EKONOMETRIA ECONOMETRICS 2(48) • 2015

ISSN 1507-3866
e-ISSN 2449-9994

Maciej Beręsewicz, Marcin Szymkowiak
Uniwersytet Ekonomiczny w Poznaniu
Urząd Statystyczny w Poznaniu
e-mails: maciej.beresewicz@ue.poznan.pl; m.szymkowiak@ue.poznan.pl

BIG DATA W STATYSTYCE PUBLICZNEJ – NADZIEJE, OSIĄGNIĘCIA, WYZWANIA I ZAGROŻENIA

BIG DATA IN OFFICIAL STATISTICS – HOPES, ACHIEVEMENTS, CHALLENGES AND RISKS

DOI: 10.15611/ekt.2015.2.01
JEL Classification: C8

Streszczenie: W artykule opisany został aktualny stan wykorzystania tzw. *big data* w statystyce oficjalnej. Przedstawione zostały doświadczenia wybranych – krajowych urzędów statystycznych w praktycznym zastosowaniu danych pochodzących od operatorów telefonii komórkowej, czujników ruchu, z portali społecznościowych czy danych transakcyjnych na potrzeby statystyki publicznej. Sformułowane zostały również szanse, wyzwania i zagrożenia, jakie stoją przed urzędami statystycznymi w wykorzystaniu tego typu informacji w nurcie statystyki publicznej.

Słowa kluczowe: *big data*, statystyka publiczna, internetowe źródła danych.

Summary: The main purpose of the article is to describe the state of the art in using big data in official statistics. The article presents selected examples of how data from mobile operators, sensors, social media or scanners are used by national statistical offices. The authors also identify chances, challenges and risks related to the use of big data in the field of official statistics.

Keywords: big data, official statistics, internet data sources.

1. Wstęp

W statystyce publicznej, prowadzonej przez krajowe urzędy statystyczne na całym świecie, od wielu dziesięcioleci wykorzystuje się dane pochodzące ze spisów i badań reprezentacyjnych. Od kilku lat na większą skalę wykorzystuje się również rejestry administracyjne.

Ze względu na rosnące oczekiwania i potrzeby odbiorców informacji statystycznej, doskonalone technologie informatyczne oraz zmieniające się wyzwania statystyki publicznej, wynikające z jej nowej roli w XXI wieku, coraz więcej uwagi

przywiązuje się do alternatywnych źródeł informacji w postaci tzw. *big data* (niestatystycznych źródeł danych), które mogą obejmować [Beręsewicz 2015; Daas i in. 2015]:

- dane pochodzące z sieci i serwisów społecznościowych czy portali internetowych (na przykład Facebook, Twitter, Google Trends czy Allegro), które są z reguły nieustrukturyzowane i pojawiają się nieregularnie;
- dane pochodzące z różnego rodzaju czujników, przekaźników, kamer, satelitów czy sieci komórkowych, które na ogół są bardziej uporządkowane i odznaczają się większą dokładnością.

Wykorzystanie *big data* przez długi czas odbywało się poza głównym nurtem statystyki publicznej, w której preferuje się w dalszym ciągu informacje pochodzące ze spisów, badań reprezentacyjnych czy rejestrów administracyjnych. Ze względu jednak na rosnące znaczenie tego nowego źródła informacji niektóre urzędy statystyczne – na przykład w Holandii, zdecydowały się przetestować ich użyteczność na potrzeby statystyki oficjalnej. Jest to zadanie niezwykle trudne, a dotyczy zarówno problematycznych kwestii prawnych, dostępności danych i sposobu ich przechowywania, zagadnień związanych z koniecznością zachowania tajemnicy statystycznej czy, co ważniejsze, problemów dotyczących obciążenia będącego konsekwencją charakteru „masowych" źródeł. Istnieje zatem ogromne zapotrzebowanie na zaadaptowanie danych pochodzących z nowych źródeł danych (*big data*) do potrzeb i celów statystyki oficjalnej, tak aby zmaksymalizować korzyści i zminimalizować zagrożenia wynikające z korzystania z tego typu źródeł informacji w statystyce publicznej [Pfeffermann 2015; Szreder 2015; Witkowski 2014].

Głównym celem referatu będzie przedstawienie możliwości wykorzystania *big data* w statystyce publicznej. Wskazane zostaną najważniejsze korzyści i szanse związane ze stosowaniem tych zbiorów informacji w statystyce oficjalnej, jak i główne zagrożenia związane z ich wykorzystaniem przez krajowe urzędy statystyczne. Rozważania na temat użyteczności *big data* zilustrowane zostaną przykładami ich użycia przez wybrane krajowe urzędy statystyczne na potrzeby statystyki oficjalnej.

2. Klasyczne źródła danych w statystyce publicznej

Zgodnie z programem badań statystycznych statystyki publicznej Główny Urząd Statystyczny może prowadzić badania na całej zbiorowości (spisy powszechne) albo jako badania reprezentacyjne na wylosowanej lub dobranej celowo próbie danej zbiorowości. Są to, w ujęciu prezentowanym w artykule, klasyczne źródła danych pozyskiwane w statystyce publicznej. Niewątpliwą zaletą spisów jest ich unikatowy charakter, umożliwiający zebranie informacji na temat wielu obszarów życia, związanych m.in. z rynkiem pracy, ubóstwem czy migracjami. Ograniczenie stanowi fakt, że dane ze spisów pozyskiwane są zazwyczaj w odstępach dziesięcioletnich, co przy obecnym zapotrzebowaniu informacyjnym na bieżące dane jest okresem za długim. W związku z powyższym statystyka publiczna w Polsce pozyskuje informacje

z innych, aniżeli spisy, źródeł danych. Są to przede wszystkim dane pochodzące z badań reprezentacyjnych związanych z określonym tematem badawczym (EU-SILC, Badanie Budżetów Gospodarstw Domowych, Badanie Aktywności Ekonomicznej Ludności itd.). Badania te mają zazwyczaj charakter reprezentacyjny i są prowadzone na wylosowanej próbie umożliwiającej uogólnianie wyników na całą zbiorowość w odpowiednich przekrojach z pewnym błędem szacunku. Dostarczają one bieżących informacji, gdyż są prowadzone bardzo często w cyklu miesięcznym czy kwartalnym, co stanowi o ich niewątpliwej zalecie. Niestety informacje pozyskiwane z tych badań agregowane są bardzo często na poziomie całego kraju czy co najwyżej województw. Istnieje jednak duże zapotrzebowanie na bieżące dane na niższych poziomach agregacji przestrzennej z dodatkowo wyznaczonymi przekrojami (na przykład poziom gminy z uwzględnieniem informacji o płci osoby). Remedium może stanowić statystyka małych obszarów, dzięki której może być możliwe pozyskanie informacji dla szczegółowo zdefiniowanych domen bądź skorzystanie z innych, aniżeli klasyczne, źródeł danych. Mowa tu przede wszystkim o rejestrach administracyjnych, lecz także o danych pochodzących z Internetu (portale społecznościowe, serwisy aukcyjne, fotoradary, kamery, GPS, zdjęcia satelitarne itd.). O ile w polskiej statystyce publicznej rejestry administracyjne są już wykorzystywane, czego przykładem może być Narodowy Spis Powszechny Ludności i Mieszkań 2011, o tyle inne – wspomniane wcześniej – źródła informacji nie są wykorzystywane w ogóle, a ich użycie może mieć miejsce w bliżej nieokreślonej przyszłości. W tym miejscu należy podkreślić, że niektóre urzędy statystyczne podjęły już pierwsze próby związane z wykorzystaniem nowych źródeł danych i dokonały ich adaptacji na potrzeby statystyki publicznej. Dotyczy to m.in. wykorzystania danych pochodzących od operatorów sieci komórkowych, z czujników natężenia ruchu czy portali społecznościowych.

Praktyczne zastosowania tych podejść na potrzeby zasilania informacyjnego w statystyce publicznej przez krajowe urzędy statystyczne zostaną opisane w dalszej części tego artykułu. Ma to umożliwić z jednej strony wskazanie „drzemiących możliwości i potencjału'' tego typu źródeł danych w statystyce oficjalnej, a z drugiej wskazać na pewne ograniczenia i zagrożenia związane z ich wykorzystaniem. Warto przy tym podkreślić, że źródła te mają „charakter niestatystyczny", tzn. nie zostały stworzone na potrzeby statystyki publicznej. Jak pokazuje jednak praktyka innych państw, mogą one zostać z powodzeniem włączone w nurt statystyki publicznej i stanowić pełnoprawne źródło informacji o charakterze oficjalnym.

3. Nowe źródła danych w statystyce publicznej

Obok klasycznych źródeł informacji, wykorzystywanych przez urzędy statystyczne i wymienionych w poprzednim podpunkcie, popularność zyskują nowe, nie mające charakteru statystycznego, źródła danych. Główną cechą je wyróżniającą jest cel ich powstania, który może dotyczyć m.in. bieżącej ewidencji wspomagającej zarządzanie

na poziomie przedsiębiorstwa, miasta, regionu czy kraju. Urzędy statystyczne, jak i instytucje naukowe coraz częściej sięgają po niestatystyczne źródła danych w celu opisu badanych zbiorowości (populacji generalnej), a zwrot ten nazywany jest w literaturze zmianą paradygmatu, tj. odejściem od danych stworzonych przez statystyków (*designed based*) do wykorzystania wszelkich istniejących źródeł (*process based*).

Omawiane źródła możemy podzielić na dwie grupy według charakteru ich powstania: **stworzone na potrzeby** (1) **sektora publicznego** oraz (2) **sektora prywatnego**. Pierwszą grupę stanowią głównie rejestry administracyjne (m.in. PESEL, REGON, PIT/CIT, ZUS, RCiWN) oraz inne dane ewidencyjne (m.in. pochodzące z fotoradarów, satelitów, pomiaru ruchu samochodowego czy giełdy), które mają za zadanie osiągnięcie celów określonych w odpowiednich ustawach. Natomiast druga grupa związana jest z działalnością prywatną (*profit* i *non-profit*), których przykładem mogą być dane transakcyjne sieci handlowych, transakcje bankowe, sieci komórkowe i Internet Rzeczy (m.in. GPS, smartfony), portale internetowe (m.in. Facebook, Twitter, Pracuj.pl, OtoDom) czy szeroko rozumiany e-commerce (np. Allegro, Ceneo).

Nowe źródła danych pojawiają się głównie pod nazwą *big data*, co ma ujmować ich charakter przez wskazanie na: (1) duży wolumen danych (*volume*), (2) dużą zmienność i dynamikę ich powstawania (*velocity*) oraz (3) dużą różnorodność oraz ich nieustrukturyzowanie (*variety*). Dodatkowo omawiane źródła pojawiają się w literaturze pod nazwą danych organicznych, które powstają wskutek ludzkiego działania, a nie zdefiniowanego i wystandaryzowanego badania statystycznego. Należy jednak zaznaczyć, że pojawiające się definicje *big data* nie mają charakteru formalnego, a jedynie są wskazaniem pewnych kluczowych elementów związanych głównie z problemami ich przetwarzania i analizowania.

W przeciwieństwie do *big data*, klasyczne źródła danych charakteryzują się na ogół niskim wolumenem, liczonym w tysiącach lub kilku milionach rekordów, i odzwierciedlają jednostki statystyczne. Zmienność i dynamika są określone przez charakter badania: częściowe są organizowane co miesiąc, kwartał czy rok, natomiast spisy zazwyczaj co dziesięć lat. Dodatkowo dane są w pełni ustrukturyzowane, posiadają odpowiednie definicje jednostek oraz zmiennych, jak i przetrzymywane są w standardowych bazach danych (np. SQL). Jedyne nieustrukturyzowanie może pojawiać się w odniesieniu do pytań otwartych znajdujących się w kwestionariuszu ankietowym towarzyszącym danemu badaniu; są one jednak rzadkością, głównie dlatego, że badania prowadzone przez statystykę publiczną mają być porównywalne w czasie oraz w przestrzeni.

4. *Big data* i inne nowe źródła danych

W celu usystematyzowania wiedzy na temat istniejących nowych źródeł danych, w tym *big data*, w kontekście możliwości ich wykorzystania na potrzeby statystyki publicznej, dokonano odpowiedniej ich klasyfikacji. Tabela 1 przedstawia pięć najważniejszych nowych grup źródeł danych, których wyróżnikiem jest sposób po-

wstawania oraz przydatność do tworzenia informacji statystycznej. Pierwszą grupę stanowią hasła oraz ich kategorie wyszukiwane w przeglądarce Google[1], a udostępnione w ramach usługi Google Trends[2]. Kolejna grupa skupia specjalistyczne portale internetowe, poświęcone głównie sprzedaży elektronicznej, pośrednictwu (na przykład na rynku nieruchomości) czy porównywaniu cen i usług. Następną grupą, związaną z wykorzystaniem Internetu, są portale społecznościowe, które służą zarówno do dzielenia się informacjami i wiadomościami przez ich użytkowników, jak i do celów marketingowych (marketingu społecznościowego). Dwie pozostałe grupy *big data* związane są ze sposobem ich powstawania wynikającym z korzystania z usług sieci komórkowych (telefonów oraz smartfonów) oraz pasywnego zbierania danych (*passive data collection*) wynikającego z korzystania z różnych urządzeń określanych mianem Internetu Rzeczy (*Internet of Things*[3]).

Tabela 1. Niestatystyczne nowe źródła danych

Źródło	Możliwe zastosowania	Przykłady
Wyszukiwania w Internecie	Prognozowanie popytu konsumenckiego, opis sytuacji na rynku pracy (np. stopy bezrobocia)	Google Trends
Specjalistyczne portale internetowe	Budowa indeksów cen, tworzenie wskaźników inflacji, ofert pracy czy wynagrodzenia	e-Commerce (na przykład Allegro, Ceneo, Skąpiec, SkyScanner), specjalistyczne portale (na przykład OtoMoto, OtoDom, Gratka.pl), pośrednictwo pracy (na przykład Pracuj.pl, Indeed.com, LinkedIn, Goldenline)
Portale społecznościowe (*Social Media*)	Badanie nastrojów społecznych, zaufania konsumenckiego, oczekiwań inflacyjnych	m.in. Twitter, Facebook, Instagram
Sieci komórkowe	Badania przepływów ludności (dojazdy do pracy), migracje, badania powiązań między użytkownikami	Dostawcy sieci komórkowych oraz Internetu (na przykład Play, Orange, T-Mobile, Plus)
Internet Rzeczy	Indeksy cen oraz inflacji, statystyki transportu, przepływy ludności i migracje, zdjęcia satelitarne, geolokalizacja (m.in. GPS), pomiar energii elektrycznej (tzw. SmartMeters)	Dane z sieci handlowych (tzw. *scanner data*) czy banków (transakcje), nawigacja GPS, satelity, Google Maps, OpenStreetMap, sensory drogowe (mierzące natężenie ruchu), RFID

Źródło: opracowanie własne.

[1] Szacuje się, że wyszukiwarka Google ma około 95% udziału w polskim rynku wyszukiwarek – na podstawie: http://ranking.pl/pl/rankings/search-engines-domains.html.

[2] Więcej informacji o Google Trends można znaleźć na stronie internetowej https://support.google.com/trends/?hl=pl#topic=4365599.

[3] „Inteligentne" urządzenia pomagające człowiekowi w wypełnianiu codziennych obowiązków, łączą się z siecią komputerową w celu komunikacji między sobą i komunikacji ze środowiskiem. Urządzenia są automatycznie aktywowane (bez konieczności działania ze strony człowieka) w celu wykonania określonego zadania (źródło: http://www.kti.ue.poznan.pl/node/1504#iot).

W dalszej części artykułu opisane zostały pokrótce najważniejsze zastosowania źródeł danych przedstawionych w tab. 1 na potrzeby statystyki publicznej przez krajowe urzędy statystyczne wybranych państw.

5. Wybrane doświadczenia urzędów statystycznych w wykorzystaniu *big data* na potrzeby statystyki publicznej

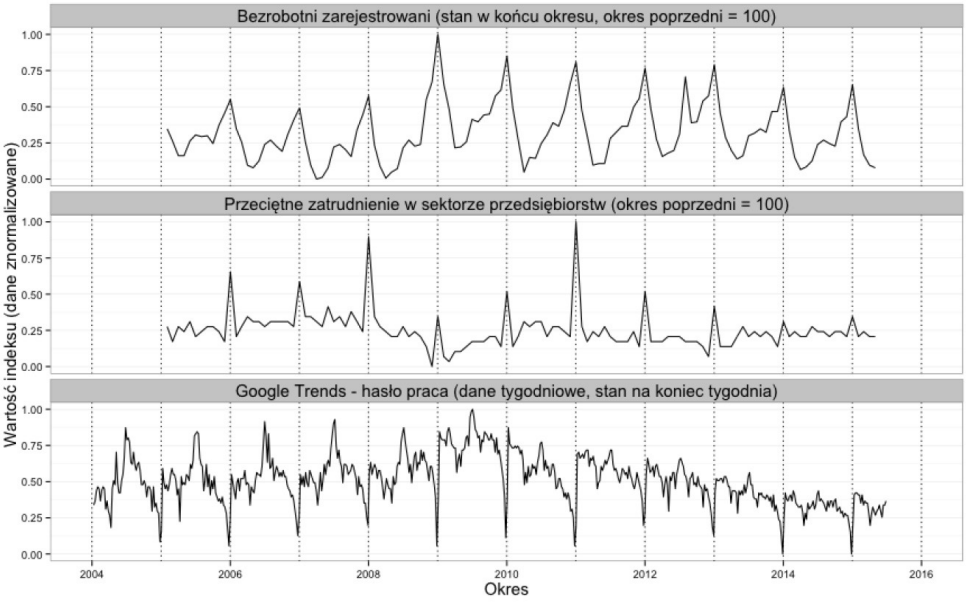
Nowe źródła danych i *big data* są już wykorzystywane przez urzędy statystyczne oraz instytucje międzynarodowe na całym świecie. Przykładem ostatnich badań prowadzonych w tym zakresie są projekty realizowane przez UNECE [UNECE 2014] oraz Eurostat [Eurostat 2014], których głównym celem jest ocena możliwości i przydatności *big data* na potrzeby statystyki publicznej. Wiele urzędów statystycznych prowadzi badania poświęcone wykorzystaniu nowych źródeł danych (m.in. ISTAT, ONS, US Census Bureau) w kontekście statystyki oficjalnej. Pionierem w tym zakresie jest jednak holenderski urząd statystyczny (Statistics Netherlands, Centraal Bureau voor de Statistiek – CBS). CBS od lat 70. XX wieku wykorzystuje rejestry administracyjne, od lat 90. współpracuje z największymi sieciami handlowymi w celu pozyskania danych transakcyjnych (*scanner data*), a w ostatnich latach nawiązał współpracę z operatorami telefonii komórkowej. CBS wykorzystuje dane pochodzące z automatycznych sieci pomiarowych ruchu drogowego czy portali społecznościowych. Część omawianych w dalszej kolejności przykładów będzie bazować na doświadczeniach holenderskich, które zostaną rozszerzone o przykłady pochodzące z innych państw. Wskazane zostaną również możliwości wykorzystania nowych źródeł danych w odniesieniu do polskiej statystyki publicznej.

5.1. Hasła wyszukiwane w Internecie – Google Trends

Google Trends (GT) jest usługą udostępnioną przez firmę Google, która umożliwia analizę popularności haseł (oraz ich kategorii) wpisywanych w wyszukiwarce Google przez jej użytkowników. Algorytm GT polega na porównaniu liczby zapytań danego hasła lub kategorii do ogólnej liczby zapytań w tym samym czasie. GT umożliwia porównywanie częstości zapytań lub kategorii w czasie oraz według ich pochodzenia, na poziomie krajów oraz regionów (na przykład województw). Niestety GT udostępnia dane bez wskazania ich kontekstu, a także brakuje informacji demograficznej, kim są osoby, które wpisują do wyszukiwarki konkretne zapytania. Jednakże mimo pewnych ograniczeń, które GT zawiera, wykorzystywane jest ono w badaniach nad prognozowaniem stopy bezrobocia [Vicente, López-Menéndez, Pérez 2015], popytu konsumenckiego [Vosen, Schmidt 2011] czy jako źródło danych pomocniczych w statystyce małych obszarów [Porter i in. 2014].

W Polsce brakuje opracowań, które umożliwiają ocenę GT jako źródła danych, m.in. na potrzeby opisu sytuacji panującej na rynku pracy. Aby zapełnić tę lukę, poniżej przedstawiono popularność wyszukiwania słowa ,,praca" w kategorii Rynek

pracy[4] i odniesiono to do rzeczywistej stopy bezrobocia rejestrowanego publikowanej przez Główny Urząd Statystyczny. Można zaobserwować, że popularność hasła „praca" w GT jest uzależniona od wysokości stopy bezrobocia w poszczególnych okresach. Widoczne jest przesunięcie czasowe polegające na tym, że jeśli w danym miesiącu zaobserwowana jest wysoka stopa bezrobocia, to w kolejnych miesiącach następuje wzrost wyszukiwań słowa „praca'' w wyszukiwarce. GT może stanowić więc cenne uzupełniające źródło informacji obrazujące nastroje panujące na rynku pracy (rys. 1). Należy jednak zaznaczyć, że nie jest znany dokładny algorytm stosowany w GT, który może błędnie kategoryzować wyszukiwane słowa. Co ważniejsze, nie jest znana rzeczywista motywacja stojąca za wpisywaniem danej kategorii słów do wyszukiwarki Google. Dlatego w obecnej chwili GT jako źródło danych na potrzeby statystyki publicznej może być wykorzystywane w ograniczonym zakresie.



Rys. 1. Zarejestrowani bezrobotni, przeciętne zatrudnienie w sektorze przedsiębiorstw a popularność hasła „praca'' w Google Trends

Źródło: opracowanie własne na podstawie danych Głównego Urzędu Statystycznego dotyczących indeksu łańcuchowego liczby zarejestrowanych bezrobotnych na koniec miesiąca i przeciętnego zatrudnienia w sektorze przedsiębiorstw oraz popularności hasła „praca" w Google Trends w ujęciu tygodniowym.

[4] Kategoria: Praca i edukacja → Rynek pracy.

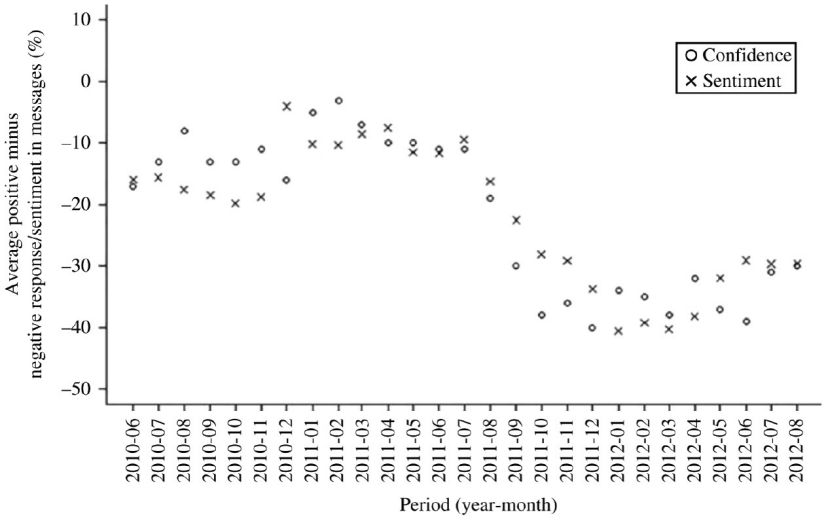
5.2. Specjalistyczne portale internetowe oraz *web-scraping*

Specjalistyczne portale internetowe są również wykorzystywane jako cenne źródło informacji przez niektóre krajowe urzędy statystyczne na potrzeby statystyki publicznej. Poniżej zostanie opisana możliwość wykorzystania tego typu źródła w ocenie wskaźnika inflacji na podstawie danych o cenach produktów ze sklepów internetowych.

Wzrost popularności handlu elektronicznego i związane z tym coraz częstsze dokonywanie zakupów przez klientów w Internecie otwierają nowe możliwości pomiaru cen. Umożliwia to również badania związane z popularnością poszczególnych produktów. Serwisy internetowe oferujące sprzedaż *on-line* stały się nowoczesną półką sklepową, która pozwala na dokładniejszą ocenę produktów i świadczonych usług. Związany z tym elektroniczny charakter prowadzonych transakcji pozwala na nieosiągalne do tej pory możliwości w pozyskiwaniu informacji bez konieczności obciążania respondentów poprzez wykorzystanie automatycznej ekstrakcji danych z sieci (*web-scraping*). Najbardziej znanym przykładem wykorzystania danych pochodzących ze sklepów internetowych jest projekt The Billion Price opracowany przez naukowców z Instytutu Technologicznego w Massachusetts (MIT) – Alberto Cavallo i Roberto Rigobona. Pierwotnym założeniem projektu była ocena publikowanych przez argentyński urząd statystyczny wskaźników inflacji [Cavallo 2013] i ich odniesienie do kilku innych krajów z Ameryki Południowej. Natomiast wraz z rozwojem projektu powstała firma PriceStats, której celem było pobieranie informacji o cenach ze sklepów internetowych z blisko 60 krajów, dostarczanie dziennych, miesięcznych i kwartalnych wskaźników inflacji oraz przeprowadzanie badań makroekonomicznych. Oprócz tego pojawiły się również pierwsze prace związane z wykorzystaniem danych do opisu zmian cen ubrań [Griffioen, de Haan, Willenborg 2014], produktów spożywczych [Swier 2015] czy rynku nieruchomości [Beręsewicz 2015] w kontekście statystyki publicznej.

5.3. Portale społecznościowe

Klasycznym przykładem *big data* są dane znajdujące się na portalach społecznościowych, które są utożsamiane głównie z portalem Facebook czy Twitter. Obydwa te portale są obecnie oceniane przez urzędy statystyczne w kontekście możliwości ich użycia na potrzeby statystyki publicznej. Wskazuje się kilka dziedzin, w których portale społecznościowe mogą być wykorzystywane – m.in. badania przepływów ludności, nastrojów społecznych i zaufania konsumenckiego. Prym w tych badaniach wiedzie wspomniany już CBS, który na podstawie Twittera oraz Facebooka mierzy opinie (*sentiment*) jako odpowiednik publikowanego przez CBS indeksu zaufania konsumenckiego (rys. 2). Indeks ten to wskaźnik, na który składają się wyniki badania gospodarstw domowych, mający na celu ocenić względną kondycję finansową, siłę nabywczą i zaufanie przeciętnego konsumenta. Na indeks zaufania konsumentów wpływają: obecna ocena sytuacji konsumentów, aktualny stan gospo-



Legenda: Kółkami oznaczone są oszacowania zaufania konsumenckiego (*confidence*) publikowanego przez CBS, natomiast krzyżykami opinie (*sentiment*) oszacowane na podstawie Facebooka i Twittera. Wskaźniki są obliczane jako różnica między średnią pozytywną a negatywną liczbą opinii. Dane prezentowane w ujęciu miesięcznym od czerwca 2010 do sierpnia 2012.

Rys. 2. Wskaźnik zaufania konsumenckiego a nastroje użytkowników portalu Twitter oraz Facebook

Źródło: wykres na podstawie [Daas i in. 2015].

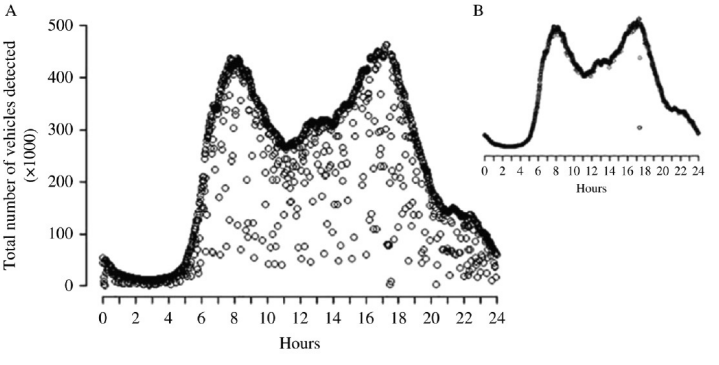
darki w opinii konsumentów oraz oczekiwania konsumentów co do rozwoju sytuacji w gospodarce w perspektywie najbliższych miesięcy. Daas i in. [2015] wskazują na silną korelację oraz kointegrację zaufania konsumenckiego z opiniami publikowanymi na Twitterze oraz Facebooku pomimo braku danych o pokryciu informacyjnym badanej populacji.

5.4. Internet Rzeczy

Przedostatnią omawianą grupą *big data* jest tzw. Internet Rzeczy. Jest to koncepcja, wedle której jednoznacznie identyfikowalne przedmioty mogą pośrednio albo bezpośrednio gromadzić, przetwarzać lub wymieniać dane za pośrednictwem sieci komputerowej. Obecnie dane pochodzące z Internetu Rzeczy (m.in. zdjęcia satelitarne, transakcje) są wykorzystywane m.in. do szacowania upraw, wykorzystania w charakterze zmiennych pomocniczych w statystyce małych obszarów (m.in. Night-time-light data) czy budowy indeksów cen. Natomiast przykładem pierwszego wdrożenia *big data* w statystyce publicznej jest wykorzystanie drogowych pętli indukcyjnych (*traffic loop detection*), które mierzą ruch samochodowy. Na terenie Holandii znajduje się 12 622 punktów pomiarowych, których pomiary przetrzymywane są w National Data Warehouse od 2010 roku. Jak podaje Daas [Daas

i in. 2015], tylko w jednym dniu, tj. 1 grudnia 2011 roku, w bazie tej znalazło się 76 milionów rekordów[5]. Wykorzystanie tych danych umożliwiło oszacowanie liczby samochodów poruszających się po drogach w tym dniu oraz określenie natężenia ruchu według minut dla wszystkich samochodów i według wielkości (małe, średnie i duże). Rysunek 3 przedstawia liczbę wykrytych samochodów 1 grudnia 2011 roku przed (A) oraz po (B) zastosowaniu procedury czyszczenia danych. Należy zauważyć nasilenie ruchu około godziny 8 spowodowane wyjazdami do pracy oraz około 17, kiedy rozpoczynają się powroty do domu.

Dane zostały poddane czyszczeniu, a pojawiające się braki danych, spowodowane głównie technicznymi usterkami urządzeń pomiarowych, zostały zaimplementowane z wykorzystaniem podejścia bayesowskiego zakładającego rozkład Poissona dla liczby wykrytych samochodów [Puts, Daas, Teenekes 2015]. Natomiast, w celu zapewnienia reprezentatywności dane zostały przeważone z uwzględnieniem informacji o lokalizacji detektorów, budowy drogi (rozkład zjazdów i wjazdów) oraz długości odcinków między kolejnymi pomiarami [Teenekes, Puts 2015].



Rys. 3. Liczba samochodów wykrytych przez drogowe pętle indukcyjne w Holandii w dniu 1 grudnia 2011 roku

Źródło: wykres na podstawie [Daas i in. 2015].

5.5. Sieci komórkowe

Kolejnym przykładem wykorzystania *big data* jest projekt Feasibility Study on the Use of Mobile Positioning Data for Tourism Statistics [Eurostat 2014] przeprowadzony przez konsorcjum stworzone przez Eurostat, Urząd Statystyczny w Estonii oraz prywatną firmę Positium. Celem projektu była ocena możliwości wykorzystania danych pochodzących z sieci komórkowych do szacowania przepływów ludno-

[5] Jak można łatwo policzyć, od 2010 do 2014 roku, przy założeniu średnio 75 mln rekordów dziennie, otrzymujemy bazę zawierającą około 137 miliardów rekordów.

ści wewnątrz i na zewnątrz Estonii. Oprócz Estonii podobne badania przeprowadzono w Holandii, Czechach, oraz Irlandii [Eurostat 2014]. Badania wskazały silną korelację między oficjalnymi danymi a tymi pozyskanymi z sieci komórkowych. Pomimo obciążeń wartości globalnych (liczby podróżujących) widoczna jest taka sama sezonowość oraz trend. Raport zawiera również podsumowanie najważniejszych zalet (m.in. możliwość estymacji bardziej szczegółowych domen, wykrywania wielokrotnych podróży, lepszego pokrycia informacyjnego dotyczącego podróży śródnocnych) oraz wad (m.in. braku informacji o celu podróży oraz o tym, jaka część podróżnych korzysta z telefonów komórkowych).

6. Szanse, wyzwania i zagrożenia w wykorzystaniu *big data* w statystyce publicznej

W tabeli 2 przedstawiono najważniejsze szanse, wyzwania i zagrożenia stojące przed potencjalnym wykorzystaniem *big data* na potrzeby statystyki publicznej, zarówno w Polsce, jak i na całym świecie. Informacje zawarte w tabeli są wynikiem rozległych prac literaturowych oraz własnych przemyśleń autorów wynikających z doświadczeń w pracy w obszarze statystyki publicznej.

Tabela 2. Szanse, wyzwania i zagrożenia stojące przed wykorzystaniem *big data* w statystyce publicznej

Szanse
1. Zmniejszenie kosztów wybranych badań (np. stosowanie automatycznego pobierania danych z Internetu) 2. Zmniejszenie obciążenia respondentów poprzez wykorzystanie już dostępnych danych 3. Możliwość wykorzystania danych z *big data* jako zmiennych pomocniczych w podejściu modelowym (na przykład w statystyce małych obszarów) 4. Uzyskanie nowych informacji niedostępnych w statystyce publicznej (np. pochodzących z Internetu Rzeczy) 5. Możliwość zastąpienia, uzupełnienia czy poprawy istniejących zbiorów danych (np. skrócenie czasu od zebrania do publikacji danych)
Wyzwania
1. Konieczność nawiązania współpracy z „dostawcami" i „producentami" dużych zbiorów danych (na przykład z operatorami telefonii komórkowej) 2. Uzyskanie dostępu do danych (na przykład z Twittera, Facebooka czy z sieci komórkowej) oraz ich integracja z istniejącym systemem statystyki publicznej 3. Zagadnienia prawne i regulacyjne (m.in. poufność danych oraz ich bezpieczeństwo) 4. Ocena jakości danych z punktu widzenia ich dokładności, przydatności, porównywalności, spójności, terminowości i punktualności 5. Ocena reprezentatywności danych oraz możliwość porównania z istniejącymi źródłami statystyki publicznej – konieczność przeprowadzenia badań 6. Kwestia zapewnienia pokrycia informacyjnego dla subpopulacji, dla których standardowo publikowane są dane w statystyce publicznej (w przekroju płci, grup wieku, stanu cywilnego itp.) oraz związane z tym problemy identyfikacji podstawowych charakterystyk demograficznych

Tabela 2, cd.

7. Problemy związane z dopasowaniem istniejącej struktury oraz metod stosowanych w statystyce publicznej do dużych danych (np. czyszczenia i edycji, imputacja, kalibracji)
Zagrożenia
1. Dostępność danych, które są najczęściej w rękach prywatnych; brak chęci współpracy z urzędami statystycznymi 2. Zagadnienia prawne (m.in. ochrona prywatności, konieczność zachowania tajemnicy statystycznej, „permanentna inwigilacja") 3. Niewystarczające pokrycie informacyjne badanej zbiorowości (m.in. ograniczenia w dostępie do Internetu, problemy z publikacją danych w różnych subpopulacjach, na niskim poziomie agregacji przestrzennej) 4. Obciążenie będące konsekwencją selektywności oraz braku reprezentatywności danych 5. Na ogół dane nie spełniają wymagań metodologicznych statystyki oficjalnej (m.in. w kontekście definicji stosowanych przez statystykę publiczną) 6. Jakość danych (m.in. błędy nielosowe na poziomie jednostki oraz źródła, pomiar w różnych odstępach czasu, dane nieustrukturyzowane) 7. Integracja z istniejącymi źródłami danych statystycznych – brak wspólnych identyfikatorów

Źródło: opracowanie własne.

7. Podsumowanie

Jednym z największych wyzwań badawczych, przed którym stoi statystyka publiczna w XXI wieku, będzie konieczność zmierzenia się ze sposobem wykorzystania nowych źródeł informacji w procesie produkcji oficjalnych danych statystycznych. Stanowi to niewątpliwie nie tylko okazję do pozyskania wielu cennych danych, które nie były do tej pory zbierane w oficjalnym nurcie statystyki uwzględniającym spisy, badania reprezentacyjne czy rejestry administracyjne, lecz przysparzać będzie również szereg problemów i kwestii spornych, które wymagać będą odpowiednich rozwiązań. Dotyczyć to będzie nie tylko znalezienia uregulowań prawnych pozwalających na wykorzystanie informacji pochodzących z *big data* w statystyce publicznej, lecz także rozwoju odpowiednich metod statystycznych umożliwiających szacowanie parametrów z użyciem danych, które z natury rzeczy zawierają błędy[6].

W związku z ogromną liczbą dostępnych źródeł oraz wieloma możliwościami ich potencjalnego użycia, zdaniem autorów, istnieje możliwość ich wykorzystania w statystyce publicznej w Polsce. W pierwszej kolejności może to być związane z:

1) **danymi transakcyjnymi z dużych sieci handlowych** (*scanner data*) – pozyskanie tych danych umożliwiłoby dokładniejszy pomiar zmian cen oraz inflacji;

[6] Może to być związane np. z tym, że informacje na Facebooku czy Twitterze umieszczane są przede wszystkim przez osoby młode. Problem uogólniania wyników na całą zbiorowość z wykorzystaniem tego typu danych związany będzie zatem głównie z brakiem reprezentatywności.

2) **danymi pochodzącymi od dostawców telefonii komórkowych** – penetracja rynku w Polsce wynosi blisko 150%, co wskazuje na bardzo wysokie pokrycie informacyjne badanej populacji i możliwości wykorzystania tych danych w zagadnieniach związanych z mobilnością ludzi czy przepływami do pracy;

3) **danymi pochodzącymi ze specjalistycznych portali internetowych** – serwisy oferujące pośrednictwo lub porównywanie cen mogą być cennym źródłem informacji o sytuacji w gospodarce naszego kraju.

Praktyczne wykorzystanie powyższych źródeł danych w polskich warunkach wymagać będzie w pierwszej kolejności stworzenia odpowiednich uregulowań prawnych, które umożliwią pełnoprawne zastosowanie w statystyce publicznej tego typu informacji. Niezależnie od tego warto zastanowić się nad eksperymentalnym wykorzystaniem w statystyce publicznej w Polsce danych pochodzących z *big data*. Jak pokazują bowiem doświadczenia innych państw, mogą one stanowić cenne źródło informacji w wielu obszarach, zwłaszcza w sytuacjach, w których klasyczne badania nie zapewniają odpowiedniego pokrycia informacyjnego.

Literatura

Beręsewicz M., 2015, *On the representativeness of Internet data sources for the real estate market in Poland*, Austrian Journal of Statistics, 4(2).

Cavallo A., 2013, *Online and official price indexes*: *Measuring Argentina's inflation*, Journal of Monetary Economics, 60(2), 152–165, doi:10.1016/j.jmoneco.2012.10.002.

Daas P.J.H., Puts M.J., Buelens B., Hurk P.A.M. van den, 2015, *Big Data as a Source for Official Statistics*, Journal of Official Statistics, 31(2), s. 249-262.

Eurostat, 2014, *Feasibility Study of the Use of Mobile Positioning Data for Tourism Statistics*, Consolidated Report Eurostat Contract No 30501.2012.001- 2012.452, 30.06.2014.

Griffioen R., de Haan J., Willenborg L., 2014, *Collecting clothing data from the Internet*, Statistics Netherlands, Den Haag.

Pfeffermann D., 2015*, Official Statistics for the Next Decade – Methodological Issues and Challenges*, referat wygłoszony na konferencji NTTS 2015, 10-12 marca 2015, Bruksela.

Porter A.T., Holan S.H., Wikle C.K., Cressie N., 2014, *Spatial Fay–Herriot models for small area estimation with functional covariates*, Spatial Statistics, 10, s. 27-42.

Puts M., Daas P., Teenekes M., 2015, *High frequency road sensor data for official statisitics*, referat wygłoszony na konferencji NTTS 2015, 10-12 marca 2015, Bruksela, dostęp online: http://www.cros-portal.eu/sites/default/files//Presentation%20S13AP4.pdf.

Swier N., 2015, *Using Web Scraped Data to Construct Consumer Price Indices*, referat wygłoszony na konferencji NTTS 2015, 10-12 marca 2015, Bruksela, http://www.cros-portal.eu/sites/default/files//Presentation%20S6AP3.pdf.

Szreder M., 2015, *Big data wyzwaniem dla człowieka i statystyki*, Wiadomości Statystyczne, Główny Urząd Statystyczny, sierpień, Warszawa.

Teenekes M., Marco P., 2015, *High Frequency Road Sensor Data for Official Statisitics*, referat wygłoszony na konferencji NTTS 2015, 10-12 marca 2015, Bruksela, http://www.cros-portal.eu/sites/default/files//Presentation%20S13AP5.pdf.

UNECE 2014, *Big Data for Official Statistics*, Technical Workshop Report, http://www1.unece.org/stat/platform/pages/viewpage.action?pageId=102664009.

Vicente M.R., López-Menéndez A.J., Pérez R., 2015, *Technological Forecasting & Social Change Forecasting unemployment with internet search data: Does it help to improve predictions when job destruction is skyrocketing ?*, Technological Forecasting & Social Change, 92, 132-139. doi:10.1016/j.techfore.2014.12.005.

Vosen S., Schmidt T., 2011, *Forecasting Private Consumption: Survey-based Indicators vs. Google Trends*, Journal of Forecasting, 578(1), s. 565-578.

Witkowski J., 2014, *Statystyka oficjalna wobec wyzwań globalnych*, Wiadomości Statystyczne, nr 4 (635), Główny Urząd Statystyczny, Polskie Towarzystwo Statystyczne.