Wiadomości Statystyczne. The Polish Statistician, 2019, vol. 64, 9, 60–74
Studia interdyscyplinarne. Wyzwania badawcze / Interdisciplinary studies. Research challenges

# Pozyskiwanie i analiza danych na temat ofert pracy z wykorzystaniem big data

**Jacek Maślankowski**[a]

**Streszczenie.** Celem artykułu jest zaprezentowanie korzyści wynikających z wykorzystania na potrzeby statystyki publicznej (rynku pracy) narzędzi do automatycznego pobierania danych na temat ofert pracy zamieszczanych na stronach internetowych zaliczanych do zbiorów big data, a także związanych z tym wyzwań. Przedstawiono wyniki eksperymentalnych badań z wykorzystaniem metod web scrapingu oraz text miningu. Analizie poddano dane z lat 2017 i 2018 pochodzące z najpopularniejszych portali z ofertami pracy. Odwołano się do danych Głównego Urzędu Statystycznego (GUS) zbieranych na podstawie sprawozdania Z-05. Przeprowadzona analiza prowadzi do wniosku, że web scraping może być stosowany w statystyce publicznej do pozyskiwania danych statystycznych z alternatywnych źródeł, uzupełniających istniejące bazy danych statystycznych, pod warunkiem zachowania spójności z istniejącymi badaniami.
**Słowa kluczowe:** big data, text mining, web scraping, rynek pracy

# The collection and analysis of the data on job advertisements with the use of big data

**Summary.** The goal of this paper is to present, on the one hand, the benefits for official statistics (labour market) resulting from the use of web scraping methods to gather data on job advertisements from websites belonging to big data compilations, and on the other, the challenges connected to this process. The paper introduces the results of experimental research where web-scraping and text-mining methods were adopted. The analysis was based on the data from 2017–2018 obtained from the most popular job-searching websites, which was then collated with Statistics Poland's data obtained from Z-05 forms. The above-mentioned analysis demonstrated that web-scraping methods can be adopted by public statistics services to obtain statistical data from alternative sources complementing the already-existing databases, providing the findings of such research remain coherent with the results of the already-existing studies.
**Keywords:** big data, text mining, web scraping, labour market

**JEL:** C18, M15

[a] Uniwersytet Gdański, Wydział Zarządzania.

Wraz z rozwojem pakietów komputerowych i narzędzi statystycznych pojawiły się nowe możliwości w zakresie pozyskiwania danych dla statystyki publicznej. Zalicza się do nich narzędzia big data. Należy zaznaczyć, że termin *big data* nie oznacza technologii. Niekiedy odnosi się do dużego zbioru danych, jednak w tym artykule występuje jako narzędzia pozwalające na przechowywanie i przetwarzanie danych, co nie byłoby możliwe przy wykorzystaniu tradycyjnych metod, np. relacyjnych baz danych (Shahin, 2016). Najbardziej znana definicja obejmuje takie cechy zbioru danych typu big data, jak duża ilość (ang. *volum*e), duża zmienność (ang. *velocity*) oraz duża różnorodność (ang. *variety*). Została ona zaproponowana przez pracowników firmy Gartner w raporcie opublikowanym na blogu firmy META Group (Douglas, 2001). Często podawane są też atrybuty dotyczące wiarygodności (ang. *veracity*) oraz wartości danych (ang. *value*). Oznacza to, że dane ze zbiorów big data powinny charakteryzować się wysoką wiarygodnością, rozumianą również jako wysoka jakość, a także stanowić wartość dla użytkowników.

Jedną z metod wykorzystywanych podczas gromadzenia danych typu big data jest automatyczne pobieranie danych z internetu, znane powszechnie jako web scraping. W tym celu wykorzystuje się programy zwane robotami internetowymi, których zadaniem jest pozyskiwanie odpowiednich informacji, w niniejszym artykule – dotyczących ofert pracy. Dodatkowo, aby uzyskać informacje na temat zawodu, wykształcenia czy kwalifikacji, zastosowano metodę text miningu, która służy do eksploracji i obróbki tekstu[1].

Celem artykułu jest zaprezentowanie korzyści wynikających z wykorzystania na potrzeby statystyki publicznej (rynku pracy) narzędzi do automatycznego pobierania danych na temat ofert pracy zamieszczanych na stronach internetowych, zaliczanych do zbiorów typu big data, a także związanych z tym wyzwań. Przedstawiono wyniki eksperymentalnych badań z wykorzystaniem metod web scrapingu oraz text miningu.

## BIG DATA JAKO ALTERNATYWNE ŹRÓDŁO DANYCH DLA STATYSTYKI PUBLICZNEJ

Zastosowaniu big data w statystyce publicznej poświęcono wiele artykułów naukowych. W literaturze podkreśla się przede wszystkim wykorzystanie tego typu zbiorów danych do obliczania wskaźnika cen towarów i usług konsumpcyjnych (ang. CPI – Consumer Price Index). Według Hackla (2016) w 2016 r. 17 krajów europejskich pracowało nad rozwiązaniami dotyczącymi big data w celu dostarczania danych na potrzeby tego wskaźnika. Listę alternatywnych źródeł danych i możliwości ich zastosowania w statystyce publicznej prezentuje zestawienie 1.

[1] https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/Main_Page (dostęp: 30.06.2018).

**ZESTAWIENIE 1. ALTERNATYWNE ŹRÓDŁA DANYCH I ICH ROLA W STATYSTYCE PUBLICZNEJ**

| Źródła danych | Obszar potencjalnego wykorzystania |
|---|---|
| Skanery kodów kreskowych | statystyka cen, statystyki ekonomiczne |
| Lokalizacja telefonów komórkowych | statystyka turystyki, statystyka ludności i migracji |
| Sensory drogowe | statystyka transportu |
| Mierniki zużycia energii | statystyka ludności, statystyka gospodarstw domowych |
| Zdjęcia satelitarne, zdalne sensory | statystyka rolnictwa, leśnictwa, rybołówstwa, statystyka środowiska naturalnego |
| Serwisy społecznościowe, internet | statystyka rynku pracy, statystyka ludności i migracji, statystyka dochodów i konsumpcji gospodarstw domowych, statystyka cen, statystyka zdrowia, statystyka społeczna |
| Strony WWW z ofertami pracy | statystyka rynku pracy |
| Ruch samolotów | statystyka transportu, statystyka ochrony środowiska |
| Strony WWW: nieruchomości, działalność e-commerce | statystyka cen |

Ź r ó d ł o: opracowanie własne na podstawie: Hackl, 2016; Kitchin, 2015.

Inne źródła danych dostarczają informacji o ruchu pojazdów, jak również dotyczą przetwarzania treści zamieszczanych w serwisach społecznościowych (Daas, Puts, Buelens i Hurk, 2015). Nazywa się je często niestatystycznymi, czyli nieutworzonymi na potrzeby statystyki publicznej (Beręsewicz i Szymkowiak, 2015), lub pozastatystycznymi. Dane z takich źródeł mają szerokie zastosowanie w statystyce publicznej. Przykładowo w ramach grantu Komisji Europejskiej ESSNet Big Data prowadzone są prace eksperymentalne, które bazują na danych pobieranych ze stron internetowych przedsiębiorstw i mają na celu identyfikację obecności przedsiębiorstwa w mediach społecznościowych czy też działalności e-commerce (wsparcie dla europejskiego badania ICT in Enterprises – ICT w przedsiębiorstwach) oraz weryfikację rodzaju działalności pod kątem zgodności z kodem w Polskiej Klasyfikacji Działalności (PKD) lub, w szerszym ujęciu, w europejskiej klasyfikacji NACE[2].

Wykorzystanie alternatywnych źródeł danych typu big data w statystyce publicznej może skutkować zmianą modelu funkcjonowania niektórych obszarów statystyki. Zgodnie z wieloletnią tradycją statystyka publiczna bazuje na danych, które pozyskuje za pomocą kwestionariuszy i sprawozdań (Braaksma i Zeelenberg, 2015). W ciągu ostatnich lat obserwuje się stopniowe odchodzenie od tradycyjnych sprawozdań na rzecz administracyjnych źródeł danych. Ma to miejsce w takich dziedzinach, jak rynek pracy czy edukacja, w której nastąpiło przejście ze sprawozdań dla szkół na System Informacji Oświatowej czy Zintegrowany System Informacji o Nauce i Szkolnictwie Wyższym POL-on[3]. Wykorzystanie

[2] https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/Main_Page (dostęp: 30.06.2018).

[3] Zob. program badań statystycznych statystyki publicznej na rok 2018, http://bip.stat.gov.pl/dzialalnosc-statystyki-publicznej/program-badan-statystycznych/pbssp-2018 (dostęp: 25.02.2019).

big data spowoduje większy nacisk na metodologię oraz sposoby analizy danych zamiast skupiania się na przygotowywaniu kwestionariuszy oraz kontroli procesu sprawozdawczości. Wzrośnie tym samym zapotrzebowanie na inżynierów danych (ang. *data scientists*) mających wiedzę z pogranicza informatyki i statystyki, niezbędną do budowy repozytoriów i przetwarzania dużych zbiorów, a w szczególności integracji różnych zbiorów danych, takich jak bazy relacyjne i typowe dla big data bazy NoSQL (Miller, 2014). Ta zmiana to jeden z elementów modernizacji statystyki publicznej, w ramach której powinny zostać wypracowane wytyczne dotyczące jakości danych, partnerstwa z firmami prywatnymi je dostarczającymi, aspektów prywatności danych, stosowanej metodologii oraz wykorzystywanych technologii (Vale, 2015). Wówczas statystyka publiczna mogłaby przyjąć dla wybranych dziedzin strategię rozwoju sterowaną danymi (ang. *data driven approach*), która może wymagać sformułowania metod gromadzenia, współdzielenia oraz ponownego wykorzystywania dużych zbiorów danych (Rousidis, Garoufallou, Balatsoukas i Sicilia, 2014). Należy jednak mieć na uwadze, że administracyjne źródła danych oraz zbiory typu big data nie zawierają wszystkich niezbędnych informacji, na które jest zapotrzebowanie użytkowników danych statystycznych, w tym uwzględnionych w aktach prawnych obligujących statystykę publiczną do ich prezentowania w ściśle określonych przekrojach i terminach.

Przyjmuje się, że w statystyce publicznej można wykorzystać źródła danych big data na pięć sposobów (Kitchin, 2015):

1. całościowe zastąpienie istniejących źródeł statystycznych (istniejące dane);
2. częściowe zastąpienie istniejących źródeł statystycznych (istniejące dane);
3. dostarczenie komplementarnej informacji statystycznej w tej samej dziedzinie statystyki, jednak z innej perspektywy (dodatkowe dane);
4. skorygowanie szacunków pochodzących z istniejących źródeł statystycznych (poprawione dane);
5. dostarczenie zupełnie nowej informacji statystycznej w danej dziedzinie (nowe alternatywne źródło danych).

Należy zwrócić uwagę, że badanie dotyczące ofert pracy jest tematem opracowań naukowych powstających od wielu lat (Gałecka-Burdziak i Pater, 2015; Kureková, Beblavý i Thum-Thysen, 2015), co pozwala na analizę porównawczą stosowanych podejść i metodologii. Badane są przede wszystkim informacje o wolnych miejscach pracy publikowane w internecie.

## ZASADY POBIERANIA DANYCH ZE STRON INTERNETOWYCH

Dane zamieszczane na stronach internetowych podlegają ochronie, nie można ich kopiować m.in. w celu udostępniania osobom trzecim. Ponadto w przypadku publikowania baz danych na stronie internetowej obowiązuje Ustawa

z dnia 27 lipca 2001 r. o ochronie baz danych, zgodnie z którą „pobieranie danych oznacza stałe lub czasowe przejęcie lub przeniesienie całości lub istotnej, co do jakości lub ilości, części zawartości bazy danych na inny nośnik, bez względu na sposób lub formę tego przejęcia lub przeniesienia". Regulacje te determinują dobór narzędzi oraz możliwości analizy dostępnych baz danych. Jednocześnie „producent bazy danych udostępnionej publicznie w jakikolwiek sposób nie może zabronić użytkownikowi korzystającemu zgodnie z prawem z takiej bazy danych, pobierania lub wtórnego wykorzystania w jakimkolwiek celu nieistotnej, co do jakości lub ilości, części jej zawartości". Należy zauważyć, że bazy danych udostępniane na stronach internetowych mają w większości charakter publiczny.

Poza kwestiami prawnymi dotyczącymi możliwości przetwarzania danych należy również wziąć pod uwagę wykorzystanie pliku *robots.txt*, który powinien znajdować się na każdej stronie internetowej. Określa on, jakiego rodzaju działania robotów są dozwolone w danym serwisie. Przykładowo wpis „User-agent: * Disallow: /" oznacza, że nie jest dozwolone pobieranie danych przez inne roboty niż wskazane w tym pliku[4]. Zasady poruszania się po stronie internetowej oraz dozwolone foldery są określane przez właścicieli serwisu.

Jeżeli działania robotów na stronie internetowej są dozwolone, wówczas możliwa staje się analiza zawartości witryn w czasie rzeczywistym. Dzięki temu nie trzeba pobierać ani przechowywać potrzebnych stron internetowych. Dobrą praktyką jest też pobieranie danych w interwałach czasowych, np. kolejnych części strony co 10 sekund, aby nadmiernie nie przeciążać serwerów właścicieli serwisu. Skutkiem nadmiernego obciążenia serwerów może być zablokowanie robota przez właścicieli serwisu. Dane na stronach internetowych występują w postaci częściowo ustrukturyzowanej, tj. zapisane są za pomocą znaczników HTML, które umożliwiają nawigację na stronie. Przykładowo znacznik  określa sekcję, a  – pojedynczą linię strony. Skojarzenie znaczników z klasami stylów pozwala nawigować na stronie narzędziom do web scrapingu, czyli zautomatyzowanego pobierania wybranych treści ze stron internetowych.

## OFERTY PRACY ONLINE – ANALIZA

Pierwsze eksperymentalne pobieranie danych ze stron internetowych zawierających oferty pracy, z myślą o wykorzystaniu utworzonego zbioru danych w polskiej statystyce publicznej, zostało przeprowadzone przez autora niniejszego artykułu w 2013 r., co pozwoliło na wstępne wskazanie problemów z jakością danych (Maślankowski, 2014). Powtarzanie tego procesu przez kolejne lata umożliwiło sformułowanie wielu wniosków dotyczących zastosowania tego

[4] http://www.robotstxt.org (dostęp: 29.01.2019).

rodzaju źródeł danych w statystyce publicznej. Odnoszą się one przede wszystkim do jakości danych rozumianej jako reprezentatywność oraz stabilność źródeł danych. Obecnie autor pobiera dane codziennie, a wynikowy zbiór danych z ofertami pracy potwierdza wnioski formułowane w latach ubiegłych. W ramach prowadzonych prac jednostką badania jest oferta pracy na terytorium Polski zamieszczona online.

W badaniu jako źródła internetowe zawierające oferty pracy wykorzystano portale: Praca.money.pl, GoWork.pl, Jobs.pl, Pracuj.pl, jak również portale branżowe BazaOgloszen.nauka.gov.pl i Nabory.kprm.gov.pl oraz portale Biuletynu Informacji Publicznej (BIP). Dane pobierano codziennie za pomocą robotów internetowych, tj. przygotowanego przez autora narzędzia do web scrapingu. W tej części artykułu posłużono się danymi z lat 2017 i 2018 według stanu na II kwartał. Odwołano się także do oficjalnych danych statystycznych zbieranych przez Główny Urząd Statystyczny (GUS) na podstawie sprawozdania Z-05 Badanie popytu na pracę (2019).

Istotne jest zdefiniowanie, jakiego rodzaju informacje można pozyskiwać ze źródeł danych big data, tj. stron internetowych zawierających oferty pracy. W tym celu odniesiono się do wybranych pozycji sprawozdania Z-05 (zestawienie 2), należy jednak podkreślić, że badanie to oferuje bardzo szeroki zakres danych, dopasowywanych do zmieniających się potrzeb odbiorców, uwarunkowanych również wytycznymi unijnymi (GUS, 2019).

**ZESTAWIENIE 2. WYBRANE INFORMACJE DOTYCZĄCE POPYTU NA PRACĘ PUBLIKOWANE PRZEZ GUS ORAZ MOŻLIWOŚCI WYKORZYSTANIA ŹRÓDEŁ DANYCH BIG DATA**

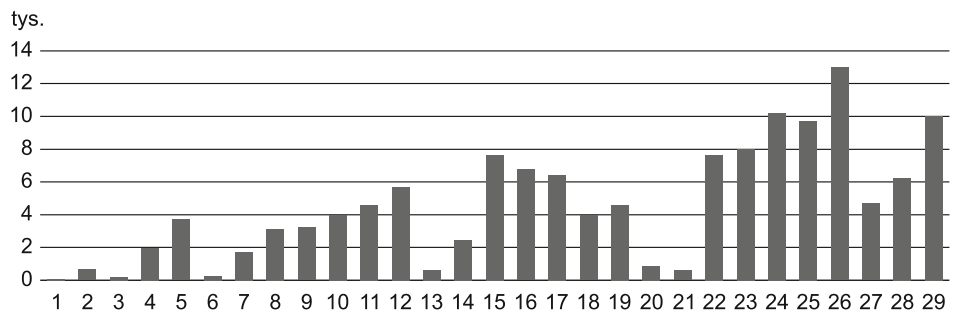
| Rodzaje informacji objętej badaniem GUS | Uwagi dotyczące wykorzystania źródeł danych big data |
|---|---|
| Wolne miejsca pracy – w końcu kwartału | możliwe do pozyskania, brak reprezentacyjności i nieznana populacja |
| Wolne nowo utworzone miejsca pracy – w końcu kwartału | możliwe do obliczenia w długim szeregu czasowym, brak reprezentacyjności i nieznana populacja |
| Nowo utworzone miejsca pracy – w kwartale | trudne do uzyskania, brak reprezentacyjności i nieznana populacja |
| Zlikwidowane miejsca pracy – w kwartale | brak możliwości pozyskania danych |

Ź r ó d ł o: opracowanie własne oraz GUS (2018).

Jak można wywnioskować z zestawienia 2, obecnie nie jest możliwe pozyskiwanie pełnej informacji na temat popytu na pracę, jak ma to miejsce w przypadku sprawozdania Z-05. Internetowe zbiory danych charakteryzują się nieznaną populacją, więc wyniki mogą znacznie odbiegać od oficjalnych danych publikowanych przez GUS.

Dalsza analiza danych dotyczy liczby ofert pracy pozyskiwanych z portali internetowych. Należy wziąć pod uwagę, że portale internetowe, na których publikowane są oferty pracy, mają zwykle wysokie standardy dotyczące aktualności zamieszczanych ofert – dane są usuwane najczęściej po upływie ok. miesiąca. Na wykr. 1 przedstawiono aktualne oferty pracy zamieszczane na wybranym portalu w kolejnych dniach maja 2017 r.

**WYKR. 1. NOWE OFERTY PRACY OPUBLIKOWANE OD 1 DO 29 MAJA 2017 R.**

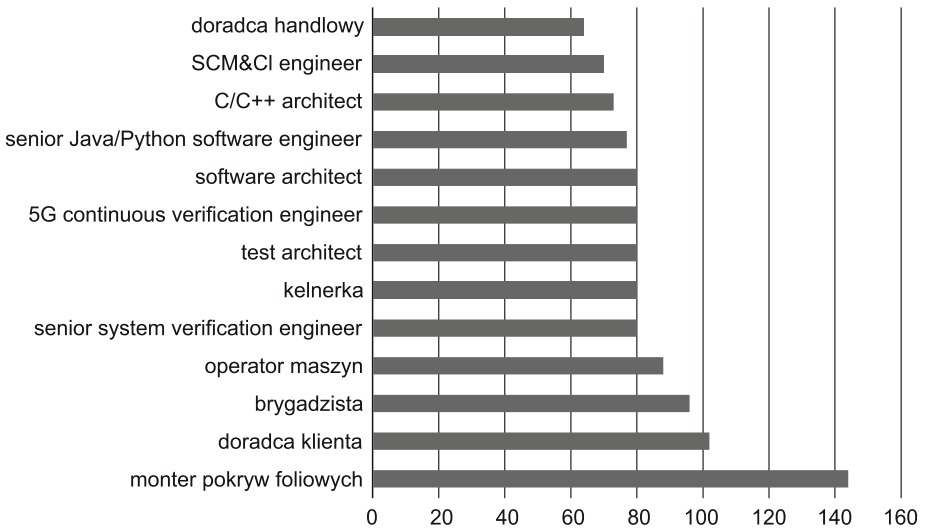



Ź r ó d ł o: opracowanie własne na podstawie wybranego portalu z ofertami pracy.

Wykres 1 przedstawia zasadę działania portali z ogłoszeniami o pracy, która polega na usuwaniu ofert pracy uznanych za nieaktualne. Uśredniając uzyskane wyniki, można zauważyć, że starszych ofert pracy jest znacząco mniej niż aktualnych. Na wykresie pominięto oferty sprzed maja 2017 r. Ich liczba ogółem wyniosła 325 (według daty publikacji oferty) w stosunku do 132575 ofert w maju 2017 r. Można również zaobserwować zjawisko polegające na codziennym umieszczaniu przez ogłoszeniodawców tych samych ofert pracy, aby były wyświetlane jako pierwsze przy domyślnym sortowaniu przeglądania ofert według daty dodania. Ten rodzaj aktywności firm na portalach z ogłoszeniami o pracy może występować kilka razy dziennie i skutkować błędami w obliczeniach. Skalę zjawiska prezentuje wykr. 2.

Wykres 2 dotyczy identycznych ofert, zawierających taki sam opis. Błędne wydaje się założenie, że firma poszukuje 80 czy 145 przedstawicieli danej profesji. Tym samym takie przykłady powinny być traktowane jako duplikaty. Aby wyeliminować tego typu problemy, opracowano metodę deduplikacji danych, która polega na wykrywaniu podobieństw w ofertach i usuwaniu duplikatów. Na podstawie analizy nazwy oferty, firmy, miejsca pracy oraz szczegółowego opisu odrzucano wszystkie oferty, które miały ten sam opis, najpierw w ramach jednego portalu (wykr. 2), a następnie dla grupy portali.

**WYKR. 2. OFERTY PRACY POWTARZAJĄCE SIĘ W MAJU 2017 R.**

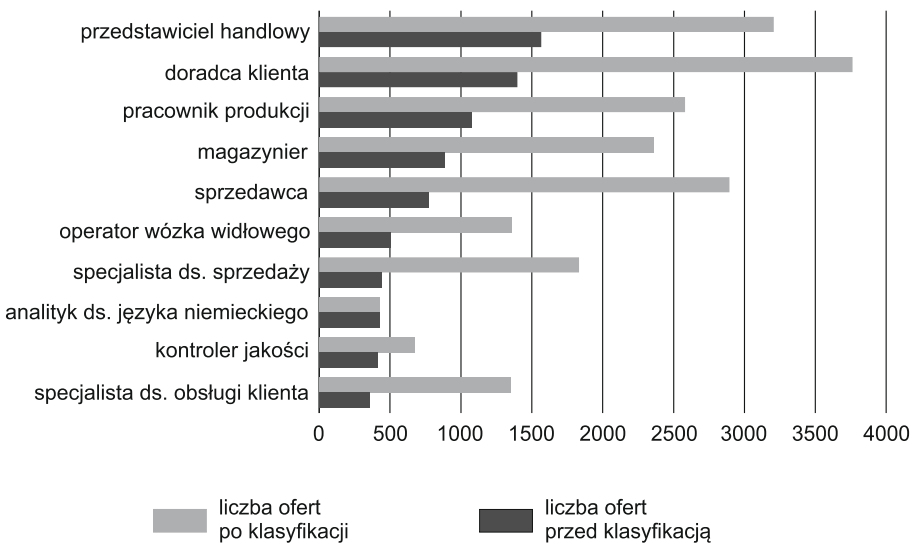



Ź r ó d ł o: opracowanie własne na podstawie wybranego portalu z ofertami pracy.

Innym wyzwaniem związanym z analizą ofert pracy jest dokonanie właściwej klasyfikacji danych. Właściciele portali internetowych zamieszczających ogłoszenia o pracy pozwalają pracodawcom na uwzględnienie dowolnych informacji w opisie ofert pracy. W związku z tym konieczne jest wykorzystanie metod text miningu, czyli metod eksploracji tekstu, w celu weryfikacji i przyporządkowania oferty pracy do właściwej kategorii według Klasyfikacji Zawodów i Specjalności (KZiS), bazującej na międzynarodowej klasyfikacji International Standard Classification of Occupation (ISCO). Metodą wspierającą właściwe odwzorowanie oferty pracy w klasyfikacji jest m.in. nadzorowane uczenie maszynowe, którego również nie można uznać za doskonałe narzędzie ze względu na dysproporcje w zbiorze treningowym w zakresie różnych zawodów. Przykładowo w zbiorze treningowym może być 100 obserwacji dla zawodu nauczyciel języka angielskiego i 5 tys. obserwacji dla zawodu programista. Oznacza to, że niektóre oferty pracy mogą mieć większą szansę na ich właściwe zaklasyfikowanie w stosunku do ofert rzadziej występujących w zbiorach treningowych. Skłania to bardziej do wykorzystywania metod deterministycznych, bazujących na wyrażeniach regularnych w text miningu, niż uczenia maszynowego. Wyrażenia regularne pozwalają na wyszukiwanie danych tekstowych według ustalonego wzorca, np. „angielsk*", co umożliwia wyeliminowanie problemów związanych z odmianą wyrazów w języku polskim.

Na wykr. 3 przedstawiono różnice pomiędzy liczbą ofert pracy przed zastosowaniem metod text miningu i po ich zastosowaniu. W celach testowych przygotowano również zbiór treningowy do uczenia maszynowego nadzorowanego w sposób ręczny, tj. przypisując nazwę oferty pracy do odpowiedniego kodu KZiS. Z przeprowadzonej analizy wynika, że nie jest możliwe uzyskanie precyzji klasyfikowania ofert prac na poziomie wyższym niż 90%.

**WYKR. 3. EFEKT ZASTOSOWANIA TEXT MININGU DO KLASYFIKACJI OFERT PRACY**




Ź r ó d ł o: opracowanie własne.

Jak zaprezentowano na wykr. 3, po zastosowaniu właściwego algorytmu klasyfikowania ofert pracy, z wykorzystaniem wcześniej wspomnianych metod text miningu i wyrażeń regularnych, można uzyskać inną liczbę ofert pracy w jednym zawodzie. Oznacza to, że niektóre oferty pracy przed zastosowaniem tych metod nie były przyporządkowane do zawodu zbliżonego do pozycji z KZiS.

Brak spójności dotyczy także wymiaru terytorialnego. Na portalach nie obowiązuje ujednolicona forma prezentacji, np. zgodna z rejestrem TERYT. Pojawiają się ogłoszenia, które nie zawierają określonej lokalizacji miejsca pracy (1–2% wszystkich ofert) lub jest ona bardzo ogólna, typu „cała Polska", „województwo pomorskie" czy „Gdańsk, Toruń, Warszawa".

Do identyfikacji zawodów w ofertach pracy wykorzystano wyrażenia regularne, które badały występowanie słów kluczowych z KZiS, np. „programista". Do-

datkowo utworzono rozwinięcie tego słownika o istotne słowa kluczowe, np. „Java" lub „Python" dla zawodu programisty. To niezbędny zabieg, gdyż w ofertach pracy nie zawsze występuje słowo „programista", a zamiast tego pojawia się np. „Java Developer". Przetwarzane dane oczyszczono wcześniej w procesie text miningu, tj. usunięto wyrazy niemające znaczenia dla analizy tekstu (ang. *stop words*). Ponadto sprowadzono słowa do postaci słownikowej, czyli do mianownika, następnie wyodrębniono rdzeń słów kluczowych dla danych zawodów, np. „programi*", który obejmuje takie formy, jak „programistów", „programista", „programiści" itd.

Najważniejsza kwestia w prowadzonym badaniu dotyczy porównywalności danych ze źródeł typu big data z danymi pozyskiwanymi w statystyce publicznej. Analiza poniższych danych ma jedynie charakter poglądowy, gdyż oferta pracy zamieszczona w internecie nie jest tożsama z wolnym miejscem pracy, które stanowi przedmiot badania GUS (tabl. 1).

**TABL. 1. WOLNE MIEJSCA PRACY WEDŁUG SPRAWOZDANIA Z-05 A OFERTY PRACY W ŹRÓDŁACH BIG DATA W II KWARTALE 2017 R.**

| Wyszczególnienie | Z-05 (wakaty) | Big data (oferty pracy) |
|---|---|---|
| | w tys. | |
| **O g ó ł e m**[a] | **122,0** | **110,0** |
| w tym: | | |
| Woj. mazowieckie | 27,9 | 23,2 |
| Pracownicy usług i sprzedawcy | 14,5 | 15,3 |

a Dane za II kwartał 2018 r. wynoszą odpowiednio: 164,7 i 131,0.

Ź r ó d ł o: opracowanie własne oraz GUS (2018).

Należy zaznaczyć, że dane w kolumnie big data mają charakter eksperymentalny i wymagają nieustannej ingerencji w metody służące do przetwarzania i analizy danych. Jest to związane z dużą zmiennością źródeł danych big data, co zostanie wyjaśnione w kolejnej części artykułu. Warto jednak zaznaczyć, że po przetworzeniu danych typu big data, polegającym m.in. na usunięciu duplikatów i zadbaniu o spójność klasyfikacji, określono skalę brakujących ofert pracy – dane porównywano grupami zawodów. Za brakujące oferty pracy uznano te, które nie są publikowane w internecie. Liczby uzyskane przy wykorzystaniu narzędzi big data są dużo niższe niż w przypadku danych pozyskiwanych na podstawie sprawozdania Z-05. Jest to znane zjawisko w analizie źródeł internetowych big data, które występuje również w innych zastosowaniach web scrapingu, np. do analizy cen produktów. Zwykle ceny produktów dostępnych w internecie są niższe niż w sklepach stacjonarnych. Odwrotnie jest w przypadku cen na rynku nieruchomości, gdzie zwykle cena ofertowa przewyższa cenę transak-

cyjną. Zaskakuje jednak to, że liczba ofert pracy dla grupy „pracownicy usług i sprzedawcy" przewyższa liczbę wolnych miejsc pracy podawaną na podstawie danych ze sprawozdania Z-05. Może to mieć związek z bardzo dużą rotacją pracowników zatrudnionych na tych stanowiskach.

Ze względu na różnicę definicji wakatu w przypadku sprawozdania Z-05 oraz oferty pracy online w przypadku źródła big data uzyskane liczby nie mogą być bezpośrednio porównywane. Jednak stosowanie źródeł typu big data umożliwi śledzenie trendów w zakresie popytu na rynku pracy, dostarczając bieżącej informacji w bardzo szybkim czasie.

## DANE SZCZEGÓŁOWE

Rozpatrując przedstawione powyżej wyniki, przede wszystkim należy wziąć pod uwagę wnioski dotyczące stabilności źródeł danych, tj. możliwości stałego pobierania danych ze stron internetowych o zbliżonym poziomie jakości. Wyniki badania pokazujące wahania w zakresie zmiany liczby ofert pracy rok do roku zaprezentowano w tabl. 2.

**TABL. 2. UŚREDNIONA LICZBA OFERT PRACY W II i III KWARTALE**

| Portale | | 2017 | 2018 |
|---|---|---|---|
| | | w tys. | |
| Ogólnopolskie: | 1 | 140 | 64 |
| | 2 | 124 | 19 |
| | 3 | 46 | 86 |
| | 4 | 39 | 54 |
| Branżowe: | 1 | 1,2 | 0,6 |
| | 2 | 0,4 | 0,8 |
| Regionalny | | 9,9 | 8,5 |

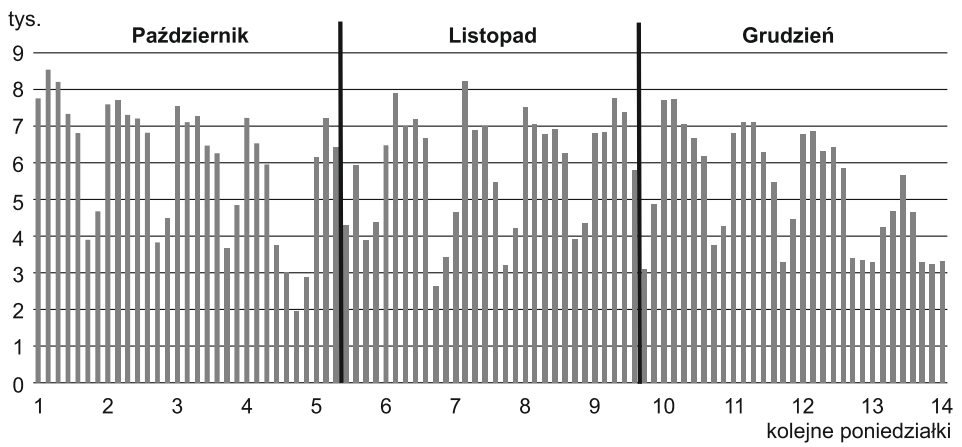
U w a g a. Cyfry w kolumnie Portale oznaczają kolejne badane portale.
Ź r ó d ł o: opracowanie własne.

Jak wynika z powyższej tabeli, zbiory danych wykazują dużą niestabilność – niektóre portale odnotowały niemal dwukrotny spadek liczby bieżących ofert pracy. Przyczyn należy upatrywać w odejściu wielu użytkowników tych portali do innych serwisów, oferujących więcej ofert. Ponadto coraz częściej praca jest oferowana poprzez media społecznościowe (np. Twitter, Facebook), bezpośrednio na stronie firmy czy na targach pracy dla profesjonalistów. Niektóre firmy zatrudniają headhunterów, którzy zajmują się poszukiwaniem odpowiednich osób poprzez portale społecznościowe, takie jak LinkedIn. W ramach tego portalu użytkownicy mogą zadeklarować swoje preferencje odnośnie do ewentualnej

zmiany pracy i otwartości na oferty headhunterów. Warto również podkreślić, że coraz więcej firm komercyjnych pobiera dane z internetu, a następnie udostępnia je w formie analiz. Istotnym przykładem wykorzystania ofert pracy online w celu budowy statystyk jest narzędzie do przeglądania ofert pracy online o nazwie Skills Online Vacancy Analysis Tool for Europe (Skills-OVATE) opracowane przez Europejskie Centrum Rozwoju Kształcenia Zawodowego (Cedefop)[5].

Brak stabilności portali internetowych, który nie pozwala uznać ich za niezmienne i stałe źródła danych, zdaje się dodatkowo potwierdzać analiza danych ujętych w układzie dziennym. Na wykr. 4 przeanalizowano dane z IV kwartału 2018 r. dotyczące jednego z większych portali internetowych zamieszczających oferty pracy.

**WYKR. 4. NOWE OFERTY PRACY OPUBLIKOWANE W IV KWARTAŁ 2018 R.**

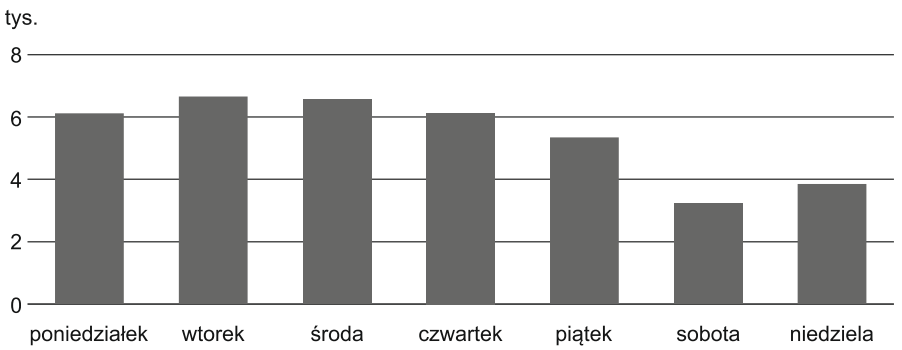



Ź r ó d ł o: opracowanie własne na podstawie danych z wybranego portalu z ofertami pracy.

Dane na wykr. 4 pokazują, że w zależności od dnia miesiąca liczba nowych ofert pracy może różnić się o ponad 50%. Dzienne zmiany dotyczące liczby ofert mogą sięgać kilkunastu procent. Oznacza to, że wyniki powinny być uśredniane w skali miesiąca/kwartału, aby w sposób wiarygodny zobrazować popyt na pracę. Istotną wartość dodaną stanowi analiza, kiedy pojawia się najwięcej ofert pracy w skali roku, kwartału czy nawet dnia tygodnia (wykr. 5), ale te informacje mogą być bardziej przydatne dla osób poszukujących pracy niż dla statystyki publicznej.

[5] https://www.cedefop.europa.eu (dostęp: 21.06.2019).

**WYKR. 5. NOWE OFERTY PRACY WEDŁUG DNIA TYGODNIA – DANE UŚREDNIONE Z IV KWARTAŁU 2018 R.**




Ź r ó d ł o: opracowanie własne na podstawie danych z wybranego portalu z ofertami pracy.

Po wyborze właściwego źródła danych zawarte w nim oferty są wstępnie przetwarzane. Podczas tego procesu następuje:

- deduplikacja danych, czyli usunięcie takich samych ofert pracy, powtarzanych wielokrotnie;
- ręczne odwzorowywanie KZiS lub wykorzystanie w tym celu uczenia maszynowego;
- dopasowanie miejscowości do województw zgodnie z rejestrem TERYT.

Trzeba pamiętać, że łączenie źródeł danych zwiększa liczbę duplikatów. Należy się zatem skoncentrować na możliwie pełnych zbiorach danych. Zagadnienie reprezentacyjności ma istotny wpływ na możliwość zastosowania metody web scrapingu w statystyce publicznej. Wykorzystanie narzędzi big data można rozpatrywać pod kątem wzbogacenia obecnie dostępnych danych statystycznych oraz uchwycenia trendów w możliwie najszybszym czasie, nawet kilka dni po zebraniu danych. W celu uniknięcia problemów z niestabilnością źródeł danych oraz precyzyjnego określenia metodologii wskazane jest nawiązanie współpracy z właścicielami serwisów internetowych.

## PODSUMOWANIE

Na podstawie przeprowadzonych badań eksperymentalnych stwierdzono, że internetowe źródła danych charakteryzują się dużą niestabilnością i niepewną jakością zamieszczanych w nich informacji. Oferty pracy publikowane w internecie mogą być nieaktualne. Dodatkowo dowolność nazewnictwa zawodu w ofercie pracy sprawia, że konieczne jest ręczne odwzorowywanie zawodu w celu uzyskania zgodności z KZiS. Jedną z metod wspierających może być nadzorowane uczenie maszynowe. Niekiedy utrudnieniem jest brak jednoznacznego przyporządkowania terytorialnego, co również wymaga zastosowania wyrażeń

regularnych oraz metod text miningu w celu ekstrakcji takiej informacji. Co więcej, zastosowanie wielu źródeł danych może prowadzić do powstawania duplikatów, gdyż te same oferty mogą być powtarzane na wielu portalach.

Co istotne, wiele krajów europejskich pracuje nad rozwiązaniami pozwalającymi analizować oferty pracy online. Jako przykład mogą posłużyć prace prowadzone w ramach grantów ESSNet czy też informacje gromadzone przez Cedefop, gdzie oferty pracy są zbierane już od wielu lat. Zaletą pokazanego w niniejszym artykule rozwiązania jest bardzo szybkie dostarczanie danych wynikowych, które mogą zobrazować trendy w zakresie zmiany liczby ofert pracy. Wciąż jednak niepewność co do źródeł danych przyczynia się do braku jednoznacznej odpowiedzi, czy tego rodzaju dane mogą być wykorzystywane przez statystykę publiczną jako oficjalne dane statystyczne.

Zaprezentowane szacunki mają wyłącznie charakter eksperymentalny i nie powinny być traktowane jako oficjalne dane statystyczne ze względu na niestabilność źródeł danych i prawdopodobną konieczność weryfikacji w przyszłości założeń do algorytmów. W artykule zwrócono uwagę na możliwości stosowania źródeł big data do przetwarzania i analizowania danych statystycznych. Należy przy tym zaznaczyć, że oferty pracy zamieszczane w internecie nie są równoznaczne z wolnymi miejscami pracy, które przedsiębiorstwa wykazują w sprawozdawczości dla GUS. Tym samym istnieje konieczność zdefiniowania nowych terminów, pozwalających rozróżnić dwa odrębne pojęcia, jakimi są oferta pracy publikowana w internecie i wolne miejsce pracy.

## BIBLIOGRAFIA


Beręsewicz, M., Szymkowiak, M. (2015). Big data w statystyce publicznej – nadzieje, osiągnięcia, wyzwania i zagrożenia. *Ekonometria*, *2*(48), 9–22. DOI: 10.15611/ekt.2015.2.01.

Braaksma, B., Zeelenberg, K. (2015). "Re-make/Re-model": Should big data change the modelling paradigm in official statistics? *Statistical Journal of the IAOS*, *31*(2), 193–202. DOI: 10.3233/sji-150892.

Daas, P. J. H., Puts, M. J., Buelens, B., van den Hurk, P. A. M. (2015). Big Data as a Source for Official Statistics. *Journal of Official Statistics*, *31*(2), 249–262. DOI: https://doi.org/10.1515/jos-2015-0016.

Douglas, L. (2001). *3D Data Management: Controling Data Volume, Velocity and Variety*. Pobrane z: http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf.

Gałecka-Burdziak, E., Pater, R. (2015). Ile jest wolnych miejsc pracy w Polsce? *Gospodarka Narodowa*, *279*(5), 171–186. DOI: https://doi.org/10.33119/GN/100855.

GUS. (2018). *Popyt na pracę w 2017 r*. Warszawa: Główny Urząd Statystyczny.

GUS. (2019). *Popyt na pracę w 2018 r*. Warszawa: Główny Urząd Statystyczny.

Hackl, P. (2016). Big Data: What can official statistics expect? *Statistical Journal of the IAOS*, *32*(1), 43–52. DOI: 10.3233/SJI-160965.

Kitchin, R. (2015). The opportunities, challenges and risks of big data for official statistics. *Statistical Journal of the IAOS*, *31*(3), 471–481. DOI: 10.3233/SJI-150906.

Kureková, L. M., Beblavý, M., Thum-Thysen, A. (2015). Using online vacancies and web surveys to analyse the labour market: a methodological inquiry. *IZA Journal of Labor Economics*, *4*(18), 1–20. DOI: 10.1186/s40172-015-0034-4.

Maślankowski, J. (2014). Data Quality Issues Concerning Statistical Data Gathering Supported by Big Data Technology. W: S. Kozielski, D. Mrozek, P. Kasprowski, B. Małysiak-Mrozek, D. Kostrzewa (red.). *Beyond Databases, Architectures and Structures* (s. 92–101) Cham: Springer.

Miller, S. (2014). Collaborative Approaches Needed to Close the Big Data Skills Gap. *Journal of Organization Design*, *3*(1), 26–30. DOI: 10.7146/jod.9823.

Rousidis, D., Garoufallou, E., Balatsoukas, P., Sicilia, M. (2014). Metadata for Big Data: a preliminary investigation of metadata quality issues in research data repositories. *Information Services &Use*, *34*(3–4), 279–286. DOI: 10.3233/ISU-140746.

Shahin, S. (2016). A Critical Axiology for Big Data Studies. *Palabra Clave*, *19*(4), 972–996. DOI: 10.5294/pacla.2016.19.4.2.

Vale, S. (2015). International collaboration to understand the relevance of Big Data for official statistics. *Statistical Journal of the IAOS*, *31*(2), 159–163. DOI: 10.3233/sji-150889.