STUDIA OECONOMICA POSNANIENSIA 2019, vol. 7, no. 4

DOI: 10.18559/SOEP.2019.4.3

Sławomir Kuźmar
https://orcid.org/0000-0002-2458-0463
Uniwersytet Ekonomiczny w Poznaniu, Katedra Makroekonomii i Badań nad Rozwojem
slawomir.kuzmar@ue.poznan.pl

BIG DATA – ISTOTA I MOŻLIWOŚCI ZASTOSOWANIA NA PRZYKŁADZIE PROGNOZOWANIA BEZROBOCIA[1]

Streszczenie: W dobie rosnącej informatyzacji, a w konsekwencji digitalizacji współczesnych społeczeństw, zagadnienia związane z wykorzystywaniem nowych niestandardowych zestawów danych, określanych współcześnie mianem „Big Data", zdają się jedną z najbardziej aktualnych kwestii dotyczących analiz empirycznych w ekonomii. Z uwagi na rosnące znaczenie, a także coraz łatwiejszy dostęp do tego typu danych, celem opracowania jest ocena przydatności danych zaliczanych do Big Data (danych dostępnych w ramach analiz Google trends) jako potencjalnie wartościowych w prognozowaniu wielkości bezrobocia w Polsce. Przeprowadzone rozważania z wykorzystaniem metod prognostycznych typu ARIMA wykazały, że dane tego typu mogą być przydatne w ocenie sytuacji na polskim rynku pracy. Wyniki prowadzonych analiz wykazały bowiem, że dodanie kolejnej zmiennej określającej liczbę konkretnych zapytań internetowych związanych z poszukiwaniem pracy podnosiło właściwości prognostyczne modelu prognozującego poziom bezrobocia w Polsce.

Słowa kluczowe: Big Data, prognozy ekonomiczne, bezrobocie, Google trends.

Klasyfikacja JEL: C22, C43, C82, E27.

[1] Projekt finansowany w ramach programu Ministra Nauki i Szkolnictwa Wyższego pod nazwą „Regionalna Inicjatywa Doskonałości" w latach 2019–2022, nr projektu 004/RID/2018/19, kwota finansowania 3 000 000 zł.

BIG DATA – DEFINITIONS AND APPLICABILITY, AN EXAMPLE OF UNEMPLOYEMENT FORECATSING

Abstract: In the era of the growing informatization of the world the term Big Data seems to be one of the most recent and popular issues concerning economic and social analysis according both to functioning economic agents, economic policy or real business. Due to the growing amount and accessibility to Big Data, the main objective of this paper is an attempt to define the term Big Data and categorize its main sources. Additionally, this paper aims to present possibilities for implementing Big Data in economic studies, particularly an assessment of Google trends Data as a useful tool for nowcasting and forecasting economic indicators of the Polish labour market (i.e. unemployment) is presented. The obtained results indicated that the addition of an additional variable determining the number of specific Internet queries related to the job seeking process raised the forecasting properties of the model forecasting the level of unemployment in Poland.

Keywords: Big Data, economic forecasting, unemployment, Google trends.

Wstęp

Istotny wzrost znaczenia analiz empirycznych w ekonomii zdaje się jedną z podstawowych cech charakteryzujących współczesny kierunek rozwoju tej dziedziny. Jednym z dowodów rosnącego znaczenia tego typu analiz jest stale rosnący odsetek publikacji naukowych zawierających elementy analiz empirycznych. D. S. Hamermesh (2013), weryfikując charakter artykułów publikowanych w czołowych czasopismach ekonomicznych z lat 1963–2011, stwierdza, że do połowy lat 80. większość opracowań miała charakter teoretyczny, pozostałe natomiast opierały się w większości na gotowych zestawach danych dostarczanych przez krajowe urzędy statystyczne czy międzynarodowe organizacje zajmujące się zbieraniem i udostępnianiem danych (m.in.: World Bank, IMF, ILO, OECD, Eurostat). W kolejnych latach liczba opracowań o charakterze empirycznym w czołowych czasopismach zaczęła istotnie wzrastać. Współcześnie opracowania te stanowią ponad 70% ogółu. Jednocześnie w znacznej części tych opracowań wykorzystuje się indywidualne dane pozyskiwane przez autorów, czy też dane pochodzące z tzw. kontrolowanych eksperymentów. Zmiana ta powodowana jest m.in. rosnącą dostępnością, a także różnorodnością danych o charakterze ekonomicznym czy quasi-ekonomicznym. Współcześnie ogólna ilość dostępnych danych (w tym także o charakterze ekonomicznym) wzrasta w zawrotnym tempie. Dane te stanowią podstawę prognoz, które uważane są

za jeden z najważniejszych obszarów praktycznej funkcji nauki (Szynkiewicz, 2017, s. 47). Podstawową przyczyną takiego stanu rzeczy jest dynamiczny rozwój Internetu i technologii teleinformacyjnych.

Obok samego wzrostu intensywności oraz liczby osób korzystających z Internetu nie bez znaczenia jest także rosnąca informatyzacja przedsiębiorstw, rozwój telefonii komórkowej i technologii GPS, a także rosnąca liczba urządzeń łączących się z Internetem (tzw. Internet rzeczy). To właśnie te dane charakteryzujące się dużą objętością i złożonością określane są współcześnie jako Big Data.

Z uwagi na rosnące znaczenie, a także coraz łatwiejszy dostęp do danych, przedmiotem niniejszego opracowania jest próba objaśnienia i skategoryzowania pojęcia Big Data, ocena jego aktualności w ramach badań naukowych oraz wskazania na możliwości wykorzystania Big Data w ekonomii. Zasadniczym celem pracy jest natomiast ocena przydatności danych zaliczanych do Big Data (danych dostępnych w ramach analiz Google trends) jako potencjalnie wartościowych w prognozowaniu wielkości bezrobocia w Polsce.

1. Dane typu Big Data

W związku z narastającym zainteresowaniem zagadnieniami związanymi z Big Data wielu naukowców, jak również przedstawicieli branży IT, a także mediów, podejmuje próby rozpoznania i opisania samego pojęcia, jak również podstawowych cech Big Data. Należy jednak podkreślić, że na gruncie teorii ekonomii, jak również w sferze biznesowej nie ukształtowała się jeszcze jedna powszechnie akceptowana definicja tego pojęcia. Jedna z pierwszych definicji Big Data została zaproponowana przez M. Coxa i D. Ellswortha. Autorzy ci traktują Big Data jako duże zestawy danych o potencjale analitycznym, których ilość należy maksymalizować w celu wydobycia wartości informacyjnych (Cox i Ellsworth, 1997, s. 1–2). Kolejną z definicji, która należy współcześnie do powszechnie cytowanych (Ward i Barker, 2013, s. 1), jest ta zawarta w tzw. Raporcie dla Gartner Group (Laney, 2001). Mimo że w raporcie tym nie przywołuje się bezpośrednio terminu Big Data, to wskazuje się w nim na trzy podstawowe cechy „3Vs", a mianowicie: objętość (*volume*), różnorodność (*variety*) oraz szybkość przetwarzania (*velocity*). Z kolei zdaniem Warda i Barkera Big Data jest pojęciem opisującym przechowywanie oraz analizowanie dużych oraz kompleksowych (wzajemnie powiązanych) zestawów danych przy wykorzystaniu szeregu technik, takich jak: NoSQL, MapReduce czy Machine Learning (Ward i Barker, 2013, s. 2).

Wobec złożoności oraz szerokiego zakresu danych zaliczanych do Big Data warto zastanowić się nad systematyzacją źródeł ich pochodzenia. Jednym z bardziej aktualnych podejść w tym zakresie jest systematyzacja zaproponowana przez zespół analityków pracujący w ramach zespołu Task Team on Big Data przy Europejskiej Komisji Gospodarczej (United Nations Economic Commission for Europe) (UNECE, 2015, s. 4–5), który to wyodrębnił trzy główne źródła tych danych, tj.: indywidualne dane digitalizowane, dane pochodzące z tradycyjnych sieci biznesowych oraz dane kreowane w ramach tzw. Internetu rzeczy.

Indywidualne dane digitalizowane obejmują informacje tworzone przez ludzi, stanowiące utrwalenie doświadczeń i wiedzy przez nich kreowanej.

Tabela 1. Zestawienie jednostek pamięci wykorzystywanych w zapisie danych cyfrowych

Processor or virtual storage	Disk storage	Przykład
1 Bit = Binary Digit	1 Bit = Binary Digit	Zapis 0 1
8 Bits = 1 Byte	8 Bits = 1 Byte	Pojedyncza litera
1024 Bytes = 1 Kilobyte	1000 Bytes = 1 Kilobyte	Połowa zapisanej strony
1024 Kilobytes = 1 Megabyte	1000 Kilobytes = 1 Megabyte	Krótka powieść
1024 Megabytes = 1 Gigabyte	1000 Megabytes = 1 Gigabyte	Przeciętny film
1024 Gigabytes = 1 Terabyte	1000 Gigabytes = 1 Terabyte	Zawartość przeciętnej biblioteki akademickiej
1024 Terabytes = 1 Petabyte	1000 Terabytes = 1 Petabyte	Zawartość wszystkich bibliotek naukowych USA
1024 Petabytes = 1 Exabyte	1000 Petabytes = 1 Exabyte	5 exabajtów – wszystkie słowa wypowiedzenie kiedykolwiek przez ludzkość
1024 Exabytes = 1 Zettabyte	1000 Exabytes = 1 Zettabyte	2 zetabajty – odpowiadają sytuacji, w której każdy mieszkaniec ziemi przez rok otrzymywałby 174 codzienne gazety
1024 Zettabytes = 1 Yottabyte	1000 Zettabytes = 1 Yottabyte	Aby pobrać hipotetyczny plik wielkości 1 jotabajta, wykorzystując łącze szerokopasmowe, potrzeba 11 bilionów lat
1024 Yottabytes = 1 Brontobyte	1000 Yottabytes = 1 Brontobyte	Odpowiada wielkości bajtów 10^{27} (27 zer)
1024 Brontobytes = 1 Geopbyte	1000 Brontobytes = 1 Geopbyte	Brak przykładów

Źródło: Opracowanie własne na podstawie: http://highscalability.com/blog/2012/9/11/how-big-is-a-petabyte-exabyte-zettabyte-or-a-yottabyte.html (dostęp: 30.11.2018 r.).

W przeszłości podstawowym nośnikiem tego typu informacji były formy pisane (głównie książki) i dzieła sztuki, następnie fotografie, zapisy dźwiękowe oraz wideo. Współcześnie prawie wszystkie informacje kreowane przez ludzi podlegają digitalizacji i są przechowywane w postaci cyfrowej na komputerach osobistych, a także coraz częściej udostępniane w sieciach społecznościowych. Ten typ danych jest stosunkowo słabo ustrukturyzowany oraz rzadko podlega zewnętrznej kontroli (Tabakow, Korczak i Franczyk, 2014, s. 139–140).

Warto także podkreślić, że wraz z dynamicznym wzrostem ilości dostępnych oraz kreowanych informacji standardowe miary, takie jak megabajt czy gigabajt, wykorzystywane do określania przestrzeni dyskowej (*disk storage*) bądź obliczeniowej lub też wirtualnej (*procesor or virtual storage*) zaczęto zastępować ich wielokrotnościami (terabajt, petabajt, exabajt itp.). Bardziej szczegółowy wykaz tych jednostek, skonstruowany na podstawie *IBM Dictionary of computing method to describe storage*, zaprezentowano w tabeli 1.

2. Wykorzystanie danych typu Big Data w ekonomii – przykłady z literatury

Obserwowany w ostatnim czasie wzrost dostępności danych zaliczanych do kategorii Big Data przekłada się na systematyczny wzrost badań i opracowań (także w ekonomii) w tym zakresie (Einav i Levin, 2014). I tak np. naukowcy z MIT (Massachusetts Institute of Technology) wraz z partnerami reprezentującymi przedsiębiorstwa typu e-commerce opracowują w ramach projektu *The Billion Prices Project* dzienny wskaźnik inflacji (Daily Price Index) (Cavallo, 2012a). Uwzględniając dane dotyczące bieżących cen szerokiego zestawu towarów i usług[2], zespół ten opracowuje aktualny wskaźnik cen towarów konsumpcyjnych odpowiadający oficjalnym danym generowanym przez urzędy statystyczne, z tą jednak różnicą, że dane są podawane z częstotliwością dzienną i bez kilkutygodniowych opóźnień. Co więcej, w ramach projektu opracowywane są również indeksy inflacji dla krajów, w których dostępność oficjalnych statystyk jest ograniczona lub ich wiarygodność może budzić istotne wątpliwości. Przykład mogą stanowić istotne różnice w danych publikowanych dla Argentyny (Cavallo, 2012b).

[2] Należy wskazać że indeks ten charakteryzuje się pewną nadreprezentacją dóbr sprzedawanych zwyczajowo przez Internet, mimo to korelacja między szerokimi indeksami cen (np. CPI) jest bardzo wysoka.

Z kolei S. Baker, N. Bloom oraz S. Davis (2013), wykorzystując informacje publikowane w wiodących dziennikach informacyjnych, szacują dzienny indeks niepewności polityki ekonomicznej (Daily Index of Economic Policy Uncertainty). W odróżnieniu od poprzedniego przykładu szacowany przez powyższych autorów indeks cechuje się nowatorskim charakterem i nie ma swojego oficjalnego odpowiednika. Nieco innaczej prezentują się analizy wykorzystujące dane gromadzone przez przedsiębiorstwa, pozwalające m.in. na ocenę funkcjonowania branż, analizę zachowań konsumentów, ocenę sytuacji na rynku pracy itp. Tego typu analizy wymagają jednak kooperacji między przedsiębiorstwami posiadającymi stosowane dane a ekonomistami wyposażonymi w odpowiednie instrumenty pozwalające na ich analizę i interpretację. Przykładem takiej kooperacji jest opracowanie autorstwa. L Einava, D. Knoepfle'a, J. Levina i N. Sundaresana (2014), w którym przy wykorzystaniu danych udostępnianych przez eBay szacowano wpływ obciążeń podatkowych na poziom zakupów on-line w poszczególnych stanach USA.

Nieco inny charakter mają analizy wykorzystujące dane gromadzone przez sektor publiczny (dane administracyjne) w ramach funkcjonowania systemów podatkowych, socjalnych czy opieki zdrowotnej. Postępująca informatyzacja pozwala na ich zbieranie w coraz szerszym zakresie oraz przechowywanie ich w formie elektronicznej. Niewątpliwą zaletą tego typu danych jest ich wysoka jakość, długoterminowy charakter (często dane te mają charakter panelowy[3]), reprezentowalność (trudna do osiągnięcia w ramach tradycyjnych badań ankietowych) oraz fakt, że nie są one obciążone problemami związanymi z niejednorodnym doborem czy też tendencyjnością próby (*sample selection bias*). Przykładem wykorzystania danych administracyjnych z systemów podatkowych jest opracowanie T. Piketty'ego i E. Saez (2014), w którym autorzy szacują udziały dochodów i majątku poszczególnych grup dochodowych, ze szczególnym uwzględnieniem osób najbogatszych. Tradycyjne podejście do tego typu badań (wykorzystanie danych ankietowych) jest utrudnione ze względu na nierównomierny zakres próby (niewielki odsetek osób najbogatszych), zaniżanie wysokości dochodów czy ograniczony zakres czasowy dostępnych danych. Dane administracyjne mogą być również użyteczne w ocenie poziomu jakości kształcenia (Rivkin, Hanushek i Kain, 2005), regionalnego zróżnicowania płac czy wydajności pracy w teoretycznie podobnych do siebie przedsiębiorstwach (Syverson, 2011; Abowd, Kramarz i Margolis, 1999).

Z uwagi na cel niniejszego opracowania za szczególnie istotne można uznać badania wykorzystujące dane dotyczące aktywności użytkowników Interne-

[3] Dostarczają informacji dotyczących poszczególnych osób przez wiele kolejnych lat.

tu, a odnoszące się m.in. do tzw. zapytań internetowych (*search query data*). Jednym z pierwszych przykładów wykorzystania w analizach ekonomicznych danych dotyczących wyszukiwanych w Internecie treści za pośrednictwem przeglądarki Google jest opracowanie H. Choi i H. Variana (2009)[4]. Autorzy ci wskazują na znaczny potencjał informacji dostarczanych przez Google trends, jednocześnie potwierdzając ich przydatność w szacowaniu bieżącej sytuacji ekonomicznej (*nowcasting*), aktywności konsumpcyjnej (sprzedaż aut, nieruchomości) czy turystycznej. J. Woźniczka (2018, s. 31) wskazał ponadto na istotną rolę Big Data w zarządzaniu i marketingu, stwierdzając wręcz, że stanowią one źródło przełomu w zarządzaniu i marketingu, a szczególnie w badaniach marketingowych.

Bieżąca dostępność tego typu danych wydaje się szczególnie użyteczna w sytuacji znacznego opóźnienia publikacji oficjalnych danych statystycznych. W kolejnych latach coraz częściej zaczęto wykorzystywać dane dotyczące aktywności w Internecie także w celu oceny bieżącej sytuacji na rynku pracy. Badania te koncentrowały się zwłaszcza na sytuacji w poszczególnych krajach. I tak np. N. Askitas i K. Zimmermann (2009) w badaniu dotyczącym Niemiec, korzystając z danych udostępnianych przez Google trends, analizowali wartość informacyjną (dla oceny aktualnego poziomu bezrobocia) zapytań (wyszukań w języku niemieckim) takich jak: *urząd pracy*, *stopa bezrobocia*, *doradca zawodowy* oraz kilku najpopularniejszych serwisów z ofertami pracy. Analizy empiryczne przeprowadzone na podstawie autoregresyjnych modeli korekty błędem (ECM) dowiodły predyktywistycznej użyteczności danych dotyczących zapytań związanych z sytuacją na rynku pracy. Z kolei N. McLaren i R. Shanbhogue (2011) w badaniu dotyczącym Wielkiej Brytanii weryfikowali potencjalne znaczenie danych Google trends dotyczących rynku pracy oraz rynku nieruchomości. W odniesieniu do rynku pracy sprawdzali przydatność zapytań, takich jak: *oferty pracy* (*jobs*), *dodatek dla szukających pracy* (*jobseeker's allowance* lub *JSA*), *zasiłek dla bezrobotnych, bezrobotny, bezrobocie*. Analizę empiryczną oparli na modelach autoregresyjnych, gdzie zmienną objaśnianą była zmiana poziomu bezrobocia, natomiast objaśniającymi zmienne dotyczące zapytań internetowych, a także dane ankietowe dotyczące przewidywanej stopy bezrobocia (*Consumer confidence question on changes in expected unemployment for the next year*) oraz przewidywanego „klimatu" ekonomicznego. Przeprowadzone analizy wykazały,

[4] Warto zauważyć, że nieco wcześniej informacje te były wykorzystywane w innych obszarach badawczych, np. w ocenie stopnia rozprzestrzeniania się wirusa grypy (Ginsberg, Mohebbi, Patel i Brammer, 2009).

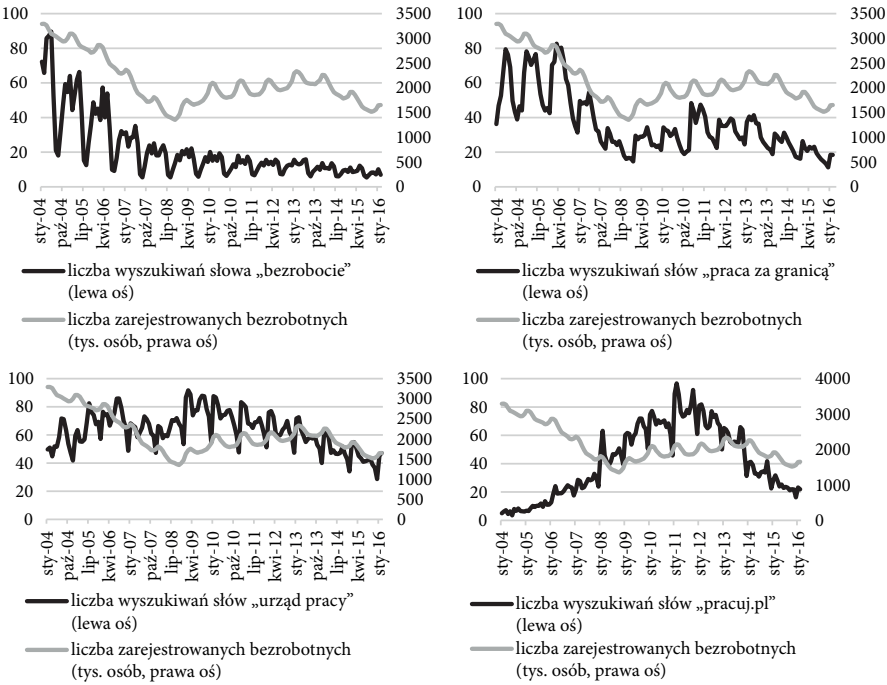
że uwzględnienie dodatkowych informacji na temat zapytań internetowych poprawiło statystyczne dopasowanie szacowanego modelu (wyższe wartości *R*-kwadrat oraz niższe wartości kryterium informacyjnego Akaikego). Podobne badania, w których również dowiedziono wartości informacyjnej danych o aktywności w Internecie w ocenie bieżącej i przyszłej sytuacji na rynku pracy, przeprowadzono także dla Włoch (D'Amuri i Marcucci, 2012), Stanów Zjednoczonych (Choi i Varian, 2012), Finlandii (Tuhkuri, 2014) i Hiszpanii (Vicente, Lopez i Perez, 2015).

3. Prognostyczne właściwości danych Google trends – analiza empiryczna

W celu oceny przydatności danych dotyczących aktywności w Internecie wykorzystano dane udostępniane przez serwis Google trends. Dane te obejmują informacje na temat liczby, pochodzenia, zależności od czasu i głównych regionów zapytań kierowanych do wyszukiwarki Google. Należy jednak pamiętać, że dane te mają charakter względny i wyznaczane są jako skalowany wolumen określonych zapytań, gdzie 100 stanowi najwyższy poziom zainteresowania danym hasłem w całym analizowanym okresie. Dane te dostępne są od 1 stycznia 2004 roku i podlegają codziennej aktualizacji. Na podstawie podejść stosowanych we wcześniejszych badaniach oraz intuicji dotyczącej funkcjonowania rynku pracy w Polsce autor zdecydował się na analizę następujących wyszukiwań: *bezrobocie*, *praca za granicą*, *urząd pracy*, *pracuj.pl*. Informacje o sytuacji na rynku pracy pozyskano natomiast z Głównego Urzędu Statystycznego w Polsce (www.stat.gov.pl). Z uwagi na chęć maksymalizacji liczby dostępnych obserwacji w opracowaniu zdecydowano się wykorzystać miesięczne dane o stopie bezrobocia rejestrowanego oraz liczbie zarejestrowanych osób bezrobotnych. Graficzną prezentację relacji występujących między analizowanymi wyszukiwaniami a liczbą osób bezrobotnych oraz stopą bezrobocia stanowi odpowiednio rysunek 1 i 2. Prezentowane dane swoim zakresem obejmują miesiące od stycznia 2004 do lutego 2016 roku[5].

Analizując kształtowanie się badanych wielkości, można dostrzec pewien stopień zależności między prezentowanymi zmiennymi, widoczny zwłaszcza

[5] Autor zdecydował się na skrócenie okresu badawczego ze względu na istotny spadek bezrobocia w kolejnych miesiącach 2016 roku i w konsekwencji coraz niższą aktywność wymaganą w celu znalezienia pracy.



Rysunek 1. Dane dotyczące wyszukiwań Google trends a liczba osób bezrobotnych w latach 2004–2016

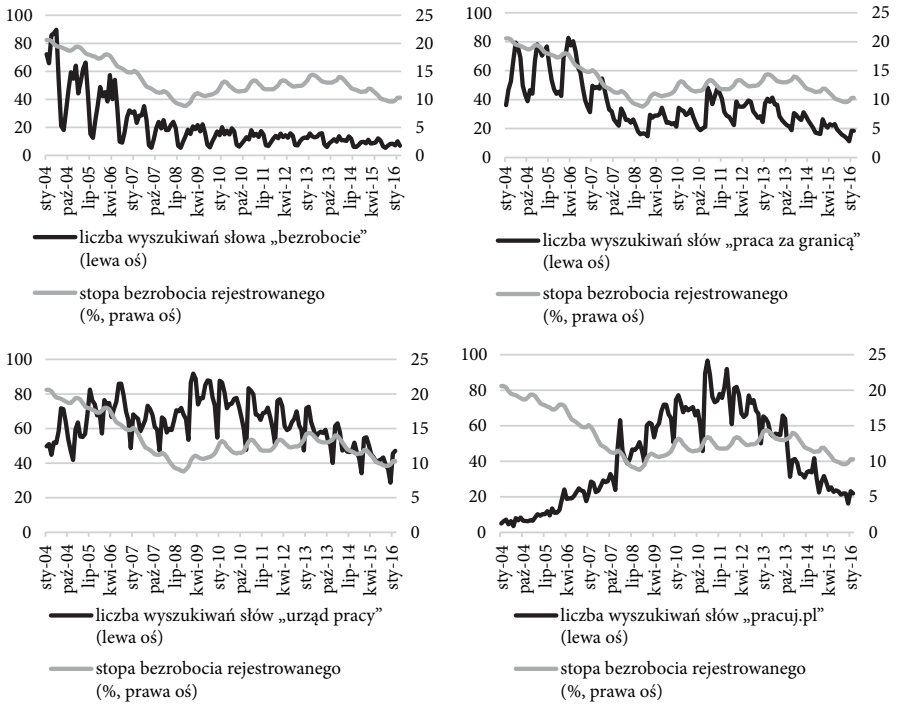
Źródło: Opracowanie własne na podstawie: Google trends oraz GUS.

Uwaga: Ponieważ dane pochodzące z Google trends publikowane są w ujęciu tygodniowym, konieczne było ich uśrednienie, tak by odpowiadały ujęciu miesięcznemu. Wyniki dotyczące wyszukiwań prezentowane są w ujęciu względnym, w zakresie 0–100, gdzie 100 oznacza największą popularność.

w przypadku wyszukań dotyczących *pracy za granicą* czy *urzędu pracy*. W celu oceny przydatności tego typu aktywności internetowej konieczna wydaje się jednak pogłębiona analiza ekonometryczna.

Do pogłębionej oceny potencjalnego znaczenia danych dotyczących liczby wyszukań zdecydowano się wykorzystać standardową procedurę prognostyczną ARIMA, posługując się podstawową oraz rozszerzoną postacią modelu stosowaną w tego typu analizach. Podstawową postać modelu można przedstawić jako:

$$\emptyset(L)y_t = \theta(L)\varepsilon_t, \tag{1}$$



Rysunek 2. Dane dotyczące wyszukiwań Google trends a stopa bezrobocia w latach 2004–2016

Źródło: Opracowanie własne na podstawie: Google trends oraz GUS.

Uwaga: jak dla rysunku 1.

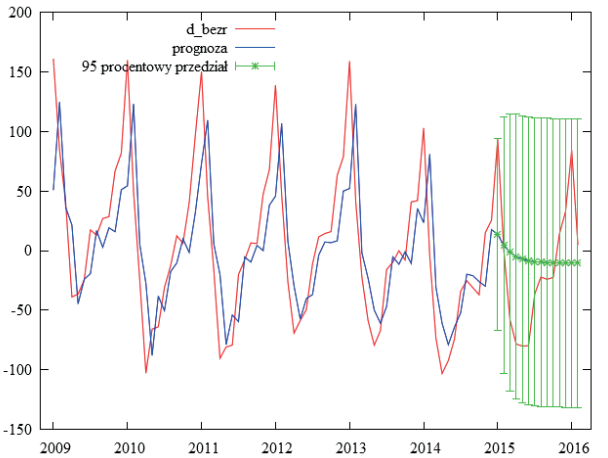
gdzie: $\emptyset(L)y_t$ oraz $\theta(U)$ są wielomianami operatora opóźnień L, który jest zdefiniowany jako $L^n\ y_t = y_{t-n}$, natomiast ε_t stanowi tzw. składnik losowy. Do modelu wprowadzono także operator opóźnień sezonowych $s = 12$ (ze względu na miesięczny charakter danych). Z uwagi na brak stacjonarności analizowanych szeregów oraz właściwości funkcji ACF oraz PACF (według procedury Boxa-Jenkinsa) ustalono odpowiednio wartości d, p i q równe 1. Ostatecznie pierwszy model przyjął postać:

$$\emptyset(L)\Phi(L^s)\ (1-L)^d\ (1\ -L^s)^D\ y_t = \theta(L)\varepsilon_t\ \Theta(L^s)\varepsilon_t, \tag{2}$$

gdzie $(1-L)^d\ y_t = \Delta^d\ y_t$ oraz $(1-L^s)^D\ y_t = \Delta^D\ y_t$ oznaczają przyrosty modelowanej zmiennej, tj. miesięczne poziomy stopy bezrobocia w analizowanym

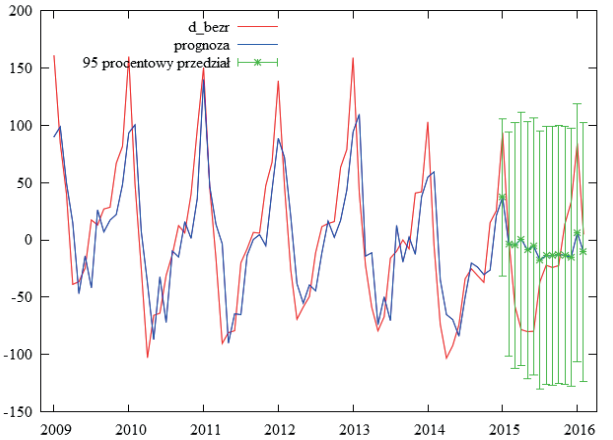
okresie. W celu oceny prognostycznych właściwości informacji dotyczących liczby zapytań internetowych związanych z bezrobociem powyższy model rozszerzono o dodatkową zmienną, dotyczącą liczby zapytań internetowych.

Wyniki prognoz statystycznych dla modelu 1 oraz 2 zaprezentowano na rysunku 3 i 4. Bardziej szczegółowa ocena jakości tych prognoz została przedstawiona w tabeli 2.



Rysunek 3. Prognoza statystyczna dla modelu 1

Źródło: Opracowanie własne: na podstawie danych jak dla rysunku 1 i 2.



Rysunek 4. Prognoza statystyczna dla modelu 2

Źródło: Opracowanie własne: na podstawie danych jak dla rysunku 1 i 2.

Tabela 2. Ocena właściwości prognostycznych modelu 1 oraz 2

Kryterium		Model 1	Model 2
Średni błąd predykcji	ME	–6,23	–6,68
Błąd średniokwadratowy	MSE	2734,80	2347,60
Pierwiastek błędu średniokwadratowy	RMSE	52,29	48,45
Średni błąd absolutny	MAE	42,99	39,59
Średni absolutny błąd procentowy	MAPE	396,86	368,53

Źródło: Opracowanie własne: na podstawie danych jak dla rysunku 1 i 2.

Dane przedstawione w tabeli 2 obrazują poziom dopasowania obu modeli w oparciu o kryteria, takie jak: średni błąd predykcji, błąd średniokwadratowy i jego pierwiastek oraz średni absolutny błąd predykcji w ujęciu absolutnym oraz procentowym. Dane prezentowane w powyższej tabeli wskazują jednoznacznie na lepsze dopasowanie prognozy skonstruowanej dla modelu 2 (niższe wartości wszystkich kryteriów oceny jakości dopasowania), a więc modelu, w którym jako zmienną objaśnianą wykorzystano dodatkowo dane dotyczące liczby wyszukań internetowych. Warto również podkreślić, że wartości prezentowanych parametrów wskazują na znaczną poprawę wartości prognostycznej drugiego modelu.

Zakończenie

Przedmiotem niniejszego opracowania była próba zdefiniowania pojęcia Big Data, a także ocena jego aktualności w kontekście prowadzonych badań naukowych. W publikacji wskazano na możliwości wykorzystania danych typu Big Data w ekonomii, a także zaprezentowano przykład wykorzystania danych dostępnych w ramach analiz Google trends jako potencjalnie wartościowych w prognozowaniu wielkości bezrobocia w Polsce.

Prowadzone rozważania wykazały, że wraz z dynamicznym wzrostem ilości samych danych typu Big Data oraz coraz łatwiejszym do nich dostępem rośnie również liczba analiz – w tym ekonomicznych – je wykorzystujących. Analizy te opierają się zwłaszcza na panelowych danych publicznych, informacjach gromadzonych w przedsiębiorstwach, a także danych dotyczących aktywności ludzi w Internecie. Te ostatnie wydają się szczególnie istotne w kontekście oceny bieżącej sytuacji gospodarczej (*nowcasting*), jak również prognozowania różnego rodzaju wielkości makroekonomicznych.

Dane te okazały się również przydatne w ocenie sytuacji na polskim rynku pracy. Prowadzone analizy empiryczne wykazały bowiem, że dodanie kolejnej zmiennej określającej poziom aktywności w Internecie, wyrażający się liczbą zapytań kojarzonych z poszukiwaniem pracy, poprawia właściwości prognostyczne modelu prognozującego poziom bezrobocia w Polsce.

Powyższe opracowanie stanowić może zatem jeden z pierwszych kroków na drodze do bardziej aktywnego wykorzystywania w Polsce danych typu Big Data dotyczących aktywności osób w Internecie w analizach *stricte* ekonomicznych, w tym także w analizach dotyczących rynku pracy.

Bibliografia

Abowd, J. M., Kramarz, F. i Margolis, D. N. (1999). High wage workers and high wage firms*, Econometrica 67*, 2, 51–333. doi:10.1111/1468–0262.00020

Askitas, N. i Zimmermann, K. F. (2009). Google Econometrics and Unemployment Forecasting*, Applied Economics Quarterly 55*(2) (2009), 107–120.

Baker, S., Bloom, N. i Davis, S. (2013). Measuring economic policy uncertainty. *Chicago Booth research paper no. 13–02*. Pobrane 12 listopada 2018 z http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2198490

Cavallo, A. (2012a*).* Scraped Data and sticky prices. *Massachusetts Institute of Technology Sloan Working Paper no. 4976–12.*

Cavallo, A. (2012b). Online and official price indexes: Measuring Argentina's inflation, *J. Monet. Econ. 60*, 152–165. doi: 10.1016/j.jmoneco.2012.10.002

Choi, H. i Varian, H. (2009). *Predicting the present with Google Trends*. Pobrane 13 listopada 2018 z http://google.com/googleblogs/pdfs/google_predicting_the_present.pdf

Choi, H. i Varian, H. R. (2012). Predicting the present with Google Trends*, Economic Record 88*(s1), 2–9.

Cox, M. i Ellsworth, D. (1997). Managing Big Data for scientific visualization. ACM SIGGRAPH '97 Course #4. *Exploring gigabyte datasets in real-time: algorithms, data management, and time-critical design*. Los Angeles.

D'Amuri, F. i Marcucci, J. (2012). *The predictive power of Google searches in forecasting unemployment* (Working Paper no. 891). Bank of Italy.

Einav, L., Knoepfle, D., Levin, J. i Sundaresan, N. (2014). Sales taxes and Internet commerce, *Am. Econ. Rev. 104*, 1–26. doi: 10.1257/aer.104.1.1

Einav, L. i Levin, J. (2014). Economics in the age of big Data*, Science 346*. doi: 10.1126/science.1243089

Hamermesh, D. S. (2013). Six decades of top economics publishing: Who and how? *J. Econ. Lit. 51,* 162–172. doi: 10.1257/jel.51.1.162

Hoff, T. (2012). *How big is a petabyte, exabyte, zettabyte, or* a *yottabyte?* Pobrane 30 listopada 2018 z http://highscalability.com/blog/2012/9/11/how-big-is-a-petabyte--exabyte-zettabyte-or-a-yottabyte.html

Ginsberg, J., Mohebbi, M. H., Patel, R. S. i Brammer L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, *457*(19).

Google Trends. Big Data. (2018). Pobrane 12 listopada 2018 z https://www.google.pl/trends/explore#q=big%20Data

GUS. (2018). Pobrane 12 listopada 2018 z http://stat.gov.pl

Laney, D. (2001). 3d Data management: Controlling datavolume, velocity and variety. *Gartner. Retrieved, 6.*

McLaren, N. i Shanbhogue, R. (2011). Using internet search Data as economic indicators, *Bank of England Quarterly Bulletin 1*, 134–140.

Piketty, T. i Saez, E. (2014). Inequality in the long run. *Science 344*, 838–843. doi: 10.1126/science.125193

Rivkin, S., Hanushek, E. i Kain J. (2005). Teachers, schools and academic achievement, *Econometrica 73*, 417–458. doi: 10.1111/j.1468–0262.2005.00584.x

Syverson, C. (2011). What determines productivity? *J. Econ. Lit. 49*, 326–365. doi: 10.1257/jel.49.2.326

Szynkiewicz, M. (2017). Przewidywanie w nauce i przewidywanie dotyczące nauki. *Studia Oeconomica Posnaniensia*, *5*(11), 35–50. doi: 10.18559/SOEP.2017.11.3

Tabakow, M., Korczak, J. i Franczyk, B. (2014). *Big* Data – definicje, wyzwania i technologie informatyczne. *Informatyka ekonomiczna Business Informatics 1*(31), 2014.

Tuhkuri, J. (2014). *Big Data: Google searches predict unemployment in Finland*. Pobrane 13 listopada 2018 z https://www.etla.fi/wp-content/uploads/tuhkuri_NTTS_2015_abstract.pdf

UEECE. (2015). Classification of Types of Big Data, Pobrane 14 listopada z http://www1.unece.org/stat/platform/display/bigData/Classification+of+Types+of+Big+Data.

Vicente, M., Lopez, A. i Perez, R. (2015). Forecasting unemployment with internet search Data: Does it help to improve predictions when job destruction is skyrocketing? *Technological Forecasting and Social Change.* doi: 10.1016/j.techfore.2014.12.005

Ward, J. S. i Barker A. (2013). Undefined by Data: A survey of Big Data definitions. *arXiv:1309.5821*, September.

Woźniczka J. (2018). Big data w marketingu: szanse i zagrożenia. *Studia Oeconomica Posnaniensia, 6*(6), 35–50. doi: 10.18559/SOEP.2018.6.3