

Wykorzystanie Google Trends do modelowania stopy bezrobocia rejestrowanego w Polsce

Mariusz Malinowski^a

Streszczenie. W artykule podjęto problematykę monitorowania stopy bezrobocia w Polsce. Celem przedstawionego badania jest sprawdzenie, czy dołączenie wybranych indeksów Google Trends do autoregresyjnego modelu stopy bezrobocia rejestrowanego poprawia trafność generowanych przez niego prognoz. Zastosowana metoda badania opiera się na technikach nowcastingu służących do oceny bieżącego stanu gospodarki. Dane za lata 2004–2019 zostały zaczerpnięte z publikacji GUS oraz serwisu Google Trends, który pozwala na śledzenie popularności terminów wyszukiwanych przez internautów. Porównano jakość dopasowania modelu do danych oraz błędy prognoz modelu podstawowego i modeli rozszerzonych o zmienne egzogeniczne. Artykuł przedstawia zarówno potencjał, jak i ograniczenia wykorzystywania nowego źródła danych w analizach makroekonomicznych dotyczących Polski. Na podstawie przeprowadzonej analizy można uznać, że indeksy Google, powszechnie wykorzystywane w literaturze anglojęzycznej, nie poprawiają trafności predykcji modelu autoregresyjnego. Zadawalające rezultaty uzyskiwane są tylko dla indeksów związanych z międzynarodową mobilnością siły roboczej.

Słowa kluczowe: nowcasting, Google Trends, bezrobocie, dane makroekonomiczne

JEL: C22, C55, C82, E27

Use of Google Trends in modelling registered unemployment rate in Poland

Abstract. The paper deals with the problem of monitoring the unemployment rate in Poland. The main aim of the article is to check whether the addition of selected Google Trends indices improves the accuracy of forecasts generated by the autoregressive model of registered unemployment rate. The research method is based on nowcasting techniques which are used to assess the current state of an economy. Data for the years 2004–2019 were retrieved from publication by Statistics Poland (GUS) and Google Trends, the latter of which allows tracking the popularity of terms searched by Internet users. The study compares the goodness of fit and forecast errors of the basic model with these of models extended with exogenous variables. Both the potential and the limitations of the utilisation of a new source of data in macroeconomic analyses concerning Poland are presented in the paper. The analysis yields a conclusion that Google indexes (commonly used in the literature written in English) do not improve the accuracy of predictions of the autoregressive model. Satisfactory results are only obtained for indices related to the international mobility of the workforce.

Keywords: nowcasting, Google Trends, unemployment, macroeconomic data

^a Uniwersytet Warszawski, Wydział Nauk Ekonomicznych / University of Warsaw, Faculty of Economic Sciences. E-mail: mariusz.a.malinowski@uw.edu.pl. ORCID: <https://orcid.org/0000-0003-4635-2960>.

1. Wprowadzenie

Cyfrowe ślady pozostawiane przez każdego użytkownika internetu spowodowały w ostatnich latach wykładniczy wzrost liczby źródeł danych, które mogą być wykorzystywane w analizach ekonomicznych (Blazquez i Domenech, 2018). Badacze życia gospodarczego mogą w swoich pracach wyjść poza konwencjonalne ankiety czy oficjalne statystyki, opracowywane i publikowane głównie przez instytucje państwowe. Aktywność ogromnej liczby indywidualnych użytkowników, firm oraz organizacji publicznych może dostarczać danych pomagających opisać ich zachowania, decyzje i motywacje, a tym samym pomóc monitorować kluczowe ekonomiczne czy społeczne zmiany i trendy.

Pod koniec lat 90. XX w. ogrom danych generowanych przez użytkowników internetu za pośrednictwem ciągle rozwijającej się technologii cyfrowej zaczął być określany terminem *big data* (Cox i Ellsworth, 1997). Na początku XXI w. termin ten zdefiniowano w kategoriach modelu 3V (Laney, 2001), na który składają się: liczba danych (*volume*), prędkość przetwarzania danych (*velocity*) oraz różnorodność danych (*variety*). W następnych latach, wraz z rozwojem analityki dużych zbiorów danych, model rozszerzono do 4V, dodając wymiar wartości (*value*). Współcześnie koncepcja *big data* zaczyna być definiowana w kategoriach modelu 5V (Bello-Orgaz i in., 2016). Kolejną jego składową jest wiarygodność (*veracity*), która odnosi się do odpowiedniego zarządzania zbiorami danych oraz polityki prywatności.

Jedno ze źródeł *big data*, które są często wykorzystywane w kontekście badań ekonomicznych, stanowią dane dotyczące słów kluczowych wyszukiwanych w sieci. Użytkownik internetu wpisuje interesujące go słowo bądź frazę do wyszukiwarki, która następnie dostarcza mu informacji możliwie najlepiej dopasowanych do danych wejściowych. Firma Google w 2000 r. zrewolucjonizowała ten proces oraz ukształtowała sposób, w jaki dzisiaj przeszukujemy sieć w celu uzyskania konkretnych informacji (Buono i in., 2017). Google wprowadziło bowiem algorytm nazywany PageRank, który podpowiada użytkownikom strony internetowe na podstawie liczby odniesień do nich występujących na innych stronach.

Od maja 2006 r. dostępna jest usługa Google Trends, która dostarcza aktualnych raportów o wolumenie zapytań o dane słowa kluczowe i frazy w różnych regionach świata oraz różnych językach, z historią wyszukiwania dostępną od stycznia 2004 r. Serwis pozwala na generowanie szeregów czasowych obrazujących zmiany zainteresowania danym tematem. Użytkownik określa słowo kluczowe bądź frazę, a Google Trends pokazuje wykres liniowy z czasem na osi poziomej oraz częstotliwością wyszukiwania na osi pionowej.

Dane z Google Trends aktualizowane są w czasie rzeczywistym, w związku z czym często wykorzystuje się je w literaturze makroekonomicznej do monitorowania bie-

żącego stanu gospodarki, które w publikacjach anglojęzycznych określane jest terminem *nowcasting* (Giannone i in., 2008). Termin ten stanowi połączenie słów *now* oraz *forecasting* i w ostatnich latach coraz częściej pojawia się w artykułach ekonomicznych, co wynika ze wzrostu zapotrzebowania na dokładne krótkoterminowe analizy i prognozy stanu gospodarki (Kapetanios i Papailias, 2018). Kluczowe miary, takie jak PKB i jego składowe, inflacja czy stopa bezrobocia, publikowane są z dość znacznym opóźnieniem, a często – po jeszcze dłuższym czasie – poddawane są rewizji. Dlatego coraz chętniej sięga się po nietypowe, ale łatwo dostępne, aktualne i wiarygodne źródła przydatne do wstępnych estymacji.

Celem badania omawianego w artykule jest sprawdzenie, czy dołączenie wybranych indeksów Google Trends do autoregresyjnego modelu stopy bezrobocia rejestrowanego poprawia trafność generowanych przez niego prognoz. W badaniu analizowana jest użyteczność indeksu praca oraz indeksów alternatywnych, wcześniej niewykorzystywanych w literaturze naukowej dotyczącej prognozowania stopy bezrobocia. Termin *praca*, z którego korzystają np. Pavlicek i Kristoufek (2015), jest bezpośrednim tłumaczeniem terminu *jobs*, najpowszechniej używanego w anglojęzycznych publikacjach poruszających tę tematykę. Warto jednak zauważyć, że *praca* w języku polskim jest stosunkowo dość często wykorzystywana w kontekście innym niż działalność zarobkowa. W języku angielskim występuje rozróżnienie na *job* i *work*, dzięki któremu wyszukiwane przez uczniów słowo kluczowe *homework* nie ma wpływu na zmienność szeregu używanego w badaniu. Natomiast w języku polskim można mówić nie tylko o pracy domowej, lecz także licencjackiej czy magisterskiej. Jest to też określenie funkcjonowania organizmu czy urządzenia, a nawet wielkość fizyczna określająca ilość energii potrzebną do przemieszczenia się obiektu.

Należy również zwrócić uwagę na charakterystyki polskiej siły roboczej, o których wspominają Pavlicek i Kristoufek (2015). Nie ulega wątpliwości, że po akcesji Polski do Unii Europejskiej w 2004 r. nasiliły się migracje mieszkańców Polski oraz zmieniła się struktura społeczno-demograficzna tych migracji. Jak wynika z *Informacji o rozmiarach i kierunkach czasowej emigracji z Polski w latach 2004–2017* (Główny Urząd Statystyczny [GUS], 2018), liczba osób przebywających czasowo za granicą zwiększyła się z 1 mln w 2004 r. do 2 mln w 2010 r. i rosła aż do 2,54 mln w 2017 r. Bartosik (2012) zwraca uwagę, że w przypadku Polski związek rynku pracy i mobilności międzynarodowej jest dwojakiego rodzaju. Z jednej strony zła sytuacja na rynku pracy (tj. wysokie bezrobocie) może stanowić czynnik wypychający (innymi słowy, sprzyjający wyjazdowi), a z drugiej – międzynarodowa mobilność siły roboczej wpływa na zmniejszanie się podaży pracy, co z kolei sprzyja spadkowi bezrobocia oraz wzrostowi płac.

W związku ze wspomnianymi problemami z terminem *praca* oraz charakterystyczną dla Polski międzynarodową mobilnością siły roboczej w badaniu wykorzy-

stywany jest również szereg czasowy *praca za granicą*. Na potrzeby artykułu przyjęto założenie, że bezrobotni najpierw szukają ogólnych opinii o możliwościach zarobkowania poza granicami kraju i najczęściej wpisują najprostsze określenie: *praca*.

Problematiczność terminu *praca* skłania do wyeliminowania tego słowa kluczowego i zastąpienia go alternatywnym – mniej wieloznacznym – terminem, który jednak dość dobrze wpisuje się w proces poszukiwania zatrudnienia. W związku z tym w niniejszym badaniu wykorzystywano szereg CV + *curriculum vitae*, przy założeniu, że najczęstszym pierwszym krokiem podejmowanym przez bezrobotnych podczas poszukiwania zatrudnienia jest właśnie przygotowanie tego dokumentu.

2. Przegląd literatury

Jeszcze zanim Google Trends stało się publicznie dostępnym serwisem, Ettredge i in. (2005) wykorzystali dane z *WordTracker's Top 500 Keyword Report*, opublikowanego przez Rivergold Associates, do analizy ekonometrycznej. Jednak ze względu na ograniczoną liczbę obserwacji w ich pracy nie można znaleźć analizy szeregów czasowych ukierunkowanej na predykcje. Autorzy badają siłę objaśniającą modeli, a jednocześnie wykazują, że istnieje pozytywna i statystycznie istotna korelacja pomiędzy słowami kluczowymi dotyczącymi poszukiwania pracy, które indywidualni internauci ze Stanów Zjednoczonych wpisują w wyszukiwarki, a oficjalnymi statystykami dotyczącymi bezrobocia.

Korzystanie z Google Trends w analizach ekonometrycznych zapoczątkowali Choi i Varian (2009), którzy badali użyteczność tego nowego źródła danych w prognozowaniu poziomów sprzedaży detalicznej, samochodów i domów w Stanach Zjednoczonych oraz liczby turystów w Hongkongu.

W tym samym roku Askitas i Zimmermann (2009) przebadali korelację pomiędzy różnymi wyszukiwanymi słowami kluczowymi a stopą bezrobocia w Niemczech w okresie od stycznia 2004 r. do kwietnia 2009 r. Wykazują oni w swoim artykule, że indeksy Google Trends z trzeciego i czwartego tygodnia poprzedniego miesiąca mogą być wykorzystywane do prognozowania stopy bezrobocia w bieżącym miesiącu. Co warto podkreślić, ta praca była pierwszą, w której zestawiono ze sobą statystyki dotyczące stopy bezrobocia i dane pochodzące z Google Trends.

D'Amuri i Marcucci (2012) przeprowadzili szczegółowe badanie siły predykcyjnej ponad 500 różnych liniowych i nieliniowych modeli prognozujących stopę bezrobocia w Stanach Zjednoczonych, przy czym w większości z nich do zbioru zmiennych objaśniających włączyli szeregi czasowe Google Trends. Ich praca jest szczególnie ważna w kontekście metody badania przedstawionej w niniejszym artykule, ponieważ wskazuje główne kroki na etapie analizowania użyteczności zmiennych Google. Autorzy przeprowadzili swoje badanie na podstawie następującego schematu: dobór

słów kluczowych (tj. wygenerowanie odpowiednich szeregów czasowych Google Trends), konstrukcja i dopasowanie modeli oraz ostateczne porównanie ich siły predykcyjnej w prognozie *out-of-sample*. Udowodnili, że modele wykorzystujące dane Google charakteryzują się większą dokładnością w przewidywaniu poziomu stopy bezrobocia.

Pavlicek i Kristoufek (2015) wykorzystali wspomniany schemat metodologiczny do analizy zasadności wykorzystywania danych Google w nowcastingu stopy bezrobocia w państwach Grupy Wyszehradzkiej. W swojej pracy dowodzą, że dołączanie zlogarytmowanych opóźnień zmiennych Google Trends poprawia siłę predykcyjną modeli dla Czech i Węgier, podczas gdy rezultaty dla Polski okazują się niekorzystne. Słowo kluczowe, którego używają autorzy w modelu dla Polski, to *praca*. Zwracają jednak uwagę, że wybór tego terminu mógł nie być najlepszy, ponieważ polscy pracownicy charakteryzują się wysokim stopniem mobilności międzynarodowej, co uwidacznia się w szczególnie częstym wyszukiwaniu w Polsce brytyjskich portali z ogłoszeniami o pracę.

Nowcasting stopy bezrobocia w Polsce z wykorzystaniem danych Google został podjęty również przez Anttonena (2018), który w badaniu wykorzystał model BVAR (*Bayesian vector autoregression*) z komponentem sezonowym i przeprowadził analizę dla poszczególnych krajów członkowskich UE. Nie jest jednak jasne, jakich terminów użył do wygenerowania szeregów czasowych wykorzystywanych w badaniu.

3. Metoda badania

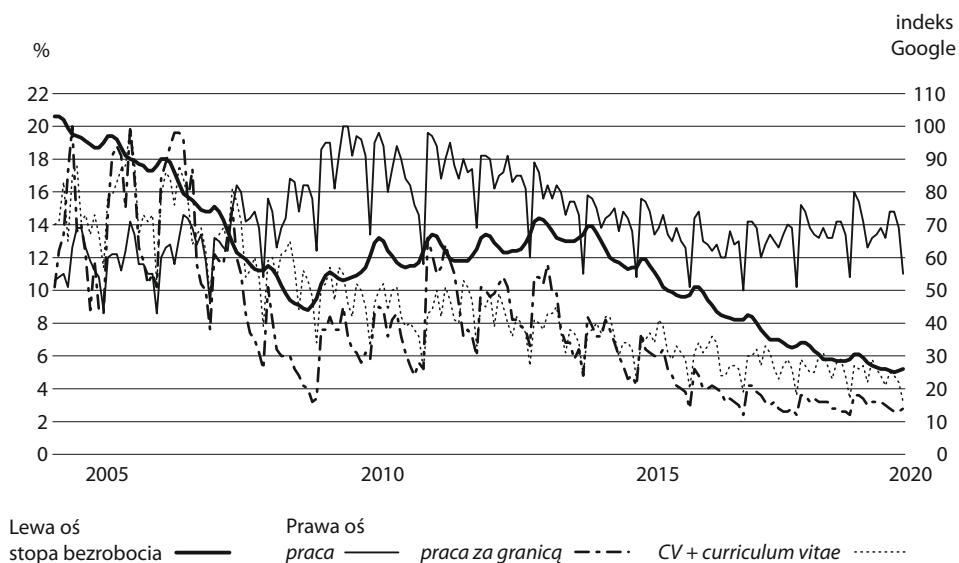
Główny Urząd Statystyczny regularnie publikuje oficjalne statystyki dotyczące stopy bezrobocia (definiowanej jako udział zarejestrowanych bezrobotnych w aktywnej zawodowo ludności cywilnej) według stanu na koniec miesiąca. Jednak statystyki te dostępne są z blisko miesięcznym opóźnieniem (dla przykładu statystyki dotyczące styczniowej stopy bezrobocia dostępne są pod koniec lutego). Innymi słowy, aktualna stopa bezrobocia nie jest znana.

Na wykresie przedstawiono dane dotyczące stopy bezrobocia rejestrowanego (GUS, 2021) w Polsce od stycznia 2004 r. do grudnia 2019 r. oraz dane Google Trends dla tego samego okresu.

Wysokość stopy bezrobocia rejestrowanego w Polsce podlegała w analizowanym okresie silnym fluktuacjom. W pierwszym kwartale 2004 r. wynosiła ponad 20%, ale wejście Polski do UE wiązało się z przyspieszeniem gospodarczym, a tym samym ze wzrostem popytu na pracę, co w konsekwencji prowadziło do szybkiego spadku bezrobocia (Bartosik, 2012). Wartość stopy bezrobocia spadała aż do października 2008 r., gdy osiągnęła poziom 8,8%. Od listopada 2008 r. następował ponowny, stopniowy wzrost bezrobocia, co wiązało się z globalnym kryzysem finansowym.

Stabilizacja sytuacji gospodarczej sprawiła, że od połowy 2014 r. bezrobocie zaczęło ponownie spadać. Rekordowo niski poziom stopy bezrobocia rejestrowanego odnotowano w październiku 2019 r., kiedy wyniosła ona 5%.

Wykres. Stopa bezrobocia rejestrowanego oraz wybrane indeksy Google Trends



Źródło: opracowanie własne z wykorzystaniem pakietu statystycznego R na podstawie: GUS (2021) i Google Trends.

Relatywnie szybki wzrost stopy bezrobocia pod koniec 2008 r. był trudny do przewidzenia. Nadzieję na uniknięcie generowania wyraźnie błędnych prognoz w przyszłości można pokładać w nowych źródłach danych, jak chociażby tych dotyczących wyszukiwania informacji w internecie.

Google Trends mierzy wolumen słów kluczowych bądź fraz, które są wyszukiwane przez użytkowników Google, a następnie porównuje ich popularność z innymi słowami kluczowymi lub frazami. Dokładne dane o samym wolumenie nie są publicznie dostępne. Zamiast tego Google Trends publikuje indeks, który wyliczany jest poprzez podzielenie liczby wyszukiwania danego terminu przez całkowitą liczbę zapytań użytkowników. Wyliczony ułamek skaluje się następnie do przedziału 0–100. Wartość 100 odpowiada maksymalnej liczbie wyszukiwań konkretnego terminu. Pozostałe wartości szeregu skalowane są zgodnie z wartością szczytową. Wynika to z prostego faktu: internet jest wciąż stosunkowo nową technologią i wzrostowy trend liczby użytkowników przekłada się na analogiczny wzrostowy trend w absolutnych wartościach większości zapytań.

Na potrzeby niniejszego badania wygenerowane zostały trzy szeregi czasowe Google Trends zobrazowane na wykresie, tj.: *praca*, *praca za granicą* oraz *CV + curriculum vitae* dla okresu od stycznia 2004 r. do grudnia 2019 r. Żadna z obserwacji dla wszystkich trzech indeksów nie wynosi 0, co oznacza, że liczba zapytań była zawsze większa od wartości granicznej. Wygenerowane indeksy Google, analogicznie do stopy bezrobocia rejestrowanego, są podane w częstotliwości miesięcznej oraz nie zostały uprzednio wyrównane sezonowo.

W celu określenia siły predykcyjnej, jaką niesie ze sobą włączenie do modelu zmiennych Google Trends, niezbędne jest w pierwszej kolejności skonstruowanie modelu, który może służyć jako benchmark. Ze względu na cel badania oraz sposób podejścia Pavlicka i Kristoufka (2015) w omawianym badaniu zbiór potencjalnych modeli, które mogłyby posłużyć jako punkt odniesienia, został ograniczony do klasy modeli autoregresyjnych (AR). Proste, jednowymiarowe modele autoregresyjne dość często wykorzystywane są jako benchmarki w literaturze z zakresu prognozowania szeregów czasowych.

W modelu autoregresyjnym (AR) zmienna zależna y_t zależy liniowo od swoich przeszłych wartości. Model autoregresyjny rzędu p można zapisać jako:

$$y_t = \mu + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \dots + \phi_p y_{t-p} + \varepsilon_t, \quad (1)$$

gdzie:

y_t – wartość zmiennej prognozowanej w momencie t ,

y_{t-i} – opóźnione w czasie wartości zmiennej prognozowanej,

μ – stała modelu,

ϕ_j – parametry modelu,

$\varepsilon_t \sim N(0, \sigma)$.

Przed przystąpieniem do określenia odpowiedniego rzędu modelu AR niezbędne jest sprawdzenie założenia o stacjonarności zmiennych. W celu zbadania stacjonarności przeprowadzone zostały: rozszerzony test Dickeya-Fullera (ADF) oraz test Kwiatkowskiego-Phillipsa-Schmidta-Shina (KPSS). Testy te mają odwrotne hipotezy zerowe, w związku z czym tworzą komplementarną parę, powszechnie wykorzystywaną do testowania stacjonarności.

Statystyki testowe obu testów wyliczone zostały dla sześciu opóźnień (analogiczne rezultaty otrzymano dla mniejszych oraz większych rzędów – maksymalny rozważany rząd wynosił 13). Wyniki rozszerzonego testu ADF wskazywały na konieczność wykorzystania pierwszych różnic dla szeregu czasowego stopy bezrobocia rejestrowanego oraz szeregu *praca*. Wyniki testu KPSS sugerowały z kolei potrzebę zastosowania różnicowania wobec wszystkich zmiennych wykorzystywanych w badaniu.

Ze względu na wyniki obu testów do procesu modelowania wykorzystano pierwsze różnice wszystkich zmiennych. Różnicowanie stopy bezrobocia wyrażonej w procentach oraz indeksu Google Trends nie pozwala na prostą interpretację parametrów estymowanych modeli, jednak – jak wspomniano we wstępie – nie jest to celem tej pracy. Warto nadmienić, że interpretacja indeksu Google Trends jest trudna, nawet jeśli nie zostaje on poddany jakiegokolwiek transformacji, co wynika ze złożoności oraz niejasności metodyki jego wyliczania.

Do określenia rzędu opóźnień w modelu $AR(p)$ wykorzystuje się m.in. funkcję autokorelacji cząstkowej (*partial autocorrelation function* – PACF). Z wykresu wygenerowanego na potrzeby badania wynika, że pierwsze opóźnienia zróżnicowanej stopy bezrobocia miały relatywnie najwyższą wartość autokorelacji cząstkowej w porównaniu z pozostałymi opóźnieniami. Autokorelacja cząstkowa stawała się statystycznie nieistotna na poziomie 5% dopiero dla opóźnień dalszych niż 13. Wykorzystanie modelu $AR(13)$ jako benchmarku nie wydaje się jednak zasadne, jeśli chodzi o zastosowanie go w nowcastingu. Włączenie parametru dla stałej wymagałoby oszacowania łącznie aż 14 parametrów, podczas gdy szereg stopy bezrobocia rejestrowanego liczy jedynie 192 obserwacje, co mogłoby prowadzić do nadmiernego dopasowania modelu do danych.

Co więcej, zbyt duża liczba parametrów byłaby konieczna do oszacowania również wówczas, gdy podstawy wyboru rzędu modelu $AR(p)$ nie stanowiłaby wizualna analiza wykresu funkcji autokorelacji cząstkowej, lecz byłoby nią kryterium informacyjne Akaike (*Akaike information criterion* – AIC), które jest jednym z formalnych sposobów wyboru pomiędzy modelami ekonometrycznymi o różnej liczbie predyktorów. Próba automatycznego dopasowania modelu autoregresyjnego do zróżnicowanego szeregu stopy bezrobocia rejestrowanego skutkowała bowiem wskazaniem modelu $AR(17)$ jako optymalnego. W mocy pozostaje zatem argument dotyczący zbyt małej liczby obserwacji w próbie.

W związku z powyższymi problemami z wyborem rzędu modelu $AR(p)$ na podstawie wykresu funkcji autokorelacji cząstkowej oraz AIC, wybór liczby opóźnień podstawowego modelu miał charakter arbitralny. Jako benchmark wykorzystano model $AR(1)$, którego użyli np. Choi i Varian (2009) w pionierskim artykule dotyczącym nowcastingu z wykorzystaniem Google Trends. Warto podkreślić, że autorzy korzystali w modelu ze zlogarytmowanego komponentu autoregresyjnego.

Montgomery i in. (1998) sugerują wybór modelu autoregresyjnego z wyłącznie jednym opóźnieniem do generowania krótkookresowych prognoz stopy bezrobocia. Prosty, tj. zawierający małą liczbę zmiennych objaśniających, model podstawowy wydaje się wystarczający do sprawdzenia, czy dane dotyczące wyszukiwanych fraz są przydatne w procesie prognozowania. Jeśli szeregi Google Trends nie okazałyby się przydatnym rozszerzeniem modelu $AR(1)$ w predykcji stopy bezrobocia, wówczas

mało prawdopodobne byłoby to, że stałyby się one cennym dodatkiem w bardziej skomplikowanych modelach ekonometrycznych. Co więcej, dane dotyczące wielkości bezrobocia w próbie z lat 2004–2019 charakteryzują się – jak wspomniano – dość dużą zmiennością. Nie jest do końca jasne, jak w świetle tej zmienności powinna być modelowana dynamika tego szeregu czasowego, ponieważ historyczne epizody mogłyby zbyt silnie oddziaływać na wartość zmiennej zależnej i tym samym zaburzać cały proces prognozowania.

Model autoregresyjny z opóźnieniem pierwszego rzędu został następnie rozszerzony o komponent sezonowy y_{t-12} . Analogiczne podejście stosują np. Choi i Varian (2009) oraz Tuhkuri (2016). Dodatkowo Montgomery i in. (1998) w badaniu stopy bezrobocia w Stanach Zjednoczonych zauważają, że charakteryzuje się ona sezonowością w długim okresie. Oprócz oparcia w literaturze dołączenie komponentu sezonowego sugerowała również wizualna analiza funkcji autokorelacji (*autocorrelation function* – ACF), która jest pomocna w identyfikacji dwóch składowych szeregów czasowych, tj. trendu oraz sezonowości. Wykres funkcji autokorelacji został wygenerowany dla zróżnicowanego szeregu stopy bezrobocia rejestrowanego, w związku z czym występowanie trendu nie było na nim widoczne, pomimo że szereg ten przed transformacją charakteryzował się trendem malejącym (o czym wspomniano przy opisie danych – bezrobocie w Polsce zaczęło szybko spadać w okresie poakcesyjnym). Wyraźne pozostały jednak wahania periodyczne o częstotliwości rocznej.

W tym miejscu należy zauważyć, że model AR(1) z komponentem sezonowym nie jest identyczny z modelem opisującym zmienność stopy bezrobocia w Polsce. W celu sprawdzenia, czy wybrany benchmark w zadowalającym stopniu opisuje proces generujący dane, przeprowadzono testy diagnostyczne. W pierwszej kolejności zweryfikowano założenie o stałości wariancji reszt, do czego wykorzystano test Breusch-Pagana (B-P). Hipoteza zerowa w tym teście mówi o homoskedastyczności składnika losowego analizowanego modelu, a hipoteza alternatywna – o jego heteroskedastyczności. Dla modelu AR(1) z komponentem sezonowym hipoteza zerowa została odrzucona przy wartości $p < 0,01$. W modelu służącym jako benchmark w tym badaniu mieliśmy do czynienia z tym samym problemem heteroskedastyczności. Kolejnym krokiem statystycznej diagnostyki modelu było przeprowadzenie testu Ljung-Boxa (L-B). Hipoteza zerowa tego testu mówi o braku autorelacji składnika losowego, a hipoteza alternatywna – o występowaniu autokorelacji. Dla podstawowego modelu hipoteza zerowa testu L-B została odrzucona przy wartości $p < 0,01$ dla rzędu opóźnień równego 3, 6, 12 oraz 24. A zatem komponent sezonowy modelu nie był w stanie uwzględnić całej sezonowości występującej w zróżnicowanym szeregu czasowym stopy bezrobocia rejestrowanego. Nie jest to szczególnie zaskakujące, jeśli wziąć po uwagę wykres ACF oraz rząd opóźnień wskazywany przez AIC.

Model AR(1) z komponentem sezonowym estymowany był w niniejszym badaniu przy wykorzystaniu metody najmniejszych kwadratów. Jednak w kontekście wyników testów B-P oraz L-B nieodzowne wydawało się skorzystanie z odpornych błędów standardowych. Niezbędną korektę wyników estymacji przeprowadzono z wykorzystaniem błędów standardowych HAC (*heteroskedasticity and autocorrelation consistent*) zaproponowanych przez Newey i Westa (1987, 1994). Pozwalają one na uniknięcie konieczności doprecyzowania natury zautokorelowanego błędu, co jest wymagane podczas korzystania z alternatywnych estymatorów o mniejszej wariancji. Warto jeszcze raz podkreślić, że restrykcje nałożone na model podstawowy były znaczne. Aby sprawdzić wrażliwość wyników jego estymacji na zmianę liczby opóźnień, dodatkowo oszacowano modele AR(2)–AR(6) z niezmiennym komponentem sezonowym. Oceniono statystyczną istotność poszczególnych zmiennych, a same modele zostały porównane na podstawie AIC oraz Bayesowskiego kryterium informacyjnego Schwarza (*Bayesian information criterion* – BIC).

Przed przystąpieniem do rozszerzenia modelu podstawowego o zmienne Google Trends przetestowano przyczynowość w sensie Grangera (1969). Sprawdzenie, czy indeksy Google mogą być przydatne do monitorowania bieżącego stanu gospodarki, za pomocą tekstu Grangera zastosował np. Suhoj (2009). Sam test polega na weryfikacji, czy opóźnione wartości jednej zmiennej w modelu autoregresji wektorowej (*vector autoregression model* – VAR) są pomocne w prognozowaniu wartości innej zmiennej.

Przed przystąpieniem do rozszerzenia modelu AR(1) z komponentem sezonowym o egzogeniczne zmienne Google Trends warto podkreślić, że są one dostępne w czasie rzeczywistym, tzn. indeksy wyszukiwania, np. ofert pracy w styczniu, są dostępne w tym samym miesiącu. Jak wspomniano, oficjalne statystyki dotyczące stopy bezrobocia rejestrowanego w Polsce są publikowane przez GUS z prawie jednomiesięcznym opóźnieniem, co sprawia, że indeksy Google mają swoistą przewagę, jeśli chodzi o ich bieżącą dostępność. Dostępność Google Trends w momencie t , podczas gdy dane dotyczące bezrobocia dostępne są dopiero w momencie $t + 1$, skłania do użycia ich do monitorowania aktualnego poziomu agregatów. Wykorzystane w badaniu model benchmarkowy oraz modele rozszerzone o poszczególne indeksy Google przedstawiają się następująco:

$$\text{model 0: } \Delta y_t = \beta_{00} + \beta_{10}\Delta y_{t-1} + \beta_{20}\Delta y_{t-12} + e_t, \quad (2)$$

$$\text{model 1: } \Delta y_t = \beta_{01} + \beta_{11}\Delta y_{t-1} + \beta_{21}\Delta y_{t-12} + \beta_{31}\Delta GT_{1,t} + e_t, \quad (3)$$

$$\text{model 2: } \Delta y_t = \beta_{02} + \beta_{12}\Delta y_{t-1} + \beta_{22}\Delta y_{t-12} + \beta_{32}\Delta GT_{2,t} + e_t, \quad (4)$$

$$\text{model 3: } \Delta y_t = \beta_{03} + \beta_{13}\Delta y_{t-1} + \beta_{23}\Delta y_{t-12} + \beta_{33}\Delta GT_{3,t} + e_t, \quad (5)$$

gdzie:

Δy_t – pierwsze różnice wartości zmiennej prognozowanej w momencie t ,

Δy_{t-i} – opóźnione w czasie wartości zmiennej prognozowanej,

$\Delta GT_{j,t}$ – pierwsze różnice indeksów Google Trends w momencie t ,

$\beta_{k,m}$ – parametry modelu,

ΔGT – zróżnicowany indeks Google Trends,

e_t – składnik losowy modelu.

Modele rozszerzone o zmienną egzogeniczną GT były estymowane analogicznie do modelu podstawowego. W celu oszacowania parametrów zastosowano metodę najmniejszych kwadratów, a niezbędną korektę wynikającą z występowania problemów związanych z heteroskedastycznością oraz autokorelacją wykonywano przy użyciu błędów standardowych HAC, zaproponowanych przez Neweya i Westa (1987, 1994).

Wszystkie cztery modele w pierwszej kolejności były estymowane na całej dostępnej do badania próbie, tj. 192 obserwacjach. Na podstawie oszacowanych w ten sposób parametrów sprawdzono jakość dopasowania modeli do danych.

Na kolejnym etapie badania zbiór danych został podzielony na część treningową oraz testową. Podobne podejście zastosowali Pavlicek i Kristoufek (2015), którzy dane z okresu od stycznia 2004 r. do grudnia 2013 r. podzielili na zbiór treningowy, liczący 96 obserwacji (tj. od stycznia 2004 r. do grudnia 2011 r.), oraz zbiór testowy, liczący 24 obserwacje (tj. od stycznia 2012 r. do grudnia 2013 r.). Ze względu na to, że długość szeregów czasowych w badaniu przedstawianym w niniejszym artykule jest większa, zdecydowano się na wykorzystanie do nowcastingu nie dwóch, lecz trzech ostatnich lat. Tym samym w celu dopasowania modelu do danych sięgnięto po 156 obserwacji (tj. od stycznia 2004 r. do grudnia 2016 r.), a predykcje wygenerowano dla 36 obserwacji (tj. od stycznia 2017 r. do grudnia 2019 r.). Zastosowanie tego podziału pozwoliło pośrednio odpowiedzieć na pytanie, które z analizowanych szeregów Google Trends mogą być przydatne w monitorowaniu bieżącego poziomu stopy bezrobocia rejestrowanego, a także ocenić wielkość oraz statystyczną istotność tej poprawy. W związku z tym nie przeprowadzono dodatkowej oceny jakości modeli oszacowanych na danych ze zbioru treningowego. Nacisk na tym etapie został położony na samo prognozowanie z wykorzystaniem wciąż relatywnie nowego źródła danych.

Trafność prognoz wygenerowanych na zbiorze testowym sprawdzono za pomocą dwóch miar błędów *ex post*, tj. średniej wartości bezwzględnej błędu (*mean absolute error* – MAE) oraz pierwiastka średniego kwadratu błędu (*root mean square error* – RMSE). Z tych samych miar odchylenia rzeczywistych wartości realizacji zmiennej

zależnej od obliczonych prognoz korzysta, w kontekście prognozowania z wykorzystaniem Google Trends, np. Önder (2017).

Porównanie dokładności, przy użyciu miar błędów *ex post*, prognoz *out-of-sample* modelu podstawowego oraz modeli rozszerzonych o zmienne Google Trends pozwala na ocenę, które z szeregów Google są przydatne w procesie monitorowania bieżącego poziomu stopy bezrobocia. Mówiąc precyzyjniej, jeśli miary wyliczone dla modeli rozszerzonych przyjmują wartości mniejsze niż miary wyliczone dla modelu AR(1) z komponentem sezonowym, to można stwierdzić, czy indeksy Google są przydatne w procesie nowcastingu oraz jak znacząca jest poprawa predykcji związana z włączeniem ich do modelu. Niemniej konieczne wydaje się sprawdzenie, czy różnica w dokładności wygenerowanych prognoz jest istotna statystycznie. W związku z tym w badaniu wykorzystano test Diebolda-Mariano (D-M) zaproponowany przez Diebolda i Mariano (2002) oraz Westa (1996). Test D-M pozwala na porównanie dwóch lub więcej alternatywnych prognoz. Hipoteza zerowa testu mówi, że różnica w dokładności predykcji jest nieistotna (Diebold i Mariano, 2002).

4. Wyniki badania

Jak stwierdzono w części dotyczącej metody badania, model podstawowy AR(1) z komponentem sezonowym jest dość restrykcyjny. Niemniej w porównaniu z modelami AR(2)–AR(6) z analogicznymi komponentami sezonowymi Δy_{t-12} jego własności statystyczne wydają się korzystne. Obliczone kryteria informacyjne AIC i BIC sugerowały wybór modelu autoregresyjnego z sześcioma opóźnieniami, chociaż dla AR(6) istotne statystycznie na poziomie 5% okazały się jedynie zmienne Δy_{t-1} oraz Δy_{t-12} , czyli składowe wykorzystanego modelu podstawowego. Testy L-B oraz B-P przeprowadzone dla AR(6) z komponentem sezonowym również nie wykazały, że mógłby on zaradzić problemom związanym z autokorelacją oraz heteroskedastycznością składnika losowego. Modele AR(2)–AR(5) cechowały się wyższymi wartościami AIC niż model podstawowy. W związku z tym – oraz z uwagi na zbyt dużą liczbę zmiennych niezbędnych do oszacowania w modelach wskazywanych przez wykres PACF oraz na automatyczny dobór modelu na podstawie AIC – za punkt odniesienia posłużył model AR(1) z komponentem sezonowym. Warto w tym miejscu nadmienić, że ten model charakteryzował się wysoką wartością współczynnika determinacji $R^2 = 0,78$. Goel i in. (2010) wskazują, że w wielu przypadkach prosty model autoregresyjny z jednym opóźnieniem oraz z komponentem sezonowym może wyjaśniać ponad 90% zmienności zmiennej zależnej. Dlatego z dużą dozą ostrożności należy podchodzić do analizy nowych źródeł danych opierającej się jedynie na jakości dopasowania modeli do danych. Większe znaczenie ma trafność predykcji generowanych przez alternatywne modele.

Hipoteza mówiąca, że szeregi czasowe Google Trends mogą zostać użyte jako rozszerzenie modelu podstawowego do prognozowania stopy bezrobocia rejestrowanego w Polsce, została zweryfikowana pozytywnie. W teście Grangera bowiem hipoteza zerowa mówiąca, że zróżnicowany indeks Google nie jest przyczyną w sensie Grangera zróżnicowanej stopy bezrobocia, została odrzucona dla wszystkich trzech szeregów (*praca*, *praca za granicą* oraz *CV + curriculum vitae*) na poziomie istotności wynoszącym 1%. Niemniej relacja kauzalna (w rozumieniu probabilistycznej koncepcji przyczynowości) nie jest w tym przypadku jednoznaczna. Hipoteza zerowa mówiąca, że zróżnicowana stopa bezrobocia nie jest przyczyną w sensie Grangera zróżnicowanego indeksu Google, również została odrzucona dla wszystkich trzech szeregów Google Trends na poziomie istotności wynoszącym 5%. Tym samym do wyników testu należy podchodzić ostrożnie.

Następnie model podstawowy AR(1) z komponentem sezonowym został rozszerzony o zmienne egzogeniczne Google. W tabelicy przedstawiono oszacowania parametrów poszczególnych modeli oraz podstawowe miary ich dopasowania do danych. Jak wspomniano, wstępne dopasowanie oraz ocena statystycznych charakterystyk wszystkich czterech modeli zostały przeprowadzone dla wartości wyliczonych na podstawie wszystkich 192 obserwacji z próby (tj. dla danych od stycznia 2004 r. do grudnia 2019 r.), a samą estymację przeprowadzono z wykorzystaniem metody najmniejszych kwadratów oraz skorygowano przy użyciu błędów standardowych HAC.

Tablica. Oszacowania parametrów modeli oraz miary ich dopasowania do danych

Zmienne objaśniające	Modele			
	0	1	2	3
<i>const.</i>	-0,0018 (0,0244)	-0,0026 (0,0254)	-0,0034 (0,0237)	-0,0018 (0,0262)
Δy_{t-1}	0,2899** (0,1070)	0,2955** (0,1063)	0,2831** (0,1049)	0,2896** (0,1044)
Δy_{t-12}	0,6653*** (0,0800)	0,6491*** (0,0862)	0,6385*** (0,0868)	0,6650*** (0,0808)
ΔGT_t	0,0019 (0,0012)	0,0034* (0,0016)	0,0001 (0,0006)
Skorygowany R^2	0,7783	0,7806	0,7844	0,7770
<i>AIC</i>	-174,0290	-174,8926	-178,0551	-172,0383
<i>BIC</i>	-161,2795	-158,9556	-162,1181	-156,1013

Uwaga. W nawiasach podano błędy standardowe. ***, **, * – zmienne istotne odpowiednio na poziomie: 1%, 5% i 10%. R^2 – współczynnik determinacji. Modele: 0 – podstawowy, 1–3 – rozszerzone o zmienne egzogeniczne odpowiednio: *praca*, *praca za granicą*, *CV + curriculum vitae*.

Źródło: opracowanie własne z wykorzystaniem pakietu statystycznego R na podstawie danych GUS (2021).

Jak można zauważyć na podstawie danych zamieszczonych w tabelicy, oszacowania parametrów przy zmiennych Google Trends okazują się większe od 0, co oznacza, że

wyszukiwane słowa kluczowe lub frazy są pozytywnie powiązane ze stopą bezrobocia rejestrowanego. Można więc stwierdzić, że poczynione we wstępie założenie, że pogarszająca się bądź zła sytuacja na rynku pracy jest główną przyczyną zwiększonej aktywności osób poszukujących zatrudnienia w internecie, okazało się trafne. Niemniej do tego stwierdzenia należy podchodzić z dużą rezerwą ze względu na to, że jedynie zmienna *praca za granicą* okazała się istotna statystycznie. Dla zmiennych *praca* oraz *CV + curriculum vitae* nie została odrzucona hipoteza zerowa testu *t*-Studenta mówiąca, że oszacowany parametr wynosi 0. Dodatkowo w modelach pozostawiono nieistotną statystycznie stałą, ponieważ jej usunięcie zasadniczo nie powodowało istotnych zmian w oszacowaniach parametrów oraz stosunków pomiędzy wyliczonymi miarami dopasowania modeli do danych. Należy jednak zauważyć, że w każdym z modeli silnie istotne statystycznie są składowe modelu AR(1) z komponentem sezonowym. Zmienna Δy_{t-12} okazała się dla wszystkich estymowanych modeli istotna statystycznie na poziomie istotności 1%, co potwierdza zasadność jej użycia. Warto również podkreślić, że wszystkie modele cechowały się wysoką wartością skorygowanego współczynnika determinacji R^2 , jednak dołączenie do modelu podstawowego szeregu *praca*, *praca za granicą* czy *CV + curriculum vitae* nieznacznie wpłynęło na jego wartość (przy czym dołączenie szeregu *CV + curriculum vitae* – negatywnie). Fakt ten można uznać za potwierdzenie tezy Goela i in. (2010), zgodnie z którą prosty model autoregresyjny z komponentem sezonowym wyjaśniał znaczną część zmienności zróżnicowanej stopy bezrobocia rejestrowanego. Warto też dodać, że wskazywały na to wyliczone wartości AIC oraz BIC, które jedynie nieznacznie różniły się pomiędzy poszczególnymi modelami 0–3, przy czym najmniejszymi wartościami obu kryteriów cechował się model ze zmienną egzogeniczną *praca za granicą*. Model ten okazał się tym samym najlepiej dopasowany do danych, a szereg Google Trends, który został do niego dołączony – jedynym istotnym statystycznie spośród trzech wykorzystanych w badaniu.

Na kolejnym etapie sprawdzono, które z modeli (dopasowanych tym razem na mniejszej liczbie obserwacji – zbiorze treningowym) pozwoliły na wygenerowanie najdokładniejszych prognoz zróżnicowanej stopy bezrobocia rejestrowanego w Polsce dla danych z okresu od stycznia 2017 r. do grudnia 2019 r. (tj. danych ze zbioru testowego). Nieznacznie mniejszymi, w porównaniu do modelu podstawowego, błędami prognoz *ex post* charakteryzował się jedynie model rozszerzony o wartości indeksu Google *praca za granicą*, zarówno jeśli chodzi o miarę MAE, jak i RMSE. Wartość MAE dla modelu AR(1) z komponentem sezonowym wynosiła 0,08, a wartość RMSE – 0,09. Z kolei dla modelu 2 – odpowiednio 0,07 oraz 0,09. Dołączenie szeregów *praca* oraz *CV + curriculum vitae* nie tylko nie poprawiło zatem miar dopasowania modelu do danych, lecz także nie przyczyniło się do zwiększenia trafności prognoz generowanych przy jego użyciu.

W teście D-M różnice w dokładności prognoz generowanych przez benchmark oraz model rozszerzony o wartości szeregu *praca za granicą* okazały się istotne statystycznie. Hipoteza zerowa mówiąca o braku tych różnic została odrzucona na 10-procentowym poziomie istotności. Z kolei dołączenie do modelu podstawowego indeksów Google Trends *praca* oraz *CV + curriculum vitae* nie wpłynęło na trafność predykcji. Wyniki testu D-M nie wskazywały bowiem na występowanie różnic pomiędzy prognozami wygenerowanymi przez benchmark a prognozami wygenerowanymi przez modele 1 i 3. Innymi słowy, modele ze zmiennymi egzogenicznymi *praca* oraz *CV + curriculum vitae* nie pogorszyły istotnie predykcji generowanych przez model AR(1) z komponentem sezonowym.

5. Podsumowanie

Celem badania przedstawionego w artykule było sprawdzenie, czy indeksy Google Trends poprawiają trafność predykcji autoregresyjnego modelu stopy bezrobocia rejestrowanego w Polsce. Zweryfikowano, czy szereg czasowy dotyczący wyszukiwania hasła *praca*, a także szeregi dotyczące wyszukiwania haseł *praca za granicą* oraz *CV + curriculum vitae* (niewykorzystywane wcześniej w literaturze) są użyteczne w procesie monitorowania poziomu tejże zmiennej makroekonomicznej.

Wyniki empiryczne pokazują, że słowo kluczowe *praca* jest mało przydatnym rozszerzeniem modeli wykorzystywanych do prognozowania stopy bezrobocia rejestrowanego w Polsce. Podobnie szereg Google Trends *CV + curriculum vitae* nie poprawia trafności predykcji generowanych przez model autoregresyjny.

Przydatny do monitorowania poziomu stopy bezrobocia jest natomiast indeks *praca za granicą*, co może wynikać z charakterystyk związanych z międzynarodową mobilnością siły roboczej w Polsce. Warto jednak zaznaczyć, że poprawa trafności okazała się niewielka. Niezbędne wydają się dalsze analizy doboru zmiennych egzogenicznych dotyczących aktywności użytkowników internetu odnoszącej się do poszukiwania zatrudnienia.

Niniejszy artykuł przedstawia zarówno ograniczenia, jak i potencjał związany z wykorzystywaniem wciąż relatywnie nowego źródła danych, jakim jest serwis Google Trends. Replikacja sposobów doboru indeksów Google, które są stosowane w literaturze anglojęzycznej z zakresu nowcastingu (a nawet szerzej – prognozowania), nie okazała się poprawnym podejściem do przeprowadzania analiz makroekonomicznych dla Polski. Jednakże uwzględnienie specyfiki języka polskiego oraz istotnych charakterystyk polskiej gospodarki może pomóc o wiele dokładniej przewidywać krajową sytuację ekonomiczną, a dzięki temu narzędzia polityki gospodarczej będą mogły być stosowane w bardziej adekwatny i skuteczny sposób.

Bibliografia

- Anttonen, J. (2018). *Nowcasting the Unemployment Rate in the EU with Seasonal BVAR and Google Search Data* (ETLA Working Papers No. 62). <http://pub.etla.fi/ETLA-Working-Papers-62.pdf>.
- Askatas, N., Zimmermann, K. F. (2009). Google Econometrics and Unemployment Forecasting. *Applied Economics Quarterly*, 55(2), 107–120. <https://doi.org/10.3790/aeq.55.2.107>.
- Bartosik, K. (2012). Popytowe i podażowe uwarunkowania polskiego bezrobocia. *Gospodarka Narodowa*, 260(11–12), 25–57. <https://doi.org/10.33119/GN/101003>.
- Bello-Orgaz, G., Jung, J. J., Camacho, D. (2016). Social big data: Recent achievements and new challenges. *Information Fusion*, 28, 45–59. <https://doi.org/10.1016/j.inffus.2015.08.005>.
- Blazquez, D., Domenech, J. (2018). Big Data sources and methods for social and economic analyses. *Technological Forecasting & Social Change*, 130, 99–113. <https://doi.org/10.1016/j.techfore.2017.07.027>.
- Buono, D., Mazzi, G. L., Kapetanios, G., Marcellino, M., Pappailias, F. (2017). Big data types for macroeconomic nowcasting. *EURONA – Eurostat Review on National Accounts and Macroeconomic Indicators*, (1), 93–145. <https://ec.europa.eu/eurostat/cros/system/files/euronaissuel-2017-art4.pdf>.
- Choi, H., Varian, H. R. (2009). *Predicting the Present with Google Trends*. https://www.google.com/googlegloss/pdfs/google_predicting_the_present.pdf.
- Cox, M., Ellsworth, D. (1997). Managing Big Data for Scientific Visualization. *ACM Siggraph*, 97, 21–38.
- D’Amuri, F., Marcucci, J. (2012). *The predictive power of Google searches in forecasting unemployment* (Bank of Italy Working Papers No. 891). https://www.bancaditalia.it/pubblicazioni/temi-discussione/2012/2012-0891/en_tema_891.pdf?language_id=1.
- Diebold, F. X., Mariano, R. S. (2002). Comparing Predictive Accuracy. *Journal of Business & Economic Statistics*, 20(1), 134–144. <https://doi.org/10.1198/073500102753410444>.
- Ettredge, M., Gerdes, J., Karuga, G. (2005). Using Web-based Search Data to Predict Macroeconomic Statistics. *Communications of the ACM*, 48(11), 87–92. <https://doi.org/10.1145/1096000.1096010>.
- Giannone, D., Reichlin, L., Small, D. (2008). Nowcasting: The real-time informational content of macroeconomic data. *Journal of Monetary Economics*, 55(4), 665–676. <https://doi.org/10.1016/j.jmoneco.2008.05.010>.
- Goel, S., Hofman, J. M., Lahaie, S., Pennock, D. M., Watts, D. J. (2010). Predicting consumer behavior with Web search. *Proceedings of the National Academy of Sciences of the United States of America*, 107(41), 17486–17490. <https://doi.org/10.1073/pnas.1005962107>.
- Główny Urząd Statystyczny. (2018). *Informacja o rozmiarach i kierunkach czasowej emigracji z Polski w latach 2004–2017*. https://stat.gov.pl/download/gfx/portalinformacyjny/pl/defaultaktualnosci/5471/2/11/1/informacja_o_rozmiarach_i_kierunkach_czasowej_emigracji_z_polski_2004-2017.pdf.
- Główny Urząd Statystyczny. (2021). *Stopa bezrobocia rejestrowanego w latach 1990–2021*. <https://stat.gov.pl/obszary-tematyczne/rynek-pracy/bezrobocie-rejestrowane/stopa-bezrobocia-rejestrowanego-w-latach-1990-2021,4,1.html?pdf=1>.
- Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3), 424–438. <https://doi.org/10.2307/1912791>.

- Kapetanios, G., Papailias, F. (2018). *Big Data & Macroeconomic Nowcasting: Methodological Review* (ESCoE Discussion Paper 2018-12). <https://escoe-website.s3.amazonaws.com/wp-content/uploads/2020/07/13161005/ESCoE-DP-2018-12.pdf>.
- Laney, D. (2001). 3D Data Management: Controlling Data Volume, Velocity, and Variety. Application Delivery Strategies. <http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-andVariety.pdf>.
- Montgomery, A. L., Zarnowitz, V., Tsay, R. S., Tiao, G. C. (1998). Forecasting the U.S. Unemployment Rate. *Journal of American Statistical Association*, 93(442), 478–493. <https://doi.org/10.1080/01621459.1998.10473696>.
- Newey, W. K., West, K. D. (1987). A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica*, 55(3), 703–708. <https://doi.org/10.2307/1913610>.
- Newey, W. K., West, K. D. (1994). Automatic Lag Selection in Covariance Matrix Estimation. *Review of Economic Studies*, 61(4), 631–653. <https://doi.org/10.2307/2297912>.
- Önder, I. (2017). Forecasting Tourism Demand with Google Trends: Accuracy Comparison of Countries vs. Cities. *International Journal of Tourism Research*, 19(6), 648–660. <https://doi.org/10.1002/jtr.2137>.
- Pavlicek, J., Kristoufek, L. (2015). Nowcasting Unemployment Rates with Google Searches: Evidence from the Visegrad Group Countries. *PLoS ONE*, 10(5), 1–11. <https://doi.org/10.1371/journal.pone.0127084>.
- Suhoy, T. (2009). *Query Indices and a 2008 Downturn: Israeli Data* (Bank of Israel Discussion Paper No. 2009.06). <https://www.boi.org.il/deptdata/mehkar/papers/dp0906e.pdf>.
- Tuhkuri, J. (2016). *Forecasting Unemployment with Google Searches* (ETLA Working Papers No. 35). <https://www.econstor.eu/bitstream/10419/201250/1/ETLA-Working-Papers-35.pdf>.
- West, K. D. (1996). Asymptotic inference about predictive ability. *Econometrica*, 64(5), 1067–1084. <https://doi.org/10.2307/2171956>.