét az adta, hogy a megyék jellemzésében kollégáink nem csak az alap mutatókat ismerhették meg. Több, mint 350 változó bemutatása által egy sokkal részletesebb és érdekesebb képet láthattak országunkról. A változók kombinálásával jobban megérthették az ott élők életvitelét, érzéseit, fogyasztási szokásait, értékrendjét. A fogyasztási szokások és az életvitel megismerése pedig hozzájárul a lapok eladásainak megértéséhez.

Példának okáért kevesen látják egyben, hogy míg Jász-Nagykun-Szolnok megyében a települések 45%-áról mondható el, hogy mind busszal, mind vonattal meg lehet közelíteni a kistérségi központot, addig Baranya megyében ez az arány csupán 11%. Baranyában a települések 86%-a csak autóbusz-összeköttetésben van a kistérségi központtal. Továbbá megtudhattuk az elemzésből, hogy Baranya megyében csupán 1524 lakos jut egy háziorvosra, míg Bács-Kiskun megyében 1824 lakos. A gyerekorvosoknál Baranya megyében 807 gyerekre jut egy gyerekorvos, míg Bács-Kiskun megyében 926 gyerekre jut egy háziorvos, Heves megyében pedig 1168 gyerekre. További érdekesség volt, hogy a távfűtésbe bekapcsolt lakások száma Baranya megyében és Komárom-Esztergom megyében volt a legmagasabb, míg Békés megyében a legalacsonyabb, és hogy a játszóterek, tornapályák, pihenőhelyek száma Komárom-Esztergom megyében a legmagasabb. A megyei szerkesz-

tőségekben dolgozók jól ismerhetik a megyéjükben élők életvitelét, azt viszont kevésbé tudhatják, hogy ez mennyire általános vagy különleges az országban. A budapesti kollégák részletes tudása az egyes megyékről úgyszintén hiányos volt. Az elemzésnek köszönhetően egy részletes térképet kaptak, melyből mindenki a munkájához kapcsolódó mutatókat használhatta fel.

Dimenziócsökkentés

Az adatbázis létrehozásának egy másik fontos, ha nem a legfontosabb célja egy besorolás volt, mely számtalan dimenziót figyelembe véve jellemzi az egyes településeket. A jellemzésre több okból is szükségünk volt. Az egyik ok az egyes eladásokra vonatkozó becslő eljárások és más elemzések támogatása volt. A fogyasztási szokásokban bekövetkező változások megértéséhez szükségünk van arra, hogy a célcsoport környezetét is ismerjük, illetve egy-egy termék bevezetésénél többek között tudnunk kell, hogy az adott településen milyen a vásárlóerő, mennyire nyitottak kulturálisan az adott termékre. A másik ok a kvalitatív kutatások támogatása volt. A kvalitatív kutatások alkalmával a vidéki helyszínek kiválasztásánál a megyeszékhelyeken túli városokat szoktuk választani. Több helyszín megjelölésénél nemcsak az országon belüli régiók vagy megyék szerinti eltérést szerettük volna figyelembe venni, hanem a kiválasztásoknál célunk volt, hogy más életvitelű, fogyasztású településeket jelöljünk ki. Feltételezésünk szerint két különböző megyében lehetnek hasonló tulajdonságokkal bíró települések. Ezeket a hasonlóságokat és különbözőségeket szerettük volna megragadni.

Különböző adatredukciós módszereket alkalmaztunk, melyek során – utólag már triviálisnak mondható – akadályokba ütköztünk. Az egyik akadály a kis lélekszámú települések szélsőséges értékei voltak. Ahol pár tíz fő él a településen, akár csak a nemüket tekintve is szélsőséges értékeket kaphatunk, melyek redukálják a modellek hatékonyságát. Ilyen esetekben például a települések többsége egy vagy két csoportba került és a szélsőséges csoportok alkottak külön csoportokat. A kutatásunk üzleti alkalmazásra készült, ezért az üzleti célokat figyelembe véve kellett mérlegelnünk. Mivel a nagyon alacsony lélekszámú települések terjesztési szempontból gyakran külön kezelt települések, ezeket ki tudtuk emelni az adatbázisból – és rájuk egy másik elemzésnél külön fókuszáltunk. A lélekszám-határt a terjesztési gyakorlatot figyelembe véve és a módszertani elvárásokat követve határoztuk meg.

A végleges adatredukciónál végül 218 változót használtunk fel, és a városokat 14 csoportba soroltuk be. A 14 csoporton belül lélekszám szerint megkülönböztettünk még négy alcsoportot, így az 5.000 fő alatti települések, az 5.001–10.000 fő közötti települések, a 10.001–20.000 fő közötti települések és a 20.001 fő feletti települések csoportját. A lélekszám szerinti megkülönböztetésre a szervezések alkalmával is szükségünk van, amikor a szűrésnél kiválasztott célcsoportok nagyságával kell kalkulálnunk. A bevont változók között szerepelnek az anyagi helyzetre, a vallási hovatartozásra, a kulturális értékekre, életvitelre, politikai beállítottságra, médiatermékek fogyasztására és az egészségügyi állapotukra vonatkozó változók is.

Végszóként

Tapasztalataink szerint sokforrású adatbázisokat érdemes összeépíteni. Az eredmények bemutatását követően ügyfeleink rendkívül sokat profitáltak az eredményekből, nagyon sok meglepő, hasznos megállapítást sikerült átadnunk számukra. Egy piaci igényeket kiszolgáló kutató számára a legnagyobb siker, ha egy-egy eredményeket bemutató prezentáció általános hivatkozási alap lesz a vállalati munkában. Ennek a kutatásnak a slide-jai a mai napig (2015 ősz) időnként előkerülnek a vezetők asztaláról, újabb esetekben is visszanyúlva hozzájuk.

Az újabb beszélgetések során viszont egyre gyakrabban felmerül egy nagyon érdekes kérdés, miszerint meddig érvényesek az elemzésben tett megállapítások, besorolások. A válasz nem lehet egységes; minden egyes alkalommal meg kell nézni az eredeti kiinduló adatbázisok érvényességét (és itt nem elsősorban az a kérdés, hogy van-e frissebb, hanem az, hogy egy esetleges új adatfelvétel esetén mennyire lennének mások az eredmények). A mi esetünkben egyes elemek ugyan kis mértékben változtak, de a legfontosabb dimenziók (utazási feltételek, vásárlási szokások, médiahasználati szokások) még nem avultak el. Ha összegezni kellene, akkor éppen a Kiadónak legfontosabb téma, a média-, és ezen belül a technológiai eszközhasználat változik a leggyorsabban. Éppen ezért szinte biztos, hogy ezt a kutatást 2016-2017 körül újra el kell készíteni abban az esetben, ha a megrendelőnek erre szüksége lesz.

IRODALOM

TNS Hoffmann & Millward Brown, Nemzeti Olvasottság Kutatás (NOK), Olvasottsági jelentés 2015/2. negyedév