



ИЗВЕСТИЯ НА БЪЛГАРСКОТО ГЕОГРАФСКО ДРУЖЕСТВО JOURNAL OF THE BULGARIAN GEOGRAPHICAL SOCIETY

Journal homepage: www.geography.bg/



Съвременни интернет-базирани източници на данни за целите на географските изследвания. Предимства и недостатъци Innovative web-based sources of data for geographical analyses. Advantages and disadvantages

Александър Коцев

Софийски Университет "Св. Климент Охридски"

Alexander Kotsev

Sofia University "St. Kliment Ohridski"

ABSTRACT

Key words:

Low-cost sensors, Twitter data, new data sources, cell phone data

The manuscript provides an overview of several prominent web-based sources of geospatial information that can be used to complement, and under certain conditions substitute traditional sources of data. Each potential data source is critically reviewed. The advantages and disadvantages outlined in the article would allow informed decisions concerning the use of those sources.

1. Въведение

Статията дава характеристика на няколко нови източника на пространствени данни, които могат да послужат като основа за географските изследвания, респ. да намерят приложение в най-различни области като регионално и градско планиране, опазване на природната среда и пр. Процесът на извличане на информация от тези данни обаче е свързан с редица предизвикателства от теоретико-методологично и чисто практическо естество, поради което критичният анализ и познаване на тези източници са много важни. В този смисъл статията има за цел да (i) характеризира три от интернет-базираните източници на данни, които според автора са с най-голяма перспектива за приложение в географските изследвания, (ii) да направи опит за систематизиране на предимствата и недостатъците на описаните източници на данните, както и (iii) да даде насоки за коректната им интерпретация. В същото време статията няма за цел да обхване всички съществуващи интернет-базирани източници на данни, нито да предостави изчерпателен отговор на широкия набор от въпроси, произтичащи от използването им. Разработката обаче може да послужи като отправна точка за дискусия в географската общност по отношение за нетрадиционни (нови) източници на данни, които съществуват в резултат на бързото развитие и все по-широкото използване на информационните и комуникационни технологии. В структурно отношение останалата част от статията дава кратка теоретико-методологична обосновка (секция 2), продължава с характеристика на избрани интернет-базирани източници на пространствена информация (секция 3), и приключва с дискусия и заключение.

2. Теоретична обосновка

В същността на пространствените данни според Fotheringham et al. (2000) е това, че освен от вариацията на определен феномен, изследователят се интересува и от местоположението му. Традиционно в географските изследвания се използва широк спектър от изходни данни от един или комбинация от следните основни източници: (i) теренна работа (вкл. картиране с помощта на GPS и/или други геодезически прибори); (ii) количествена и качествена информация, събрана и обработена с помощта на статистическа извадка (обикновено представителна); (iii) вече съществуващи картографски материали (вкл. ГИС-базирани); (iv) показания от дистанционни изследвания и пр. Според Попов (2012), в областта на географските изследвания "... , вече не представлява трудност с помощта на съвременните информационни технологии да се натрупа огромно количество емпирични данни, необходими за изследователски и други цели". Бързото технологично развитие предоставя възможност за комбиниране и в някои, засега относително редки случаи, дори замяна на тези данни, с такива, произхождащи от качествено нови източници. Пример за такива източници са социалните медии (напр. Facebook, Twitter, Foursquare и др.), както и такива, създавани от все по-големия брой мобилни устройства, свързани по дефиниция в Интернет (напр. смартфони и таблети). В частност по отношение на свързаността на мобилни устройства в Интернет трябва да имаме предвид, че се очаква броят им да достигне 50 милиарда до 2020 (Swan, 2012). Основната разлика между двата описани по-горе подхода, ако вземем за пример социално-икономическите и демографски данни, се състои в това, че ако при традиционното събиране

на данни по дефиниция, респондентите са наясно, че предоставят данните си за определена цел (напр. преброяване на населението и жилищния фонд), то при използването на социални медии и/или мобилни устройства като източник на данните, самото събиране може да е абсолютно независимо от респондентите. Нещо повече, самите те често не са наясно, че данните, които споделят в Интернет, или когато ползват мобилния си телефон (без значение дали разговарят или не), генерират много голям по обем набор от данни. При всички положения обаче, използването на такъв тип данни предоставя много възможности за географите, тъй като може да доведе до качествено нови изследвания (напр. в областта на анализа на достъпността при евентуално проследяване на населението в реално време).

3. Нови интернет-базирани източници на данни

3.1. Социални медии

А) Обща характеристика

Използвайки социалните медии като Facebook, Twitter и пр., потребителите създават съдържание, което освен за целите на собствениците на съответните платформи (напр. маркетинг, пр.) могат да се използват и за научно-изследователска дейност. Примери в това отношение са използването на Twitter за идентифициране на горски пожари (Osterman and Spinsanti, 2012), идентифициране на динамиката и промените в земята използването на базата на изображения от FourSquare (Syratos et. al., 2016) и пр.

Б) Предимства

Данните в социалните медии се променят в реално време и поради това могат да послужат за много бърза индикация относно пространственото и/или времево проявление на даден феномен, напр. в случаите на природен катаклизъм или на опасност за националната сигурност. Изходната информация при социалните медии е с потенциално много голям брой създатели, като по този начин последните играят ролята на “сензори” (по Goodchild, 2007).

В) Недостатъци

Основните предизвикателства по отношение на използването на тези данни за целите на научните изследвания се изразяват в (i) достъпът до съдържанието, (ii) механизмите за извличане на информация и отсяване на съществуващия “шум”, (iii) различната “география” на потребителите, както и затрудненото геореферирание на данните (iv).

Конкретно, по отношение на достъпа (i) трябва да имаме предвид, че лицензионният режим за данните в по-голямата част от случаите силно ограничава начините на ползването им. Нещо повече, с изключение на Twitter, където е налично отворен приложно-програмен интерфейс (API) за извличане на данни, за по-голямата част от социалните платформи има достъп само до публичното съдържание. Във връзка с начините за извличане на данни и отстраняване на шума (ii) трябва да имаме предвид, че методите за обработка на тези данни са много различни от традиционните, като изискват широко използване на методи в областта на машинната обработка на текст като тези, описани от Manning и Schütze (1999). По отношение на “географията” на потребителите (iii) трябва да се отчете, че различни социални медии са популярни в отделните държави, напр. потребителите на Twitter в Източна Европа са относително малко на брой за разлика от други европейски държави, където тази социална медия е много масово използвана. Във връзка с това, при необходимост от създаване на трансгранично приложени на база на такива данни, трябва да се отчитат спецификите в отдел-

ните държави. Не на последно място, особено имайки предвид темата на настоящата статия, трябва да се има предвид, че в голямата си част тези данни не са геореферирани (напр. по-малко от 5% от потока данни от Twitter са геокодирани на база на GPS показанията на устройството, с помощта на което са създадени).

3.2. Любителски сензорни мрежи

А) Обща характеристика

Събирането на данни от специално изградена мониторингова мрежа, състояща се от специализиран хардуер, мониторингови прибори, софтуер, процедури и пр. е традиционно в обхвата на дейността на дадени административни структури (напр. Изпълнителна агенция по опазване на околната среда, басейнови дирекции, НИМХ, пр.). С намаляването на цените на хардуера (вкл. сензорите), и разработката на софтуер със свободен код, който е в състояние лесно и бързо да публикува данните в уеб се формират алтернативи на традиционните способи за събиране на информация. Проекти като RaspberryPI, Arduino, Waspmote и др. набират все по-голяма популярност и формират общности от ентузиаста, които са ангажирани в изграждане и поддържане на сензорни мрежи с определена тематична насоченост.

Б) Предимства

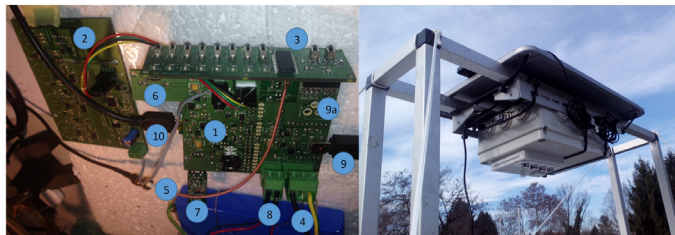
Информацията, събирана от любителски сензорни мрежи, по дефиниция е достъпна в уеб среда, като обикновено не съществуват ограничения по отношение на ползването ѝ, т.е. тя може да се окаже единственият безплатен и наличен източник за дадена територия в условия, подобни на българските, където организациите, съхраняващи данните (напр. НИМХ), все още осигуряват платен достъп до публичния ресурс, който управляват и поддържат. Нещо повече, за общностите, ангажирани с изграждането на любителски сензорни мрежи, самото изграждане на мрежите има ясно изразен образователен елемент, като в процеса те научават повече, наред с технологичните и организационни аспекти и относно явлението, което е обект на изследване. Не на последно място данните от любителските мрежи са налични обикновено в реално време, като се характеризират ако не с висока пространствена, то поне с висока темпорална резолюция (напр. наблюдение на всяка минута и пр.).

В) Недостатъци

Основният недостатък на този тип информация е свързан с качеството на изходните данни, което в голяма част от случаите е ниско, или не може да бъде установено. Докато при официалния източник на информацията определена държавна структура гарантира за събирането на качеството, калибриране на приборите, следване на определена методика на наблюдението, то при любителските мрежи нищо от това не е гарантирано. Нещо повече - лицата, ангажирани с изграждането, в много голямата част от случаите не са специалисти и техните стимули са много различни от тези при официалния източник.

По отношение на качеството съществуват инициативи, които целят да подобрят качеството на наблюдението при любителските сензорни мрежи. Например проектът AirSensEUR (фиг. 1-а, 1-б) създава оперативно съвместима платформа за качество на въздуха с нискобюджетни компоненти, които са подбрани по такъв начин, че да се осъществи относително по-високо качество на наблюдението в сравнение с аналогични в ценово отношение платформи (Kotsev et. al., 2016). Качеството на данните в този конкретен случай се осъществява с помощта на (i) подбор на сензори въз основа

на емпирична работа по установяване на качеството на данните им, и (ii) процес на калибриране в (вкл. в реално време) въз основа на официален източник на данни. Въпреки това все още съществуват редица предизвикателства относно осигуряване на качество на изходните данни от подобен тип изследвания.



Фигура 1. (а) AirSensEUR хардуер (б) AirSensEUR инсталиран в гр. Испра, Италия.

Figure 1. (a) AirSensEUR hardware (b) AirSensEUR installed in Ispira, Italy.

3.3. Мобилни телекомуникационни мрежи

А) Обща характеристика

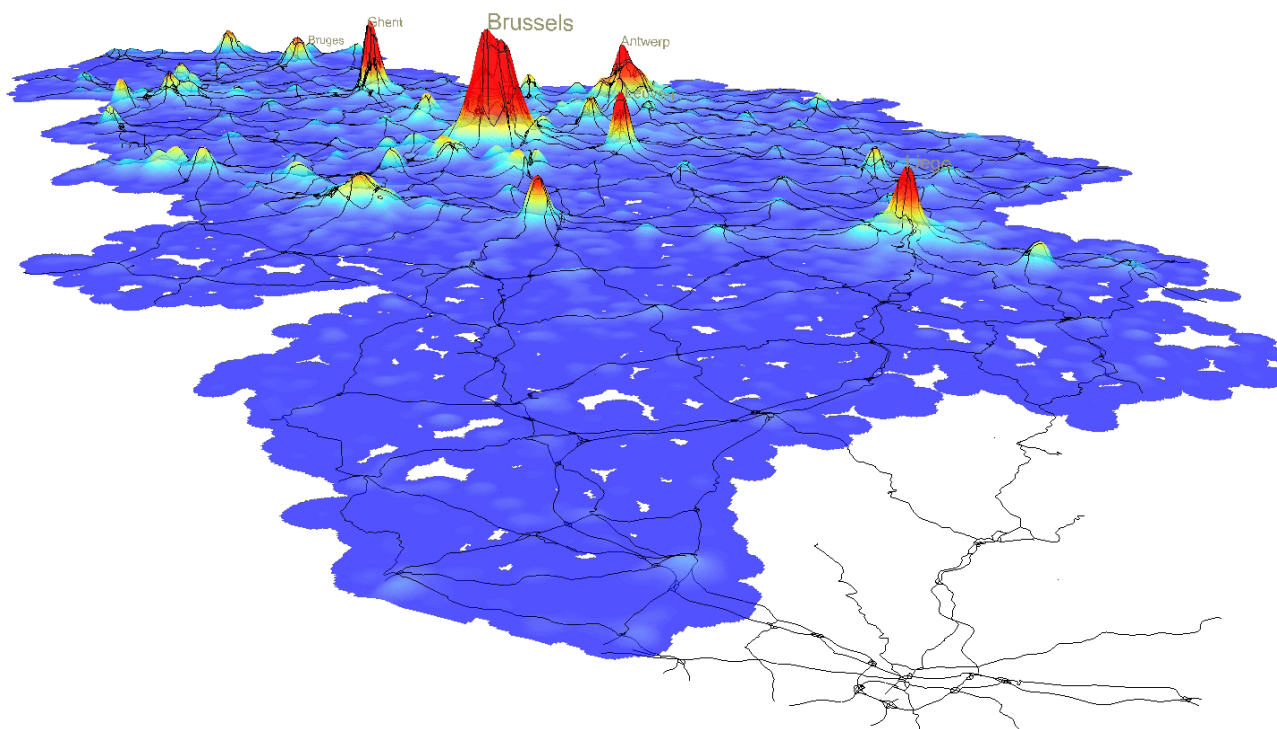
Операторите на мобилни мрежи разполагат с данни в реално време за абонатите, на които предоставят платен достъп до инфраструктурата си. Във всеки момент е известно към коя мобилна клетка е присъединен всеки потребител, а това е достатъчно условие за геореферирание на данните с позиционна точност, която е функция от гъстотата на клетките на мобилната мрежа, т.е. е значително по-висока в градска среда, тъй като броят клетки там е логично много висок с цел обслужване на големия брой абонати.

Б) Предимства

Предимствата на такъв тип данни са очевидни. Като се има предвид много големият брой на абонатите на телекомуникационните оператори, тази информация е с пространствено-времева резолюция, която е немислима за официалната статистика. Технологично, при комбиниране на данни от всички мобилни оператори в дадена държава, би било относително лесно да се изчисли сравнително точно каква е пространствено-времевата конфигурация на населението (ползващо мобилната мрежа) за всяка минута от всяко денонощие от годината. Фигура 2 например показва гъстота на устройствата, включени в мрежата на Proximus (Белгия) в 08:01 на 16.11.2015. Данните са получени въз основа на полигони на Тисен при които центроиди са отделните клетки в мрежата на оператора. Броят на присъединените устройства е наличен за всяка минута от денонощието, т.е. гъстота може да се изчисли за всяка минута. Нещо повече, този тип информация дава възможност да се анализира пространствено-времеовото поведение на населението, като става ясно какви са миграционните потоци и поведението на индивидите. Аналитичните възможности, напр. за анализ на достъпността на населението до услуги, на заболяемостта, градско планиране и т.н. са огромни.

В) Недостатъци

Въпреки описаните по-горе предимства, за използването на такъв тип данни съществуват редица проблеми, които са от чисто методологично, но и от правно, етично и т.н. естество. По отношение на методите за използване на данните трябва да се има предвид, че за разлика напр. от преброяването на населението, в случая с мобилните оператори не е ясна структурата на населението. Също така приравняването на население към мобилни устройства



Фигура 2. Гъстота на устройствата в мрежата на Proximus (Белгия) в 08:01 на 16.11.2015¹.

Figure 2. Device density of the Proximus Network (Belgium) at 08:01 on 16.11.2015¹.

¹Данни, предоставени от мобилния оператор. Визуализация - Ал. Коцев, 2016 / Data provided by the mobile operator. Preview - Al. Kotsev, 2016

в мрежата на операторите не е еднозначно. Ако даден потребител ползва два или повече мобилни телефона, това остава невидимо и от гледна точка на данни се равнява на двама души. По отношение на движението в пространството пък съществуват случаи, когато даден мобилен телефон е на границата между две мобилни клетки, като постоянно преминава от една към друга. Без съпътстваща информация и подходящи алгоритми за “отсяване” на такива случаи, ако данните бъдат въведени в ГИС, потребителите ще останат с впечатление, че мобилното устройство се движи в пространството.

Не на последно място трябва да е ясно, че такъв тип данни са много трудно достъпни и засега се предоставят за целите на научните изследвания относително рядко и на базата на добра воля от мобилните оператори. Без правна рамка, която да ги задължава да предоставят такива данни, напр. на националните статистики, е особено трудно да се получи достъп. Освен това евентуално изтичане на подобен тип данни от един оператор към друг е крайно нежелано за самите оператори, тъй като представлява маркетингова тайна. Също така технологично би било възможно въз основа на такива данни да се идентифицират отделни лица, което поражда етични въпроси, както и такива, свързани със защита на личните данни, с националната сигурност и пр.

4. Заключение

Настоящата статия дава характеристика на няколко съвременни интернет-базирани източника на географска информация, заедно с авторското виждане относно техните предимства и недостатъци. Примерите, с които са характеризирани тези източници, произтичат от практиката на автора и нямат амбицията да бъдат изчерпателни. Въпреки това те могат да се използват за дискусия в географската общност, посветена на адаптирането ѝ към новите реалности, произтичащи от процесите на навлизане на интернет-базираните технологии в ежедневието. В тази връзка, според Попов (2012), съвременните източници на данни се отличават се с все по-големи обеми и разнообразие, но в същото време полезната информация и полезното знание остават “скрити” и съвсем не е лесно да бъдат открити. Това важи с пълна сила и за източниците, описани в настоящата статия. Трябва да имаме предвид също така, че традиционните алгоритми за обработка и анализ на данни в географията вероятно не са изцяло пригодени за новите условия.

Големият въпрос за географите, във връзка с описаните източници на данни, е как и до каква степен те следва да бъдат използвани в техните изследвания. При всички случаи отговорът на този въпрос не трябва да бъде краен, т.е. новите източници не бива да се игнорират, тъй като по този начин се пропускат много възможности за получаване на ново познание. От друга страна, използването на данните без критичен анализ по отношение на тяхното качество, представителност и пр. е еднакво опасно и трябва да се избягва. Едва ли съществува универсално решение по отношение на използването на който и да е източник на данни, т.е. те винаги трябва да бъдат разглеждани в съответствие с конкретния контекст и поставената цел.

Използвана литература

- Попов, А. Географски Информационни Системи, изд. Анубус, 2012
- Kotsev, A., Schade, S., Craglia, M., Gerboles, M., Spinelle, L., & Signorini, M. (2016). Next Generation Air Quality Platform: Openness and Interoperability for the Internet of Things. *Sensors*, 16(3), 403.
- Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2000). *Quantitative geography: perspectives on spatial data analysis*. Sage.
- Ostermann, F., & Spinsanti, L. (2012). Context analysis of volunteered geographic information from social media networks to support disaster management: A case study on forest fires. *International Journal of Information Systems for Crisis Response and Management (IJISCRAM)*, 4(4), 16-37.
- Spyrtos, S., Stathakis, D., Lutz, M., & Tsinaraki, C. (2016). Using Four-square place data for estimating building block use. *Environment and Planning B: Planning and Design*, 0265813516637607.
- Goodchild, Michael F. "Citizens as sensors: the world of volunteered geography." *GeoJournal* 69, no. 4 (2007): 211-221.
- Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing* (Vol. 999). Cambridge: MIT press.
- Swan, M. (2012). Sensor mania! the internet of things, wearable computing, objective metrics, and the quantified self 2.0. *Journal of Sensor and Actuator Networks*, 1(3), 217-253.