Földrajzi Közlemények 2015. 139. 2. pp. 118–131.

A WEBRŐL GEOKÓDOLT TARTALMAK TÉRBELI ELOSZLÁS-VIZSGÁLATA: TERÜLETI EGYENLŐTLENSÉGEK ÚJ NÉZŐPONTBÓL

JAKOBI ÁKOS

EXAMINING SPATIAL DISTRIBUTION OF GEOCODED WEB CONTENT: NEW ASPECTS OF REGIONAL INEQUALITIES

Abstract

Although contents of the internet are basically placeless, there is still the possibility to geographically identify web content. This study introduces a crawler methodology of collecting large amount of content data that could be connected to addresses with known geographical position. By analysing geocoded web content completely new aspects of regional inequalities of the information age appear. Queries of keywords reflect that geocoded web content is not spatially random but follows certain geographical characteristics of the society. On the other hand, visual interpretations of results revealed new inequality patterns and served as evidences of presumed but not yet tested assumptions.

Keywords: web content, big data, regional inequalities, geocoding, cybergeography

Bevezetés

Napjaink rohamos információtechnológiai fejlődése olyan szolgáltatásokat és alkalmazásokat hívott életre, amelyek merőben új eszközöket és lehetőségeket kínálnak a területi kutatók számára is. Az új lehetőségek leginkább abból fakadnak, hogy az információs és kommunikációs technológiák (IKT) mára szinte mindenhová beszivárogtak, s a számítógépes megoldások már-már az élet minden szféráját áthatották. Az angol nyelvű szakirodalomban „pervasive computing" vagy „ubiquitous computing" kifejezés (SATYANARAYANAN, M. 2001; FRIEDEWALD, M.RAABE, O. 2011; WEISER, M. 1991) hátterében pedig a térbeli információs eszközhasználat szétterjedése és a térbeli információrobbanás is megtalálható (JIANG, B. – YAO, X. 2006; GALLOWAY, A. 2004, ZOOK, M.A. et al. 2004). A területi folyamatokat kutatók számára pedig éppen abban nyílik új potenciál, hogy ezek a térbeli információk egyre szélesebb körben válnak hozzáférhetővé és a társadalom mind sokszínűbb térbeli működési mechanizmusába kínálnak bepillantást vagy legalábbis jobb rálátást.

A társadalom térbeli folyamatainak megértéséhez minden eddiginél nagyobb minták állnak rendelkezésre. A szakmai körökben „big data" (szabad fordításban „óriási adathalmaz") néven ismert kifejezés arra a hatalmas adatmennyiségre utal, amely információs világunkban nagy sebességgel és folyamatosan keletkezik, s amelynek feldolgozása a hagyományos kapacitásokkal és eljárásokkal operáló módszerekkel már-már megoldhatatlan kihívást jelent. A big data emellett ugyanakkor nagy lehetőségeket is kínál. A sokáig csak virtuális melléktermékként számon tartott napi információhalom ugyanis épp akkor válik értékessé, amikor a különböző adatokat sikerül összekötni, köztük összefüggéseket, felismerhető mintázatokat találni, s mindebből értékelhető következtetéseket levonni. A világhálón közzétett strukturált vagy strukturálatlan térbeli információtartalom vagy az egyre szélesebb körben terjedő, térbeli információkat is használó alkalmazások révén pedig a big data a társadalom térbeli működésének megértéséhez is megszámlálhatatlan mennyiségben kínál új forrásokat.

Amiket az ún. big data környezetben a területi kutatók haszonnal vizsgálhatnak, azok az úton útfélen hagyott térbeli tartalommal is rendelkező direkt vagy indirekt digitális nyomok. Az efféle adatokra épülő adatbázisok közvetlen módon például az okostelefonok különböző helyalkalmazásaihoz kötődően keletkeznek (ezeket végső soron a felhasználó állíthatja be), vagy például egyes honlapok célzott geotagekkel, azaz földrajzi azonosító kódokkal való ellátásakor. Ám ennél is jóval érdekesebbek a geoinformációkat tartalmazó digitális nyomok indirekt halmazai, melyek nem szándékoltan, de mégis nagy számban keletkeznek. Példaként említhetők azok az elektronikus közlekedési kártyák vagy megfigyelő rendszerek, amelyek rögzítik a közlekedési rendszerbe való belépés és kilépés helyét és idejét, lehetőséget adva – elméletileg – a közlekedési térpályák, szokások stb. vizsgálatára. Digitális nyomokat hagyunk továbbá akkor is, amikor egy-egy weboldalt meglátogatunk, hiszen (általában) beazonosítható az az IP-cím, s ezáltal az a földrajzi hely is, ahonnan a világháló szolgáltatásait igénybe vettük. A digitális nyomok indirekt felhasználására, elemzésére persze számos más példa is említhető (például Girardin, F. et al. 2008; 2009; Järv, O. 2012; Naaman, M. 2011), melyek mind a „melléktermékként" keletkező digitális adatok vizsgálatával hozzák meg következtetéseiket.

Annak ellenére, hogy a világhálóra felkerülő tartalmak alapvetően térfüggetlenek, mégis akadnak olyan megoldások, amelyek az egyes weboldalak földrajzi azonosítását is elősegítik (erről részletesebben lásd Jakobi Á. 2014). Nemcsak arra lehetünk kíváncsiak, hogy kik és hol tesznek közzé információkat az internetes világban, de arra is, hogy miről, avagy mely helyekről közölnek tipikusan információkat a felhasználók. Természetesen az, hogy szemantikai értelemben milyen tartalmú információ kerül az internetre, területi szempontból általánosságban nehezen vizsgálható és nem is könnyen értelmezhető, ellenben az, hogy az egyes információk hol keletkeznek, illetve az egyes információtípusokra hol kíváncsiak, már elemezhető. Új lehetőségként az online tartalmak szövegbányászati módszerekkel történő feldolgozása említhető, ami lehetővé teszi például, hogy számszerűsített módon is meghatározható lehessen egyes helyek és terek online reprezentációja, avagy adott helyhez kötődő tudattartalmak minősége vagy nagysága.

Jelen tanulmány egy a fentihez hasonló technikát alkalmazva tesz kísérletet arra, hogy a webes tartalmak földrajzi azonosításával a társadalom térbeli differenciáira következtessen. Feltevésünk szerint a webes tartalmak térbeli mintázatai nem véletlenszerűek, hanem többé-kevésbé a „tradicionális földrajzi tér" társadalmi térbeli sajátosságait követik. Ugyanakkor fordított módon az is igaz lehet, hogy a webről geokódolt tartalmak térképi vizualizációjával és földrajzi értelmezésével a társadalom területi szerveződésének új aspektusai is megismerhetőkké válnak. Az eddig nem látott, vagy csak sejtett, de tapasztalati bizonyítékokkal alá nem támasztott területi evidenciák megfogalmazásában ezért új lehetőségként használjuk ki a kialakított big data adathalmazokat.

Az eddigi eljárások néhány tapasztalata

A témával foglalkozó legtöbb eddigi vizsgálat a legnépszerűbb közösségi oldalak vagy nagy forgalmú gyűjtőoldalak gigantikus adathalmazainak geotag adatait elemezte. A geotag adatok a weboldalakon a közzétett információhoz kapcsolódva jelentek meg, ami azt eredményezte, hogy ezzel az amúgy térfüggetlen információk is térbelivé váltak (éppen ez kínált nagy lehetőséget az újszerű elemzésekre). Másrészt viszont a fent említet adatok az adott oldalakon közzétett információk immanens részeiként rögzültek, tehát az információk a közzétételkor már eleve rendelkeztek valamiféle térbeli azonosítási lehetőséggel, nem volt szükséges azok utólagos geokódolása.

Graham és Zook (2011) például a Wikipedia közismert oldalainak geotaggel ellátott bejegyzéseit elemezte ilyen formában. Munkájuk során összegyűjtötték és területi adatbázisba rendezték az egyes Wikipedia oldalak HTML forrásában fellelhető geotageket, majd ezekből térinformatikai eljárásokkal térképeket készítettek. A kapott ábrák már érdemi új eredményekkel szolgáltak az információs társadalom térbeli szerveződésének megértéséhez, jól szemléltetve a közzétett tartalmak és közvetve a felhasználók területi eloszlásának egyenlőtlenségeit.

Számos kutató vizsgálta és vizsgálja még ma is a Twitter közösségi hálóján publikált tartalmakat földrajzi szempontból, kihasználva azt, hogy a közzétett információk itt gyakorta térbeli azonosítókkal együtt rögzülnek (pl. Graham M. et al. 2013; Cuevas R. et al. 2014). Leetaru és munkatársai (2013) tanulmányukban a georeferált Twitter bejegyzések sokszínű elemzési lehetőségeit mutatják be, minden esetben a bejegyzésekhez kapcsolt földrajzi metaadatokra építve megállapításaikat. Az elérhető geoadatok köre itt kétféle: egyrészt lehet településekre vonatkozó, amit a Twitter felhasználók manuálisan állítanak be egy menürendszer segítségével, másrészt lehet pontos földrajzi lokációt jelölő koordináta-pár, melyet általában a GPS és egyéb celluláris helymeghatározó alkalmazások szolgáltatnak. A település megjelölését a felhasználók a Twitter által felkínált listából választják ki, főleg akkor, ha a közösségi portált asztali vagy fix helyzetű eszközön keresztül használják. Ezt a helymegjelölést a felhasználó manuálisan kell, hogy frissítse, így az utazáskor küldött bejegyzések esetenként csak a legutóbb választott lokáció szerint rögzülnek. Ezzel ellentétben a pontos helykoordinátákat közlő mobilalkalmazásokat használó változatban a felhasználónak semmi dolga sincs, hogy a rá vonatkozó helyinformációkat frissítse; ez tehát automatikusan megtörténik. A felhasználók aktuális helyzetét a bejegyzések közzétételekor négy tizedes pontosságú koordinátaértékekkel rögzítik, ami lehetővé teszi, hogy az alkalmazást használók helyzetének pontos utca, házszám, vagy épület szintű beazonosítása is lehetséges legyen (személyiségi jogi kockázatok miatt a felhasználóknak engedélyezniük kell persze az ilyen pontosságú térbeli azonosítást). Egy átlagos napon, Leetaru és munkatársai (2013) kutatásai szerint, a bejegyzések 2,02%-a tartalmaz földrajzi metaadatokat, 1,8% települési megjelöléssel, 1,6% pontos helykoordinátával, de előfordul, hogy egy bejegyzés mindkettővel rendelkezik. Mindezen adatkör ugyanakkor elégségesnek és megfelelően nagynak tűnik így is, hogy a társadalom térbeli működésének sajátosságait, sőt esetleg részleteit is megismerhessük (lásd pl. Graham M. et al. 2013).

Eric Fisher (2013) munkájában az ugyancsak rendkívül nagyszámú napi adatot produkáló Flickr fotómegosztó oldal bejegyzéseit vizsgálta. Itt a közzétett fotókhoz kapcsolt geotag adatok kínáltak lehetőséget arra, hogy a publikált tartalmak térbeli eloszlását vizsgálni lehessen. Ráadásul a fotókhoz vagy a felvételeket feltöltő felhasználókhoz kapcsolódó egyéb attribútumok (például hogy helyi vagy nem helyi az illető) alapján már tematikus vizsgálatok elvégzésére is lehetőség nyílt. A végeredményként kapott vizuális ábrázolások nemcsak érdekesek, de a webes szolgáltatások ezen speciális változatát használók társadalmi térbeli sajátosságairól is új információkkal szolgáltak.

A fentebb ismertetett példák mindegyike tehát a felhasználói információk hátterében eleve rögzített geoadatok feldolgozására épült, következésképpen nem alkalmazott önálló geokódolási mechanizmusokat. A világhálón azonban számos olyan tartalom lelhető fel, amely földrajzilag azonosítható ugyan, de geotag, azaz eleve hozzárendelt földrajzi azonosító nem kapcsolódik hozzá. Ilyen esetekben önálló és utólagos geokódolási feladatok elvégzésére van szükség. A következő fejezetek egy ilyen eljárás eredményeit ismertetik.

A vizsgálat módszere

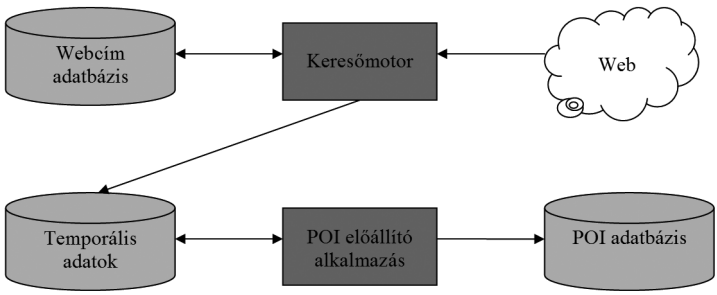
A weben található gigantikus információmennyiség teljes feldolgozása lényegében lehetetlen feladat. Célzott megoldásokkal, jól strukturált alkalmazásokkal azonban a web viszonylag nagy részéről lehet sikerrel értelmezhető adatokhoz jutni. Az adatbányászat ilyen formája irányulhat kifejezetten a földrajzi helyzetet jelölő adatok keresésére is, amiket végül adatbázisba rendezve lehetőség nyílik a kapcsolódó tartalmak térbeli elemzésére. Vizsgálatunk célja tehát az volt, hogy a lehető legtöbb földrajzi azonosítást lehetővé tevő információt összegyűjtsük a világhálóról (legalábbis annak egy definiáltan nagy részéről), majd ezt követően annak meghatározása, hogy az egyes oldalakon közzétett webtartalmak a Föld mely térbeli pontjaihoz köthetően értelmezhetők. Itt tehát nem arra voltunk kíváncsiak, hogy az adott weboldalt hol jegyezték be (erre bőven léteznek különféle statisztikák), hanem arra, hogy a közzétett *tartalom* mely térbeli pontra vonatkozik.

A vizsgálati eljárás során a weboldalakon található geoadatok (jellemzően címek, postacímek) szolgáltak segítségül az adott oldalon közzétett tartalom térbeli relevanciájának meghatározásában. Egy kiválasztott weboldalon ugyanis nagy eséllyel feltételezhető, hogy az ott közölt cím az oldalon közzétett információkhoz kapcsolódik. Ha egy cég például a termékeit vagy szolgáltatásait ismerteti a honlapján, akkor gyakorta találjuk meg a cég elérhetőségét is a weboldalon, így a honlapon fellelhető tartalom rögtön a cég megadott címéhez köthetően is értelmezhetővé válik a földrajzi térben. Példaként említhetők az olyan weboldalak is, ahol egy intézmény, egy szervezet, egy étterem stb. mutatkozik be, vagy amelyen csak hírek, információk kerülnek közzétételre az egyes vonatkozó címek mellett. Természetesen, ha egy weblapon nem található semmiféle címinformáció, akkor annak tartalma nem is válik a térben beazonosíthatóvá, másrészt, ha egy weboldalon több cím is található, akkor a tartalom több térbeli ponthoz köthetően is rögzíthető.

A webes tartalmak geokódolását az ESRI Magyarország Kft. G-Search technológiájával végeztük, ami egy térkép-alapú keresőmotorra épül. A keresőmotor (crawler) általánosságban egy olyan alkalmazás, ami bizonyos feltételeknek (többnyire egy szónak vagy kifejezésnek) megfelelő információkat keres számítógépes környezetben. A webes keresőmotor az internet állományait gyűjti össze és rendszerezi automatikus módon. Az adatgyűjtés során a kereső-folyamat kiolvassa a hivatkozásokat a letöltött állományokból, mellyel ezután további állományokat tud letölteni, így a folyamat automatikusan halad, nem igényel manuális munkát. A térkép-alapú keresőmotor ezen túlmenően nemcsak linkeket és metaadatokat gyűjt, hanem földrajzi címeket is. A címek kinyerése azonban önmagában nem elégséges, a kapott adatoknak ún. geokódolási folyamaton is át kell esniük, azaz koordinátákká kell konvertálni azokat, hogy térképre lehessen helyezni a találatokat.

Az adatgyűjtési folyamat során *(1. ábra)* először egy (magyarországi) webcím-adatbázist volt szükséges kialakítani. Ehhez kezdetként egy, az ESRI Magyarország Kft. által korábban készített kiindulási webcím-adatbázist alkalmaztunk, amelyből a keresőmotor elindíthatta az újabb webcímek felkeresését (a legfontosabb magyarországi webcímek tára általában az adatszolgáltatóktól megvásárolható). A meglátogatott weboldalakról a keresőmotor újabb linkeket, továbbá földrajzi címeket, koordinátákat, meta-adatokat nyert ki és tárolt el egy átmeneti (szaknyelven temporális) adatbázisba. A kapott adatokat ezek után végül az ESRI Magyarország Kft. alkalmazása segítségével POI adatbázisba rendeztük (POI: points of interest, hasznos helyek, érdekes pontok). Ez az alkalmazás intelligens algoritmusok futtatásával címpontokat vont össze, továbbá POI nevet és releváns szöveges tartalmat állapított meg az egyes földrajzi címekhez és mentett el a POI adatbázisba (lásd www.gsearch.hu). Végső soron ez az adatbázis lett az, amelyből – egyszerű vagy komp-

lex lekérdezésekkel – a webes tartalmak területi elemzéséhez, térképi megjelenítéséhez szükséges adatok végül kinyerhetővé váltak.



1. ábra A vizsgálati adatbázis kialakításának folyamata
Figure 1 The process of creating the research database

A webes tartalmak földrajzi beazonosítása természetesen nem lehet mindig tökéletes, sőt előfordulhat az is, hogy egy honlapon közzétett cím egy téves földrajzi pontot azonosít be, vagy legalábbis bizonytalan annak eldöntése, hogy a közölt tartalom tényleg a weboldalon megjelölt címhez rendelhető-e. Ennek meghatározása, tisztázása rendkívül bonyolult feladat, s csak komplex intelligens algoritmusok alkalmazásával végezhető el, több-kevesebb sikerrel. Jelen vizsgálat ugyancsak alkalmazott ilyen korrekciós mechanizmusokat, melyekkel legalább a találatok földrajzi azonosításának megbízhatóságát sikerült meghatározni. Eredményként minden találat mellett egy megbízhatósági értéket (score-t) is eltárolt az adatbázis.

A fenti eljárással kialakított adatbázis térképi és területi elemzések elvégzésére közvetlenül még nem alkalmas, már csak a tárolt adatok mennyisége miatt sem. A célzott térbeli elemzésekhez megfelelően szűkített lekérdezésekkel tudtunk adattáblákat létrehozni, avagy kulcsszavak alapján kikerestük azokat a találatokat, amelyek a számunkra fontos információt tartalmazták.

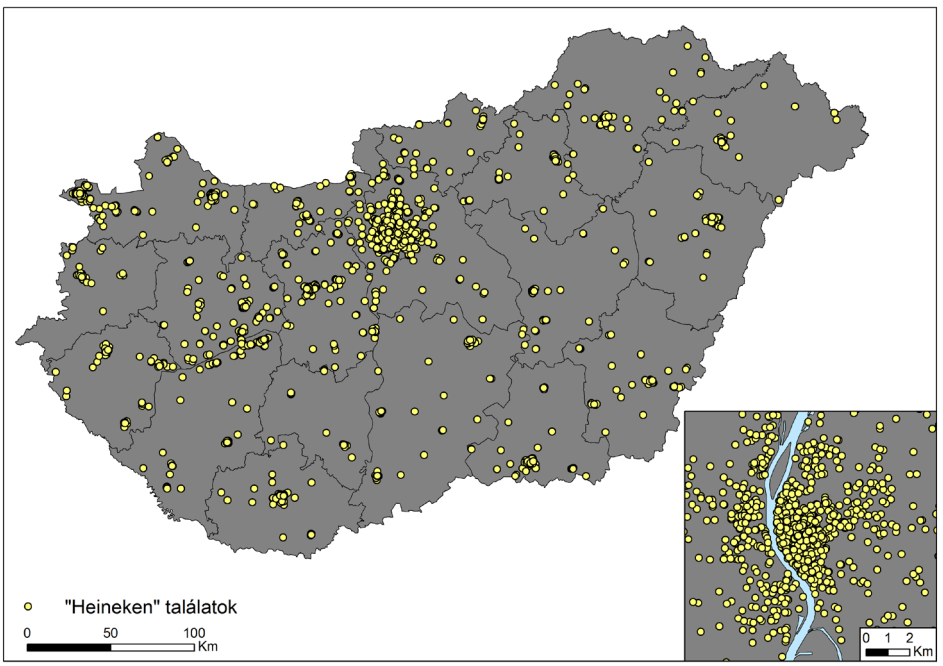
A következőkben néhány egyszerűbb kulcsszó keresési eredményeinek példáján mutatjuk be a webes tartalmak térbeli eloszlásának elemzési lehetőségeit, tanulságait. A keresés során egyrészt konkrét (például jól meghatározott termékhez kötődő), másrészt általános kulcsszavak térbeli előfordulásait is megvizsgáltuk.

A geokódolt tartalmak térbeli eloszlásának gyakorlati vizsgálata

Egy-egy kulcsszó geokódolt webes előfordulásainak lekérdezésekor pontszerű adatok halmaza formájában kaptuk meg az eredményeket. A ponthalmaz a térinformatika klaszszikus eljárásaival ezt követően könnyű szerrel térképre vihető volt, ami így lehetővé tette, hogy egy-egy vizsgált jelenség a webes tartalomból leszűrt kulcsszó formájában térbeli eloszlásviszonyait tekintve is meghatározható és elemezhető lehessen.

A *2. ábra* egy egyszerű lekérdezés térképi eredményeit mutatja. Vizsgálatunkban arra voltunk kíváncsiak, hogy egy termék (jelen esetben jól beazonosítható márkanévvel) milyen térbeli gyakorisággal fordul elő az ország egyes területein. Lehetőség lenne természetesen az adott terméket árusító összes cég, étterem, bolt stb. földrajzi helyzetét feltüntetni az ábrán, de itt alapvetően nem erről van szó. A térképek a virtuális tér földrajzi leképeződései, azaz a *weben közzétett tartalmak* térbeliségét tükrözik, a térképen ábrázolt pontok (térképi jelek)

tehát azokat a helyeket azonosítják, amelyekhez az interneten talált tartalom (a weben fellelt kulcsszó) köthető. Az ugyan feltételezhető, hogy a keresett kifejezés a terméket árusító helyek honlapján is nagy valószínűséggel megtalálható, így az árusítópontok a térképen is nagy eséllyel megjelennek, de emellett számos olyan földrajzi helyet is beazonosított az eljárás, melyek azon weboldalakhoz köthetők, ahol csak megemlítették a terméket. Ha ez utóbbi weboldalakat sikerül a földrajzi térben azonosítani, akkor az is megtudható, hogy hol tettek közzé információt az adott termékről, az adott kulcsszóval kapcsolatban, s ezáltal végső soron a termék virtuális térbeli elterjedtsége is meghatározható.

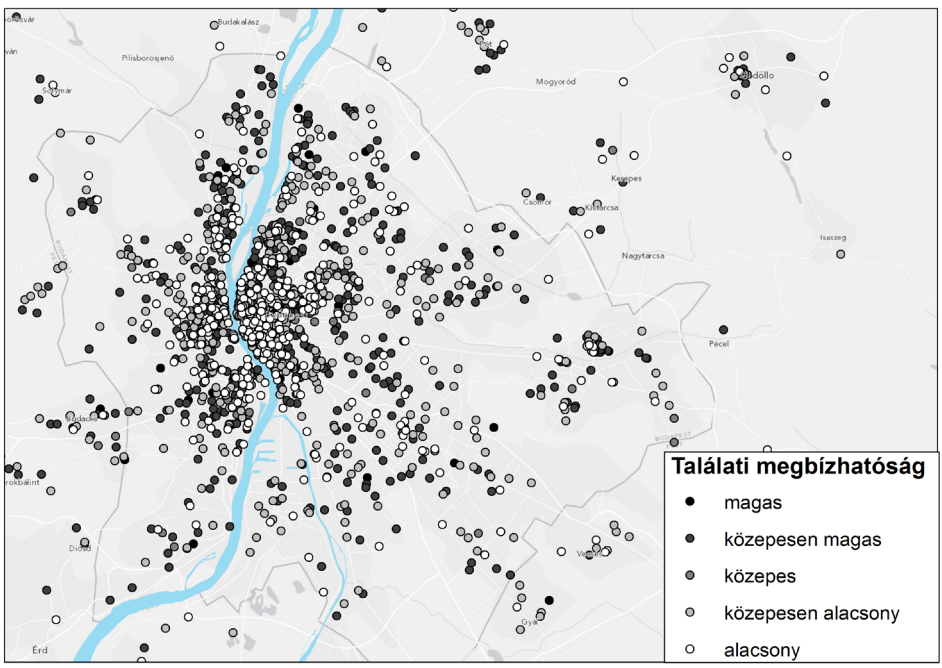


2. *ábra* A „Heineken" kulcsszó előfordulásai a weben Magyarországon (N = 52 646)
Figure 2 The occurrences of the "Heineken" keyword on the web in Hungary (N = 52 646)

A geokódolt tartalmakat ábrázoló térkép közvetlen és közvetett tanulságokkal szolgál. A kapott térképekről közvetlen módon leolvasható az ábrázolt tartalomhoz köthető honlapok földrajzi helye. Ennél azonban talán fontosabb, hogy a térkép mintázatából, a ponteloszlás (pontosabban a térképi jelek eloszlásának) vizsgálatából közvetett módon az adott webes tartalomnak a helyi internethasználó társadalmon belüli ismertségére, népszerűségére, előfordulásának gyakoriságára is következtetni lehet. A sűrűsödő területek feltételezhetően azok a térségek, ahol az információs térben fontosabbnak tartják az adott kifejezés közzétételét, a kifejezés többször, több helyen jelenik meg. Másrészt az is kijelenthető, hogy a virtuális térbeli tartalmak nagyobbrészt ott keletkeznek, ahol az internethasználók abszolút száma is nagyobb. Nem meglepő, hogy az ábra sűrűsödéseiből részben visszakövethető a magyar településszerkezet is, de persze ettől komoly eltérések is megfogalmazhatók.

A *2. ábrán* látható „Heineken" keresési kulcsszó előfordulásai egyértelműen a főváros környékén a leggyakoribbak, ahol amúgy is nagyobb a potenciális találati valószínűség az internethasználók és a bejegyzett honlapok nagyobb abszolút száma miatt. A melléktérképen

ugyanakkor az is látható, hogy a főváros ebből a szempontból nem egy homogén terület, a pesti és a budai belváros sűrűbb területeivel ellentétben a találati gyakoriság a külsőbb részeken jóval alacsonyabb (az ábrán találathiányos sávként a Duna vonala is kivehető). Gyakorlottabb térképolvasók a pontok kisebb sűrűsödéseinek helyén vidéki városainkat azonosíthatják be (például Pécs, Szeged, Debrecen vagy Győr esetében), ami ismételten azt támasztja alá, hogy a virtuális térbeli találati gyakoriság a valódi térbeli abszolút lakosságszámmal is kapcsolatban van. Ugyanakkor ettől eltérő eredmények is látszanak: a térképen sűrűbb területként sejlik a Balaton térsége, ahol közvetve az adott termék iránti valamelyest nagyobb helyi kereslet, közvetlen módon pedig a kulcsszóhoz kapcsolódó webes tartalmak intenzívebb jelenléte állhat feltételezhetően a háttérben.



3. ábra A „borsodi” kulcsszó előfordulásai a találatok megbízhatósága szerint Budapesten
Figure 3 The occurrences of the “borsodi” keyword in Budapest according to the reliability of hits

Mint azt fentebb említettük, nem számíthattunk arra, hogy minden weboldal tökéletesen és egyértelműen beazonosítható lesz a földrajzi térben. A térbeli azonosítás hol sikeresebb, hol kevésbé sikeres volt az utólagos finomító algoritmusok eredményességének függvényében. A találatokhoz ezért megbízhatósági értéket is rendeltünk. A *3. ábra* a pontok térbeli eloszlásán túl a pontokhoz tartozó megbízhatósági értékek (score-ok) kategóriáit is mutatja. Aktuális vizsgálatunkban látható, hogy a fővárosi találatok többsége a város belső részein tömörül. Ugyanakkor az is nyilvánvaló, hogy a találatok egy igen jelentős része csak kis megbízhatósággal volt meghatározható, különösen a nagy pontsűrűségű belvárosi körzetekben. A kevésbé vagy fokozottabban megbízható találatok térbeli sűrűsödése mögött több dolog rejlik. Egyrészt az alacsony vagy közepesen alacsony megbízhatóságú találatok részaránya általában minden lekérdezésnél igen jelentős, így várható az is, hogy a térképi megjelenésük is intenzív lesz. Ugyanakkor ezek a megbízhatósági szintű értékek sem teljesen véletlenszerűen szóródnak a térben, valamilyen szinten a vizsgált fogalom általános

térbeli eloszlásához igazodnak. A vizsgálati algoritmus az egyes weboldalakat tehát be tudta azonosítani, de mivel esetleg a weboldal struktúrájában a földrajzi lokalizáció (cím) nem egyértelműen kapcsolódott a tartalomhoz, így a kapott eredményt is csak bizonyos óvatossággal fogadhatjuk el. Az azonban, hogy a magas megbízhatóságú és a közepesen vagy kevésbé megbízható találatok térbeli elrendeződése között kapcsolat van, feltételezhető. Mindezt az egyes kategóriákba eső pontok kvadrátanalízissel számított sűrűségi viszonyainak korrelációs elemzésével vizsgáltunk. Az analízis során a ponteloszlásra fektetett elemi területegységeken, más néven kvadrátokon vagy cellákon belüli pontgyakoriságokat értékeltük (a módszert részletesebben lásd THOMAS R. W. 1977, LLOYD C. D. 2011).

1. táblázat –Table 1

A különböző megbízhatóságú találati kategóriák térbeli eloszlásának korrelációs viszonyai („borsodi" kulcsszó-lekérdezéssel)
The correlation between the spatial distribution of points according to the categories of reliability (query of the "borsodi" keyword)

		Alacsony t.m.	**Köz. alacsony t.m.**	**Közepes t.m.**	**Köz. magas t.m.**	**Magas t.m.**
Alacsony találati megbízhatóság	Pearson korrelációs eh.	1	,037	,408**	,193**	,347**
	Szig. (2 oldali)		,094	,000	,000	,000
	N	1029	2024	2024	2024	2024
Közepesen alacsony találati megbízhatóság	Pearson korrelációs eh.	,649**	1	,189**	,024	,037
	Szig. (2 oldali)	,000		,000	,283	,094
	N	1029	1029	2024	2024	2024
Közepes találati megbízhatóság	Pearson korrelációs eh.	,826**	,767**	1	,361**	,201**
	Szig. (2 oldali)	,000	,000		,000	,000
	N	1029	1029	1029	2024	2024
Közepesen magas találati megbízhatóság	Pearson korrelációs eh.	,554**	,485**	,582**	1	,041
	Szig. (2 oldali)	,000	,000	,000		,065
	N	1029	1029	1029	1029	2024
Magas találati megbízhatóság	Pearson korrelációs eh.	,715**	,614**	,667**	,444**	1
	Szig. (2 oldali)	,000	,000	,000	,000	
	N	1029	1029	1029	1029	1029

Megjegyzés: ** A korreláció szignifikáns a 0,01-es szinten
A főátló alatt a 10x10 km-es kvadrátokat alkalmazó országos analízis, a főátló felett az 1x1 km-es kvadrátokat alkalmazó fővárosi analízis eredményei láthatók

A különböző megbízhatóságú pontok eloszlásának korrelációs viszonyait átfogóbb országos és részletesebb fővárosi szinten is vizsgáltuk *(1. táblázat)*. A magasabb aggregáltsági szintű, 10x10 km-es kvadrátokat alkalmazó országos vizsgálat minden megbízhatósági találati kategória térbeli eloszlása között szignifikáns összefüggést mutatott ki (ezt jelzik a táblázatban jelzett 2 oldali szignifikancia teszteredmények, az ún. p-értékek 0,000 körüli eredményei). Az 1029 vizsgált területi cellában tehát nagyjából hasonló, vagy legalábbis közepesen hasonló mértékben találunk magasabb vagy alacsonyabb megbízhatóságú találatokat. A térfelosztást sűrűbbre szabva az egyes cellákba eső találattípusok varianciája határozottan nagyobb lesz, következésképpen az egyes kategóriák területi eloszlásviszonyai között már csak gyengébb hasonlóságok lesznek mérhetők. A részletesebb területi bontású fővárosi

vizsgálatban 1x1 km-es kvadrátokat, összesen 2024 darab cellát használtunk ennek igazolására. A kapott eredmények ebben az esetben a várakozásoknak megfelelően gyengébb összefüggésekre utaltak: a Pearson korrelációs együtthatók értéke bár továbbra is pozitív maradt, de egyértelműen alacsonyabbnak és kevésbé szignifikánsnak mutatkozott (amit a táblázatban jelzett 2 oldali szignifikancia teszteredmények relatíve magasabb értékei jeleznek).

A webről geokódolt tartalmak területi eloszlását tekintve az is feltételezhető, hogy azok a nagyobb népességű, eleve sűrűbben lakott körzetekben jelennek meg intenzívebben. Települési szintű aggregációban nézve a találatokat ez a sejtés azonban nem mindig igazolódik *(2. táblázat)*. Az abszolút találatszám ugyan erős korrelációt mutat az abszolút népességszámmal, de semmiféle összefüggés nem látszik a népsűrűség nagyságával. A népsűrűséggel való bármiféle kapcsolat települési szinten nem, legfeljebb csak alacsonyabb mikrokörzet léptékben (pl. városrészek, vagy háztömbök szintjén) lehet feltételezhető. A fajlagos (népességarányos és területarányos) találatszámok egymással szoros, az abszolút találatszámmal és a népességszámmal viszont közepes, illetve gyenge, de szignifikáns összefüggésben vannak. A területarányos (10 km^2-re jutó) találatszám általában közepesen erős, de határozott korrelációs viszonyban áll a vizsgált jelzőszámokkal, ami arra enged következtetni, hogy ezen indikátor használata átfogó értelemben más társadalmi-gazdasági vizsgálatok esetén is hasznos, illetve hasznosítható lehet.

A találatok térbeli sűrűségének értelmezése – mint az fentebb látható volt – jelentős mértékben függ az alkalmazott területi aggregáció mértékétől, a vizsgálat területi szintjétől. Ez utóbbi szempont kevésbé lényeges akkor, amikor elemzésünket a térfelosztástól

2. táblázat – Table 2

A különböző találati indikátorok települési szinten mért korrelációs viszonyai („borsodi" kulcsszó-lekérdezéssel)

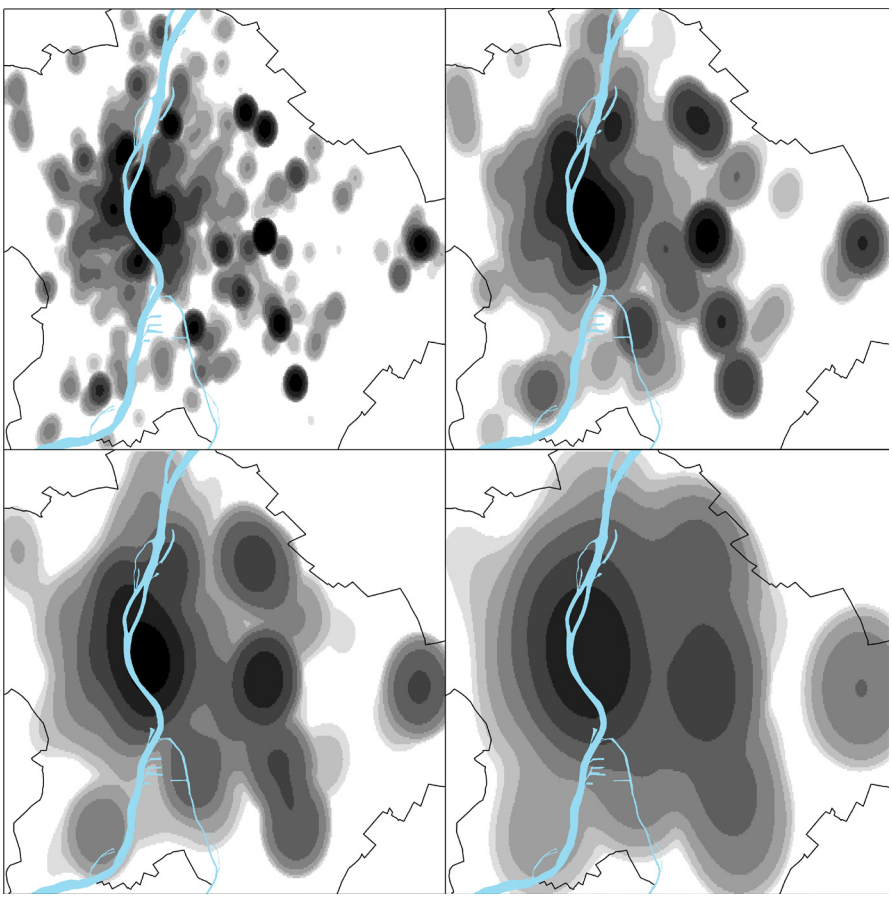
The correlation between different indicators of hits measured on the level of settlements (query of the "borsodi" keyword)

		Abszolút találat-szám	**Népesség-szám**	**Nép-sűrűség**	**Ezer főre jutó talá-latszám**	**10 km^2-re jutó talá-latszám**
Abszolút találatszám (db)	Pearson korrelációs eh.	1				
	Szig. (2 oldali)					
	N	3152				
Népesség-szám (fő)	Pearson korrelációs eh.	,966**	1			
	Szig. (2 oldali)	,000				
	N	3152	3152			
Népsűrűség (fő/km^2)	Pearson korrelációs eh.	,001	,000	1		
	Szig. (2 oldali)	,967	,982			
	N	3152	3152	3152		
Ezer főre jutó találatszám (db/ezer fő)	Pearson korrelációs eh.	,146**	,048**	–,001	1	
	Szig. (2 oldali)	,000	,008	,953		
	N	3152	3152	3152	3152	
10 km^2-re jutó találatszám (db/10 km^2)	Pearson korrelációs eh.	,574**	,471**	,000	,843**	1
	Szig. (2 oldali)	,000	,000	,988	,000	
	N	3152	3152	3152	3152	3152

Megjegyzés: ** A korreláció szignifikáns a 0,01-es szinten

független interpolált sűrűségmodellekkel végezzük. Ugyanakkor ez az elemzési mód is tartalmaz szubjektív elemeket akkor, amikor az interpolációs paramétereket beállítjuk, vagy az eljárásokat kiválasztjuk.

A pontelemek sűrűségének következő vizsgálatához az ún. kernel-interpoláció módszerét használtuk, melynek során a pontelemek előfordulási gyakorisága szerint izovonalas felületmodelleket alakítottunk ki (a módszer leírását lásd pl. SILVERMAN, B. W. 1986, BOWMAN, A. W. – AZZALINI, A. 1997). Az elkészült sűrűségtérképek az alapparaméterek beállításaitól függően részletesebb vagy elnagyoltabb képet rajzolnak elénk *(4. ábra)*. A térképek a „Dreher" kulcsszó találati előfordulási adatainak felhasználásával készültek folyamatosan növelt kernelnagyságok mellett. A kernelek mérete 0,01; 0,02; 0,03; illetve 0,05 fok volt, a modellekben súlyként nem pusztán a találatok darabszáma, hanem az azokhoz kapcsolódó találati megbízhatósági értékek szerepeltek. Az ábra mintázata a kernelméret növelésével egyre kiegyenlítettebbnek adódott, egyre összefogottabban szemléltetve a vizsgált kulcsszó területi eloszlásának főbb vonásait. Az ábra sötétebb területein nagyobb, míg világosabb részein kisebb kulcsszó-előfordulási sűrűséget mértünk.

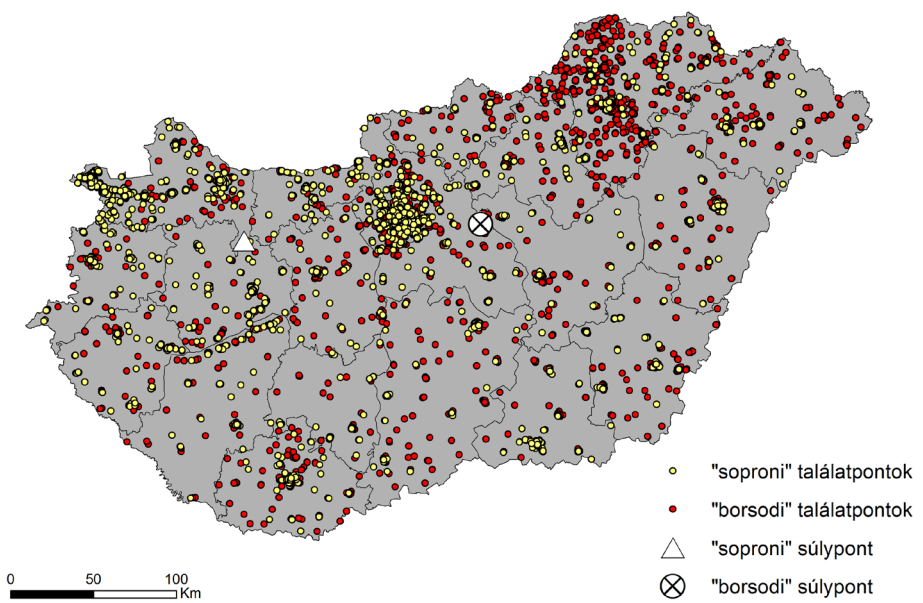


4. ábra A „Dreher" kulcsszó-előfordulások különböző interpolált modelljei Budapesten
(kernel 0,01; 0,02; 0,03; 0,05 fok, súly = találati megbízhatóság, a sötétebb színek magasabb találatsűrűséget jelölnek)
Figure 4 Different interpolation models of the occurrences of the "Dreher" keyword in Budapest
(kernel 0,01; 0,02; 0,03; 0,05 degrees, weight = reliability of hits, darker colours indicate higher density of hits)

Összehasonlító területi sűrűségvizsgálatok

Az egyedi lekérdezések mellett különösen érdekes eredményekre számíthatunk akkor, amikor kettő vagy több kulcsszó webes előfordulásainak területi képét hasonlítjuk össze (lásd pl. SHELTON, T. 2010). Ilyenkor lehetőség van egymástól teljesen független kereső-szavak, vagy egymást kiegészítő dichotóm kifejezések, esetleg konkurens terméknevek találati különbségeinek feltárására. Ez utóbbi párosítás kifejezetten alkalmas az egyes termékek vagy márkák közötti piaci verseny vizsgálatára, a kibertérbeli jelenlétük egyenlőtlenségeinek felmérésére, illetve esetünkben a geokódolt tartalmak révén közvetett módon a földrajzi elterjedtség vagy ismertség eltéréseinek meghatározására is. Az effajta összehasonlító vizsgálatok korábban csak nagymintás kérdőíves felmérések segítségével voltak kivitelezhetők, itt azonban a virtuális térből nyert strukturált big data állományok adhatnak választ némely feltett kérdésre, vagy használhatók a termékek vagy márkák közti csaták (lásd „brand wars", LOOSLEY, R. et al. 2012) felmérésére.

A ponteloszlások, illetve a találati sűrűségviszonyok összehasonlítására több lehetőség is adódik, de már az adatok egyszerű térképi megjelenítése is jól szemléltetheti a differenciákat. Az *5. ábra* a „soproni" és a „borsodi" kulcsszavak térbeli előfordulásainak eltéréseit szemlélteti. A vizsgálat egyrészt tükrözheti a két söripari márkanév területi versenyét, de az eredmény itt csalóka lenne, mivel ezek a kulcsszavak földrajzi térségeket is jelölnek. Ez részben látszódik is a „soproni" találatok Sopron környéki sűrűsödésében, vagy a „borsodi" találatok Borsod-Abaúj-Zemplén megyei gyakoribb előfordulásaiban. Mindamellett mindkét kereső-kifejezés találatai között szép számmal akadnak olyanok, melyek nem a fent említett térségekhez köthetők. Különösen szembeötlő a találatok főváros környéki sűrűsödése, ami a korábbi megállapítások fényében nem meglepő.

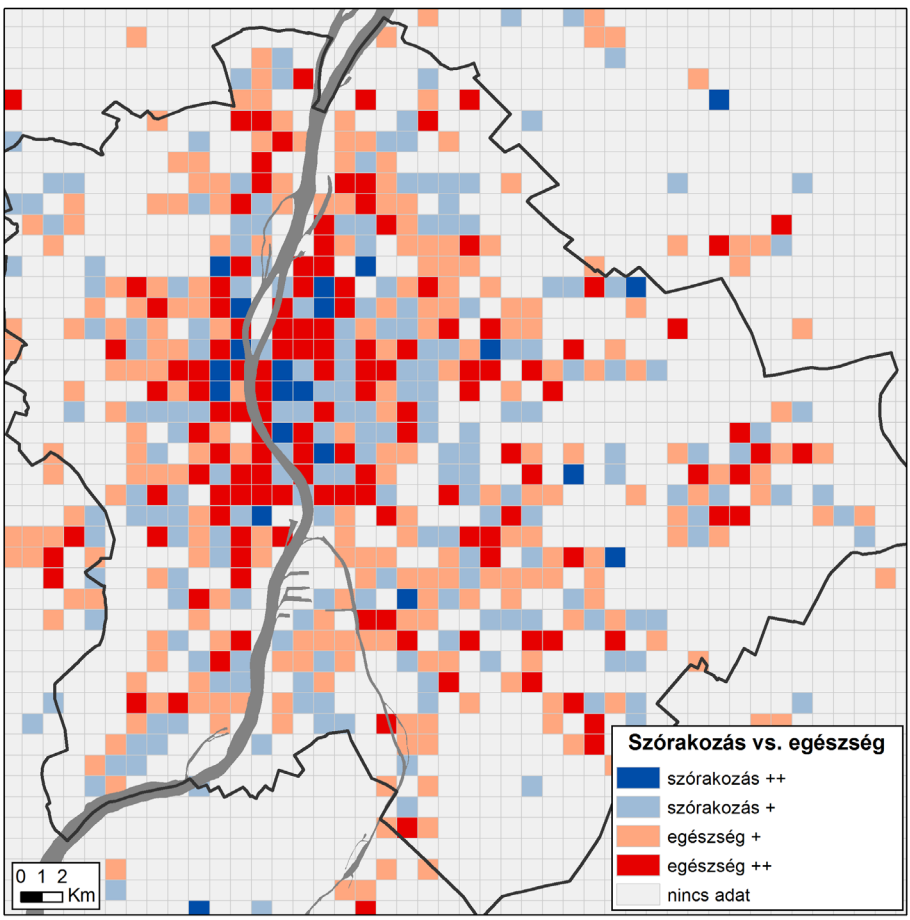


5. ábra A „soproni" (N=35755) és a „borsodi" kulcsszavak (N=50733) előfordulásai a weben Magyarországon, valamint a találatok súlypontjai
Figure 5 The occurrences of the "soproni" (N=35755) and "borsodi" (N=50733) keywords on the web in Hungary and the center of gravity of occurrences

A két ponteloszlás térbeli egyenlőtlenségeit a kapott találatpontok területi súlypontjainak összevetésével is vizsgálhatjuk. Eredményeink szerint a „soproni” találatok súlypontja egyértelműen nyugatabbra, míg a „borsodi” találatok súlypontja valamelyest keletebbre esik, bár ez utóbbi a fővároshoz (és az ország geometriai középpontjához, Pusztavacshoz) kissé közelebb található. A „borsodi” kifejezés az országban valamelyest szélesebb körben elterjedt, bár az északkeleti országrész dominanciájával, míg a „soproni” kifejezés határozottabban jellemző a dunántúli területekre.

Összefoglalás

A webről geokódolt tartalmak fenti ábrái a söripar néhány meghatározó kulcsszavának példáján keresztül mutatták be a módszerünkkel kapott eredmények elemzési lehetőségeit. A terméknevek elterjedtségének vizsgálatán túl persze általánosabb kifejezések is összevethetők. A *6. ábra* a „szórakozás” és az „egészség” kulcsszavak webes előfordulá-



6. ábra A „szórakozás” és az „egészség” kulcsszavak előfordulásainak dominancia-viszonyai a főváros környékén (1x1 km-es körzetekben)
Figure 6 The dominance of the occurrences of “leisure” and “health” keywords in Budapest (in 1x1 km cellsize)

sait hasonlítja össze a főváros térségében, ami jó példázza, hogy mily széles értelemben kínálkoznak új lehetőségek a társadalmi területi egyenlőtlenségek feltárására.

A tanulmányban körvonalazott eljárás csak egy lehetőség a sok közül arra, hogy az interneten jelen lévő óriási információhalmaz földrajzi motívumait meghatározhassuk, beazonosíthassuk. A területi kutatóknak ugyanakkor lépést kell tartaniuk az új kor kihívásaival, azaz nem hagyhatják kihasználatlanul azt az esélyt, amelyet a „big data korszak" új adatforrásai kínálnak. Ehhez igazodva állítható, hogy a webes tartalmak területi dominancia-viszonyainak meghatározása a társadalom működése megértésének egyik újszerű eszköze lehet.

Köszönetnyilvánítás

A tanulmány a Bolyai János Kutatási Ösztöndíj támogatásával készült. A szerző köszönetét fejezi ki továbbá az ESRI Magyarország Kft-nek az adatelemzésben nyújtott támogatásért.

Jakobi Ákos
ELTE TTK Regionális Tudományi Tanszék, Budapest
jakobi@caesar.elte.hu

IRODALOM

Bowman, A. W. – Azzalini, A. 1997: Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations. Oxford Science Publications, Oxford University Press, Oxford.

Cuevas, R. – Gonzalez, R. – Cuevas, A. – Guerrero, C. 2014: Understanding the locality effect in Twitter: measurement and analysis. Personal and Ubiquitous Computing 18. (2.) pp. 397–411.

Fischer, E. 2013: Locals and Tourists map. Gnip, MapBox project. Elérhető: http://mapbox.com/labs/twitter-gnip/locals

Friedewald, M. – Raabe, O. 2011: Ubiquitous computing: An overview of technology impacts. Telematics and Informatics 28. pp. 55–65.

Galloway, A. 2004: Intimations of everyday life: Ubiquitous computing and the city. – Cultural Studies 18. (2) pp. 384–408.

Girardin, F. – Calabrese, F. – Fiore, F.D. – Ratti, C. – Blat, J. 2008: Digital Footprinting: Uncovering Tourists with User-Generated Content. Pervasive Computing, IEEE 7. (4.) pp. 36–43.

Girardin, F. – Vaccari, A. – Gerber, A. – Biderman, A. – Ratti, C. 2009: Quantifying urban attractiveness from the distribution and density of digital footprints.International Journal of Spatial Data Infrastructures Research 4. pp. 175–200.

Graham, M. – Zook, M. 2011: Visualizing Global Cyberscapes: Mapping User-Generated Placemarks. Journal of Urban Technology 18. (1.) pp. 115–132.

Graham, M. – Hale, S. A. – Gaffney, D. 2013: Where in the World are You? Geolocation and Language Identification in Twitter. Professional Geographer, (előkészületben). http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2224233

Jakobi Á. 2014: Újszerű területi statisztikai adatgyűjtési lehetőségek az információs világ egyenlőtlenségeinek kutatásában. Területi Statisztika 54 (1) pp. 34–52.

Järv, O. – Ahas, R. – Saluveer, E. – Derudder, B. – Witlox, F. 2012: Mobile Phones in a Traffic Flow: A Geographical Perspective to Evening Rush Hour Traffic Analysis Using Call Detail Records. PLoS ONE 7 (11) e49171. doi:10.1371/journal.pone.0049171

Jiang, B. – Yao, X. 2006: Location-based services and GIS in perspective. – Computers, Environment and Urban Systems 30. pp. 712–725.

Leetaru, K.H. – Wang, S. – Cao, G. – Padmanabham, A. – Shook, E. 2013: Mapping the Global Twitter Heartbeat: The Geography of Twitter. First Monday 18. (5-6).

Loosley, R. – Thevarajaha, S. – Patel, P. 2012: Brand Wars in Cyberspace. Technology, Media and Telecommunications Bulletin. 28 June 2012., Fasken – Martineau. http://www.fasken.com/files/Publi-

cation/1544dfef-6622-479a-9b67-7e13acb710eb/Pre sentation/PublicationAttachment/78167655-40a9-4367-8a0f-0913b72ffa24/Trademark _Protection_Bulletin_-_RJL.pdf
LLOYD, C. D. 2011: Local models for spatial analysis. CRC Press. Boca Raton, USA.
NAAMAN, M. 2011: Geographic information from georeferenced social media data. SIGSPATIAL 3 (2) pp. 54–61.
SATYANARAYANAN, M. 2001: Pervasive computing: vision and challenges. – Personal Communications, IEEE 8. (4.) pp. 1017.
SHELTON, T. 2010: What do church, bowling, firearms and strip clubs have in common? Floatingsheep, 20. Jan. 2010., http://www.floatingsheep.org/2010/01/what-do-church-bowling-firearms-and.html
SILVERMAN, B. W. 1986: Density Estimation for Statistics and Data Analysis. Chapman and Hall, London.
THOMAS, R. W. 1977: An introduction to quadrat analysis. Concepts and Techniques in Modern Geography, 12, Geo Abstracts Ltd., University of East Anglia, Norwich.
WEISER, M. 1991: The computer for the 21st century. – Scientific American, 265. (3.) pp. 94–104.
ZOOK, M.A. – DODGE, M. – AOYAMA, Y. – TOWNSEND A. 2004: New Digital Geographies: Information, Communication, and Place. In. BRUNN, S.D. – CUTTER, S.L. – HARRINGTON, J.W. (eds.): Geography and Technology. Kluwer Academic Publishers. pp. 155-176.