Rozwój Regionalny i Polityka Regionalna 46: 67–89
2019

Michał Rzeszewski[1], Olga Rodak[2]

[1]Uniwersytet im. Adama Mickiewicza w Poznaniu
Instytut Geografii Społeczno-Ekonomicznej i Gospodarki Przestrzennej
e-mail: mrzeszewski@gmail.com
[2]Akademia Leona Koźmińskiego
Katedra Zarządzania w Społeczeństwie Sieciowym
e-mail: orodak@kozminski.edu.pl

Czy więcej znaczy lepiej? Badania ilościowe w geografii społeczno-ekonomicznej ery Big Data

Zarys treści: Wielkie zbiory danych typu Big Data są obecnie nieodłącznym elementem badań w wielu dziedzinach nauki – w tym w geografii społeczno-ekonomicznej. Znajdują one szereg zastosowań, zarówno wnosząc możliwość analizy nowych danych w klasycznych problemach badawczych, jak i same w sobie będąc nowym przedmiotem badań oraz pozwalając na badanie cyfrowych geografii. Jednocześnie Big Data krytykuje się za niejednorodność, brak reprezentatywności, nierówność reprezentacji czy też problemy etyczne. Z tego powodu zwracamy uwagę na szereg dylematów związanych z obecnością Big Data w praktyce badawczej, sugerując istotną rolę krytycznego podejścia do ich zastosowania w praktyce badawczej.

Słowa kluczowe: Big Data, geografia społeczno-ekonomiczna, badania ilościowe, Data Science

Wprowadzenie

Upowszechnienie się w ostatnich latach wielu nowych technologii, takich jak Internet, Internet rzeczy (IoT – *Internet of Things*), media mobilne i wszelkiego typu technologie lokalizacyjne, spowodowało lawinowy wzrost ilości produkowanych danych. Każda nowa technologia wydaje się wzmacniać ich strumień – media społecznościowe, smartfony, usługi lokalizacyjne, inteligentne zegarki czy opaski fitness przyzwyczaiły nas do stałej produkcji osobistych danych, czy to w sposób pośredni w tle, czy też jako wynik świadomego dążenia użytkownika w ramach *sousveillance* (Mann i in. 2003). Wiele nowych urządzeń, nawet tak zwyczajnych jak pralka, lodówka czy odkurzacz, ma możliwość zbierania i przetwarzania

danych, a każde kolejne „smart" rozwiązanie zbliża nas, jak się wydaje, do antyutopii *dataveillance (*Clarke 1994). Rośnie nie tylko ilość zbieranych danych, ale też arsenał ich źródeł. Przykładowo coraz popularniejsze aplikacje wirtualnej i rozszerzonej rzeczywistości przesuwają granicę intymności zbieranych danych, umożliwiając twórcom serwisów takich jak Facebook Spaces gromadzenie danych biometrycznych poprzez czujniki zestawów VR. Przestrzeń naszych nowoczesnych miast również dawno przestała być biernym bytem, „o którym" zbiera się dane. To raczej przestrzeń, naszpikowana sensorami, kamerami i oprogramowaniem, aktywnie śledzi nas i samą siebie. Kiedy korzystamy z karty miejskiej lub z aplikacji do zakupu biletu, zostawiamy ślad swojej podróży. Kiedy wychodzimy z domu, nasz automatyczny robot sprzątający skanuje przestrzeń mieszkania i wysyła jego obraz do swoich twórców – w celu, który trudno nam sobie nawet wyobrazić. Nierzadko dane zbierane są niejako „na zapas", tylko dlatego, że jest ku temu okazja i (stale malejące) przyzwolenie społeczne lub po prostu brak świadomości funkcjonowania tego procesu.

Jednocześnie fakt istnienia tego typu danych stwarza ogromne szanse badawcze. Fenomen powszechności danych bywa określany jak „powódź danych" (np. Anderson 2008) – niosąca ze swoim nurtem obietnicę „bogatych, szczegółowych, wzajemnie połączonych, aktualnych i tanich danych" (Kitchin 2013). Obietnica ta od kilku lat energetyzuje przedstawicieli nauk społecznych. Jeszcze do niedawna żyjący w świecie dyscyplin ubogich w dane, dość nagle, na przestrzeni kilku lat, znaleźli się w sytuacji konieczności zmiany paradygmatów badawczych (Gonzalez-Bailon 2013, Ruppert 2013, Housley i in. 2014). Strumienie Big Data są istotne dlatego, że pozwalają na bezpośrednie uchwycenie czasoprzestrzennej dynamiki zjawisk w wielu wymiarach jednocześnie, bez polegania na mocno ograniczonych „migawkach". Zbierane w sposób ciągły dane mają potencjał odzwierciedlenia zarówno codziennej rutyny, jak i zjawisk niezwykłych (Miller, Goodchild 2015, s. 451). W tym aspekcie nauki społeczne mogą dołączyć do dziedzin od dawna wykorzystujących Big Data, takich jak np. meteorologia i klimatologia (Li i in. 2017).

Geografia społeczno-ekonomiczna jest jedną z nauk, dla których rewolucja Big Data ma szczególne znaczenie. Duża część produkowanych danych ma bowiem komponent lokalizacyjny, a co za tym idzie – możliwe jest ich właściwie bezpośrednie wykorzystanie do analizy zjawisk przestrzennych. Niektórzy autorzy widzą w tym szansę na swoisty renesans informacji geograficznej oraz geografii jako dyscypliny (Hudson-Smith i in. 2009). Wszechobecność danych otwiera też zupełnie nowe geograficzne pola dociekań. Big Data nie tylko bowiem opisują geografie, ale równocześnie zmieniają je i tworzą. Celem artykułu jest wskazanie szans i przede wszystkim zagrożeń związanych z modą na zastosowanie Big Data w badaniach, ze szczególnym uwzględnieniem geografii społeczno-ekonomicznej. Zwracamy uwagę na wyzwania poznawcze i etyczne oraz proponujemy zestaw wskazówek przydatnych dla badaczy i badaczek rozważających użycie Big Data we własnej praktyce naukowej. Rozważania zaczniemy jednak od przyjrzenia się kwestii niejako podstawowej, czyli samej definicji Big Data, a także charakterystyce Data Science jako projektu naukowego, będącego odpowiedzią na ich wyzwanie.

Big Data i Data Science – ustalenia definicyjne

Termin Big Data pojawił się już w latach 1990., natomiast pierwszą jego definicję zaproponował Douglas Laney, analityk danych biznesowych, we wpisie blogowym z 2001 r. (Kitchin, McArdle 2016). Według tej wpływowej definicji, Big Data to zbiór danych, który można opisać za pomocą trzech cech: *volume, velocity, variety*. Pierwsza cecha odnosi się do olbrzymiej objętości, nieporównywalnej z dostępnymi wcześniej zbiorami, wyrażonej w terabajtach lub petabajtach. Druga cecha desygnuje prędkość przyrastania danych w tempie rzeczywistym. Natomiast trzecie określenie oznacza różny stopień ustrukturyzowania danych (Laney 2001)[1].

Z czasem następni autorzy dodawali do tej puli kolejne cechy Big Data, jednak większość odnosiła się do problemów kojarzonych z tym rodzajem danych (np. prywatność) niż do ich faktycznych cech odróżniających je od tradycyjnych ilościowych zbiorów danych, przede wszystkim danych ankietowych i danych administracyjnych (Kitchin, McArdle 2016). Jeszcze inne cechy dotyczyły technicznych aspektów nowych danych, które „stanowiły wyzwanie dla konwencjonalnych technik statystycznych i mocy obliczeniowej koniecznej do wyciągnięcia z nich wniosków" (Kitchin, McArdle 2016, s. 2). Na podstawie istniejących definicji Kitchin zaproponował taksonomię Big Data, na którą składa się siedem najważniejszych cech odróżniających te zbiory danych od ich poprzedników. Do wyżej wymienionych trzech klasycznych cech dodał: kompletność (w odróżnieniu od próby) (*exhaustivity*), wysoką „rozdzielczość" (*resolution and indexicality*), relacyjność umożliwiającą łączenie zbiorów danych (*relationality*), a także łatwość rozbudowy i skalowalność danych (*extensionality and scalability*) (Kitchin 2013). Podobna próba rozbudowy definicji podjęta została przez zespół Li i in. (2016) w odniesieniu do danych geoprzestrzennych – dodali oni do trzech klasycznych „V" kolejne trzy: wiarygodność (*veracity*), możliwość wizualizacji (*visualisation*) oraz widoczność (*visibility*). Mimo tych prób termin Big Data nadal daleki jest od jednoznaczności, a definicje różnią się pomiędzy różnymi obszarami nauki i praktyki (Chen i in. 2014).

Nieusatysfakcjonowani rozrastającymi się, lecz nieostrymi definicjami Big Data, Kitchin i McArdle (2016) zastosowali rozbudowaną definicję Kitchina (2013) do analizy 26 rodzajów danych uznawanych za Big Data przez pracujących na nich badaczy. Ich celem była precyzyjna eksploracja jakościowych cech oraz wyodrębnienia różnych „gatunków" Big Data w obliczu braku definicji, która koncentrowałaby się na faktycznej ontologii nowych danych. Autorzy wykazali, że niektóre z analizowanych zbiorów nie spełniają wszystkich kryteriów albo że kryteria same w sobie nie są adekwatne, bo mogą być zastosowane również do opisu tradycyjnych zbiorów danych. W ich opinii o specyfice Big Data – dzięki której oferują one nowe możliwości poznawcze – decydują takie cechy, jak kompletność populacji i prędkość przyrastania danych w momencie ich produkcji i publikacji. Wcześniej cechy te charakteryzowały tylko niektóre rodzaje danych, takie jak

[1] Warto podkreślić, że w tym ujęciu nie wszystkie duże zbiory danych to Big Data, natomiast Big Data z konieczności są dużymi zbiorami danych.

dane giełdowe czy dotyczące pogody; dzisiaj takie dane wytwarzane są w odniesieniu do potencjalnie każdego problemu dzięki opisanym wyżej technologiom.

Badanie literaturowe Kitchina i McArdle'a oparte było na spostrzeżeniu, że w praktyce badawczej funkcjonuje specyficzne podejście do nazywania zbiorów danych mianem Big Data, które nie jest oparte na przyporządkowywaniu ich według istniejących klasyfikacji. Wyznacznikiem jest tu raczej pewien rodzaj uniwersalnego konsensusu, który mówi o tym, że Big Data są to ustrukturyzowane bądź nieustrukturyzowane zbiory danych o ogromnej objętości, które nie mogą być w łatwy sposób zbierane, przechowywane, manipulowane, analizowane i prezentowane za pomocą tradycyjnego sprzętu, oprogramowania i technologii bazodanowych (Li i in. 2016). I słowo „łatwy" wydaje się tu znaczące, sugerując kluczową subiektywność tej oceny. Batty (2013, s. 274) żartobliwie opisuje ten fenomen jeszcze inaczej – Big Data to „każde dane, które nie mieszczą się w tabelce Excela". Choć wszyscy autorzy wymienionych przez nas poniżej przykładowych prac nawiązują w swoich tekstach do fenomenu Big Data, to jednak często wykorzystywane przez nich zestawy danych są stosunkowo niewielkie i właściwie nie spełniają warunków definicyjnych Big Data – poza Hochmannem i Manovichem (2013) dotyczy to na przykład wszystkich wymienionych badaczy korzystających z Instagrama. Można zauważyć tutaj powtarzający się często, subiektywny sposób myślenia, w którym za Big Data uznajemy to, co dla nas nowe, trudne w obróbce, przekraczające nasze, skromne przecież często, możliwości sprzętowe.

Nietypowy charakter Big Data – zarówno ich fizyczna forma, wymuszająca korzystanie z technologii komputerowej, jak i cechy znacząco odróżniające je od tradycyjnych, nawet dużych zbiorów danych, tworzonych zwykle z myślą o konkretnym pytaniu badawczym – staje się dźwignią innowacji w zakresie metod zarządzania danymi oraz ich analizy (Kitchin 2014). Oznacza to modyfikację istniejących technik statystycznych oraz sięganie do innych dyscyplin po zasoby metodologiczne, takie jak przetwarzanie języka naturalnego, uczenie maszynowe czy analiza sieciowa (Gorman 2013). Metody te polegają też w dużej mierze na wizualizacji danych (Kitchin 2014).

Ontologia nowych danych wymusza jednak nie tylko użycie komputerów i nowych metod, ale w ogóle przemyślenie epistemologii nauk społecznych, a zatem również geografii społeczno-ekonomicznej. Data Science to projekt tworzący się na naszych oczach; zdaniem niektórych, czwarty paradygmat nauk społecznych, w którym to metody badawcze są podporządkowane dostępnym danym, a nie na odwrót (Hey i in. 2009). Niedostateczna refleksja nad tym przesunięciem oznacza niebezpieczeństwo błędów poznawczych: naiwnego empirycyzmu czy też pseudopozytywizmu. Geografia jako dyscyplina, w której nie jest rzadkością „ukryty pozytywizm" (Kitchin 2006), może mieć tendencje do popadania w te koleiny – i polska geografia nie jest tu wyjątkiem (Wójcik, Suliborski 2014).

Z drugiej strony, wielu badaczy argumentuje, że to właśnie geografia społeczno-ekonomiczna wnosi wiele w rozwój Data Science. Geografia poprzez swoje bliskie relacje z systemami informacji geograficznej (GIS) już wiele dekad wcześniej radziła sobie ze zbiorami danych generowanymi dzięki technologii, których objętość przekraczała moc obliczeniową komputerów. Działo się tak na przykład

w przypadku technologii teledetekcji satelitarnej (Miller, Goodchild 2015) czy chociażby w początkach historii GIS związanych z zarządzaniem zasobami środowiska i analizą spisów powszechnych, będących bez wątpienia zbiorami Big Data – nawet jeśli w latach 70. XX w. termin ten nie był jeszcze znany. Korzenie GIS są związane raczej z „twardym", pozytywistycznym podejściem, jednak już od początku swojego istnienia technologie i metody systemów informacji geograficznej spotykały się z ostrą krytyką ze strony nauk społecznych (w tym geografów społeczno-ekonomicznych), dotyczącą między innymi bezrefleksyjnego podejścia do kwestii reprezentacji i obiektywizmu oraz fetyszyzmu technologicznego (Harley 1988, Pickles 1995, Schuurman 2000, Crampton 2010). Rozejm między uczestnikami tej ognistej nieraz dyskusji, jaki nastąpił w 1993 r. (Wilson 2017), był zapalnikiem wielu owocnych przedsięwzięć badawczych w bogatym nurcie krytycznego GIS-u (*critical GIS*), takich jak partycypacyjny i feministyczny GIS, a w ostatnich kilkunastu latach również jakościowy GIS, przestrzenny nurt cyfrowej humanistyki czy GIS w połączeniu z *non-representational theory* (Wilson 2014). Ta postpozytywistyczna krytyka znacząco zwiększyła świadomość badaczy na temat filozofii nauki, a także jej społecznych uwarunkowań i konsekwencji. Badacze doszukują się daleko idących podobieństw między początkiem użycia technologii w geografii a obecnym zachłyśnięciem się Big Data w innych naukach (Barnes 2013). Ze względu na to doświadczenie geografia społeczno-ekonomiczna i GIS powinny być dobrze przygotowane do wypracowywania świadomości na temat uwarunkowań nowej praktyki naukowej opartej na Big Data (Graham, Shelton 2013).

Zastosowanie Big Data w geografii społeczno-ekonomicznej

Powszechność zasobów Big Data odzwierciedla się również w wielości sposobów wykorzystania tego potencjału w szeroko rozumianej działalności badawczej. Geografia społeczno-ekonomiczna – tak jak cała geografia – nie pozostaje w tym względzie w tyle. Można wręcz powiedzieć, że jest to dla niej naturalna nisza, chociażby z uwagi na fakt, że, jak podają niektórzy autorzy, aż 80% zasobów Big Data ma komponent przestrzenny[2]. Poniżej prezentujemy przykłady zastosowania Big Data w geografii społeczno-ekonomicznej, proponując jednocześnie ich podział na dwie grupy ze względu na sposób, w jaki Big Data wykorzystywane są w praktyce badawczej. Po pierwsze, są to nowe metody w starych problemach. Ta kategoria obejmuje prace, w których zastosowano Big Data do próby sformułowania nowego podejścia metodologicznego i uzyskania nowych odpowiedzi na klasyczne problemy badawcze geografii społeczno-ekonomicznej (tab. 1). Po drugie, są to całkowicie nowe pola badawcze, które powstały jako efekt przetwarzania przez Big Data istniejących i produkowania nowych geografii (tab. 2).

[2] Choć pochodzenie tego powszechnie przytaczanego szacunku nie jest do końca znane (patrz: dyskusja pod adresem http://www.gislounge.com/80-percent-data-is-geographic/), a sama wartość procentowa dyskusyjna, to jednak nie ma większych wątpliwości, że jest to udział znaczący.

Tabela 1. Wybrane przykłady zastosowania Big Data do klasycznych problemów geografii społeczno-ekonomicznej

Nowe metody – stare problemy	
Problem badawczy	Przykłady zastosowania Big Data
Obszary i regiony	Shelton, Poorthuis 2019, van Meeteren, Poorthuis 2018, Rae, Singleton 2016, Hollensten, Purves 2010, Miller 2010
Demografia	Deville i in. 2014, Ruggles 2014, Burrows 2013
Obraz miasta	Zasina 2018, Salesses i in. 2015, Gali, Donaire 2015, Hochman, Schwartz 2012, Hochman, Manovich 2013
Zachowania przestrzenne	Birkin 2019, Gadziński 2018, Miah i in. 2017, Wu i in. 2016, Shaw i in. 2016, Rzeszewski 2015a, Yin i in. 2015, Hawelka i in. 2014, Orellana i in. 2012
Transport	Gadziński 2017, Tao i in. 2014
Turystyka	Zajadacz 2017, Majewska i in. 2016, Sun i in. 2013

Tabela 2. Wybrane przykłady nowych problemów badawczych związanych z Big Data w geografii społeczno-ekonomicznej

Nowe problemy	
Problem badawczy	Przykłady badań
Smart cities – inteligentne miasta	Zook 2017, Colleta, Kitchin 2017, Piskorz-Ryń 2017, Hao i in. 2015, Hancke i in. 2013, Batty 2013
LBS, prywatność i komodyfikacja danych	Huang, Gartner 2018, Rzeszewski, Luczys 2018, Thatcher 2017
Geografie mediów społecznościowych	Huang, Wong 2016, Robertson, Feick 2016, Wilson 2015, Stephens, Poorthuis 2015, Shelton i in. 2014, Kulshrestha i in. 2012
Dane wytwarzane przez użytkowników (UGC) i informacja geograficzna z wolontariatu (VGI)	Rzeszewski, Luczys 2017, Burns 2015, Haklay 2010, Connors 2012, Goodchild 2007

Przedstawiony tutaj podział na nowe metody i nowe problemy nie jest i nie może być podziałem rozłącznym. Podajemy szereg przykładów w obydwu kategoriach, przyznając jednocześnie, że ich klasyfikacja mogłaby wyglądać inaczej. Nowe metody stosowane do analizy od dawna funkcjonujących problemów badawczych mają bowiem również potencjał transformacji tych problemów i ich rekonceptualizacji. I odwrotnie – wyniki zastosowania nowej metody mogą zmienić podejście do niej i przeformułować jej założenia, często nazbyt optymistyczne. Przywołane przez nas przykłady w żadnej mierze nie wyczerpują też tej klasyfikacji. Zostały one dobrane w taki sposób, by unaocznić zróżnicowanie możliwych podejść i metodologii i pozwolić na ich krytyczną ocenę.

Nowe metody w starych problemach

Jedne z pierwszych prób zastosowania nowego rodzaju danych dotyczyły wyznaczania za pomocą metod ilościowych granic obszarów i regionów, które przedtem były trudno uchwytne i rozmyte. Przy czym, jak zauważają Miller (2010) oraz Rae i Singleton (2015), choć dla geografii regionalnej nowe zasoby danych są wyjątkową szansą, to z ich wykorzystaniem wiążą się poważne wyzwania metodologiczne. Można zauważyć, że najpopularniejszym źródłem danych w tego

typu badaniach są dane z mediów społecznościowych – stosunkowo łatwo dostępne i mogące zastąpić brakujące informacje o rozmieszczeniu przestrzennym badanych zjawisk. Przykładem może być badanie Hollenstein i Purves (2010), w którym użyli oni 8 mln zdjęć z serwisu Flickr do przestrzennej delimitacji centrów sześciu dużych europejskich i amerykańskich miast. Delimitacja polegała na określaniu zasięgu występowania tagów[3] wskazujących na centrum miast bądź poszczególne dzielnice lub sąsiedztwa. Miejskie sąsiedztwa analizowane były też przez Sheltona i Poorthuisa (2019), którzy pokazali dzięki analizie danych z mediów społecznościowych ich płynny, relacyjny charakter w kontraście do administracyjnych, historycznych wydzieleń. Ciekawym przykładem może być praca van Meeteren i Poorthuis (2018), której autorzy podejmują się ambitnego zadania dyskusji z teorią ośrodków centralnych Christallera w świetle danych uzyskanych z serwisów Foursquare i Twitter.

Kolejnym obszarem, w którym wykorzystywane są zasoby Big Data, jest demografia. W ostatnich latach badacze zyskali potężne narzędzia w postaci zdigitalizowanych danych pochodzących ze spisów powszechnych (Ruggers 2014), jednak nie jest to jedynie nowe źródło danych. Jako przykład może posłużyć wykorzystanie danych z sieci komórkowych do konstrukcji dynamicznego obrazu rozkładu populacji w mieście (Deville i in. 2014) czy też danych z komercyjnych systemów geodemograficznych do identyfikacji „bogatych" dzielnic i sąsiedztw (Burrows 2013).

Następnym problemem badawczym, który zyskał nowe podejście dzięki Big Data, jest szeroko rozumiany obraz miasta. Chociaż od czasów Kevina Lyncha (1960) „obraz miasta" (*image of the city*) rozpalał na wiele sposobów wyobraźnię badaczy, to właśnie wielkie zbiory danych pozwoliły spojrzeć na niego niemalże zupełnie na nowo. Media społecznościowe są ogromną skarbnicą subiektywnych opisów przestrzeni miejskich, co pozwoliło na zastosowanie nowych metod badawczych w nieznanej dotychczas skali. Szczególną uwagę poświęca się właśnie wątkowi informacji wizualnej i analizie treści zdjęć w serwisach typu Flickr, Instagram, itp. Przykładem takiej analizy z naszego rodzimego podwórka jest praca Zasiny (2018), w której wykorzystał on koncepcje Lyncha do przedstawienia cyfrowego, „instagramowego" obrazu Łodzi. Inny przykład podejścia do tego zagadnienia to klasyczne już niemal prace Hochmanna i Schwartza (2012) i Hochmana i Manovicha (2013), w których setki tysięcy zdjęć, także z Instagrama, zostały cyfrowo przetworzone w celu pokazania „wizualnych rytmów" (*Visual Rhytms*) – nowego sposobu na wizualny opis miasta i jego kulturowej treści, zarówno w czasie, jak i w przestrzeni. Łatwość uzyskania dużej ilości zdjęć pozwala na próby kwantyfikacji i ilościowego zbadania tak nieuchwytnego do tej pory elementu obrazu miasta, jakim jest percepcja przestrzeni miejskiej (Gali, Donaire 2015, Salesses i in. 2015). Interesującym wątkiem na marginesie naszej klasyfikacji jest fakt istnienia sprzężenia zwrotnego między obrazem miasta zapisanym w danych a praktykami w przestrzeni. Używając przestrzeni miejskiej,

[3] Tagi to krótkie słowa kluczowe używane powszechnie w mediach społecznościowych na dookreślenie treści i usytuowanie jej w pożądanym kontekście. Umożliwiają innym użytkownikom łatwiejsze wyszukiwanie i obserwowanie interesujących ich treści.

produkujemy nieustannie dane lokacyjne i jednocześnie opieramy nasze przemieszczenia i działania na wskazówkach wygenerowanych na podstawie tych danych (Leszczynski, Crampton 2016).

Praktyki przestrzenne oraz inne zagadnienia związane z przemieszczaniami i ludzką mobilnością są jednym z pól badawczych, w których era Big Data została powitana ze szczególnym entuzjazmem. Ze swojej natury bowiem pole to zajmuje się problemami wymagającymi analizy ogromnej ilości danych, których pozyskanie jeszcze do niedawna nastręczało wiele problemów związanych z koniecznością polegania na wysoce nieprecyzyjnych i niedokładnych metodach. Powszechność urządzeń lokalizujących użytkownika pozwoliła te bariery przezwyciężyć. W chwili obecnej badanie zachowań ludzkich można oprzeć na: analizie śladów z urządzeń GPS (Rzeszewski 2015b) – co wymaga aktywnego udziału badanych (Orellana i in. 2012, Rzeszewski 2015b), pomiarze aktywności urządzeń podłączonych do sieci komórkowych – zarówno w sposób aktywny (Gadziński 2018), jak i bierny (Deville i in. 2014), czy wreszcie na śledzeniu aktywności w mediach społecznościowych – co znalazło największy użytek w badaniach turystów (Miah i in. 2017). Wartość danych z mediów społecznościowych dla badań mobilności wynika między innymi z faktu, że pozwalają na badanie zjawisk w dużych skalach przestrzennych, np. globalnych migracji (Hawelka i in. 2014). Zagadnieniem powiązanym z mobilnością jest też transport miejski, w badaniu którego możliwość śledzenia realnych przemieszeń pasażerów, chociażby na podstawie śladów z sieci komórkowych czy kart miejskich, ma liczne praktyczne zastosowania (Tao i in. 2014, Gadziński 2017). Podobnie dane tego typu mogą być wykorzystane do analizy zachowań turystycznych (Sun i in. 2013, Majewska i in. 2016, Zajadacz 2017). Ostatnie publikacje wskazują również na wartość oraz na trudności wynikające z wykorzystywania w badaniach mobilności danych konsumenckich (*consumer data*), to znaczy danych powstałych w wyniku interakcji klienta i organizacji biznesowej, która dostarcza produkty i serwisy związane z mobilnością (Birkin 2019).

Nowe pola badawcze

Istnienie zbiorów Big Data stwarza również nowe pola badawcze, wynikające z ich zwrotnego wpływu na przestrzeń społeczną. Spośród tych nowych pól kilka zasługuje naszym zdaniem na szczególne wyróżnienie. Pierwszym z nich jest inteligentne miasto – smart city. Pojawienie się paradygmatu inteligentnego miasta i jego pochodnych, takich jak zarządzanie oparte na danych (*data-driven governance*) (Zook 2017) czy też zarządzanie za pomocą algorytmów (*algorithmic governance*) (Coletta, Kitchin 2017), jest stałym wyzwaniem dla badaczy z wielu dziedzin nauki. Geografowie, szczególnie ci związani z nurtem badań miejskich, zauważają istotność tego zagadnienia zarówno od strony teorii, jak i praktyki badawczej, jako możliwość uzyskania danych dotyczących wielu aspektów życia miast. Smart city i Big Data to bowiem terminy nierozłącznie ze sobą powiązane (Al Nuaimi i in. 2015) Analiza danych i ich wykorzystanie przy podejmowaniu decyzji jest tym, co umożliwia funkcjonowanie i samo istnienie inteligentnego miasta. Zaś

nowoczesna tkanka przestrzeni miejskiej, naszpikowana sensorami i oprogramowaniem, produkuje stały strumień danych przestrzennych. W tym wzajemnie napędzającym się systemie jest miejsce na nowe spojrzenia i nowe problemy badawcze – tym bardziej że, jak zauważa Miller (2010), większość tego, co wiemy o miastach, pochodzi z ery nad wyraz skąpej w dane. Te nowe spojrzenia mają często praktyczny charakter, taki jak na przykład możliwości wykorzystania czujników RFID (*Radio-Frequency Identification*) oraz technologii NFC (*Near-Field Communication*) do kontroli usług miejskich (Hancke i in. 2013), planowanie transportu publicznego za pomocą danych z systemu kart magnetycznych (Batty 2013) czy też lepsze planowanie przestrzeni miejskiej (Hao i in. 2015) dzięki wykorzystaniu między innymi opisanych wyżej metod analizy zachowań przestrzennych. Istnieje ponadto cała gama aspektów miejskiego życia, w których analiza Big Data jest postrzegana jako remedium na rozliczne problemy (Al Nuaimi i in. 2015) – również te, które są klasycznymi problemami miast sprzed ery cyfrowej.

Natomiast specyficzne dla obserwowanej w tej chwili powodzi wszelkiego rodzaju Big Data są rozważania na temat samej charakterystyki tych danych – ich pochodzenia, przydatności, możliwości wykorzystania, etyki, reprezentatywności i reprezentacji. Jednym z problemów badawczych są przykładowo kwestie wykorzystania osobistych danych lokalizacyjnych, czyli danych pozwalających na lokalizacje konkretnych osób. Są one podstawą działania tzw. usług opartych na lokalizacji (LBS – *Location Based Services*) (Huang, Gartner 2018). LBS opierają się na dostarczaniu użytkownikom treści zależnych od ich lokalizacji. Generuje to popyt na tego typu bazy danych i zjawisko komodyfikacji samej lokalizacji (Thatcher 2017), do czego odnoszą się z kolei pytania o etyczne aspekty tej sytuacji i związane z nią ryzyko utraty prywatności (Dobson, Fisher 2007, Abbas i in. 2014). Użycie LBS powoduje również zmiany w percepcji przestrzeni (Evans 2011) – istotne jest więc badanie cyfrowych warstw palimpsestu miejsca. Badacze zastanawiają się nad ich charakterem, sposobem powstawania i nastawieniem do nich mieszkańców miast (Michael, Michael 2011, Rzeszewski, Luczys 2018).

Osobnym zjawiskiem jest fenomen treści generowanych przez użytkowników (UGC – *User Generated Content*). W odróżnieniu od osobistej informacji lokalizacyjnej, która powstaje najczęściej niejako przy okazji korzystania z LBS, są to informacje tworzone świadomie (przynajmniej w założeniu) przez użytkowników. Przyjmują one różnorodną postać – od wpisów w Wikipedii, przez posty na Facebooku i filmy na YouTube, do aktualizacji treści na mapach internetowych. W geografii najbardziej interesująca jest Informacja Geograficzna z Wolontariatu (VGI – *Volunteered Geographic Information*) (Goodchild 2007, Elwood 2008) – czyli świadomie wytworzona i udostępniona informacja przestrzenna. Jest ona podstawą wielu projektów z nurtu tzw. neogeografii (Turner 2006), do których można zaliczyć Open Street Map (OSM) – alternatywę dla komercyjnych zasobów mapowych typu Google Maps, czy też Ushahidi – platformę crowdmappingową będącą emanacją zjawiska cyfrowego humanitaryzmu (*digital humanitarism*), w którym Big Data odgrywają niemałą rolę (Zook 2010, Burns 2015). W tym przypadku ciekawość badaczy wzbudza zarówno możliwość wykorzystania takiej informacji, jak i sam sposób, w jaki powstaje – procesy społeczne za tym stojące i efekty,

jakie pojawienie się takiej informacji powoduje np. w kontekście relacji władzy w obrębie wytwórstwa oficjalnych kartografii. Badacze VGI zajmują się zagadnieniami związanymi z jakością tego typu zbiorów danych (Haklay 2010, Flanagin, Metzger 2010) lub, jak w przypadku pracy Stephens (2013), uwarunkowaną płcią stronniczością głównych producentów treści przenoszącą się na dobór powstającej treści. Sama natura VGI też poddawana jest analizie. Do VGI bezrefleksyjnie zalicza się bowiem nader często wszystkie dane z mediów społecznościowych, a w tym przypadku ciężko mówić o świadomym tworzeniu informacji przestrzennej – jest to proces raczej przypadkowy i przebiegający w tle głównej aktywności (Stefanidis i in. 2013).

W ten sposób dochodzimy do głównego, jak się wydaje w chwili pisania tego artykułu, źródła Big Data w badaniach geografii społeczno-ekonomicznej, czyli mediów społecznościowych. W wielu przywoływanych tutaj przykładach badacze wykorzystują media społecznościowe jako źródło danych, pozwalające na uzyskanie odpowiedzi na wiele pytań (Huang, Wong 2016). Jednak na dane z mediów społecznościowych można patrzeć też z zupełnie innej perspektywy – jako na odrębne zjawisko, rządzące się swoim prawami, a nie będące jedynie odzwierciedleniem procesów w „prawdziwym" świecie (Wilson 2015). Z punktu widzenia geografii najbardziej interesujący – a przynajmniej najczęściej badany – wydaje się fakt, że media społecznościowe mają swoje geografie. Geografie te są mniej lub bardziej związane z materialną rzeczywistością (Kulshrestha i in. 2012), różnią się pomiędzy regionami i poszczególnymi serwisami (Rzeszewski, Beluch 2017), a nawet w obrębie jednego miasta tworzą wysoce nierównomierny rozkład przestrzenny (Robertson, Feick 2016). Z drugiej strony, geografie mediów społecznościowych przynajmniej częściowo odtwarzają materialne relacje społeczne i przestrzenne (Stephens, Poorthuis 2015).

Wyzwania Big Data

Trudności z wykorzystaniem Big Data w analizach ilościowych

Opisany wyżej ontologiczny charakter Big Data pociąga za sobą wyzwania w bezpośredniej aplikacji tradycyjnego podejścia ilościowego. W związku z tymi wyzwaniami odwróceniu ulec musi klasyczna kolejność procedury badawczej: zamiast dobierać dane do pytania badawczego, badacze starają się ustalić, na jakie pytania dostępne dane mogą dać odpowiedź (Miller, Goodchild 2015, Brooker i in. 2016).

Pierwsza trudność dotyczy reprezentatywności danych. W klasycznych badaniach statystycznych określona uprzednio „niewielka" – w porównaniu ze zbiorami, o których mowa – ilość danych zbierana jest, by umożliwić udzielenie odpowiedzi na postawione pytanie badawcze. Zebrana próba ma określone cechy socjodemograficzne, a populacja, z której została wyprowadzona, powinna być znana. Tymczasem zbiory Big Data nie są próbą określonej populacji, ale same w sobie tworzą populacje. Choć niweluje to wiele problemów charakterystycznych dla tworzenia prób losowych, jednocześnie rodzi nowe. Populacje te składają

się z osób, które użytkują daną technologię (na przykład populacja użytkowników Twittera albo populacja użytkowników Twittera, których tweety są lokalizowane przestrzennie), więc będą się one charakteryzowały dużą niekiedy nadreprezentacją określonych grup, której nie zniweluje nawet duża ilość przypadków. Z kolei brak albo szczątkowa charakterystyka demograficzna jednostek ujętych w próbie czyni niemożliwym odniesienie grupy uchwyconej w danych do jakiejś większej populacji albo wyjaśnienie otrzymanych wyników – na przykład przestrzennego wymiaru korzystania z telefonii mobilnej – w odniesieniu do kontekstu socjoekonomicznego (Liu i in. 2016).

Co więcej, informacje na temat każdego przypadku w populacji mogą być liczne i granularne, ale nie oznacza to, że dają pełny obraz tego przypadku. Jedną z głównych przyczyn tego stanu rzeczy jest fakt, że charakter danych zależy od tego, jak zaprojektowana jest technologia do ich gromadzenia i przetwarzania. Projekt z kolei naznaczony jest nie tylko możliwościami technicznymi, ale również „czynnikiem ludzkim", czyli założeniami i interesami dostarczycieli technologii na temat tego, jakie aspekty rzeczywistości powinny być uchwycone i w jaki sposób. Przez to umykają praktyki, interakcje i znaczenia nieprzewidziane przez projektantów technologii albo trudne w uchwyceniu (Graham, Shelton, 2013). Kluczowe znaczenie ma tu fakt, że technologie, o których mowa, przeważnie mają prywatny, komercyjny charakter, co oznacza, że zbierane są przede wszystkim te dane, które mają wartość ekonomiczną i które można w opłacalny sposób zgromadzić, przechować i przetworzyć (Dalton i in. 2016).

Drugą przyczyną niepełności danych jest niepozostawianie ich po sobie przez użytkowników. Dobrze widać to na przykładzie mediów społecznościowych, gdzie większość użytkowników tworzy znikomy ślad w postaci publicznych postów, komentarzy czy ikon emocji. Dość powiedzieć, że wiele kont na Twitterze nie udostępnia swojej lokalizacji za pomocą funkcji GPS albo podaje w swoim profilu fałszywą, fikcyjną bądź nieprecyzyjną lokalizację. Co więcej, użytkownicy wybiórczo udostępniają informację na własny temat, koncentrując się raczej na tych aspektach swojego doświadczenia, które uznają za odpowiednie do upublicznienia (Rui, Stefanone 2013, Miller, Goodchild 2015). Udostępniona informacja nie musi zresztą w pełni odzwierciedlać rzeczywistego doświadczenia; na przykład powszechną praktyką jest oznaczanie lokalizacji *post factum*.

Uzależnienie danych od prywatnych technologii powoduje kolejne istotne wyzwanie dla badań ilościowych, a mianowicie niedostatek wiedzy o charakterze danych i zbiorów. Technologie gromadzące dane funkcjonują jak czarne skrzynki: firmy zwykle nie ujawniają informacji, na jakiej zasadzie zbierają dane, jak skonstruowane są algorytmy danej technologii oraz jak często te ostatnie są zmieniane. Można jednak założyć, że zbieraniu danych przyświeca logika zysku, a nie spełniania wymogów naukowości. Przykładem może być Google Flu Trend Index, stosowany do przewidywania epidemii grypy na podstawie słów-kluczy wpisywanych do wyszukiwarki, krytykowany przez badaczy za błędy wynikające ze zmian algorytmu wyszukiwania (Liu i in. 2016). Z kolei badacze korzystający na przykład z danych z Twittera po przekroczeniu limitów dostępu narzuconych przez dostarczyciela platformy nie dysponują wiedzą na temat charakteru próby otrzymanych tweetów.

Następny problem to niejednoznaczność danych, czyli możliwość nieadekwatnego zinterpretowania znaczenia śladów pozostawionych przez użytkowników technologii. I znowu media społecznościowe są dobrym przykładem: trudno jednoznacznie orzec, jakie intencje użytkowników stoją za użyciem poszczególnych funkcjonalności serwisów, na przykład ikon emocji albo funkcji *retweet* (Macskassy, Michelson 2011). Narzędzia do interpretacji treści internetowych, takie jak programy do analizy sentymentu, nie są jeszcze w stanie do końca trafnie skojarzyć użycia słów i ich znaczenia emocjonalnego (Tomanek 2014). Inaczej mówiąc, znaczenie danych nie wynika z ich formy i funkcjonalności przypisanej im przez twórców technologii; wymaga problematyzacji na podstawie kontekstu, a to jest trudne, biorąc pod uwagę przepastność tych danych. Markham opisuje dysonans pomiędzy wskazaniami danych a wygraną Donalda Trumpa w wyścigu do fotela prezydenta Stanów Zjednoczonych właśnie w kategorii błędnej ich interpretacji (Markham 2018).

Ostatnim z wyzwań, które poruszymy, jest nierównomierność reprezentacji, zwłaszcza w wymiarze przestrzennym. Znów jest to bezpośredni skutek mediacji technologii. Dane są z natury przestrzenne nie tylko dlatego, że zawierają komponent lokalizacji, ale również dlatego, że ich tworzenie i analiza uzależnione są od przestrzeni, w której ma ono miejsce. W niektórych przestrzeniach występuje znacznie większe zagęszczenie technologii zdolnych gromadzić dane. Porównajmy chociażby bogactwo i różnorodność śladu, jaki zostawia po sobie użytkownik telefonu komórkowego w Londynie i w Mauretanii (Dalton i in. 2016). Z kolei dane dotyczące użycia telefonów komórkowych reprezentują nie miejsce, w którym znajduje się użytkownik, ale pozycję wieży telekomunikacyjnej, dlatego adekwatność wyników zależy od gęstości infrastruktury w danej przestrzeni (Liu i in. 2016). Co więcej, reprezentacja danych zależy również od statusu ekonomicznego, pozwalającego na korzystanie z technologii, oraz kompetencji w użytkowaniu technologii: nie tylko technicznych, ale też wynikających z kapitału kulturowego. Problem nierównomiernej reprezentacji dotyczy również serwisów opartych na crowdsourcingu, których użytkownicy, ponieważ jest wśród nich nadreprezentacja określonej grupy, mają tendencję do koncentrowania się na wybranych przestrzeniach albo problemach (Miller, Goodchild 2015).

Z drugiej strony, kluczowe znaczenie ma geografia procesu przetwarzania danych. Zwykle dokonują tego badacze w centrach krajów rozwiniętych, posiadający niewystarczająco kontekstualną wiedzę, by właściwie zinterpretować dane napływające z krajów peryferyjnych. Dane te mogą też być pomijane jako nieistotne albo z powodu nieumiejętności ich odczytania. W tym sensie Big Data może odtwarzać dawne nierówności epistemiczne, i to mimo większej dostępności danych (Dalton i in. 2016).

Ocena jakości danych

Big Data przeważnie zbierają się niejako samoczynnie, bez dokumentacji, i nie są kontrolowane w momencie zbioru. W tej sytuacji badacze mogą albo potraktować takie nieuporządkowane dane jako materiał do eksploracji i stawiania nowych

hipotez, albo starać się je zweryfikować (Miller, Goodchild 2015). Goodchild i Li (2012) wyróżnili trzy strategie czyszczenia i weryfikowania danych: dzięki internetowemu tłumowi (*crowd solution*), społeczności (*social solution*) oraz wiedzy (*knowledge solution*), poprzez odniesienie danych do istniejących teorii i faktów albo w drodze empirycznej. Ta ostatnia strategia wymaga wypracowania sposobów formalizowania wiedzy geograficznej, co może być trudne zwłaszcza w odniesieniu do nieostrych pojęć geograficznych, takich jak na przykład „kraje rozwijające się". Ponadto wymagałoby to wypracowania sposobów komputerowej reprezentacji wiedzy geograficznej, zarówno formalnych modeli, jak i pojęć zbudowanych w drodze interpretacji (Miller, Goodchild 2015).

Empirycyzm i pseudopozytywizm

Innowacyjny charakter Big Data powoduje konieczność wypracowania nowych metod, a nawet przemyślenia filozofii nauk społecznych. Niewystarczająca refleksja w tym zakresie może zwieść badaczy na manowce poznawcze. Wraz ze zwiększeniem zainteresowania Big Data popularność zyskiwał pogląd, który został potem przez komentatorów skojarzony z empirycyzmem, czyli z filozoficznym założeniem, że dane same w sobie stanowią fakty i wystarczą, aby wyciągać z nich wnioski, bez odwołania do teorii. W 2008 r. Chris Anderson ogłosił na łamach „Wired", że zalew danych przyniesie „koniec teorii", która nie będzie już konieczna do wyjaśniania znaczenia danych – wystarczy, że algorytmy statystyczne wykryją korelację w zbiorze danych (Anderson 2008).

Pogląd ten zaprzecza klasycznemu podejściu do rozwoju nauki poprzez budowanie hipotez na podstawie teorii. Oparty jest on na mylnym założeniu, że analiza wielkich zbiorów danych daje obraz całości zjawiska oraz że jest ona wolna od – w założeniu wadliwych – ludzkich założeń badawczych, a także że znaczenie wyników istnieje samo w sobie w oderwaniu od założeń teoretycznych i wiedzy wytworzonej w ramach danej dyscypliny. Efektem ubocznym wybuchu zainteresowania Big Data może być więc marginalizacja studiów typu *small data* wynikająca z niezrozumienia, że bez względu na wielkość i charakter zbioru danych, podlegają one różnego rodzaju skrzywieniom poznawczym i wymagają interpretacji w odniesieniu do kontekstu i wytworzonej już wiedzy. Przykładem mogą być badania fizyków, którzy poszukują „praw" procesów przestrzennych i prognozują rozwój miast bez odniesienia do wiedzy o społecznych, kulturowych czy ekonomicznych czynnikach kształtujący tenże (Kitchin 2013).

Wielu aktorów ma interes w przedstawianiu Big Data jako nowego zjawiska, ponieważ sugeruje to, że Big Data wolne są od ograniczeń innych zbiorów danych (Dalton i in. 2016). Tymczasem geografowie argumentują, że Big Data dzieli wiele problematycznych założeń z ilościową rewolucją lat 1950., kiedy to „unaukowiono" geografię poprzez adaptację pozytywistycznej filozofii i metod wypracowanych dla nauk przyrodniczych. Barnes (2013) nazywa nawet Big Data „über wersją ilościowej rewolucji". Założenia te były krytykowane od lat 1970. i choć krytyka ta nie przyjęła się w głównym nurcie dyscypliny (Panelli 2009), to jej podstawy mają ciągle istotne znaczenie. Pierwsze zagrożenie to „fetyszyzacja technik

i liczb". Istnieje niebezpieczeństwo wyłączenia z równania tych informacji, które nie dają się skwantyfikować, takich jak na przykład specyfika kontekstu geograficznych danych. Ponadto wzory danych mogą zastąpić interpretację czy szukanie przyczyn zaobserwowanych zależności. Kolejne problemy to niedostrzeganie faktu, że liczby nie są neutralne, a raczej ucieleśniają założenia instytucji mających władzę, dzięki której realizują one określone interesy. W końcu pozytywistyczna geografia jest zajęta badaniem tego, jak jest, zamiast problematyzować to, jak powinno być. Olson w 1974 r. wykazał, że szwedzcy geografowie, próbując wyznaczyć lokalizację instytucji dobra publicznego, opierali się na danych z cenzusu na temat interakcji przestrzennych. W rezultacie podjęte decyzje odzwierciedlały zastany porządek społeczny, zamiast poprawiać sytuację społeczną poprzez lepszą lokalizację przestrzenną potrzebnych instytucji (Barnes 2013).

Kitchin zwraca uwagę, że pozytywizm został zaadaptowany w naukach geograficznych powierzchownie, bez uświadomienia sobie filozofii stojącej za procedurami badawczymi, tak jakby był to jedyny naukowy, a zatem neutralny i oczywisty sposób badania rzeczywistości. Niestety „ukryty pozytywizm" wciąż dominuje w geografii mimo krytyki, a Big Data może przypieczętować jego pozycję. Może to prowadzić do ignorowania faktu, że algorytmy do analizy danych zostały wymyślone przez badaczy, którzy mają mniej lub bardziej uświadomione założenia teoretyczne i epistemologiczne (Kitchin 2006).

Problemy etyczne

Paradoks nowych danych wytworzonych dzięki technologii polega na tym, że są to głównie ślady aktywności człowieka, od których został on wyalienowany poprzez pozbawienie prawa własności danych, kontroli nad nimi i zdolności do ich analizy (Andrejevic 2014). Charakterystyka nowych danych tworzy nowego rodzaju zagrożenia dla integralności badanych podmiotów, które wymagają szczególnego przemyślenia i działań. Fakt dostępności danych czy też ich publiczny status (na przykład większość tweetów ma publiczny charakter) decyduje o legalności ich wykorzystania, co nie jest jednak równoznaczne z etycznością. Jako że w przypadku wielkich zbiorów danych nie jest możliwe uzyskanie świadomej zgody wszystkich osób, które pozostawiły po sobie ślad, badacze powinni rozważyć kierowanie się zasadą nieczynienia szkody i każdorazowo przemyśliwać skutki prezentowania wycinków danych, zestawień danych lub zagregowanych wyników, które mogły umożliwić identyfikację tożsamości badanych albo prezentować pewne zależności. Przykładem mogą być dane o nielegalnych migracjach (Dalton i in. 2016). Szczególnej kontroli powinny podlegać procedury przechowywania, obróbki i publikacji danych. Akademickie instytucje stojące na straży etyki badań niestety niedostatecznie szybko wypracowują standardy badań z Big Data (Liu i in. 2016).

Również epistemologiczne błędy mogą być przyczyną etycznych nadużyć. Bezkrytyczna akceptacja Big Data jako nieproblematycznej reprezentacji rzeczywistości społecznej może prowadzić do sytuacji, w której decyzje podejmowane są na podstawie zbioru danych o niejasnym charakterze, których wielkość i specyfika uniemożliwiają ich weryfikację przez nieprzygotowanego odbiorcę,

analizowanych w komercyjnych pakietach wymagających technicznej wiedzy. Samo odproblematyzowanie danych ilościowych nie jest zresztą niczym nowym, a geografia jako nauka ma swój wkład w opisanie tego procesu i jego konsekwencji. Graham i Shelton (2013) przytaczają dyskusje na temat relacji pomiędzy danymi, finansjeryzacją i urbanizacją, by ukazać, jak prezentowanie danych jako neutralnych „faktów" służyć może odpolitycznieniu decyzji, które w istocie nie biorą pod uwagę interesów grup o mniejszych zasobach władzy. Odwołując się do uwagi, że władza w definiowaniu statusu danych oraz ich przetwarzaniu ma przestrzenny charakter, należy zauważyć, że w przypadku Data Science mamy do czynienia z dynamiką centrum/peryferie: wartość z danych – zarówno kapitał ekonomiczny, jak i wiedzy – przechwytują podmioty ulokowane w krajach rozwiniętych (Dalton i in. 2016).

Dylematy stosowania Big Data w geografii społeczno-ekonomicznej

Wydaje się, że Big Data na dobre zadomowiły się w geografii społeczno-ekonomicznej – na wiele różnych sposobów. Jednak czy faktycznie stanowią czynnik transformujący proces badawczy? Według Kitchina (2014) akurat w przypadku nauk humanistycznych i społecznych mało prawdopodobne jest wyłonienie się nowego paradygmatu w związku z upowszechnieniem się tych zbiorów danych. Wynika to zarówno z faktu jednoczesnego współistnienia w tych naukach wielu różnych paradygmatów, jak i z ograniczonej przydatności zastanych zbiorów danych do rodzajów pytań przez te nauki stawianych. Big Data to raczej nowe dane, które wzbogacą istniejące studia[4]. Kitchin argumentuje też, że zamiast odrzucać Big Data, naukowcy społeczni mogą wnieść krytyczną, postpozytywistyczną perspektywę do analizy wielkich zbiorów danych, tak jak miało to miejsce w przypadku *critical GIS* czy *radical statistics* (Kitchin 2014)[5]. Dodatkowo wykorzystany przez nas podział zastosowań Big Data pokazuje jeszcze jedną drogę mariażu Big Data i geografii społeczno-ekonomicznej – badanie geografii tychże danych. Wielkie zbiory danych transformują i tworzą przestrzeń i same w sobie mają aspekt przestrzenny, co czyni je przedmiotem badań geografii. Otwartych pozostaje jednak szereg pytań o metodologiczne aspekty prowadzenia zarówno badań Big Data, jak i badań z wykorzystaniem Big Data. Poniżej prezentujemy krótką listę najważniejszych naszym zdaniem pytań, na jakie należy sobie odpowiedzieć, mierząc się z danymi tego typu w praktyce badawczej.

---

[4] Warto tutaj zaznaczyć, że za czynnik o równie wysokim potencjale transformującym może uchodzić pojawienie się otwartych źródeł danych (Open Data) (Arribas-Bel 2014) – jest to zagadnienie powiązane z Big Data, jednak nie jest ono przedmiotem naszych rozważań.

[5] Jak zawsze w takim przypadku istnieje duże prawdopodobieństwo marginalizowania tego typu czasochłonnych i niejednoznacznych badań. Jednak ich rola, choć mniej „parametrycznie" zyskowna, jest rolą niezmiernie istotną i podejmowanie badań krytycznych powinno być częścią akademickiego etosu.

1. **Czy chcę wykorzystać Big Data jako nowy rodzaj danych w istniejącym problemie czy też interesują mnie same dane i problemy badawcze, jakie stwarzają?** Jest to pytanie bodajże podstawowe, warunkujące dobór metodologii badawczej oraz stawianych pytań i hipotez. W pierwszym przypadku są one już najprawdopodobniej znane. W drugim może istnieć konieczność stworzenia nowej metodologii, a część ograniczeń nie ma takiego znaczenia (np. reprezentacja i reprezentatywność).
2. **Czy interesuję się daną problematyką tylko z tego powodu, że istnieją dane?** To częsta pułapka, w jaką wpadają niedoświadczeni badacze Big Data. Powszechność danych nie powinna być czynnikiem inicjującym badanie jakiegoś zagadnienia. Prowadzi to bowiem do chęci zignorowania podstawowych etapów postępowania badawczego i odwrócenia całej procedury badawczej (*HARKing* – patrz niżej).
3. **Czy mogę postawić odpowiednie hipotezy lub pytania badawcze przed analizą danych?** Postawienie pytań i hipotez przed rozpoczęciem analizy jest jeszcze ważniejszym etapem procedury badawczej w przypadku Big Data niż przy zbiorach danych o „standardowych" rozmiarach. Skala Big Data powoduje, że łatwo jest sformułować hipotezę pasującą do obserwowanych w danych prawidłowości (*HARKing – Hypothesizing After the Results are Known*), w których na pierwszy rzut oka wysoka istotność statystyczna może być jedynie dziełem przypadku. Generowanie zestawu hipotez powinno zawsze poprzedzać analizę danych – dodatkowo ogranicza to ilość testowanych zmiennych i możliwych korelacji między nimi i pozwala uniknąć strategii „gotowania oceanu"[6].
4. **Czy analiza dużych zbiorów danych pozwoli mi uzyskać wystarczająco wnikliwy obraz zagadnienia?** Jakkolwiek rozległa analiza Big Data może dostarczyć szerokiego spektrum informacji o badanym zjawisku, część informacji może być dla tej metodologii niedostępna. Przykładem mogą być dane o przemieszczeniach turystów. Możemy dzięki nim zmierzyć i zbadać, którędy przemieszczają się poszczególne osoby, jednak nie zrozumiemy przyczyn ich działań.
5. **Czy dysponuję odpowiednim zapleczem technicznym i analitycznym?** Nie bez powodu w przytaczanych przez nas w pierwszej części tego artykułu definicjach pojawiają się kwestie ogromnej objętości zbiorów, wykraczającej poza możliwości standardowych komputerów czy nawet stacji roboczych. W pewnej skali potrzebne są rozwiązania przetwarzania równoległego czy strategie typu MapReduce, wymagające odpowiedniej wiedzy i umiejętności.
6. **Czy dysponuję pełną wiedzą o pochodzeniu i metodologii zbierania danych?** Wielkie zbiory danych to różnorodność metod zbierania i mechanizmów kontroli jakości. Badacz nie zawsze dysponuje pełną wiedzą na ten temat. W przypadku danych z mediów społecznościowych API (*Application Programming Interface*) kwestie dostępu do danych są często pobieżnie opisane, tak aby nie ujawniać tajemnicy przedsiębiorstwa. Innym problemem mogą być

[6] Określenie „Boil the Ocean" odnosi się do mało efektywnej strategii polegającej na próbie uzyskania jakichkolwiek wyników przy braku konkretnie określonych problemów.

zestawy będące skutkiem agregacji wielu różnych strumieni informacji – często z różnych regionów geograficznych. Najlepszą rekomendacją wydaje się tutaj odpowiednio krytyczne podejście, próba opracowania własnych wskaźników jakości i przede wszystkim odpowiedni opis wyników.

7. **Czy jestem pewien swoich praw do korzystania z nich?** Nawet jeśli znane są doskonale metody zbierania i pozyskania danych – nadal pozostaje kwestia praw i warunków korzystania z serwisu. Szczególnie dotyczy to mediów społecznościowych, takich jak Facebook, Instagram czy Twitter. Jeszcze stosunkowo niedawno nie było problemem uzyskanie niskim kosztem niezależnych zbiorów danych przy wykorzystaniu natywnych API czy też web-scrapingu. I choć nadal jest to technicznie możliwe, to jednak od jakiegoś już czasu umowy licencyjne nie pozwalają na automatyczne pozyskiwanie danych dowolną metodą i właściwie jedynym sposobem legalnego dostępu do nich jest współpraca z daną firmą. Nawet kosztowny zakup gotowych zbiorów może być obwarowany wieloma ograniczeniami.
8. **Jaka populacja reprezentowana jest przez posiadane dane?** Często powtarzanym sloganem jest ten mówiący o Big Data jako o zbiorach reprezentujących całą populację, a nie tylko jej próbkę. Nawet jeśli tak jest w istocie (co jest dyskusyjne – patrz pkt 6), to należy dążyć do uzyskania wiedzy o charakterze tej populacji. Jest to często trudne w świecie, w którym źródłem aktywności mogą być chociażby boty. Trzeba też pamiętać, że każdy błąd i obciążenie będzie w zbiorach Big Data wielokrotnie wyolbrzymione.
9. **Jakie problemy etyczne mogą wiązać się z wykorzystaniem danego zbioru danych?** Na badaczu ciąży o wiele większa odpowiedzialność etyczna niż na korporacjach ery Big Data. Istnieje wiele przypadków, kiedy wstrzymanie się od analizy może być bardziej pożądane. Warto pamiętać, na co zwracają uwagę głosy krytyczne wobec Big Data, że rzadka jest sytuacja, w której informacje, szczególnie personalne, udostępniane są przy pełnej świadomości badanych podmiotów.
10. **Czy Big Data są mi potrzebne?** Istnieje cały szereg pytań, na które analiza Big Data nie potrafi, z uwagi na swoje ograniczenia, udzielić dobrej odpowiedzi. Więcej nawet, może doprowadzić do odpowiedzi mylnych, złudnie uwiarygodnionych mirem obiektywnej analizy ilościowej. Być może istnieją inne metody, które można wykorzystać w danym przypadku z większą pewnością, rezygnując z modnej etykietki na rzecz solidności warsztatu?

W każdym indywidualnym przypadku badacz winien samodzielnie udzielić sobie odpowiedzi na powyższe pytania. Lista ta może być jedynie drogowskazem. Poszczególne studia Big Data są różnorodne i wielowymiarowe, tak jak i ich podmioty – o czym też nie należy zapominać.

Tytułem zakończenia

Żywimy nadzieję, że zaproponowane przez nas rekomendacje okażą się pomocne dla badaczy chcących zagłębić się w świat wielkich zbiorów danych typu Big Data

i przede wszystkim skierują ich do odpowiednich źródeł wiedzy na ten temat. Na koniec chcemy zasygnalizować jeszcze dwa istotne zagadnienia, będące niejako pokłosiem istnienia zbiorów Big Data w geografii społeczno-ekonomicznej. Są to sprawy związane z polityką prowadzenia badań oraz z wątkiem geografii cyfrowej.

Kwestie polityczne dotyczą przede wszystkim konieczności zabezpieczenia autorytetu poznawczego i zasobów dla dyscypliny w związku z żywotnym zainteresowaniem instytucji przyznających środki na badania nad zbiorami Big Data. Podobnie jak w przypadku innych nauk społecznych, zwłaszcza socjologii, geografowie zauważają, że wraz z dostępnością Big Data następuje przejmowanie problemów typowych dla geografii przez przedstawicieli innych dyscyplin, którym brak przygotowania z zakresu teorii nauk społecznych i którzy w swoich interpretacjach opierają się głównie na wzorach wykrytych w danych. Grozi to nie tylko dostarczeniem wadliwych wniosków, ale również utratą przez geografię jej autorytetu w określonych dziedzinach (Kitchin 2013). Kitchin w 2013 r. prognozował, że z powodu powolności nauk społecznych w dostosowywaniu swoich metod do potrzeb nowych danych inne nauki, szczególnie takie, jak informatyka, matematyka czy fizyka, mogą je wyprzedzić w wyścigu o dostęp do danych i środki na badania. Osobnym problemem są też próby przypisania etykiety badań Big Data przedsięwzięciom niespełniającym żadnego z wymogów definicyjnych. W naszym odczuciu może to prowadzić do odwrotnego niż zamierzony efektu i osłabienia legitymacji dyscypliny w świecie badań opartych na danych.

Drugą z kwestii jest geografia cyfrowa lub raczej – geografie cyfrowe. W świecie, w którym algorytmy namacalnie wpływają na przestrzeń, a cyfrowe dane są jedną z warstw budujących percepcję przestrzeni i kształtujących ludzkie w niej zachowania, dane okazują się językiem, za pomocą którego możemy ten świat badać i próbować zrozumieć. I będąc powszechne, zbiory danych stają się wielkie. By je analizować, potrzebne są nowe podejścia, metodologie i praktyki badawcze. Istotnym uwarunkowaniem dla badaczy cyfrowego świata jest również fakt, że o ile w przypadku części zjawisk będących przedmiotem badań geografii społeczno-ekonomicznej Big Data nie są dobrym metodologicznym wyborem, to geografie cyfrowe niemal wymuszają użycie wielkich zbiorów danych. To właśnie Big Data ukrywają prawidłowości funkcjonowania kapitalizmu kognitywnego, nowych mediów, zarządzania opartego na algorytmach czy też kształtu cyfrowych podziałów i wykluczeń. Zgadzając się z poglądem o braku konieczności promowania powstania nowej subdyscypliny – geografii cyfrowej (Ash i in. 2018) – widzimy jednocześnie potencjał geografii społeczno-ekonomicznej jako dyscypliny wiodącej w badaniach cyfrowych geografii. W sposób unikalny łączy ona bowiem w swoim rdzeniu wiedzę na temat metod analizy danych geograficznych z krytyczną wrażliwością nauk społecznych. Jak staraliśmy się w tym artykule pokazać, obie te rzeczy są naszym zdaniem niezbędne w celu poprawnej interpretacji prawideł funkcjonowania przekształconego cyfrowo świata.

Literatura

Abbas R., Michael K., Michael M.G. 2014. The regulatory considerations and ethical dilemmas of location-based services (LBS): A literature review. Information Technology & People, 27: 2–20.

Al Nuaimi E., Al Neyadi H., Mohamed N., Al-Jaroodi J. 2015. Applications of big data to smart cities. Journal of Internet Services and Applications, 6: 25.

Anderson Ch. 2008. "Essay: The Data Deluge Makes the Scientific Method Obsolete". Wired Magazine, 10–12.

Andrejevic M. 2014. The Big Data Divide. International Journal of Communication 8(1): 1673–1689.

Arribas-Bel D. 2014. Accidental, open and everywhere: Emerging data sources for the understanding of cities. Applied Geography, 49: 45–53.

Barnes T.J. 2013. Big Data, Little History. Dialogues in Human Geography, 3(3): 297–302.

Batty M. 2013. Big data, smart cities and city planning. Dialogues in Human Geography, 3: 274–279.

Birkin M. 2019. Spatial data analytics of mobility with consumer data. Journal of Transport Geography, 76: 245–253.

Brooker P., Barnett J., Cribbin T., Sharma S. 2016. Have We Even Solved the First 'Big Data Challenge?' Practical Issues Concerning Data Collection and Visual Representation for Social Media Analytics. [W:] H. Snee, Ch. Hine, Y. Morey, S. Roberts, H. Watson (red.), Digital Methods for Social Science. An Interdisciplinary Guide to Research Innovation. Palgrave Macmillan, New York, s. 17–33.

Bruns A., Burgess J. 2016. Methodological Innovation in Precarious Spaces: The Case of Twitter. [W:] H. Snee, Ch. Hine, Y. Morey, S. Roberts, H. Watson (red.), Digital Methods for Social Science. An Interdisciplinary Guide to Research Innovation, Palgrave Macmillan, New York, s. 17–33.

Burrows R. 2013. The new gilded ghettos: The geodemographics of the super-rich. Sociology, 41(5): 885–899.

Burns R. 2015. Rethinking big data in digital humanitarianism: Practices, epistemologies, and social relations. GeoJournal, 80(4): 477–490.

Coleman D.J., Georgiadou Y., Labonte J. 2009. Volunteered geographic information: The nature and motivation of produsers. International Journal of Spatial Data Infrastructures Research, 4: 332–358.

Coletta C., Kitchin R. 2017. Algorhythmic governance: Regulating the 'heartbeat' of a city using the Internet of Things. Big Data & Society, 4: 2053951717742418.

Connors J.P., Lei S., Kelly M. 2012. Citizen science in the age of neogeography: Utilizing volunteered geographic information for environmental monitoring. Annals of the Association of American Geographers, 102(6), 1267–1289.

Dalton C., Taylor L., Thatcher J. 2016. Critical Data Studies: A Dialog on Data and Space. Big Data & Society, June: 1–9.

Dalton C., Thatcher J. 2015. Inflated Granularity: Spatial 'Big Data' and Geodemographics. Big Data & Society, December: 1–15.

Deville P., Linard C., Martin S., Gilbert M., Stevens F.R., Gaughan A.E., Blondel V.D., Tatem A.J. 2014. Dynamic population mapping using mobile phone data. Proceedings of the National Academy of Sciences, 111: 15888–15893.

Dijck J. van 2014. Datafication, Dataism and Dataveillance: Big Data between Scientific Paradigm and Ideology. Surveillance and Society, 12(2): 197–208.

Dobson J.E., Fisher P.F. 2007. The Panopticon's Changing Geography. Geographical Review, 97: 307–323

Elwood S. 2008. Volunteered geographic information: key questions, concepts and methods to guide emerging research and practice. GeoJournal, 72: 133–135.

Evans L. 2011. Location-based services: transformation of the experience of space. Journal of Location Based Services, 5: 242–260.

Flanagin A.J., Metzger M.J. 2008. The credibility of volunteered geographic information. GeoJournal, 72: 137–148.

Gadziński J. 2017. Wykorzystanie telefonów komórkowych w badaniach zachowań transportowych ludności. Prace Komisji Geografii Komunikacji PTG, 20(4): 7–19.

Gadziński J. 2018. Perspectives of the use of smartphones in travel behaviour studies: Findings from a literature review and a pilot study. Transportation Research Part C: Emerging Technologies, 88: 74–86.
Galí N., Donaire J.A. 2015. Tourists taking photographs: the long tail in tourists' perceived image of Barcelona. Current Issues in Tourism, 18: 893–902.
Goodchild M.F. 2007. Citizens as sensors: the world of volunteered geography. GeoJournal, 69: 211–221.
Gorman S.P. 2013. The Danger of a Big Data Episteme and the Need to Evolve. Geographic Information Systems, 3(3): 285–291.
Graham M., Shelton T. 2013. Geography and the Future of Big Data, Big Data and the Future of Geography. Dialogues in Human Geography, 3(3): 255–61.
Haklay M. 2010. How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. Environment and Planning, B, Planning & Design, 37: 682.
Hancke G.P., de Carvalho e Silva B., Hancke G.P, Jr. 2013. The role of advanced sensing in smart cities. Sensors, 13(1), 393–425.
Hawelka B., Sitko I., Beinat E., Sobolevsky S., Kazakopoulos P., Ratti C. 2014. Geo-located Twitter as proxy for global mobility patterns. Cartography and Geographic Information Science 41, 260–271.
Hey T., Tansley S., Tolle K. 2009. The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft Research, Redmond.
Hochman N., Manovich L. 2013. Zooming into an Instagram City: Reading the local through social media. First Monday, 18.
Hochman N., Schwartz R. 2012. Visualizing instagram: Tracing cultural visual rhythms. [W:] Proceedings of the Workshop on Social Media Visualization (SocMedVis) in Conjunction with the Sixth International AAAI Conference on Weblogs and Social Media (ICWSM-12), s. 6–9.
Hollenstein L., Purves R. 2010. Exploring place through user-generated content: Using Flickr tags to describe city cores. Journal of Spatial Information Science, 21–48.
Huang H., Gartner G. 2018. Current Trends and Challenges in Location-Based Services. ISPRS International Journal of Geo-Information, 7: 199.
Huang Q., Wong D.W.S. 2016. Activity patterns, socioeconomic status and urban spatial structure: what can social media data tell us? International Journal of Geographical Information Science, 30: 1873–1898.
Iwasiński Ł. 2016. Społeczne zagrożenia danetyzacji rzeczywistości. [W:] B. Sosińska-Kalata, N. Przastek (red.), Nauka o Informacji w okresie zmian. Informatologia i humanistyka cyfrowa. Wydanictwo SBP, s. 135–146.
Kitchin R. 2006. Positivistic Geographies and Spatial Science. [W:] C.A. Stuart, G. Valentine (red.), Approaches to Human Geography. Thousand Oaks, CA, Sage, s. 20–29.
Kitchin B. 2013. Big Data and Human Geography: Opportunities, Challenges and Risks. Dialogues in Human Geography 3(3): 262–67
Kitchin R. 2014. Big Data, New Epistemologies and Paradigm Shifts. Big Data & Society, 1(1): 205395171452848.
Kitchin R., McArdle G. 2016. What Makes Big Data, Big Data? Exploring the Ontological Characteristics of 26 Datasets. Big Data & Society, 3(1): 205395171663113.
Kulshrestha J., Kooti F., Nikravesh A., Gummadi P.K. 2012. Geographic Dissection of the Twitter Network. [W:] Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media.
Krzysztofek K. 2012. Zmiana permanentna? Refleksje o zmianie społecznej w epoce technologii cyfrowych. Studia Socjologiczne, 4(207): 7–39.
Laney D. 2001. 3D data management: Controlling data volume, velocity and variety. [w:] Gartner Blog Network (https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf).
Lapowsky I. 2019. How Cambridge Analytica Sparked the Great Privacy Awakening. Wired 2019 (https://www.wired.com/story/cambridge-analytica-facebook-privacy-awakening/).
Leszczynski A., Elwood S. 2015. Feminist geographies of new spatial media: Feminist geographies of new spatial media. The Canadian Geographer/Le Géographe canadien, 59: 12–28.

Li Z., Hu F., Schnase J.L., Duffy D.Q., Lee T., Bowen M.K., Yang C. 2017. A spatiotemporal indexing approach for efficient processing of big array-based climate data with MapReduce. International Journal of Geographical Information Science, 31(1): 17–35.
Li S., Dragicevic S., Castro F.A., Sester M., Winter S., Coltekin A., Pettit C., Jiang B., Haworth J., Stein A., Cheng T. 2016. Geospatial big data handling theory and methods: A review and research challenges. ISPRS Journal of Photogrammetry and Remote Sensing, 115: 119–133.
Liu J., Jie L., Li W., Wu J. 2016. Rethinking Big Data: A Review on the Data Quality and Usage Issues. ISPRS Journal of Photogrammetry and Remote Sensing, 115: 134–142.
Liu Y., Sui Z., Kang C., Gao Y. 2014. Uncovering patterns of inter-urban trip and spatial interaction from social media check-in data. PloS one, 9(1), e86026.
Macskassy S., Michelson M. 2011. Why Do People Retweet ? Anti-Homophily Wins the Day! [W] Proceedings of the Fifth International Conference on Weblogs and Social Media – ICWSM '11, s. 209–216.
Majewska J., Napierała T., Adamiak M. 2016. Wykorzystanie nowych technologii i informacji do opisu przestrzeni turystycznej. Folia Turistica, 41: 309–339.
Markham A. 2018. Troubling the Concept of Data in Qualitative Digital Research. [W:] W. Flick (red.), The SAGE Handbook of Qualitative Data Collection. Sage Publications, s. 511–523.
Miah S.J., Vu H.Q., Gammack J., McGrath M. 2017. A Big Data Analytics Method for Tourist Behaviour Analysis. Information & Management, Smart Tourism: Traveler, Business, and Organizational Perspectives, 54: 771–785.
Michael K., Michael M.G. 2011. The social and behavioural implications of location-based services. Journal of Location Based Services, 5: 121–137.
Miller H.J. 2010. The data avalanche is here. Shouldn't we be digging? Journal of Regional Science, 50(1): 181–201.
Miller H.J., Goodchild M.F. 2015. Data-Driven Geography. GeoJournal, 80(4): 449–461.
Orellana D., Bregt A.K., Ligtenberg A., Wachowicz M. 2012. Exploring visitor movement patterns in natural recreational areas. Tourism Management, 33: 672–682.
Panelli R. 2009. Social Geography. [W:] R. Kitchin, N. Thrift (red.), International Encyclopedia of Human Geography. Elsevier, s. 185–194.
Piskorz-Ryń A. 2017. Inteligentne miasta jako wyzwanie dla samorządu terytorialnego. Przedsiębiorczość i Zarządzanie, s. 23–33.
Rae A., Singleton A. 2015. Putting big data in its place: a Regional Studies and Regional Science perspective. Regional Studies, Regional Science, 2: 1–5.
Robertson C., Feick R. 2016. Bumps and bruises in the digital skins of cities: unevenly distributed user-generated content across US urban areas. Cartography and Geographic Information Science, 43: 283–300.
Robinson A.C., Demšar U., Moore A.B., Buckley A., Jiang B., Field K., Kraak M.-J., Camboim S.P., Sluter C.R. 2017. Geospatial big data and cartography: research challenges and opportunities for making maps that matter. International Journal of Cartography, 3:sup1, 32–60.
Rodak O. 2017. Twitter jako przedmiot badań socjologicznych i źródło danych społecznych: perspektywa konstruktywistyczna. Studia Socjologiczne, 3(226): 209–236.
Ruggles S. 2014. Big Microdata for Population Research. Demography, 51: 287–297.
Rui J.R., Stefanone M.A. 2013. Strategic Image Management Online: Self-Presentation, Self-Esteem and Social Network Perspectives. Information Communication and Society 16 (8): 1286–1305.
Rzeszewski M. 2015a. Cyberpejzaż miasta w trakcie megawydarzenia: Poznań, Euro 2012 i Twitter. Studia Regionalne i Lokalne, 123–137.
Rzeszewski M. 2015b. Systemy lokalizacji satelitarnej w analizie zachowań przestrzennych użytkowników miasta. Rozwój Regionalny i Polityka Regionalna, 111–121.
Rzeszewski M., Luczys P. 2018. Care, Indifference and Anxiety – Attitudes toward Location Data in Everyday Life. ISPRS International Journal of Geo-Information, 7: 383.
Salesses P., Schechtner K., Hidalgo C.A. 2013. The Collaborative Image of The City: Mapping the Inequality of Urban Perception. PLOS ONE 8, e68400.
Shaw S.L., Tsou M.H., Ye X. 2016. Human dynamics in the mobile and big data era. International Journal of Geographical Information Science, 30(9): 1687–1693.
Shelton T., Poorthuis A., Graham M., Zook M. 2014. Mapping the data shadows of Hurricane Sandy: Uncovering the sociospatial dimensions of 'big data.' Geoforum, 52: 167–179.

Shelton T., Poorthuis A. 2019. Atlanta's Neighborhood Planning Unit system. Annals of the American Association of Geographers, 1–21.
Stefanidis A., Crooks A., Radzikowski J. 2013. Harvesting ambient geospatial information from social media feeds. GeoJournal, 78: 319–338.
Stephens M. 2013. Gender and the GeoWeb: divisions in the production of user-generated cartographic information. GeoJournal, 78: 981–996.
Stephens M., Poorthuis A. 2015. Follow thy neighbor: Connecting the social and the spatial networks on Twitter. Computers, Environment and Urban Systems, 53: 87–95.
Sun Y., Fan H., Helbich M., Zipf A. 2013. Analyzing Human Activities Through Volunteered Geographic Information: Using Flickr to Analyze Spatial and Temporal Pattern of Tourist Accommodation. [W:] J.M. Krisp (red.), Progress in Location-Based Services. Springer, Berlin, Heidelberg, s. 57–69.
Tao S., Corcoran J., Mateo-Babiano I., Rohde D. 2014. Exploring Bus Rapid Transit passenger travel behaviour using big data. Applied Geography, 53: 90–104.
Thatcher J. 2017. You are where you go, the commodification of daily life through 'location.' Environment and Planning, A 49: 2702–2717.
Tomanek K. 2014. Analiza sentymentu – metoda analizy danych jakościowych: przykład zastosowania oraz ewaluacja słownika RID i metody klasyfikacji bayesa w analizie danych jakościowych. Przegląd Socjologii Jakościowej, 10(2): 118–36.
Tsou M.H. 2015. Research challenges and opportunities in mapping social media and Big Data. Cartography and Geographic Information Science, 42(1): 70–74.
Turner A. 2006. Introduction to neogeography. O'Reilly Media, Inc.
Waszewski J., Gurtowski M. 2015. Cyfrowy rasizm? Zautomatyzowane techniki nadzoru jako narzędzie segregacji i dyskryminacji. Transformacje, 1/2(84/85): 88–107.
Wilson M.W. 2015. Morgan Freeman is dead and other big data stories. Cultural Geographies, 22: 345–349.
Wójcik M., Suliborski A. 2014. Geografia społeczna w Polsce – geneza, koncepcje i zróżnicowanie problemowe, ze szczególnym uwzględnieniem studiów geograficzno-miejskich w ośrodku łódzkim. [W:] A. Suliborski, M. Wójcik (red.), Dysproporcje Społeczne i Gospodarcze w Przestrzeni Łodzi. Czynniki, Mechanizmy, Skutki. Wydawnictwo Uniwersytetu Łódzkiego, Łódź, s. 17–48.
Wu W., Wang J., Dai T. 2016. The geography of cultural ties and human mobility: Big data in urban contexts. Annals of the American Association of Geographers, 106(3): 612–630.
van Meeteren M., Poorthuis A. 2018. Christaller and "big data": recalibrating central place theory via the geoweb. Urban Geography, 39: 122–148.
Yin L., Cheng Q., Wang Z., Shao Z. 2015. 'Big data' for pedestrian volume: Exploring the use of Google Street View images for pedestrian counts. Applied Geography, 63: 337–345.
Zajadacz A. 2017. Dyssatysfakcja w przestrzeni turystycznej. Negatywne opinie użytkowników TripAdvisor na temat głównych atrakcji turystycznych wybranych miast w Polsce. Prace i Studia Geograficzne, 62(3): 63–88.
Zasina J. 2018. The Instagram Image of the City. Insights from Lodz, Poland. Bulletin of Geography. Socio-economic Series, 42: 213–225.
Zook M., Graham M., Shelton T., Gorman S. 2010. Volunteered geographic information and crowdsourcing disaster relief: a case study of the Haitian earthquake. World Medical & Health Policy, 2: 7–33.
Zook M. 2017. Crowd-sourcing the smart city: Using big geosocial media metrics in urban governance. Big Data & Society, 4: 205395171769438.

Finansowanie: Tekst powstał dzięki badaniom realizowanym w ramach projektu Narodowego Centrum Nauki UMO-2018/31/B/HS4/00059. Artykuł jest również wynikiem badań zrealizowanych w ramach grantu Narodowego Centrum Nauki DEC-2012/05/E/HS4/01498

Does more mean better? Quantitative research in the socio-economic geography in the age of Big Data

Abstract: Big Data are inseparable part of the current research methodology in many scientific disciplines – including socio-economic geography. They are being used as a new data in old research problems, as a new research objects and as a means to investigate digital geographies. Big Data are being criticized for lack of representativeness and problematic representation, heterogeneity, and ethical issues. Despite these problems, their presence as a part of the geographical research methodology is increasingly frequent. We propose a set of recommendations that in our view allows a researcher to critically assess the validity of utilizing Big Data in a given research project.

Key words: Big Data, socio-economic geography, quantitative methods, Data Science