UNIWERSYTET EKONOMICZNY W POZNANIU

Wydział Informatyki
i Gospodarki Elektronicznej

Greta Białkowska
Magdalena Maślak i Krzysztof Marcinkowski

Internetowe źródła danych o popycie na pracę w Polsce

Internet data sources on the demand for labour in Poland

Praca Magisterska

Promotor: dr Maciej Beręsewicz

Pracę przyjęto dnia:

Podpis Promotora

Kierunek: Informatyka i ekonometria

Specjalność: Analityka gospodarcza

Poznań 2019

Spis treści

Wstęp

Polska gospodarka w ostatnich latach bardzo przyspieszyła. Według najnowszych prognoz Organizacji Współpracy Gospodarczej i Rozwoju (ang. Organisation for Economic Co-operation and Development; OECD) Polska w 2019 roku będzie czwarta w rankingu najszybciej rozwijających się krajów świata. Wyższe tempo wzrostu PKB posiadać będą jedynie 3 kraje: Indie (7,16%), Chiny (6,2%), i Indonezja (5,07%) (OECD, 2018). Wysokie tempo rozwoju gospodarki sprawia, iż z jednej strony bezrobocie w Polsce cały czas maleje, a z drugiej na rynku znajduje się znaczna liczba wakatów zarówno dla obywateli Polski jak i cudzoziemców.

Analiza wolnych miejsc pracy pozwala na zidentyfikowanie zawodów oraz branż, które charakteryzują się największym popytem, a także problemem w znalezieniu odpowiednich pracowników. Kluczowy w tym aspekcie jest bieżący monitoring ofert pracy, który umożliwi rządzącym, administracji publicznej, a także szkołom i uczelniom wyższym na identyfikację zawodów deficytowych oraz kompetencji wymaganych przez pracodawców. W Polsce jednym ze źródeł, które dostarcza informacji w tym zakresie jest badanie Popyt na Pracę realizowane przez Główny Urząd Statystyczny. Z drugiej strony pracodawcy korzystają różnego typu kanałów, aby znaleźć odpowiednich kandydatów przez m.in. Powiatowe i Wojewódzkie Urzędy Pracy czy portale internetowe.

W pracy poruszona zostanie problematyka pomiaru rynku pracy z wykorzystaniem niestatystycznych źródeł danych, w szczególności ofert pracy zgłoszonych do Powiatowych i Wojewódzkich Urzędów Pracy (PUP; WUP) oraz portalu OLX. W związku z nielosowym charakterem tych źródeł nie można ich bezpośrednio wykorzystać do pomiaru popytu i zachodzi potrzeba ich oceny ze względu na jakość danych, wykorzystane definicje czy reprezentatywność (błędy pokrycia, błędy nielosowe).

Praca składa się ze wstępu, pięciu rozdziałów oraz zakończenia. Rozdział pierwszy jest rozdziałem teoretycznym, w którym uwaga została skierowana na ogólną problematykę pomiaru rynku pracy – opisane zostały podstawowe definicje dotyczące rynku pracy oraz, przede wszyst-

kim, zaprezentowane źródła danych o popycie na pracę. Przedstawiono zarówno źródła statystyczne, jak i niestatystyczne oraz dokonano ich wstępnego porównania. Rozdział ten został przygotowany przez wszystkich autorów niniejszej pracy.

W rozdziale drugim poruszona została problematyka reprezentatywności internetowych źródeł danych jako jednego z kanałów wykorzystywanych przez pracodawców. Głównym źródłem informacji wykorzystanym do analizy jest Badanie Przedsiębiorców przeprowadzone w ramach cyklicznego, reprezentacyjnego Badania Kapitału Ludzkiego (BKL) realizowanego na zlecenie Polskiej Agencji Rozwoju Przedsiębiorczości (PARP). Dane wykorzystane w tym rozdziale pochodzą z lat 2010-2014. Rozdział składa się z części teoretycznej i empirycznej. W części teoretycznej przedstawiony został problem reprezentatywności oraz opis modeli Item Response Theory (pol. teoria odpowiedzi na pytania testowe; IRT)[1], wykorzystanych w drugiej części rozdziału. Model ten nie był do tej pory wykorzystywany do badania reprezentatywności, więc jego aplikacja na potrzeby niniejszego badania jest pewnym novum wniesionym przez autorkę tego rozdziału – Magdalenę Maślak. Natomiast część empiryczna zawiera wyniki modeli IRT oszacowanych przy użyciu pakietu `LTM` (Rizopoulos, 2006) oraz pakietu `MIRT` (Chalmers, 2012) w języku R. Wyniki przedstawione zostały w rozbiciu na różne grupy pracodawców. Charakterystykami wykorzystanymi do oceny reprezentatywności były: sekcje PKD, wielkość firmy oraz województwo.

Rozdział trzeci ma na celu przeanalizowanie niestatystycznego źródła danych o popycie na pracę jakim jest Centralna Baza Ofert Pracy (CBOP), która zawiera większość ofert pracy spływających do PUP i WUP. Dane jednostkowe (na poziomie oferty pracy) zostały pozyskane od Ministerstwa Rodziny, Pracy i Polityki Społecznej w ramach zapytania o informację publiczną. Zgodnie z wiedzą autorki – Grety Białkowskiej – jest to pierwsze, naukowe wykorzystanie tych danych w Polsce, a w szczególności na potrzeby opisu rynku pracy. W pierwszej kolejności skupiono się na porównaniu CBOP do badania Popytu na Pracę (GUS) a następnie zweryfikowanie czy zastosowanie metod ważenia danych (kalibracji) może zredukować błędy nielosowe wynikające z charakteru tych danych. Rozdział dzieli się na część teoretyczną oraz empiryczną. W części teoretycznej opisane zostały błędy nielosowe oraz metody ich korekcji takie jak imputacja oraz kalibracja. W części praktycznej przeanalizowane zostały dane pochodzące z CBOP. Zostały wybrane najbardziej istotne zmienne, poprawiono błędy nielosowe, które znajdowały

[1] W pracy używany będzie anglojęzyczny termin ponieważ polskie tłumaczenie, w odbiorze autorów niniejszej pracy, nie do końca precyzyjnie oddaje założenia i cel tej metody. Dodatkowo w polskiej literaturze zwykle spotyka się określenie anglojęzyczne.

się w danych. Za pomocą współczynnika V Cramera zbadano korelacje pomiędzy zawodem a poszczególnymi zmiennymi, które mogłyby wpływać na umieszczanie w bazie ofert z danej kategorii zawodu. Następnie dane porównano do danych statystycznych pochodzących z Badania Popytu na Pracę, zbadano rozkłady liczebności ofert według sekcji Polskiej Klasyfikacji Działalności, wybrano odpowiednie zmienne, by w kolejnym kroku zastosować kalibrację do skorygowania braków danych w zbiorze CBOP. Podczas kalibracji wykorzystano dwie zmienne jako zmienne pomocnicze, a kalibracja wag nastąpiła w trzech wariantach. Na końcu przedstawiono wnioski.

W czwartym rozdziale poruszany jest aspekt uczenia maszynowego. Celem rozdziału jest stworzenie modelu który będzie w stanie zaklasyfikować oferty pracy do kategorii *Klasyfikacji Zawodów i Specjalności* wykorzystywanej w badaniu Popytu na Pracę oraz CBOP. W tym celu napisano autorski program do pobierania danych z portalu OLX. Jako zbiór danych uczących wykorzystano badanie ofert pracy realizowane w ramach Badania Kapitału Ludzkiego z 2014 roku oraz CBOP za 2017 rok. Wynikowe dane poddane poddane zostały przetworzeniu z wykorzystaniem technik *text mining*. Następnie na ich podstawie stworzone zostaną modele uczenia maszynowego, 1) regresję LASSO oraz 2) algorytm Naiwnego Bayesa przy założeniu wielomianowego rozkładu badanej cechy (zawodów). Obliczenia wykonano w infrastrukturze *InnoUEP*, specjalnej platformy umożliwiającej wykorzystanie komputerów o dużej mocy na potrzeby badań naukowych i prac dyplomowych. W celu wykorzystania pełni możliwości tejże infrastruktury modele zostaną zbudowane w dwóch pakietach 1) `H2O` (LeDell i in., 2019) oraz 2) `glmnet` (Friedman, Hastie & Tibshirani, 2010). Pakiety te dostarczają paletę możliwości związanej z uczeniem maszynowym. Model wielomianowej regresji logistycznej LASSO zbudowany zostanie zarówno w pakiecie `glmnet` jak i `H2O`, model Naiwnego Bayesa zbudowany zostanie w pakiecie `H2O`. Zbudowane modele posłużą do klasyfikacji ofert z portalu internetowego OLX. W celu oceny klasyfikacji wylosowana zostanie próbka, wynosząca 1%, co wynosi około 4000 obserwacji. Dane zostaną manualnie zaklasyfikowane do kategorii *Klasyfikacji Zawodów i Specjalności*. Następnie wyniki modeli zostaną porównane z automatyczną klasyfikacją. Ocena modeli będzie oparta o macierze klasyfikacji oraz takie miary jak precyzja czy też miara F1. Zgodnie z wiedzą autora – Krzysztofa Marcinkowskiego – wykorzystanie i ocena danych z portalu OLX na potrzeby popytu na pracę nie była wcześniej tak dogłębnie przedstawiana w polskiej literaturze.

Rozdział piąty został przygotowany przez wszystkich autorów niniejszej pracy i zawiera w so-

bie podsumowanie, czyli porównanie wykorzystanych w pracy zbiorów danych: danych pochodzących z Badania Popytu na Pracę, Centralnej Bazy Ofert Pracy oraz z portalu OLX. W tym rozdziale skupiono się przede wszystkim analizie porównawczej rozkładów ofert pracy ze względu na kwartał, województwo oraz zawód dla każdego ze źródeł danych. Ponadto zostały zbadane korelacje pomiędzy źródłem a zwodem oraz źródłem a województwem za pomocą współczynnika V Cramera, korelacje pomiędzy odsetkami oraz obciążenie i relatywne absolutne obciążenie. Przedstawione zostały wnioski z powyższych analiz.

Pracę kończy podsumowanie oraz przedstawienie dalszych kroków badań.

Rozdział 1

Nowe źródła danych o popycie na pracę (Greta, Magdalena, Krzysztof)

1.1 Problematyka pomiaru popytu na pracę

1.1.1 Podstawowe definicje

Omówienie źródeł danych o popycie na pracę należy zacząć od definicji, które stoją za ich powstaniem. Warto zacząć od najbardziej ogólnej, czyli od definicji rynku pracy. Otóż rynek niezależnie od swojej formy jest miejscem, w którym odbywają się transakcje kupna i sprzedaży różnych produktów, zatem rynek pracy można definiować jako *miejsce, gdzie dochodzi do konfrontacji podaży oraz popytu na pracę*.

Poprzez podaż pracy rozumiemy siłę roboczą, czyli ludność aktywną zawodowo, na którą składają się osoby pracujące oraz osoby szukające zatrudnienia. Z kolei popyt na pracę można wyrazić poprzez liczbę osób, którą pracodawcy są w stanie zatrudnić na określonych zasadach - ogólnie rzecz biorąc jest to liczba miejsc pracy oferowanych przez gospodarkę.

Poprzez liczbę miejsc pracy rozumiemy zarówno zagospodarowane (liczba pracujących), jak i wolne miejsca pracy.

Definicja 1.1. Na potrzeby Badania Popytu na Pracę (o którym szerzej w następnej sekcji), definiuje się wolne miejsca pracy jako z miejsca pracy wykreowane w rezultacie ruchu osób zatrudnionych, lub też nowo utworzone, które spełniły jednocześnie trzy warunki (Główny Urząd Statystyczny, 2015):

- stanowiska w dniu sprawozdawczym były nieobsadzone,

- pracodawca starał się znaleźć osoby chętne do podjęcia stanowiska,
- w przypadku znalezienia odpowiednich osób, pracodawca byłby gotów do natychmiastowego ich przyjęcia.

Próba pomiaru rynku pracy nie jest łatwym zadaniem, ze względu na specyfikę oferowanych ”produktów”. Podczas badania popytu na pracę, głównym źródłem informacji są oferty pracy.

Należy zaznaczyć, że Główny Urząd Statystyczny publikuje również statystyki oparte na danych z powiatowych urzędów pracy (PUP), do których należy m.in. liczba osób bezrobotnych zarejestrowanych w PUP czy liczbę ofert pracy zgłoszonych do PUP. W odniesieniu do tej ostatniej używana jest następująca definicja.

Definicja 1.2. Oferta pracy to zgłoszone przez pracodawcę do powiatowego urzędu pracy (Główny Urząd Statystyczny, 2019b):

- co najmniej jedno wolne miejsce zatrudnienia lub inna praca zarobkowa,
- miejsce aktywizacji zawodowej, przyjęte do realizacji.

W przypadku internatowych źródeł danych (o których szerzej w kolejnej sekcji) można oprzeć się na definicjach wykorzystanych w regulaminach korzystania z usług danego serwisu. Na przykład, portal OLX (Grupa OLX, 2019) wykorzystuje następującą definicję

Definicja 1.3. Ogłoszenie – sporządzone przez Użytkownika ogłoszenie dotyczące sprzedaży (zaproszenie do zawarcia umowy sprzedaży), zamiany, pracy, oferowanych usług itd., zamieszczane w Serwisie, na warunkach przewidzianych w Regulaminie.

Czy ogłoszenie o pracę jest równoznaczne z ofertą pracy? Nie, na podstawie powyższej definicji za ofertę pracy można uznać jedynie oferty zgłaszane do Powiatowych Urzędów Pracy. Definicja ta nie uwzględnia ofert czy ogłoszeń o pracę zamieszczanych w innych miejscach, na przykład w Internecie. Stąd należy zadać pytanie: jak można definiować oferty pochodzące z prasy bądź portali internetowych, które również kształtują popyt na pracę? Aby odpowiedzieć na to pytanie trzeba w pierwszej kolejności zrozumieć jak poszczególne źródła danych powstają oraz jakie są ich charakterystyki.

1.2 Źródła danych o popycie na pracę

1.2.1 Źródła danych w statystyce

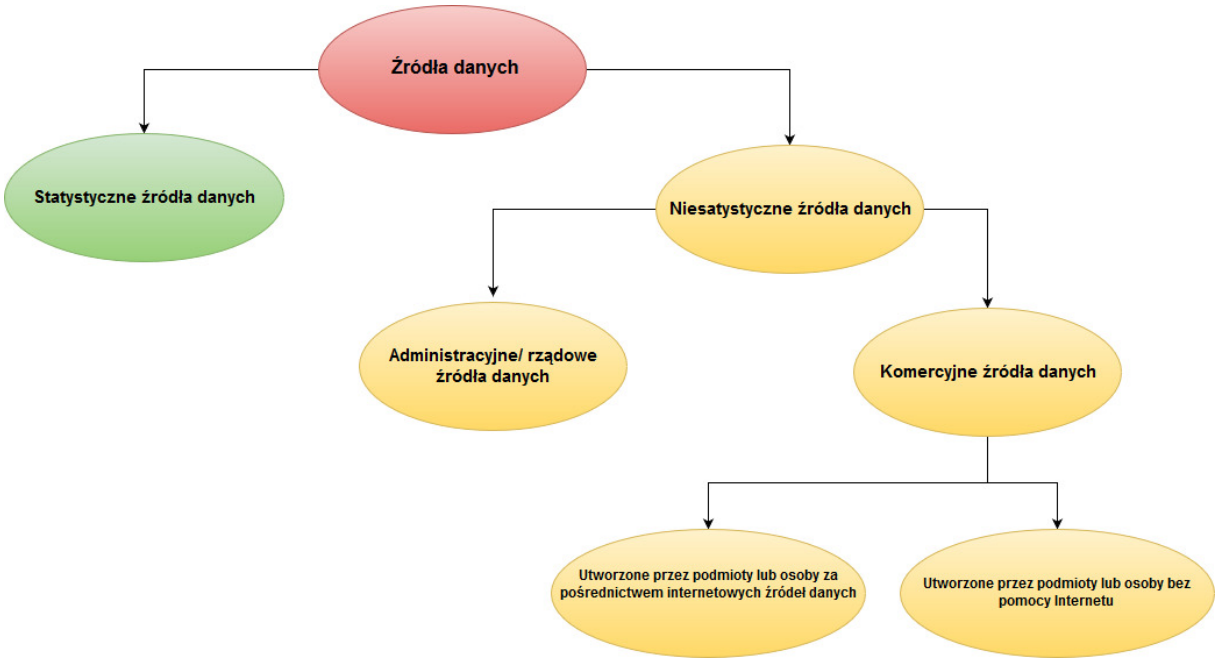
Podczas przeprowadzania różnego rodzaju badań, bez względu na ich tematykę, głównym czynnikiem, który decyduje o rzetelności wyników, jest jakość wykorzystywanych danych. Do badania popytu na pracę wykorzystać można szereg źródeł, które mogą mieć różny charakter wynikający z celu w jakim owe źródła powstały.

Źródła danych w statystyce podzielić można na dane statystyczne, jak i niestatystyczne. Źródła statystyczne to takie, które powstały do celów statystycznych, do badań i wnioskowania. Przykładem takich danych są wszelkie badania prowadzone przez urzędy statystyczne czy spisy narodowe. Natomiast dane niestatystyczne to wszystkie dane powstające ”przy okazji”, są to źródła, których pierwotnym celem nie było dostarczenia informacji statystycznej. Przykładem takich zbiorów danych są rejestry administracyjne, które stworzone są do zarządzania, czy dane pochodzące z Internetu, które tworzone są w różnym celu ale wykorzystywane przez firmy do celów biznesowych. Oba te przykłady mogą zostać wykorzystywane do celów statystycznych jednak dopiero po ich dogłębnej ocenie.

Niestatystyczne źródła danych dzielimy na dwie kategorie. Pierwszą z nich są dane rządowe/administracyjne, czyli dane, które powstają w celu zachowania ładu i porządku w codziennym funkcjonowaniu państwa. Przykładem takich danych są wspomniane powyżej rejestry administracyjne, dane medyczne powstające w szpitalach czy informacje pochodzące z radarów drogowych. Drugą kategorią danych niestatystycznych są dane komercyjne – prywatne, które dzielą się na dane pochodzące z Internetu oraz nie. Przykładem danych internetowych są dane pochodzące z portali społecznościowych (tzw. social media), aukcyjnych, sklepów internetowych itd., natomiast przykładem niestatystycznych źródeł danych spoza internetu to dane pochodzące z handlu tradycyjnego, banków czy Internetu Rzeczy (Beręsewicz, 2016).

Dane statystyczne dzielą się również na dane pierwotne oraz dane wtórne. Źródła pierwotne to źródła statystyczne – statystyka publiczna, natomiast dane wtórne to dane, które powstają w sposób niekontrolowany, czyli dane niestatystyczne, do których zaliczamy dane pochodzące z Internetu.

Badania prowadzone przez urzędy statystyczne często powstają z pewnym opóźnieniem za obecnymi zmianami. Skutkuje to rosnącą przepaścią informacyjną między stanem faktycznym a klasycznymi źródłami danych w niektórych branżach, zwłaszcza tych najszybciej się rozwija-



Rysunek 1.1. Klasyfikacja źródeł danych w statystyce

Źródło: Opracowanie własne na podstawie: (Beręsewicz, 2016)

jących. Ponadto uzyskanie danych za pomocą statystycznych metod wiąże się z stosunkowymi kosztami. Sami badani coraz częściej odmawiają udziału w badaniach. Niestatyczne źródła danych rysują się jako właśnie takie które mogą zminimalizować, zarówno koszty jak i rozwiązać problem z odmowami z strony osób badanych. Same dane mogą być uzyskiwane na różne sposoby, najprostszym, pod względem technicznym, sposobem jest uzyskanie ich od twórców konkretnych stron. Inną, najczęstszą, metodą jest utworzenie specjalnych algorytmów które będą automatycznie sczytywać dane z stron internetowych (tak zwany *web scraping*).

Zwrot urzędów statystycznych i instytucji naukowych w kierunku innych źródeł informacji w celu opisu badanej zbiorowości nazywany jest w literaturze zmianą paradygmatu. Opiera się on na odejściu od danych stworzonych przez badaczy (ang. *designed-based*) do wykorzystania istniejących źródeł informacji (ang. *process-based*). Ważnym przykładem jest tutaj z pewnością holenderski urząd statystyczny, który jest pionierem w tym zakresie. Należy zaznaczyć, że Główny Urząd Statystyczny (GUS) również prowadzi w tym zakresie badania m.in. zostało stworzone Centrum Kompetencji Big Data w administracji publicznej [1].

Nierozerwalnie z tym zjawiskiem związane jest pojęcie *big data*. Opisuje ono charakter opisywanych danych, a przede wszystkim ich duży wolumen, dużą zmienność i tempo ich po-

[1] Por. https://stat.gov.pl/prezentacje/gus-centrum-kompetencji-bigdata-w-administracji-publicznej-,75,1.html

wstawania, różnorodność oraz ich niestrukturyzowanie. Dane te powstają na skutek ludzkiego działania, nie są wynikiem z góry założonych badań. Dlatego też w literaturze nazywane one są danymi organicznymi. Klasyczne źródła danych często cechują się niskim wolumenem, mają określoną dynamikę oraz są ustrukturyzowane (Beręsewicz & Szymkowiak, 2015).

Warto w tym miejscu zdefiniować to czym właściwie są internetowe źródła danych, w literaturze można przeczytać, że internetowe źródła danych są samodzielnie wybraną próba, stworzoną przy użyciu internetu i utrzymywaną poza urzędami statystycznymi oraz niepodlegającą ich metodologii.

Definicja ta porusza kilka ważnych aspektów tego źródła informacji. Po pierwsze, pomimo często dużej ilości danych, to wciąż dane internetowe powinny być traktowane jako próba. Ponadto, należy uwzględnić fakt, że dane te są wynikiem samodzielnego wyboru konkretnych jednostek, co skutkuje deterministycznym charakterem internetowych źródeł danych. Warto zaznaczyć że to źródło danych różni się od innych rodzajów danych. Zwłaszcza w kontekście trafności estymacji, aktualności czy też porównywalności. Internetowe źródła danych wciąż nie są jednoznacznie ocenione pod względem możliwości ich użycia w statystyce (Beręsewicz, 2017).

1.2.2 Źródła danych o popycie na pracę w Polsce

1.2.2.1 Badanie Popyt na Pracę

W Polsce głównym źródłem informacji o popycie na pracę jest badanie *Popyt na Pracę* realizowane przez GUS i ma na celu dostarczenie informacji o popycie zrealizowanym oraz niezrealizowanym, czyli o pracujących, wolnych i nowo utworzonych miejscach pracy. Jest to badanie oparte na losowej próbie i prowadzone kwartalnie.

Populacją badania są podmioty gospodarki narodowej o liczbie zatrudnionych osób 1 i więcej. Badanie obejmuje podmioty zaliczane do wszystkich rodzajów działalności. Dobór jednostek do próby odbywa się w dwóch oddzielnych etapach:

1. pierwszy to losowanie jednostek, które zatrudniają powyżej dziewięciu osób,
2. drugi to losowanie przedsiębiorstw zatrudniających do dziewięciu pracowników

GUS motywuje ten podział różnymi celami dla każdej części. W przypadku pierwszej grupy, czyli jednostek zatrudniających więcej osób, za cel przyjęto uzyskanie informacji na temat poszczególnych rodzajów działalności, sekcji Polskiej Klasyfikacji Działalności (PKD), w podziale na województwa. W tej części zbiorowości wyodrębnione zostały 304 oddzielne populacje.

Założona liczebność próby wynosi około 50 tys. jednostek (w każdym roku). Podział próby pomiędzy poszczególne subpopulacje (stratyfikacja) przeprowadzany jest w taki sposób, by wyniki badania (w przybliżeniu) osiągnęły identyczną precyzję dla populacji generalnej. W każdej populacji jednostki sortowane są ze względu na liczbę pracujących w malejący ciąg. Największe jednostki z każdej populacji włącza się do badania bez losowania. Wykorzystując metodę optymalizacji numerycznej założoną próbę rozdziela się między populacje, natomiast w każdej z 304 populacji ustala się próg liczby pracujących. Podmioty o liczbie pracujących większej niż ustalony próg włącza się do badania bez losowania, natomiast pozostała część dzielona jest na warstwy o jednakowych rozmiarach, a następnie losowana jest próba o założonej wcześniej liczebności z zastosowaniem losowania proporcjonalnego (Główny Urząd Statystyczny, 2015).

Dla jednostek o liczbie pracujących do dziewięciu osób zasadniczym celem badania było w głównej mierze uzyskanie precyzyjnych wyników dla poszczególnych rodzajów działalności – 19 sekcji PKD. W badaniu tym zastosowana została również "Klasyfikacja zawodów i specjalności" wprowadzona Rozporządzeniem Ministra Pracy i Polityki Społecznej. Alokacja liczebności próby (50 tys.) dokonywana jest, w tym przypadku, pomiędzy poszczególne sekcje PKD tak, by oczekiwana precyzja dla owych sekcji była jednakowa. Jednostki w sekcjach warstwowane są według województw. Próba losowana jest zgodnie ze schematem losowania warstwowego proporcjonalnego (Główny Urząd Statystyczny, 2018b).

Dane wynikowe publikowane są w następujących przekrojach:

- sekcje i działy PKD,
- zawody,
- sektory własności,
- wielkość jednostek według liczby pracujących,
- regiony,
- województwa.

1.2.2.2 Bilans Kapitału Ludzkiego

Kolejnym źródłem danych o popycie na pracę jest projekt Bilansu Kapitału Ludzkiego (BKL). Badanie przeprowadzone zostało przez Polską Agencję Rozwoju Przedsiębiorczości (PARP) wraz z Centrum Ewaluacji i Analiz Polityk Publicznych Uniwersytetu Jagielońskiego (Antosz, 2014).

Głównym celem badania było rozpoznanie zasobów kompetencyjnych na polskim rynku pracy. Zebrane dane pozwalają określić jakich pracowników poszukują pracodawcy dodający ogłoszenia oraz jakie kompetencje posiadają osoby poszukujące pracy. Badanie umożliwia więc wskazanie potencjalnych luk kompetencyjnych w różnych branżach, zawodach czy też regionach (Antosz, 2014).

Projekt Bilansu Kapitału Ludzkiego realizowany był w latach 2010-2014 i przez 5 lat dostarczył wiele cennych informacji na temat sytuacji na polskim rynku pracy. Dla przejrzystości projektu, został on podzielony na 4 różne moduły, które skupiały się następująco na:

- badaniu pracodawców,
- badaniu ofert pracy,
- badaniu ludności,
- badaniu instytucji szkoleniowych.

Każde z przeprowadzonych badań dostarcza informacje przedstawiające sytuacje na polskim rynku pracy. Jednak w kolejnych rozdziałach szczególna uwaga zostanie zwrócona na dwa pierwsze moduły Badania Kapitału Ludzkiego.Jednym z wspomnianych wyżej modułów jest Badanie Ofert Pracy.

Badanie ofert pracy polegało na zbieraniu danych na temat ogłoszeń pochodzących z różnych źródeł. Początkowo do tego celu wylosowano 160 urzędów pracy. Pierwotne założenia projektu zakładały, iż oferty pobierane będą z Centralnej Bazy Ofert Pracy (CBOP), wybranych stron urzędów oraz portalu internetowego Careerjet.pl. Portal Careerjet.pl. to ogólnokrajowy portal internetowy umożliwiający wyszukiwanie różnych ofert pracy. Korzystając z tego źródła ankieterzy dokonywali zrzutów ekranu każdej oferty a następnie wprowadzali wybrane części oferty do gotowej formatki (Antosz, 2014).

Poniżej przedstawiony został rozkład wykorzystania źródeł danych w poszczególnych edycjach badania.

Na podstawie powyższej tabeli można zauważyć, iż struktura źródeł danych wykorzystywanych w badaniu ulegała zmianie. Na przełomie lat 2010 - 2012 około 10 % danych wykorzystywanych w badaniu pochodziła z powiatowych urzędów pracy. Natomiast w III i IV edycji dane z tego źródła stanowiły jednie 3% wielkości badania. Warto jednak zauważyć, iż we wspominanym okresie wzrósł odsetek danych pochodzących z bazy CBOP. Może to świadczyć o coraz

Tabela 1.1. Wielkość próby w poszczególnych badaniach oraz jakość kodowania

Rok	2010	2011	2012	2013	2014
Dzień	10 września	28 marca	26 marca	25 marca	28 marca
PUP	N/A	2 012	2 812	382	696
CBOP	8 198	5 004	4 440	5 232	7 846
Careerjet.pl	11 811	13 618	14 342	14 467	12 914
Ogółem	20 009	20 634	21 594	20 081	21 456
Kodowanie	0.72	0.89	0.96	0.96	0.963

Źródło: Opracowanie własne na podstawie Badania Kapitału Ludzkiego 2010-2014.

lepszym działaniu tej bazy. Co ciekawe ilość danych pochodzących z portalu Careerjet.pl z roku na rok zachowywała trend rosnący, a w 2013 roku stanowiła nawet 72% całego badania. Jednak w V edycji przeprowadzanej w 2014 roku, wykorzystanie tego źródła stanowiło już 60% wielkości próby. Może to być skutkiem większej ilość ogłoszeń publikowanych w CBOP (wykorzystanie tego źródła wzrosło aż o 11 % w porównaniu do roku 2013) lub mniejszej puli ofert zamieszczanych na portalu Carrerjet.pl, który mógł stracić swoją popularność na rzecz innych portali, m.in. OLX lub Pracuj.pl.

Kolejnym badaniem szerzej wykorzystywanym w dalszej części pracy jest badanie przedsiębiorców. Badanie to polegało na pozyskaniu informacji od podmiotów gospodarczych zatrudniających przynajmniej jednego pracownika i funkcjonujących w momencie prowadzenia badania (Antosz, 2014). Każdy z przedsiębiorców biorący udział w badaniu, miał możliwość uczestnictwa w wywiadzie który dostarczał informacji m.in. na temat:

- miejsc umieszczania ogłoszeń,
- wielkości firmy,
- rodzaju placówki,
- branży do której kierowane były ogłoszenia pracy.

Problemem napotkanym podczas pozyskiwania informacji od przedsiębiorców okazała się niechęć respondentów do uczestnictwa w badaniu. W 2014 roku aż 68% zaplanowanych wywiadów nie zostało zrealizowanych. Do głównych powodów ich niezrealizowania zaliczone zostały:

- odmowa (30%),

- nieobecność w trakcie realizacji badania (15%),
- nieistniejący numer (7%),
- badanie w firmie już zrealizowano (1%),
- inne przyczyny (15%).

Jednak wielkość próby sprawiła, iż mimo licznych odmów, wywiady które udało się zrealizować wciąż dostarczają wielu cennych informacji na temat przedsiębiorców zamieszczających ogłoszenia w różnych źródłach przekazu. Dane pochodzące z badania pozwalają więc przeprowadzić segmentacje pracodawców ze względu na wybór miejsc zamieszczania ogłoszeń.

Bilansu Kapitału Ludzkiego nie można w sposób jednoznaczny przyporządkować do źródeł statystycznych lub niestatystycznych. BKL jest na pograniczu definicji obu źródeł danych, z jednej strony jego celem jest zbadanie rynku, natomiast z drugiej zebrane dane nie są standardowe - oferty pracy zbierane z portali internetowych. Jednak ze względu, iż Bilans Kapitału Ludzkiego jest przykładem statystycznego wykorzystania danych niestatystycznych, w tej pracy został zaklasyfikowany jako źródło statystyczne.

1.2.2.3 Powiatowe Urzędy Pracy i Centralna Baza Ofert Pracy

Szczególną uwagę należy zwrócić na portal CBOP, który wcześniej został wymieniony jako jedno z niestatystycznych źródeł danych na temat rynku pracy. CBOP jest wspólną bazą urzędów pracy, w której przechowywane są informacje na temat aktualnych ofert pracy, praktyk oraz staży. Baza ta prowadzona i udostępniana jest przez Ministerstwo Rodziny, Pracy i Polityki Społecznej. Dużą zaleta portalu jest możliwość zamieszczania ogłoszeń za pośrednictwem e-formularza. Podczas pobierania danych z CBOP, ankieterzy musieli zmierzyć się z problemem licznych braków oraz małej ilości aktualnych ogłoszeń. Zdecydowali się więc na dodatkowe pozyskiwanie danych poprzez bezpośredni kontakt z wybranymi urzędami pracy. Ankieterzy kontaktowali się urzędami telefonicznie bądź pojawiali się w nich osobiście. Jednym z powodów wykorzystania portalu CBOP w badaniu jest fakt, iż zamieszczane tam ogłoszenia charakteryzują się jednakowym formatem i opisem bardziej szczegółowym niżeli informacje dostępne na stronach internetowych poszczególnych urzędów pracy (Antosz, 2014; Ministerstwo Rodziny, Pracy i Polityki Społecznej, 2019).

Urzędy Pracy to instytucje zajmujące się badaniem oraz analizowaniem rynku pracy, a ponadto pośrednictwem zawodowym i udzielaniem informacji osobom bezrobotnym. W Polsce

funkcjonują zarówno powiatowe, jak i wojewódzkie urzędy pracy. Jednak pod względem podległości urzędy te nie są ze sobą powiązane oraz realizują różny zakres zadań (Wikipedia, 2019).

Jeśli pracodawca chciałby zgłosić ofertę pracy do urzędu pracy powinien wybrać jeden urząd ze względu na siedzibę pracodawcy i miejscy wykonywania pracy bądź nie. Następnie powinien wypełnić formularz zgłoszeniowy, w którym znajdują się między innymi dane identyfikacyjne i teleadresowe pracodawcy, rodzaj działalności, miejsce wykonywania pracy czy dane dotyczące wykonywanej pracy. Pracodawca musi również podać informację czy w ciągu ostatnich 365 dni przed złożeniem oferty był karany - urząd pracy nie może przyjąć oferty od pracodawcy, który był karany ani oferty, w której pracodawca zawarł wymagania naruszające zasadę równego traktowania w zatrudnieniu (Publiczne Służby Zatrudnienia, 2019).

1.2.2.4 Internetowe źródła danych

Wykorzystanie portali internetowych jako źródła danych dotyczących sytuacji na polskim rynku pracy i zestawienie ich z ofertami zamieszczanymi w CBOP bądź PUP otworzyło również wiele możliwości analizowania Badania Ofert Pracy oraz Badania Pracodawców pod kątem miejsca zamieszczania poszczególnych ogłoszeń.

W literaturze przedmiotu można przeczytać o następujących zaletach badania ofert pracy bazując na Internecie (Pater, 2017):

- to pracodawca jest zainteresowany umieszczaniem w ogłoszeniu swoich wymagań, co daje przewagę nad badaniami ankietowymi,
- dowolność przekrojów,
- możliwość przechowywania danych i porównywanie ich z sobą w czasie,
- niskie koszty, zwłaszcza w porównaniu z badaniami ankietowymi.

Jednakże należy wskazać również minusy:

- pobieranie ofert pracy może być określone przez badaczy. Istnieje więc możliwość że nie zostaną uwzględnione wszystkie oferty pracy publikowane w danych miesiącu. Spowodowane może to być zróżnicowaną i nieznaną długością istnienia danego ogłoszenia w Internecie. Oczywiście odpowiedzią może być tutaj inna częstotliwość z czytywania danych,

- ten sposób badania nie uwzględnia przypadków w których pracodawca szuka pracowników z wykorzystaniem innych metod niż ogłoszenia w Internecie,
- zdarza się, że ogłoszenia zawierają kilka ofert pracy, w tym przypadku zakłóca to poprawność obliczeń.

Przykładowymi portalami, które mogą dostarczać danych dotyczących popytu na pracę są: olx.pl, pracuj.pl, goldenline.pl czy linkedin.com, z czego najbardziej popularnym portalem tego typu w Polsce jest olx.pl.

Kolejnym źródłem danych jest Barometr Ofert Pracy (BOP), który powstaje przy współpracy Katedry Makroekonomii Wyższej Szkoły Informatyki i Zarządzania w Rzeszowie oraz Biura Inwestycji i Cykli Ekonomicznych. BOP bazuje na danych o unikalnych w skali kraju ofertach pracy. Obliczany jest on w formie indeksu, w punktach procentowych, w którym podstawą jest średnia z 2010 roku. Indeks ten podawany jest zarówno dla kraju jak i poszczególnych województw. Początkowo, opierał on się na ofertach prasowych zamieszczanych w Gazecie Wyborczej. Od 2008 roku zainteresowanie badaczy zostało skierowane na internetowe źródła danych.

Dane o ofertach agregowane są do miesiąca, samo zbieranie danych następuje 20. dnia każdego miesiąca. Pozbawiane są one czynnika sezonowego, zapewnia im to porównywalność pomiędzy konkretnymi porami roku, ponadto eliminuje wpływ zmian ofert prac sezonowych, czyli części ogłoszeń dotyczących pracy tymczasowej. Następnie badacze sprawdzają oferty które się powtarzają i biorą pod uwagę tylko unikalne ogłoszenia w skali kraju. Biorąc pod uwagę zarówno zmiany bezrobocia, jak i Barometr Ofert Pracy istnieje możliwość uzyskać informacje w jakiej fazie znajduje się aktualnie rynek ofert pracy. Jak można przeczytać na oficjalnej stronie Biura Inwestycji i Cykli Ekonomicznych, BOP często posiada właściwości wyprzedzające w porównaniu do zmian zatrudnienia i bezrobocia (Biuro Inwestycji i Cykli Ekonomicznych, 2019).

O ile zastosowanie statystycznych źródeł danych jest powszechnie uznane, to właśnie niestatystyczne źródła danych będą obiektem szczególnego zainteresowania w tej pracy. Dane pochodzące z innych źródeł, zwłaszcza internetowych, mogą być odpowiedzią na rosnące oczekiwania i potrzeby społeczeństwa XXI wieku. Badania prowadzone przez jednostki administracyjne często ograniczają się do konkretnego, im wyznaczonego regionu. Opierając się na statystykach Głównego Urzędu Statystycznego, w 2017 roku przynajmniej jeden komputer w gospodarstwie domowym miało prawie 82%, a dostęp do internetu szerokopasmowego to 78%. Z tych gospodarstw domowych, które nie posiadają dostępu do Internetu, aż dwie trzecie wskazało brak potrzeby korzystania z niego jako przyczynę (Główny Urząd Statystyczny, 2017).

Szerszy opis korzystania ze źródeł internetowych znajdzie się w rozdziale 2, który poświęcony jest reprezentatywności tychże danych.

1.3 Porównanie wybranych cech badanych źródeł danych

Tabela 1.2 przedstawia porównanie wymienionych wcześniej źródeł danych. Skupia się ona na 4 wspomnianych wcześniej źródłach danych. Są nimi: Popyt na pracę, BKL, CBOP oraz źródła internetowe.

Po pierwsze, cel powstania poszczególnych źródeł różni się od siebie. Popyt na pracę oraz BKL powstały w celu badania rynku pracy. Dane z Powiatowych Urzędów Pracy powstają na wskutek pośrednictwa jakim te placówki się zajmują, ostatecznie dane Internetowe funkcję badawczą pełnią dopiero w drugiej kolejności, ponieważ oryginalnie służą do pośrednictwa na rynku pracy, będącego często źródłem przychodu dla przedsiębiorstwa.

Populacją w przypadku badania Popytu na pracę są podmioty gospodarki narodowej, jednakże tylko te których liczba zatrudnionych to 1 lub więcej. BKL łączy ze sobą zarówno dane pochodzące z Powiatowych Urzędów Pracy jak i dane internetowe. Ostatecznie PUPy bazują na ofertach pracy do nich zgłaszanych. Dane internetowe z kolei to pobrane z wybranych stron internetowych oferty pracy. Jednostką badania są za każdym razem pojedyncze elementy badanej zbiorowości, czyli przedsiębiorstwa.

Aspektem mocno różnicującym badane źródła jest ich koszt. W przypadku badania Popytu na pracę oraz BKL jest on stosunkowo wysoki, zwłaszcza jeżeli porównać go to kosztów PUP oraz internetowych źródeł danych. Równie mocno zróżnicowany jest rozmiar danych w zależności od pochodzenia.

Badania takie jak POP czy PUP operują na stosunkowo niedużych wielkościach danych, które są dodatkowo ustrukturyzowane. Internetowe źródła danych natomiast, bazują na dużych wolumenach informacji, często nieustrukturyzwanych. Jednakże dużą zaletą Internetowych źródeł danych jest ich terminowość, w przeciwieństwie do pozostałych, pozyskiwanie danych internetowych nie wiążę się z dużym opóźnieniem.

Dane reprezentatywne powinny odzwierciedlać populację i jej cechy. Próba powinna być zaprojektowana w taki sposób, aby była ”miniaturą populacji”. Temat reprezentatywności zostanie szerzej opisany w kolejnych podrozdziałach. Badanie Popytu na Pracę oraz Bilans Kapitału Ludzkiego to badania zaprojektowane, gdzie dane to próba losowa - z tego względu będą one

Tabela 1.2. Porównanie wybranych źródeł danych o popycie na pracę

Obszar porównania	Popyt na Pracę	BKL	CBOP/PUP	Internet
Cel powstania	Badanie rynku pracy	Badanie rynku pracy	Badanie i pośrednictwo na rynku pracy	Komercyjne pośrednictwo na rynku pracy
Populacja	Podmioty gospodarki narodowej o liczbie zatrudnionych osób 1 i więcej	Powiatowe Urzędy Pracy oraz źródła internetowe	Oferty zgłoszone do Powiatowych Urzędów Pracy	Dane pobrane z stron internetowych
Jednostka	Pojedynczy element zbiorowości, przedsiębiorstwo	Pojedynczy element zbiorowości: Przedsiębiorstwo	Pojedynczy element zbiorowości: Przedsiębiorstwo	Pojedynczy element zbiorowości: Przedsiębiorstwo
Sposób uzyskania	zbierane od respondentów	pochodzące z portalu CBOP i Careerjet.pl oraz od respondentów (przedsiębiorcy, ankiety)	baza urzędów pracy zbierająca oferty pracy	pobierane automatycznie
Koszt Badania	1 751 798 zł	Wysoki	Potencjalnie niski	Potencjalnie niski
Rozmiar	małe/średnie zbiory danych	duże zbiory danych	średnie/duże zbiory danych	duże zbiory danych
Dane statystyczne	Tak	Tak	Nie	Nie
Terminowość	Powstają z opóźnieniem	Powstają z opóźnieniem	Powstają z opóźnieniem	Brak opóźnień
Strukturyzowanie	Tak	Tak	Tak	Nie
Reprezentatywność	próba reprezentatywna	próba reprezentatywna	dane nierepreprezentatywne	dane nierepreprezentatywne

Źródło: Opracowanie własne na podstawie raportów poświęconych badaniom: Popyt na Pracę, Bilans Kapitału Ludzkiego oraz Centralnej Bazy Ofert Pracy.

reprezentatywne, natomiast oferty pobrane z portalu internetowego czy dane pochodzące z Internetu nie będą odzwierciedlać populacji, ponieważ są to dane, które powstają ”dodatkowo”, nie są tworzone w celu zbadania danego zjawiska.

Terminowość danych mówi o tym, czy powstają one z opóźnieniem czy uzupełniane są na bieżąco. W przypadku zbiorów danych przedstawionych w pracy jedynie dane internetowe powstają bez opóźnień, ponieważ są to dane, które pobierane są automatycznie i na bieżąco.

Dane internetowe mogą być publikowane oraz analizowane z większą częstotliwością, niż dane z planowanych badań, takich jak na przykład Badanie Popytu na Pracę, które jest badaniem kwartalnym. W badaniu tym najpierw zbierane są informacje dla danego kwartału i dopiero po tym czasie można uzyskać pełny zbiór danych za określony kwartał.

1.4 Podsumowanie

W powyższym rozdziale podjęto próbę usystematyzowania wiedzy z zakresu źródeł o popycie na pracę. Skupiono się na omówieniu definicji, cech wspólnych oraz czynników różniących poszczególne źródła. Aby określić na ile źródła poza statystyczne mogą zostać wykorzystane na potrzeby opisu rynku pracy istotne jest określenie ich reprezentatywności. W kolejnym rozdziale zostanie omówiona kwestia pokrycia, definicji reprezentatywności oraz wskazane główne charakterystyki przedsiębiorców wpływające na wybór określonych kanałów pozyskiwania pracodawców.

Rozdział 2

Reprezentatywność internetowych źródeł danych w świetle modelu Item Response Theory (Magdalena Maślak)

2.1 Cel rozdziału

Ciągły rozwój technologii i cyfryzacji sprawia, że niektóre klasyczne metody rozwiązywania problemów odchodzą w zapomnienie. Internet odgrywa coraz większą role w życiu człowieka, ponieważ umożliwia szybkie i wygodne rozwiązania. Pozyskiwanie danych dotyczących oferty pracy na polskim rynku za pomocą Internetu jest również wygodne i tanie. Nie wymaga ono bowiem zaangażowania dużej ilości ankieterów lub przepisywania wielu formularzy. W rozdziale drugim poruszony zostanie temat reprezentatywności internetowych źródeł danych o polskim rynku pracy. Ogromna liczba ofert oraz ich szeroka grupa odbiorców sprawia, że z pewnością jest to jedno z ważnych źródeł informacji na temat rynku pracy. Warto jednak zastanowić się czy to oznacza, że wraz z rozwojem technologii i innowacji, urzędy pracy i inne klasyczne formy poszukiwania pracowników faktycznie zaczęły wygasać na rzecz Internetu. Budowa i analiza teorii odpowiadania na pozycje testowe, znanego pod angielskim pojęciem *Item Reponse Theory* (IRT)[1] pozwoli scharakteryzować przedsiębiorstwa zamieszczające ogłoszenia w Internecie na podstawie Badania Kapitału Ludzkiego 2010-2014, w szczególności modułowi poświęconemu pracodawcom. Charakterystyka podmiotów zamieszczających ogłoszenia w Internecie umoż-

[1] W pracy korzystać będziemy z pojęcia angielskiego oraz skrótu IRT, które w dokładniejszy sposób oddaje specyfikę wykorzystanych modeli.

liwi ocenę reprezentatywności tego źródła.

2.2 Problematyka reprezentatywności

2.2.1 Definicje reprezentatywności

Każdego dnia na świecie generowana jest ogromna ilość danych. W ciągu ostatnich dwóch lat wyprodukowanych zostało więcej danych niżeli od początku istnienia ludzkości (Forbes, 2017). Niektóre z nich przechowywane są w różnego rodzaju rejestrach, archiwach, aplikacjach lub na stronach internetowych. Z biegiem czasu zwiększyła się więc świadomość społeczeństwa dotycząca korzyści z efektywnej analizy produkowanych informacji.

Rosnąca ilość generowanych danych udostępnia szeroki wachlarz źródeł z których można czerpać informacje dotyczące badanych populacji. Jednak podczas próby przeprowadzenia konkretnych badań decydującym czynnikiem wpływającym na poprawność otrzymanych wyników jest jakość danych. Warto więc zastanowić się z jakich źródeł informacji należy korzystać aby przeprowadzona analiza była wiarygodna.

Pojęciem wykorzystywanym do oceny jakości badania jest reprezentatywność. Próby reprezentatywne to takie które pozwalają na możliwie najtrafniejsze zbadanie cechy ogółu populacji. Reprezentatywność można więc zdefiniować jako pewną ”miniaturkę”większej grupy, której dotyczy badanie. Jest to pewnego rodzaju reprezentacja danych.

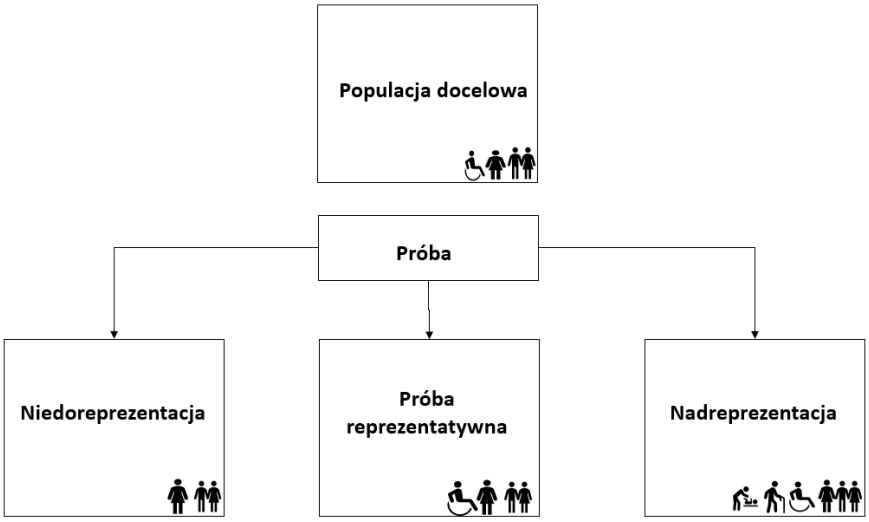
Reprezentatywności jest pojęciem szeroko dyskutowanym w środowisku badawczym, ponieważ często uznawany jest za niejasny termin, który należy sprecyzować.

Nie ma więc jednej prostej definicji tego terminu badawczego. Trudność stworzenia jednej ogólnej definicji, która zawierałaby wszystkie interpretacje tego słowa, zainspirowała Kruskala i Mostellera w 1979 do stworzenia listy pojęć, pojawiających się podczas próby definiowania terminu reprezentatywności. Według Kruskala i Mostellera przez reprezentatywność można rozumieć (Beręsewicz, 2016):

- ogólne, nieuzasadnione określenie dla danych,
- brak nielosowych mechanizmów selekcji (np. autoselekcji),
- miniaturę populacji,
- typowe lub idealne przypadki,

- pokrycie populacji,
- niejasny termin, który należy sprecyzować,
- określony sposób (losowego) doboru próby,
- mechanizm pozwalający na nieobciążone szacunki,
- wystarczające do określonego celu.

Reprezentatywność można więc rozumieć na wiele różnych sposobów. Jednak każde z tych pojęć dotyczy trafności doboru próby badawczej. Za badanie reprezentatywne można więc uznać taką próbę, która w najlepszy sposób zobrazuje cechy wybranej populacji. Pokrycie populacji nie jest jednak łatwym zadaniem. Podczas przeprowadzania różnego rodzaju badań, naukowcy muszą brać pod uwagę wystąpienia zjawiska „nadreprezentacji" (ang. *overcoverage*) lub „niedoreprezentacji" (ang. *undercoverage*).



Rysunek 2.1. Problem nadreprezentacji i niedoreprezentacji w badaniach statystycznych

Źródło: Opracowanie własne.

Problem niedoreprezentacji polega na nieuwzględnieniu lub pominięciu jednostek (lub całych subpopulacji) badanej populacji. Przykładowo badając rynek pracy można opierać się na oficjalnych danych dotyczących bezrobocia na podstawie listy osób zarejestrowanych w urzędach pracy jako bezrobotni. Niestety pominięte zostają wtedy osoby, które nie pracują i nie zarejestrowały się jeszcze w urzędzie pracy lub po prostu tej pracy nie szukają.

W dalszej części rozdziału rozwinięte zostaną rozważania dotyczące reprezentatywności internetowych źródeł danych w ocenie sytuacji panującej na rynku pracy.

2.3 Teoretyczne podstawy modelu Item Response Theory

Badanie wyboru miejsca zamieszczania ogłoszeń jest złożonym procesem. Nie polega on bowiem jedynie na analizie zamieszczonych ofert pracy. Aby zrozumieć mechanizmy kierujące pracodawcami podczas poszukiwania pracowników, warto rozpatrywać cały proces trochę szerzej.

Ocena zachowań behawioralnych często związana jest z próbą zagłębienia się w sposób myślenia respondenta. Jest to związane z koniecznością wykorzystania narzędzi umożliwiających pomiar psychologiczny. Analizując proces doboru źródła zamieszczania ogłoszeń przez wybrane podmioty gospodarcze warto zastanowić się czy istnieje relacja między decyzją wyboru źródła a teoretycznymi czynnikami, które na nią wpływają (Kleka, 2012). Poniżej przedstawiona została rycina ukazująca proces doboru miejsca zamieszczenia ogłoszeń. Na podstawie poniższego wykresu można zauważyć, iż na decyzje pracodawcy może wpływać wiele różnych czynników.

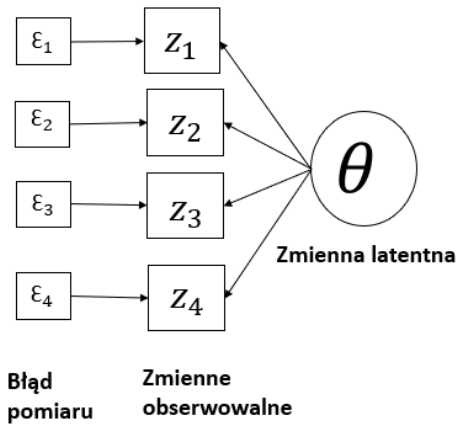
Rysunek 2.2. Czynniki wpływające na wybór źródła zamieszczenia ogłoszenia

Źródło: Opracowanie własne.

Wpływ na wybór miejsca zamieszczania ogłoszeń może mieć zarówno wielkość firmy jak i branża której dotyczy ogłoszenie. Są to tzw. zmienne obserwowalne, których wartość odczytać można na podstawie odpowiedzi w kwestionariuszach. Warto jednak zastanowić się jaki wpływ na decyzje mają skłonności pracodawców do zamieszczania ogłoszeń. Wartość takiej zmiennej nie jest łatwa do zdefiniowania, ponieważ nie ma jednej miary określającej skłonności osób do poszczególnych wyborów. Można ją jednak wyrazić poprzez inne wskaźniki, które określają

pośrednio skłonności osób do rożnych decyzji. Zmienne składające się z grupy innych czynników nazywane są zmiennymi ukrytymi, inaczej latentnymi (od angielskiego pojęcia *latent*). W pracy używane będą te dwa pojęcia wymiennie.

Zmienne latentne są to zmienne, które posiadają znacznie empiryczne jedynie poprzez obserwowalne konsekwencje tych zmiennych. Obserwowalne konsekwencje zmiennych ukrytych nazywamy wskaźnikami, pozycjami testowymi, lub zadaniami testowymi (Konarski, 2004b). Na rysunku 2.3 za pomocą wykresu ścieżkowego przedstawiona została relacja między zmienną ukrytą a jej wskaźnikami.



Rysunek 2.3. Wykres ścieżkowy relacji między zmienną ukrytą a jej wskaźnikami

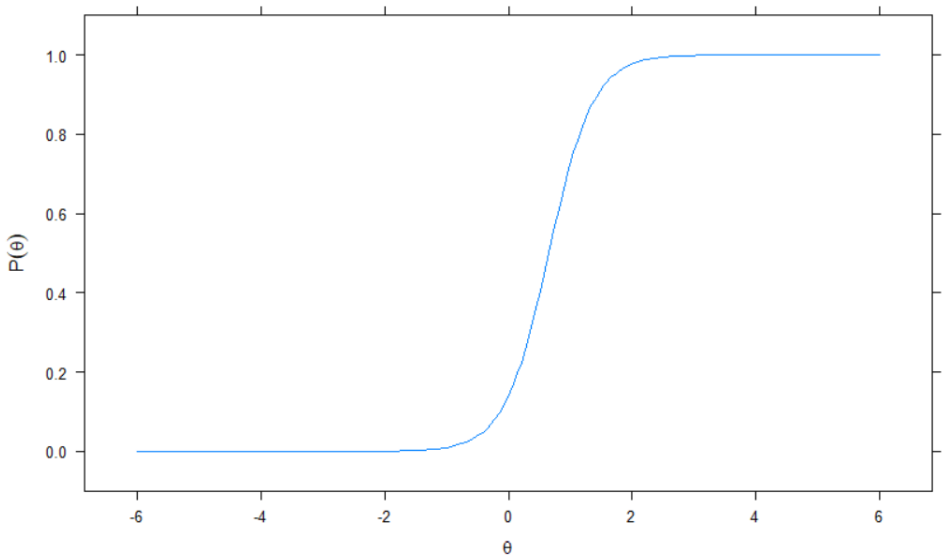
Źródło: Opracowanie własne na podstawie (Konarski, 2004b)

Modelami, które zajmują się badaniem wpływu nieobserwowalnych cech na konkretne odpowiedzi w testach są modele IRT. Już w pierwszej połowie XX wieku Derrick N. Lawley postulował, iż wynik w testach spowodowany jest wartością nieobserwowalnej cechy i każda pozycja testowa jest z nią związana w niepowtarzalny sposób (Kleka, 2012).

Ideą budowania modeli IRT było stworzenie modelu statystycznego, który określa rozkład odpowiedzi na konkretną pozycje testową w ramach pewnej zmiennej latentnej, reprezentującej poziom mierzonej cechy (Brzezińska, 2016).

Jednym z głównych form prezentujących rozkład badanej cechy jest ilustracja modelu za pomocą krzywych charakterystycznych (ang. *Item Characteristic Curve*; ICC). Krzywe obrazują zależności pomiędzy prawdopodobieństwem udzielenia poprawnej odpowiedzi, a rożnymi wartościami zmiennej ukrytej. Rozkłady tworzone są podczas modelowania prawdopodobieństwa

udzielenia konkretnej odpowiedzi z wykorzystaniem modelu logistycznego (Brzezińska, 2016). Poniżej przedstawiony został wykres 2.4 obrazujący rozkład przykładowej zmiennej ukrytej.



Rysunek 2.4. Rozkład prawdopodobieństwa udzielenia poprawnej odpowiedzi na wybraną pozycję testową w zależności od poziomu wartości zmiennej ukrytej.

Źródło: Opracowanie własne.

Wykres krzywej charakterystycznej obrazuje w jaki sposób zmienia się prawdopodobieństwo udzielania odpowiedzi w zależności od wartości zmiennej latentnej, która w modelach IRT często nazywana jest ”zdolnością” (ang. *ability*). Stromość i kształt omawianych krzywych często zależy od rodzaju modeli IRT, które posiadają różne założenia np. w stosunku do poziomu dyskryminacji poszczególnych pozycji testowych.

W literaturze naukowej wyróżnia się kilka typów modeli IRT. Modele te różnią się głównie funkcją matematyczną, która wyjaśnia prawdopodobieństwo uzyskania poprawnej odpowiedzi (Konarski, 2004a). Do najbardziej znanych i wykorzystywanych modeli pochodzących z rodziny IRT zaliczają się:

- Jednoparametryczny model logistyczny (1PLM) – równoznaczny z modelem Rascha,
- Dwuparametryczny model logistyczny (2PLM),
- Trójparametryczny model Birnbauma (3PLM).

Jednoparametryczny model logistyczny (model Rascha) jest modelem przedstawiającym rozkład prawdopodobieństwa zmiennej latentnej przy wykorzystaniu informacji o poziomie jej

trudności. Warto również wspomnieć, że model Rascha zakłada, iż poziom dyskryminacji każdej zmiennej jest równy 1. Poniżej przedstawione zostało równanie opisujące omawiany model:

$$P(\theta) = \frac{1}{1 + e^{-(\theta - b)}} \tag{2.1}$$

gdzie b jest parametrem stopnia trudności natomiast θ określa poziom zmiennej ukrytej. W modelu Rascha wystepuje również parametr dyskryminacji oznaczany literą a, który zawsze przyjmuje wartość jeden.

Aby lepiej zrozumieć założenia modelu oraz wykorzystywane w nim oznaczenia warto wyjaśnić znaczenie merytoryczne poszczególnych zmiennych. Jak już zostało wyżej wspomniane model Rascha zakłada, iż poziom dyskryminacji każdej zmiennej jest równy 1. Parametr dyskryminacji wpływa na stromość krzywej ICC. Im wyższa jest jego wartość, tym silniejsze nachylenie krzywej w punkcie przegięcia. Warto również zauważyć, iż większa stromość krzywej charakterystycznej w wybranym punkcie zmiennej latentej, poprawia zdolność pozycji testowej do rozróżniania poziomu badanej cechy wśród respondentów znajdujących się po dwóch stronach tego punktu (Kondratek & Pokropek, 2013).

Parametr dyskryminacji pokazuje więc zdolności pozycji testowej do rozróżniania poziomu badanej cechy (Kondratek & Pokropek, 2013). W założeniach modelu Rascha poziom dyskryminacji każdego pytania jest jednakowy i przyjmuje wartość 1. Skutkuje to tym, że wszystkie krzywe ICC są względem siebie równoległe. Takie podejście posiada zarówno wady jak i zalety. Ustalenie równego poziomu dyskryminacji dla każdej zmiennej może usztywnić model, powodując przy tym gorsze jego dopasowanie do danych. Z drugiej strony omawiany model Rascha wyróżnia się wieloma przydatnymi właściwościami matematycznymi m.in. jego wyniki sumaryczne w teście są wystarczające do oszacowania poziomu zmiennej latentnej. W przypadku pozostałych modeli, aby oszacować poziom zmiennej ukrytej, potrzebne są wartości całego wektora odpowiedzi (Kondratek & Pokropek, 2013).

Kolejnym parametrem wykorzystywanym w rodzinie modeli IRT jest parametr poziomu trudności, w równaniach oznaczany zmienną b. Jest to punkt na skali wartości zmiennej latentej, w którym prawdopodobieństwo uzyskania poprawnej odpowiedzi wynosi 50%. Teoretycznie zakłada się, że wartości tej zmiennej pochodzą z przedziału $(-\infty \leq b \leq \infty)$. Jednak typowe wartości omawianej zmiennej mieszczą się w przedziale $(-3 \leq b \leq 3)$ (Baker & Kim, 2017).

Następnym modelem z rodziny IRT jest dwuparametryczny model logistyczny (2PL). Jest to

przykład modelu, który w przeciwieństwie do modelu Rascha uwzględnia różne poziomy dyskryminacji zmiennych. Oznacza to, że w tym modelu wpływ na prawdopodobieństwo uzyskania właściwej odpowiedzi na konkretne pytanie testowe zależy zarówno od poziomy trudności pytania jaki i parametru dyskryminacji. Poniżej przedstawione zostało równanie dwuparametrycznego modelu logistycznego :

$$P(\theta) = \frac{1}{1 + e^{-a(\theta - b)}} \tag{2.2}$$

gdzie a jest parametrem dyskryminacji, b jest parametrem stopnia trudności, a θ określa poziom zmiennej ukrytej.

Można więc zauważyć, iż dwuparametryczny model logistyczny jest pewnego rodzaju uszczegółowionym modelem Rascha. Uwzględnia on bowiem więcej zmiennych podczas szacowania prawdopodobieństwa udzielenia konkretnej odpowiedzi.

Do rodziny modeli IRT należy również trójparametryczny model Birnbauma (3PL), który w swoim równaniu dodatkowo uwzględnia wpływ zjawiska zgadywania na prawdopodobieństwo uzyskania trafnej odpowiedzi w teście. Faktem jest, iż istnieje wiele testów z pytaniami, na które osoba wypełniająca zupełnie przez przypadek może udzielić poprawnej odpowiedzi. Przykładem takich testów są testy jednokrotnego wyboru. Respondent posiadający cztery możliwe odpowiedzi do wyboru, już na wstępie uzyskuje 25% szans udzielenia poprawnej odpowiedzi. Trójparametryczny model Birnbauma uwzględnia fakt, iż na niektóre pozycje testowe wpływać może również szczęście respondenta. Poniżej przedstawiony został model, który uszczegóławia pozostałe modele IRT o parametr zgadywania:

$$P(\theta) = c + (1 - c)\frac{1}{1 + e^{-a(\theta - b)}} \tag{2.3}$$

gdzie a jest parametrem dyskryminacji, b jest parametrem stopnia trudności, c jest parametrem zgadywania i θ określa poziom zmiennej ukrytej.

Parametr c określa prawdopodobieństwo udzielenia poprawnej odpowiedzi w teście tylko i wyłącznie dzięki zjawisku zgadywania. Jego teoretyczne wartości mieszą się w przedziale $(0 \leq c \leq 1)$. Jednak w praktyce wartości powyżej 0,35 są nieakceptowalne (Baker & Kim, 2017).

Modele IRT wykorzystywane są do badań w wielu dziedzinach. Swoje zastosowanie mają zarówno w badaniach psychologicznych jak i marketingowych. Dodatkowo modele IRT mogą być wykorzystywane do badań medycznych lub różnych analiz testów edukacyjnych. Wachlarz

zastosowań rodziny modeli IRT jest więc ogromnym. W dalszej części rozdziału modele IRT posłużą jako narzędzie do analizy ankiety wypełnianej przez pracodawców biorących udział w Badaniu Przedsiębiorców. Motywacja do wykorzystania modelu IRT wynika z konstrukcji kwestionariusza pracodawców w Badaniu Kapitału Ludzkiego, w którym pracodawcy mogli określać kanały wykorzystywane do poszukiwania pracowników. Pracodawcy mogli wskazać więcej niż jedną odpowiedź i należy spodziewać się, że korzystanie z poszczególnych kanałów będzie ze sobą skorelowanie. W takich przypadkach model IRT pozwala na określenie zmiennej latentnej odnoszącą się do skłonności pracodawcy do wykorzystywania danego kanału, a otrzymane wyniki będą cennym źródłem wiedzy o pracodawcach poszukujących nowych pracowników.

2.4 Badanie mechanizmu selekcji z wykorzystaniem modelu Item Response Theory

2.4.1 Opis źródła informacji o podmiotach gospodarczych zamieszczających ogłoszenia w Internecie

W ramach badania BKL, pracodawcy biorący udział w badaniu mieli okazje wypełnić ankietę, która gromadziła informacje o przedsiębiorstwie i samym respondencie. Ankieta zawierała pytania dotyczące m.in. wielkości badanego przedsiębiorstwa, rodzaju placówki lub branży. Jednym z ważnych pytań okazała się również pozycja testowa dotycząca miejsca zamieszczania ogłoszeń (pytanie wielokrotnego wyboru). Treść tego pytania brzmiała następująco ”Pracowników do pracy można poszukwiać na różne sposoby. Czy szukał Pan(i) ich do swojej firmy poprzez...”:

Respondent jako odpowiedz miał do wyboru takie źródła jak

1. Państwowe Urzędy Pracy;
2. Prywatne biura pośrednictwa pracy;
3. Head hunter;
4. Szkoła i akademickie ośrodki kariery;
5. Prasa;
6. Internet;

7. Ogłoszenia rozwieszane w obrębie firmy;

8. Polecenie rodziny i znajomych;

9. Targi pracy.

Pracodawca biorący udział w ankiecie mógł wskazać dowolną liczbę źródeł. Możliwe jest bowiem, że osoby poszukujące pracowników nie ograniczały się do jednego miejsca zamieszczania ogłoszeń.

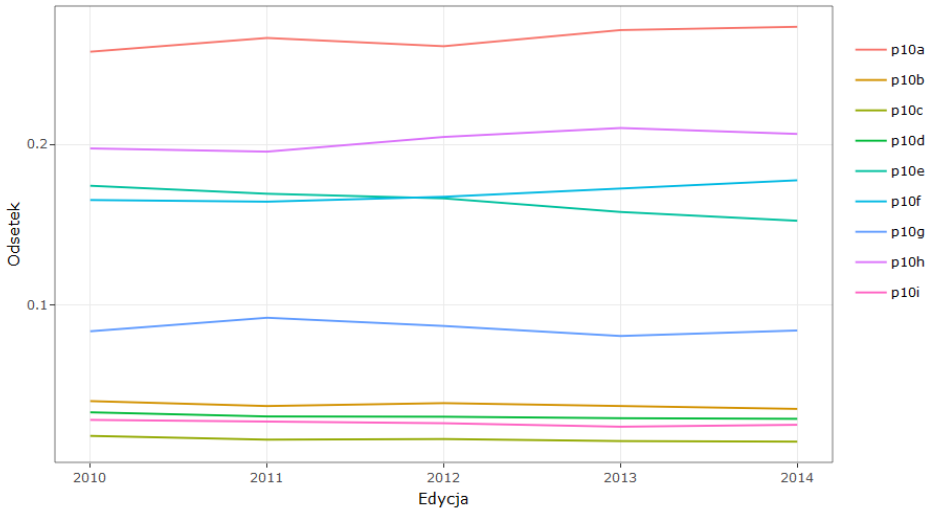
2.4.2 Opis danych wykorzystanych do deklaracji modelów IRT

Aby przeanalizować wpływ skłonności pracodawców do zamieszczania ogłoszeń na wybór ich źródła konieczne jest wykorzystanie modeli bazujących na zmiennych latentnych. Dlatego w dalszej części rozdziału to właśnie modele IRT pozwolą zbadać wpływ skłonności pracodawców do zamieszczania wielu ogłoszeń na wybór miejsca dodawania ofert.

Zbiór danych wykorzystany do szacowania modeli IRT zawiera informacje pochodzące z badań przedsiębiorców przeprowadzonych w latach 2010-2014. Posiada 80 018 obserwacji i 592 zmienne. Jest to duży zbiór, który musiał zostać zmodyfikowany aby być poddanym analizie.

Początkowym etapem modyfikacji był proces wyboru odpowiednich zmiennych i stworzenie mniejszego podzbioru. Podzbiór zawierał zmienne z informacjami o wielkości firmy, województwie, sekcji PKD, rodzaju placówki oraz miejscu zamieszczania ogłoszeń. Pytanie dotyczące źródła dodawania ofert pracy zostało podzielone na 9 zmiennych (od p10a do p10i). Każda zmienna reprezentowała poszczególne miejsce dodawania ogłoszeń. Jeżeli podczas szukania nowych pracowników respondent korzystał z konkretnego źródła, wartość takiej zmiennej była równa 1 (”tak”). Gdy pracodawca zaprzeczał korzystania z danej formy dodawania ogłoszeń, wartość zmiennej wynosiła wówczas 2 (”nie”). Warto zauważyć, iż modele IRT działają na danych z wartościami binarnymi, dlatego w dalszej części procesu przygotowywania danych konieczna była zamiana wartości 2 na liczbę 0. Dopiero dobrze zmodyfikowany i przygotowany zbiór mógł zostać poddany dalszej analizie. Poniżej przedstawiony został wykres obrazujący popularność wskazań poszczególnych źródeł w latach 2010- 2014.

Na podstawie wykresu 2.5 można zauważyć iż popularność większości źródeł jest stała w czasie. Na przestrzeni 5 lat najpopularniejszym miejscem dodawania ogłoszeń były państwowe urzędy pracy. Warto jednak zaznaczyć, iż wykorzystanie Internetu z roku na rok sukcesywnie rosło. Odsetek wskazań konkretnych źródeł pokazuje również, że pracodawcy biorący



Rysunek 2.5. Odsetek wskazań poszczególnych źródeł w różnych edycjach

Źródło: Opracowanie własne na podstawie danych Badania Kapitału Ludzkiego 2010-2014.
Oznaczenia: p10a = Państwowe Urzędy Pracy, p10b = Prywatne biura pośrednictwa pracy, p10c = Head hunter, p10d = Szkoła i akademickie ośrodki kariery, p10e = Prasa, p10f = Internet, p10g = Ogłoszenia rozwieszane w obrębie firmy, p10h = Polecenie rodziny i znajomych oraz p10i = Targi pracy.

udział w badaniu najchętniej szukali nowych pracowników wykorzystując głównie 5 następujących źródeł : Państwowe Urzędy Pracy, polecenia, Internet, prasę lub ogłoszenia zamieszczane w obrębie firmy. Pozostałe miejsca zamieszczania ogłoszeń były znacznie rzadziej wskazywane.

2.4.3 Szacowanie wartości zmiennej latentnej w pakiecie LTM

Jednym z pakietów wykorzystywanych do szacowania modeli IRT w środowisku R jest pakiet LTM. Posiada on szereg funkcji pozwalających szacować m.in. modele Rascha lub dwuparametryczny model logistyczny. Zaimplementowane w nim funkcje pozwalają zbadać poziom zmiennej ukrytej przy różnych założeniach. Warto jednak zauważyć, iż użytkownik korzystający z gotowych funkcji w pakiecie LTM (Rizopoulos, 2006) nie widzi sposobu i procesu szacowania wartości zmiennych latentnych. Korzystając z zaimplementowanych w pakiecie funkcji warto poznać metodologie ich działania. Jedną z przydatnych funkcji szacującej wartości zmiennej ukrytej, dla każdej sekwencji odpowiedzi jest funkcja FACTOR.SCORE z pakietu LTM (Rizopoulos, 2006). Estymowane w niej wartości są sumarycznymi miarami rozkładu a posteriori $P(z|x)$, gdzie z oznacza wektor zmiennych ukrytych, natomiast x ukazuje wektor zmiennych obserwowalnych.

Oceny czynników przypisywane są na podstawie rozkładów posteriori maksymalnie osza-

cowanych prawdopodobieństw. Empiryczne oszacowania Bayesa i powiązanej z nimi wariancji są właściwymi miarami rozkładu posteriori jeżeli $p \to \infty$, gdzie p jest liczbą pozycji testowych. Omawiane oceny czynników opierają się na wyniku poniższego równania:

$$p(z|x) = p(z|x;\hat{\theta})(1+O(1/p)), \tag{2.4}$$

gdzie $\hat{\theta}$ jest wektorem parametrów oszacowanym metodą największej wiarygodności (ang. MLE – Maximum Likelihood Estimate), O liczbą obserwacji, natomiast p liczbą pozycji testowych (Rizopoulos, 2006).

Tabela 2.1 prezentuje wycinek ramki danych utworzonej podczas obliczania wartości zmiennych latentnych dla poszczególnych sekwencji odpowiedzi.

Tabela 2.1. Przykładowe wyniki dla kilku sekwencji wyboru miejsca zamieszczania ogłoszenia

	a	b	c	d	e	f	g	h	i	OBS	exp	z1	se.z1
1	0	0	0	0	0	0	1	1	0	23	10.57	-0.97	0.66
2	0	0	0	0	0	1	0	0	0	39	54.31	-0.70	0.65
3	0	0	0	0	0	1	0	1	0	32	49.39	-0.51	0.64
4	0	0	0	0	0	1	0	1	1	1	6.52	-0.03	0.60
5	0	0	0	0	0	1	1	0	0	27	23.23	-0.49	0.64
6	0	0	0	0	0	1	1	0	1	2	3.03	0.03	0.59

Źródło: Opracowanie własne, gdzie pytania były numerowane od p10a do p10i ale tutaj użyto wyłącznie ostatnich liter. Oznaczenia: a = Państwowe Urzędy Pracy, b = Prywatne biura pośrednictwa pracy, c = Head hunter, d = Szkoła i akademickie ośrodki kariery, e = Prasa, f = Internet, g = Ogłoszenia rozwieszane w obrębie firmy, h = Polecenie rodziny i znajomych, i = Targi pracy; OBS = liczba obserwacji, exp = liczba obserwacji jaką zakładał model, z1 = wartość oczekiwana zmiennej latentnej, se.z1 = błąd standardowy

Wynik przedstawione w tabeli 2.1 są rezultatem pojawiającym się po wywołaniu funkcji `factor.score` na przykładowym zbiorze danych. Wartości przedstawione w tabeli pokazują częściowy proces tworzenia modeli IRT. Każda ze zmiennych od p10a do p10i odpowiada za poszczególną pozycje testową. W odniesieniu do bazy dotyczącej badania przedsiębiorców, źródła internetowe oznaczone zostały literą F. Podczas tworzenia modeli, algorytm grupuje wszystkie unikalne sekwencje wyboru źródeł, zlicza ich ilość, a następnie szacuje dla nich poszczególne wskaźniki. Warto więc zauważyć, iż pakiet LTM (Rizopoulos, 2006) przy obliczaniu wartości zmiennej latentenej dla poszczególnej pozycji testowej bierze pod uwagę wszystkie sytuacje, w których dana pozycja została wybrana.

Zmienna OBS pokazuje dokładnie ilu pracodawców wybierało konkretne źródła w takich

samych kombinacjach, natomiast zmienna EXP określa ile obserwacji takich sekwencji przewidywał model. Zmienna z1 jest oczekiwanym poziomem umiejętności (wartością zmiennej latentnej) dla wybranej sekwencji odpowiedzi. Można przez to rozumieć, że do każdej sekwencji odpowiedzi przypisany zostaje pewien oczekiwany poziom umiejętności. Warto również zauważyć, iż zdarzają się takie sekwencje, w których wartość oczekiwana umiejętności(w tym przypadku skłonność pracowników do zamieszczania ogłoszeń) jest wyższa od poziomu pozostałych sekwencji mających więcej poprawnych odpowiedzi. Przykładem takiej sytuacji jest pierwsza i druga obserwacja w tabeli 2.1.

Pierwszy wiersz obrazuje sytuacje, w której wybrane zostały dwa miejsca zamieszczania ogłoszenia - ogłoszenie w obrębie pracy (10g) oraz polecenie znajomych (10h). Natomiast drugi wiersz przedstawia sytuację kiedy respondenci zamieścili swoje ogłoszenie tylko w Internecie. Pomimo, że pracodawcy zaliczający się do pierwszej grupy wybrali więcej ogłoszeń, posiadają niższy poziom skłonności do zamieszczania ofert pracy w wielu miejscach. Zaistniałe różnice wynikają z założeń modeli, które analizując wszystkie sekwencje odpowiedzi, zwracają uwagę na wpływ wyboru jednego źródła na resztę odpowiedzi.

W środowisku R występuje wiele pakietów szacujących modele IRT.Pakiety opierają się na złożonych algorytmach, które w szybki sposób potrafią przeanalizować duży zbiór danych na podstawie charakterystyki poszczególnych sekwencji odpowiedzi. Można więc zauważyć, iż wykonują ogromną ilość obliczeń, z którą człowiek bez użycia maszyny nie byłby w stanie sobie poradzić. Złożoność algorytmów pozwala więc ocenić skłonności osób nie tylko na podstawie liczby wybranych źródeł. Wzięcie pozostałych czynników pod uwagę umożliwia faktyczne zagłębienie się w proces wyboru miejsca zamieszczenia ogłoszeń. W dalszej części rozdziału do szacowania modeli IRT oprócz pakietu `LTM` (Rizopoulos, 2006) wykorzystany zostanie również pakiet `MIRT` (Chalmers, 2012).

2.4.4 Wyniki modelu dla całej próby badawczej

Przed podzieleniem zbioru na mniejsze grupy warto spojrzeć na wyniki modeli dla całej bazy pracodawców. W tym celu wykorzystany został model Rascha, który zakłada, że poziom dyskryminacji każdej pozycji testowej jest równy jeden. Oznacza to, że zamieszczanie ogłoszeń w konkretnych miejscach ma jednakową zdolność rozróżniania skłonności pracodawców do poszukiwania pracowników w wielu źródłach. W dalszej części pracy wyniki modeli przedstawiane będą w tabelach, które zawierać będą następującą numeracje źródeł zgodnie tabelą 2.2.

Tabela 2.2. Oznaczenia źródeł na potrzeby pracy

	Zmienna	Źródło
1	p10a	Urzędy Pracy
2	p10b	Prywatne biura pośrednictwa pracy
3	p10c	Head hunter
4	p10d	Szkolne i akademickie centra kultury
5	p10e	Prasa
6	p10f	Internet
7	p10g	Ogłoszenia w obrębie miejsca pracy
8	p10h	Polecenie
9	p10i	Targi pracy

Poniżej przedstawione zostały wyniki modelu Rascha dla wszystkich pracodawców biorących udział w badaniu.

Tabela 2.3. Wskaźniki trudności modelu Rascha dla całego zbioru danych

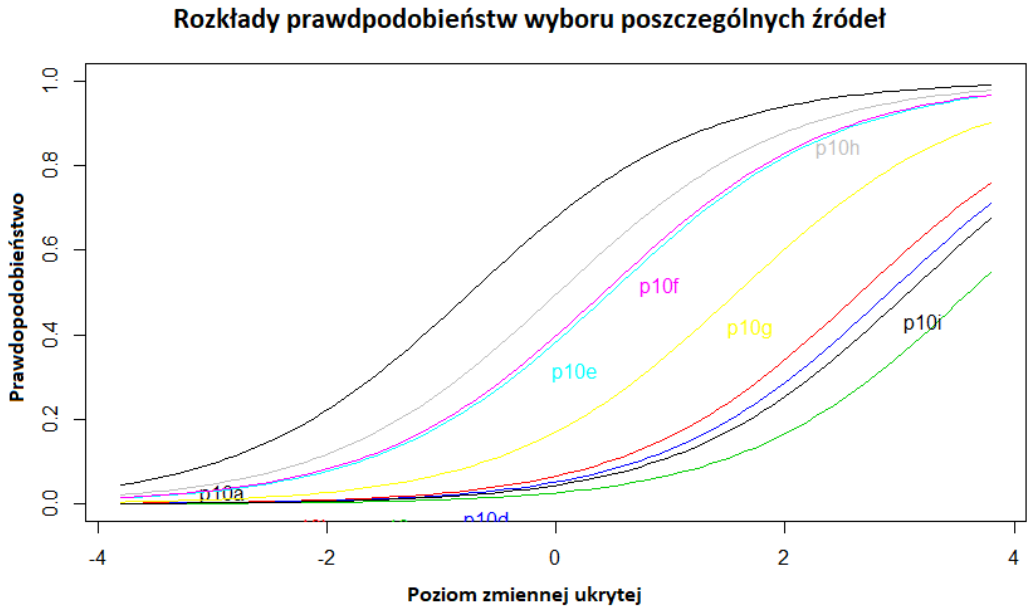
	p10a	p10b	p10c	p10d	p10e	p10f	p10g	p10h	p10i
Poziom trudności	-0.75	2.65	3.61	2.90	0.47	0.41	1.58	0.02	3.07

Źródło: Opracowanie własne na podstawie danych pochodzących z badania przedsiębiorców

Wartości przedstawione w tabeli określają poziom trudności pozycji testowej w modelu Rascha. Oznacza to, że jeżeli pracodawca posiada bardzo małą skłonność do zamieszczania ogłoszeń w wielu miejscach to pierwszym źródłem jakie wybrałby do poszukiwania pracowników jest źródło o najniższym wskaźniku poziomu trudności. Im niższa jest jego wartość tym silniejsze przywiązanie pracodawcy do danego źródła. Wskaźnik trudności wskazuje bowiem przy jakim poziomie skłonności pracodawców do dodawania wielu ogłoszeń, prawdopodobieństwo wyboru konkretnego źródła wynosi 50%. Według wyników modelu Rascha prawdopodobieństwo zamieszczenia ogłoszenia w Internecie wynosi 50 % jeżeli poziom skłonności pracodawcy do dodawania wielu ogłoszeń jest na poziomie 0.41. Można więc zauważyć, że w momencie kiedy poziom trudności wybranej pozycji testowej jest wysoki,to pracodawca najczęściej korzystał z tego źródła jako dodatkowej opcji poszukiwania pracowników. Z tabeli 2.3 można więc wyczytać z jakich źródeł najczęściej korzystali pracodawcy biorący udział w badaniu.

Rysunek 2.6 obrazuje rozkład prawdopodobieństwa wykorzystania wybranego źródła ze względu na poziom skłonności pracodawców do dodawania wielu ogłoszeń.

Na podstawie tabeli 2.3 oraz wykresu 2.6 można zauważyć, iż osoby o najmniejszych skłonnościach do zamieszczania wielu ogłoszeń najchętniej szukały nowych pracowników w urzędach

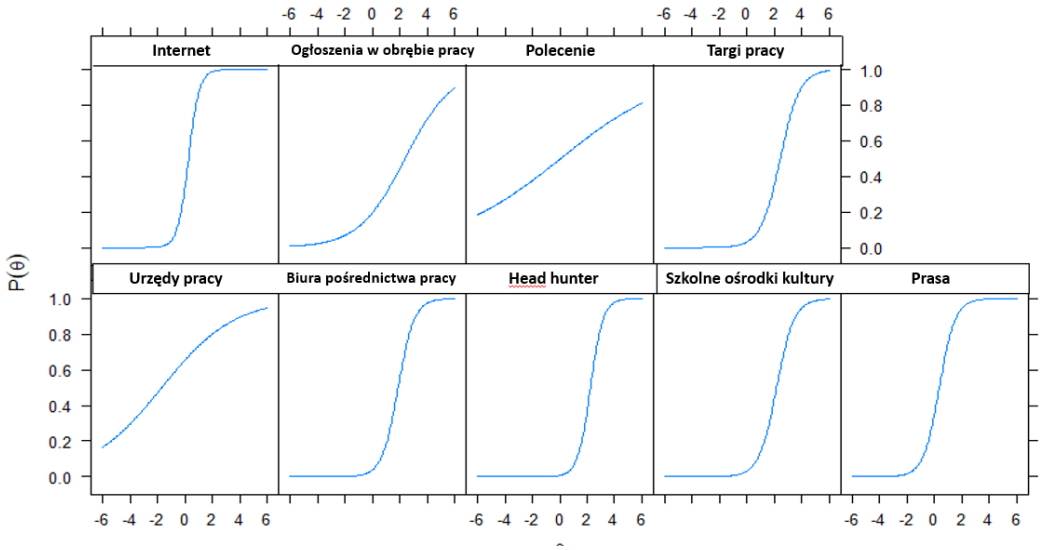


Rysunek 2.6. Rozkład prawdopodobieństw wyboru poszczególnych miejsc zamieszczania ogłoszeń

Źródło: Opracowanie własne na podstawie danych Badania Kapitału Ludzkiego 2010-2014.
Oznaczenia: p10a = Państwowe Urzędy Pracy, p10b = Prywatne biura pośrednictwa pracy, p10c = Head hunter, p10d = Szkoła i akademickie ośrodki kariery, p10e = Prasa, p10f = Internet, p10g = Ogłoszenia rozwieszane w obrębie firmy, p10h = Polecenie rodziny i znajomych oraz p10i = Targi pracy.

pracy lub poprzez polecenia rodziny i znajomych. Internet (p10f) plasował się na 3 miejscu wśród najchętniej wykorzystywanych źródeł. Warto jednak zastanowić się czy wybór konkretnego źródła jednakowo wpływał na zdolność rozróżniania skłonności pracodawców. Aby zbadać jak duży wpływ może mieć wybór każdego ogłoszenia na poziom skłonność respondentów do szukania pracowników w wielu miejscach zastosowany został dwuparametryczny model logistyczny. W środowisku R pakietem wykorzystywanym do tworzenia takich modeli jest pakiet `MIRT` (Chalmers, 2012), który w ciekawy sposób obrazuje rozkład prawdopodobieństwa każdej pozycji testowej. Rysunek 2.6 przedstawia rozkłady prawdopodobieństwa wykorzystania poszczególnych źródeł w zależności od poziomu skłonności pracodawców.

Na podstawie powyższego wykresu można zauważyć, iż niektóre źródła mają inna zdolność do rozróżniania poziomu skłonności przedsiębiorców. Krzywa przedstawiająca rozkład prawdopodobieństwa wykorzystania Internetu jest dosyć stroma. Oznacza to, że osoby ze skłonnościami do zamieszczania wielu ogłoszeń, z dużym prawdopodobieństwem dodałyby swoje oferty w Internecie. Świadczy to również o tym, że jeżeli ktoś wybrał inne formy poszukiwa-



Rysunek 2.7. Rozkłady prawdopodobieństwa wyboru poszczególnych źródeł zamieszczania ofert pracy w zależności od poziomu skłonności pracodawców

Źródło: Opracowanie własne na podstawie danych Badania Kapitału Ludzkiego 2010-2014.

nia pracowników to z dużym prawdopodobieństwem korzystał również z Internetu. Z drugiej strony, niezamieszczenie ofert pracy w Internecie może świadczyć o bardzo niskim poziomie skłonności pracodawców. Z wykresu 2.7 wynika również, że bez względu na poziom nieobserwowalnej zmiennej modelu, szukanie pracowników poprzez polecenia lub urzędy pracy jest cały czas jest wysoce prawdopodobne. Jednakże wybór tych źródeł nie pozwala jednoznacznie rozróżnić poziomu skłonności respondentów do zamieszczania ogłoszeń w wielu miejscach.

2.4.5 Wyniki modelu według sekcji PKD

Jedną z charakterystyk, która pozwala rozróżnić pracodawców zamieszczających ogłoszenie jest branża, której dotyczy oferta. W celu zbadania wpływu skłonności przedsiębiorców na wybór miejsca dodawania ofert pracy wykorzystany został model Rascha. Założone zostało więc, że wybór konkretnego źródła jednakowo rozróżnia poziom skłonności pracowników do dodawania wielu ofert jednocześnie. Dane pochodzące z badania przedsiębiorców posiadały informacje o pracodawcach z 19 sekcji PKD. Niestety w badaniu nie wzięły udziału przedsiębiorstwa z 3 następujących sekcji :

- A - Rolnictwo, leśnictwo, łowiectwo i rybactwo

- T - Gospodarstwa domowe zatrudniające pracowników; gospodarstwa domowe produkujące wyroby i świadczące usługi na własne potrzeby
- U - Organizacje i zespoły eksterytorialne

Poniżej przedstawiona została tabela 2.4 obrazująca wielkość prób wykorzystanych podczas tworzenia poszczególnych modeli Rascha.

Tabela 2.4. Wielkość próby z podziałem na sekcje PKD w Badaniu Kapitału Ludzkiego 2010-2014

PKD	Opis sekcji	Liczba
B	Górnictwo i wydobywanie	165
C	Przetwórstwo przemysłowe	8 352
D	Wytwarzanie i zaopatrywanie w energię elektryczną	290
E	Dostawa wody, gospodarowanie ściekami i odpadami	991
F	Budownictwo	4 402
G	Handel hurtowy i detaliczny, naprawa pojazdów	8 561
H	Transport i gospodarka magazynowa	1 960
I	Działalność związana z zakwaterowaniem i usługami gastronomicznymi	1 268
J	Informacja i komunikacja	589
K	Działalność finansowa i ubezpieczeniowa	589
L	Działalność związana z obsługą rynku nieruchomości	1 215
M	Działalność profesjonalna, naukowa, techniczna	2 379
N	Działalność w zakresie usług administrowania	968
O	Administracja publiczna i obrona narodowa	3 671
P	Edukacja	8 446
Q	Opieka zdrowotna i pomoc społeczna	2 883
R	Działalność związana z kulturą, rozrywką i rekreacją	1 045
S	Pozostała działalność usługowa	344

Źródło: Opracowanie własne.

Podzielenie zbioru na kilka grup pozwala ocenić stopień wykorzystania Internetu jako miejsca poszukiwania pracowników w poszczególnych branżach. Poniżej przedstawiona została tabela przedstawiająca poziom zmiennej określającej trudność poszczególnych pozycji testowych w modelu Rascha.

Wartości przedstawione w tabeli określają poziom trudności pozycji testowej w modelu Rascha. Z tabeli 2.5 można więc wyczytać z jakich źródeł najczęściej korzystali pracodawcy z różnych branży. Źródła internetowe największą popularnością cieszą się w sekcji J, reprezentującej pracodawców z branży Informacji i telekomunikacji. Wskaźnik trudności dla źródeł internetowych w tej grupie jest najniższy. Może to być związane z tym, że to właśnie Internet jest jednym z głównych źródeł komunikacji społecznej i dostarczania wszelkich informacji.

Tabela 2.5. Wyniki modelu Rascha dla poszczególnych sekcji PKD

	p10a	p10b	p10c	p10d	p10e	p10f	p10g	p10h	p10i
B	-0.95	2.16	3.35	2.55	0.01	**0.28**	1.22	-0.21	2.63
C	-1.11	1.79	2.81	2.46	-0.12	**0.07**	1.30	-0.70	2.42
D	-0.28	2.60	3.20	2.65	0.39	**0.48**	1.29	0.71	3.00
E	-1.28	3.71	4.29	3.57	0.73	**0.65**	1.61	0.62	3.42
F	-0.45	2.63	3.79	3.26	0.26	**0.57**	2.15	-1.04	3.13
G	-0.41	2.53	3.54	3.39	0.47	**0.61**	1.35	-0.53	3.54
H	-0.38	2.53	3.53	3.42	0.21	**0.19**	1.76	-0.62	3.19
I	-0.78	2.83	4.38	2.98	0.08	**-0.03**	1.20	-0.60	3.08
J	0.23	1.22	1.99	1.19	-0.20	**-1.10**	2.07	-0.50	2.18
K	-0.32	2.20	2.58	2.27	0.71	**0.42**	1.72	-0.23	2.78
L	-0.58	3.41	4.66	3.49	0.46	**0.65**	1.46	0.44	3.47
M	-0.02	2.57	3.36	2.14	0.62	**0.15**	2.09	-0.16	3.27
N	-0.88	2.00	3.24	2.39	-0.32	**-0.51**	1.47	-0.76	2.06
P	-0.77	4.44	5.72	3.37	1.83	**0.96**	2.28	1.57	3.65
Q	-0.92	4.42	5.74	3.40	1.58	**0.80**	2.00	1.35	3.68
R	-0.94	3.37	4.44	2.96	0.63	**0.04**	1.27	0.69	3.58
S	-0.86	3.32	4.77	3.17	0.85	**0.39**	1.26	0.75	3.71
T	-0.64	2.88	3.84	3.43	0.26	**0.43**	1.77	-0.57	3.30

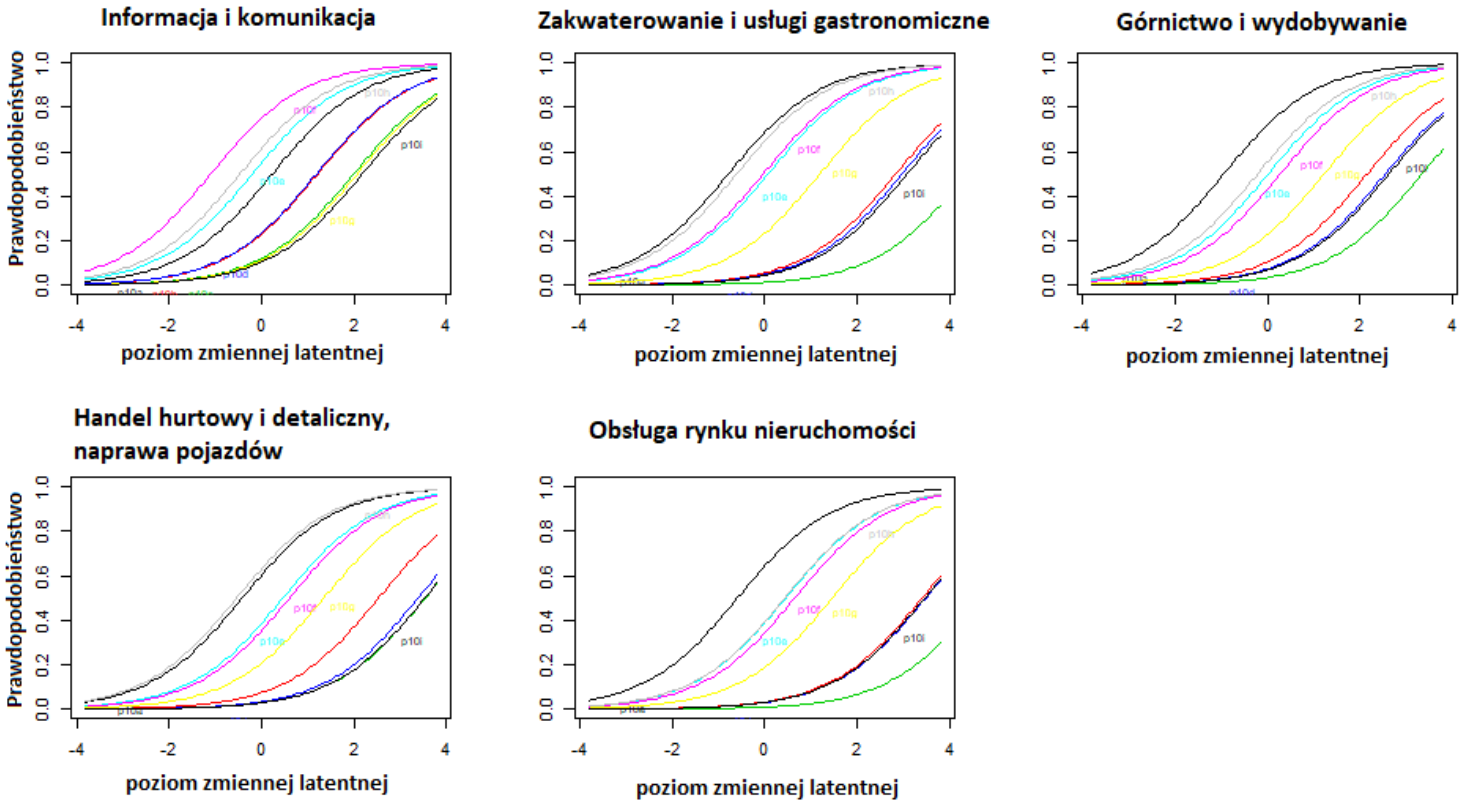
Źródło: Opracowanie własne na podstawie danych Badania Kapitału Ludzkiego 2010-2014.
Oznaczenia: p10a = Państwowe Urzędy Pracy, p10b = Prywatne biura pośrednictwa pracy, p10c = Head hunter, p10d = Szkoła i akademickie ośrodki kariery, p10e = Prasa, p10f = Internet, p10g = Ogłoszenia rozwieszane w obrębie firmy, p10h = Polecenie rodziny i znajomych oraz p10i = Targi pracy.

Kolejną grupą pracodawców najchętniej wybierających źródła internetowe jest branża I zawierająca podmioty gospodarcze zajmujące się zakwaterowaniem i usługami gastronomicznymi. Wpływ na to może mieć fakt, że większość takich ofert dotyczy prac sezonowych, które często podejmowane są przez osoby niezamieszkałe w rejonie miejsca pracy. Przykładem takich ogłoszeń mogą być wakacyjne oferty pracy w kurortach nadmorskich do których zgłaszają się osoby z całej Polski. Szeroki zasięg ogłoszenia zwiększa prawdopodobieństwo dotarcia do dużej liczby osób niekoniecznie zamieszkałej na stałe w miejscu, którego dotyczy oferta. Branżami, które rzadziej szukają nowych pracowników przez Internet są m.in. sekcje B, G oraz L . Sekcja B skupia oferty pracy z górnictwa i wydobycia. Jest to grupa ogłoszeń, które odnoszą się do osób zamieszkałych w konkretnych rejonach Polski. Specyfika branży pokazuje, iż nowi pracownicy zdecydowanie częściej poszukiwani są przez polecenia rodziny i znajomych oraz urzędy pracy. Natomiast sekcja G dotyczy handlu detalicznego oraz naprawy samochodów. W wybranej grupie pracodawców również największą popularnością cieszą się polecenia znajomych i rodziny.

Podobnie może świadczyć to o silnej potrzebie zaufania do sprawdzonych pracowników z okolicznych miejscowości. Sektor podmiotów pochodzących z branży związanej z obsługa rynku nieruchomości również częściej wybiera Urzędy Pracy lub polecenia znajomych podczas poszukiwania nowych pracowników. Branżą, w której poziom trudności pozycji Internetu w teście jest najwyższy okazała się być edukacja (sekcja P). Mimo, iż Internet jest drugim najchętniej wybieranym źródłem w tej grupie, to różnica pomiędzy nim a urzędami pracy jest dosyć spora.

Rysunek 2.8 przedstawia rozkład krzywych charakterystycznych dla omówionych 5 sekcji PKD. Na podstawie wyników modelu Rascha można zauważyć iż wykorzystanie Internetu jako miejsca dodawania ogłoszeń, różni się w poszczególnych branżach. Pracodawcy częściej szukają nowych pracowników przez Internet jeżeli stanowisko dotyczy technicznych stanowisk wiążących się z obsługą komputera. Przedsiębiorcy znacznie częściej zamieszczają ogłoszenia pracy w Internecie, jeżeli mają one trafić do szerokiego grona odbiorców. Przykładem takich stanowisk są oferty prac sezonowych. Branże, które są silnie związane z miejscem ich występowania np. górnictwo, znacznie rzadziej korzystają z Internetu niżeli z poleceń lub lokalnych urzędów pracy.

Rysunek 2.8. Rozkłady prawdopodobieństw wyboru miejsca zamieszczenia ogłoszenia dla 5 sekcji PKD



Źródło: Opracowanie własne na podstawie danych Badania Kapitału Ludzkiego 2010-2014. Oznaczenia: p10a = Państwowe Urzędy Pracy, p10b = Prywatne biura pośrednictwa pracy, p10c = Head hunter, p10d = Szkoła i akademickie ośrodki kariery, p10e = Prasa, p10f = Internet, p10g = Ogłoszenia rozwieszane w obrębie firmy, p10h = Polecenie rodziny i znajomych oraz p10i = Targi pracy.

2.4.6 Wyniki według wielkości

Kolejnym czynnikiem, który może mieć wpływ na skłonności pracodawców do zamieszczania ogłoszeń jest wielkość firmy. Liczba zatrudnionych pracowników oraz struktura firmy może istotnie wpływać na preferencje pracodawców co do sposobu poszukiwania nowej siły roboczej. W badaniu przedsiębiorców wzięły udział firmy o wielkości z następujących przedziałów:

- od 1 do 10 pracowników
- od 10 do 50 pracowników
- od 50 do 250 pracowników
- od 250 do 1000 pracowników
- powyżej 1000 pracowników

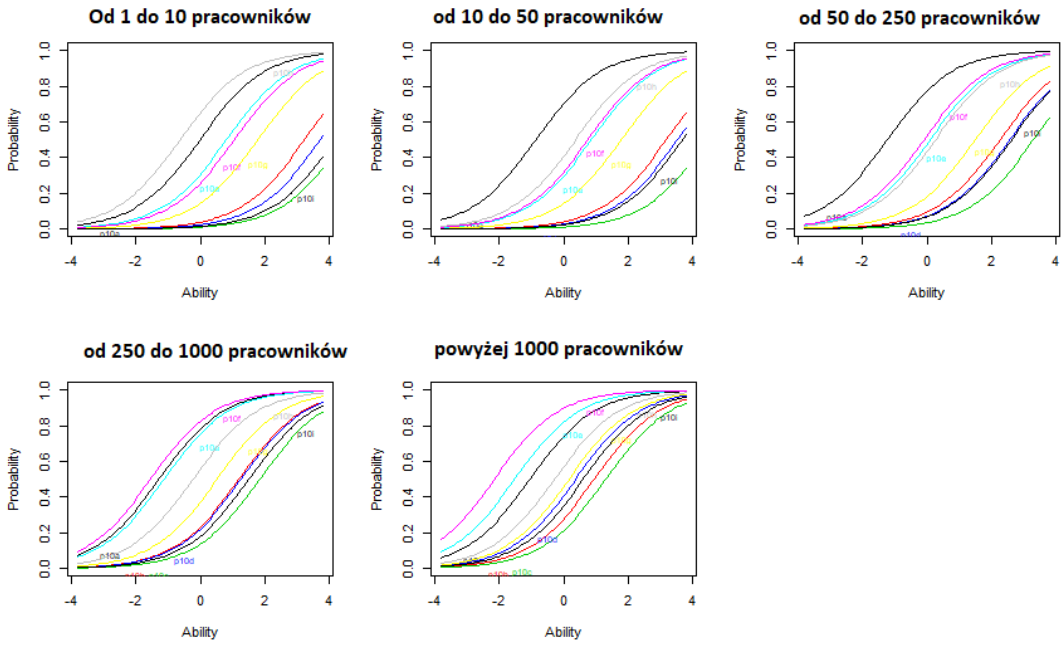
Poniżej przedstawiona została tabela prezentująca wskaźniki trudności z modelu Rascha w wyżej wymienionych grupach.

Tabela 2.6. Wyniki modelu Rascha dla firm w podziale ze względu na wielkość

	p10a	p10b	p10c	p10d	p10e	p10f	p10g	p10h	p10i
Od 1 do 10	-0.02	3.23	4.47	3.72	0.81	**1.05**	1.78	-0.63	4.19
Od 10 do 50	-0.87	3.18	4.46	3.53	0.86	**0.75**	1.78	0.36	3.69
Od 50 do 250	-1.23	2.25	3.30	2.56	0.09	**-0.10**	1.49	0.25	2.61
Od 250 do 1 000	-1.29	1.19	1.84	1.25	-1.13	**-1.55**	0.51	-0.24	1.50
powyżej 1 000	-1.05	0.96	1.30	0.37	-1.52	**-2.16**	0.16	-0.29	0.60

Źródło: Opracowanie własne na podstawie danych Badania Kapitału Ludzkiego 2010-2014.
Oznaczenia: p10a = Państwowe Urzędy Pracy, p10b = Prywatne biura pośrednictwa pracy, p10c = Head hunter, p10d = Szkoła i akademickie ośrodki kariery, p10e = Prasa, p10f = Internet, p10g = Ogłoszenia rozwieszane w obrębie firmy, p10h = Polecenie rodziny i znajomych oraz p10i = Targi pracy.

Zawarte w tabeli 2.6 wartości wskaźnika trudności konkretnych pozycji testowych pokazują przy jakim poziomie skłonności pracowników do dodawania wielu ogłoszeń, prawdopodobieństwo wyboru konkretnego źródła wynosi 50%. Wyniki modelu Rascha pokazują, iż wielkość firmy ma duży wpływ na wybór miejsca zamieszczania ogłoszeń. Pracodawcy posiadający małe firmy zdecydowanie częściej zamiast Internetu wybierają urzędy pracy lub polecenia. Wraz ze wzrostem wielkości firmy, rośnie popularność umieszczania ogłoszeń w Internecie.



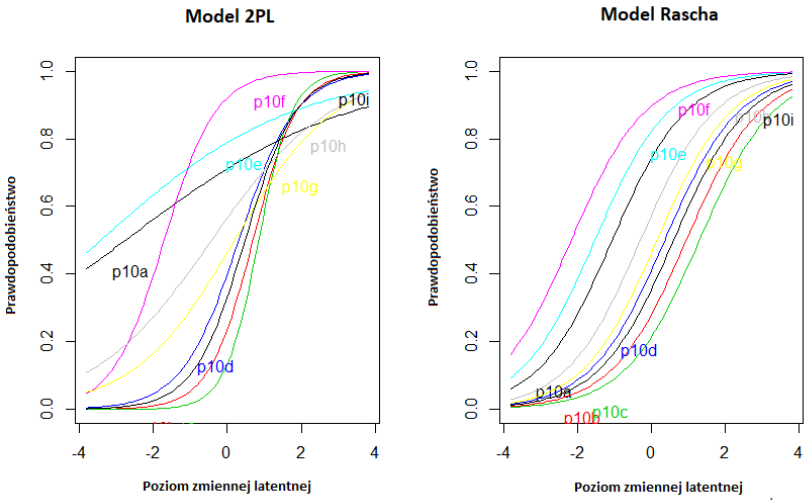
Rysunek 2.9. Wyniki modelu Rascha ze względu na wielkość firmy

Źródło: Opracowanie własne na podstawie danych Badania Kapitału Ludzkiego 2010-2014.
Oznaczenia: p10a = Państwowe Urzędy Pracy, p10b = Prywatne biura pośrednictwa pracy, p10c = Head hunter, p10d = Szkoła i akademickie ośrodki kariery, p10e = Prasa, p10f = Internet, p10g = Ogłoszenia rozwieszane w obrębie firmy, p10h = Polecenie rodziny i znajomych oraz p10i = Targi pracy.

Na wykresie 2.9 przedstawione zostały rozkłady prawdopodobieństwa wyboru poszczególnych miejsc dodawania ogłoszeń w zależności od poziomu zmiennej ukrytej. Przedstawione wartości pokazują, iż w dużych firmach wykorzystanie Internetu jest największe. Warto jednak sprawdzić jak kształtuje się poziom dyskryminacji poszczególnych źródeł w tej grupie pracodawców. Poniżej przedstawione zostały wyniki porównania modelu Rascha do dwuprametrycznego modelu logistycznego, zakładającego inny poziom dyskryminacji dla każdego źródła. Poniżej przedstawiony został wykres porównujący model Rascha z dwuparametrycznym modelem logistycznym dla firm zatrudniających powyżej 1000 osób.

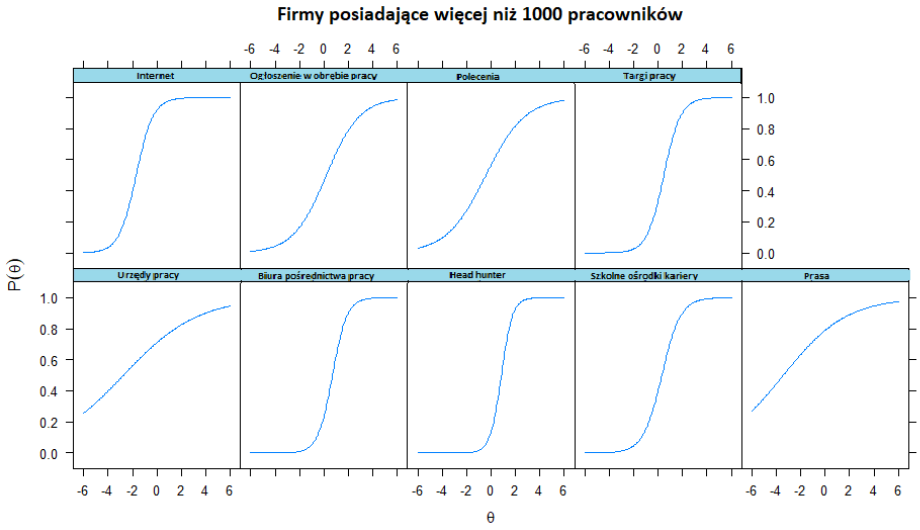
Na podstawie wykresu 2.10 można zauważyć, iż niektóre źródła niezależnie od poziomu skłonności pracodawców do zamieszczania ogłoszeń są chętnie wykorzystywane podczas procesu rekrutacji nowych pracowników. Aby przyjrzeć się dokładniej rozkładom kluczowych źródeł warto spojrzeć na nie każdemu z osobna. Poniżej przedstawione zostały rozkłady prawdopodobieństw wyboru konkretnych źródeł.

Na wykresie 2.11 krzywa charakterystyczna obrazująca rozkład prawdopodobieństwa wyboru Internetu jako źródła poszukiwania nowych pracowników jest dosyć stroma. Świadczy to



Rysunek 2.10. Porównanie wyników dwuparametrycznego modelu logistycznego z modelem Rascha

Źródło: Opracowanie własne na podstawie danych Badania Kapitału Ludzkiego 2010-2014.
Oznaczenia: p10a = Państwowe Urzędy Pracy, p10b = Prywatne biura pośrednictwa pracy, p10c = Head hunter, p10d = Szkoła i akademickie ośrodki kariery, p10e = Prasa, p10f = Internet, p10g = Ogłoszenia rozwieszane w obrębie firmy, p10h = Polecenie rodziny i znajomych oraz p10i = Targi pracy.



Rysunek 2.11. Krzywe charakterystyczne każdego źródła dla firm zatrudniających powyżej 1000 pracowników

Źródło: Opracowanie własne na podstawie danych Badania Kapitału Ludzkiego 2010-2014.
Oznaczenia: p10a = Państwowe Urzędy Pracy, p10b = Prywatne biura pośrednictwa pracy, p10c = Head hunter, p10d = Szkoła i akademickie ośrodki kariery, p10e = Prasa, p10f = Internet, p10g = Ogłoszenia rozwieszane w obrębie firmy, p10h = Polecenie rodziny i znajomych oraz p10i = Targi pracy.

o dużych zdolnościach omawianej pozycji testowej do rozróżniania poziomu skłonności pracodawców. Warto jednak zwrócić uwagę na to,iż wybór urzędów pracy oraz prasy w przedsiębiorstwach powyżej 1000 osób znacznie słabiej pomaga zróżnicować poziom badanej cechy. Świadczy to o braku zależności pomiędzy preferencjami pracodawców, a wyborem źródła zamieszczenia ogłoszenia. Na podstawie powyższej analizy można więc stwierdzić, iż wysoki poziom skłonności przedsiębiorców do dodawania ogłoszeń daje bardzo duże prawdopodobieństwo umieszczenia wielu ofert dopiero po dodaniu ogłoszenia w Internecie. Przedsiębiorcy o bardzo niskich skłonnościach do dodawania ogłoszeń zamiast Internetu wybierają prasę lub urzędy pracy, jako jedyną formę poszukiwania pracowników.

Na podstawie omówionych modeli można zauważyć duży wpływ wielkości firmy na wybór Internetu jako miejsca dodawania ogłoszeń. W małych firmach ogłoszenia zamieszczane są w Internecie dopiero po nieudanych próbach znalezienia nowych pracowników w urzędach pracy lub braku poleceń ze strony znajomych. Im większa firma tym wykorzystanie Internetu wzrasta. Firmy zatrudniające wielu pracowników często borykają się z dużą rotacją siły roboczej. Ciągłe poszukiwanie nowych pracowników zmusza ich do zamieszczania ogłoszeń w wielu miejscach m.in. w Internecie, który posiada ogromny zasięg wśród społeczeństwa.

2.4.7 Wyniki według województw

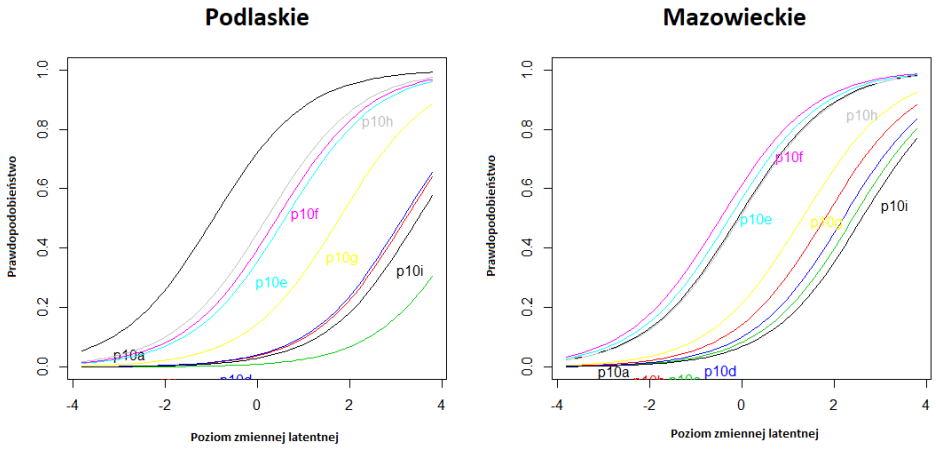
Jak już zostało wcześniej wspomniane przy okazji analizowania różnych sekcji PKD, na wybór miejsca zamieszczenia ogłoszenia wpływać może specyfika miejsca stanowiska pracy. Warto zastanowić się czy wykorzystanie poszczególnych źródeł zamieszczania ogłoszeń jest równomierne w każdym województwie. W tabeli 2.7 zawarte zostały wskaźniki trudności modelu Rascha dla poszczególnych województw.

Na podstawie tabeli 2.7 można zauważyć, iż osoby z różnych województw posiadają inne preferencje dotyczące miejsca zamieszczania ogłoszeń. Województwem, w którym wykorzystanie Internetu podczas poszukiwania pracowników jest najczęstsze okazało się być województwo Mazowieckie. Zapewne jest to związane z silnie rozwiniętą gospodarką i dużym popytem na pracę w Warszawie i okolicach. Ogromna liczba ofert skłania pracodawców do poszukiwania pracowników wykorzystując wiele źródeł jednocześnie. Warto więc zauważyć, iż Internet jest źródłem, który trafia do najszerszego grona odbiorców i wykorzystywany jest w miejscach z dużym popytem na pracę. Według danych zawartych w tabeli wynika, że Internet najczęściej wykorzystywany jest w województwach z dużą liczbą wolnych miejsc pracy.

Tabela 2.7. Wyniki modelu Rascha w podziale na województwa

	p10a	p10b	p10c	p10d	p10e	p10f	p10g	p10h	p10i
Dolnośląskie	-0.81	2.41	3.24	2.70	0.32	**0.13**	1.44	-0.24	2.83
Lubelskie	-0.93	3.13	4.24	3.06	0.94	**1.02**	1.62	0.17	3.24
Lubuskie	-0.97	2.90	4.16	3.36	0.74	**0.79**	1.59	-0.03	3.59
Łódzkie	-0.62	2.63	3.65	2.95	0.37	**0.63**	1.65	-0.06	3.43
Małopolskie	-0.66	2.62	3.64	2.74	0.69	**0.33**	1.64	0.02	3.02
Mazowieckie	-0.09	1.79	2.40	2.19	-0.27	**-0.46**	1.29	-0.05	2.60
Opolskie	-0.89	2.85	4.17	3.38	0.83	**0.78**	1.61	0.13	3.14
Podkarpackie	-1.08	3.19	4.39	3.28	0.92	**0.74**	1.76	0.23	3.26
Podlaskie	-0.95	3.23	4.62	3.16	0.59	**0.42**	1.77	0.19	3.49
Pomorskie	-0.51	2.60	2.58	2.83	0.27	**0.07**	1.62	-0.15	2.67
Śląskie	-0.71	2.22	3.21	2.54	0.28	**-0.01**	1.52	-0.06	2.77
Świętokrzyskie	-0.80	3.32	4.31	3.47	0.74	**0.89**	1.64	0.21	2.96
Warmińsko - Mazurskie	-1.01	3.35	4.47	3.64	0.78	**0.71**	1.71	0.01	3.37
Wielkopolskie	-0.56	2.41	3.26	3.63	0.08	**0.24**	1.39	-0.07	3.02
Zachodnipomorskie	-0.82	2.67	4.01	3.19	0.39	**0.37**	1.49	-0.05	3.38

Źródło: Opracowanie własne na podstawie danych Badania Kapitału Ludzkiego 2010-2014.
Oznaczenia: p10a = Państwowe Urzędy Pracy, p10b = Prywatne biura pośrednictwa pracy, p10c = Head hunter, p10d = Szkoła i akademickie ośrodki kariery, p10e = Prasa, p10f = Internet, p10g = Ogłoszenia rozwieszane w obrębie firmy, p10h = Polecenie rodziny i znajomych oraz p10i = Targi pracy.



Rysunek 2.12. Porównanie wyników dla województw z różnym popytem na prace

Źródło: Opracowanie własne na podstawie danych Badania Kapitału Ludzkiego 2010-2014.
Oznaczenia: p10a = Państwowe Urzędy Pracy, p10b = Prywatne biura pośrednictwa pracy, p10c = Head hunter, p10d = Szkoła i akademickie ośrodki kariery, p10e = Prasa, p10f = Internet, p10g = Ogłoszenia rozwieszane w obrębie firmy, p10h = Polecenie rodziny i znajomych oraz p10i = Targi pracy.

Według opublikowanego badania GUS w końcu II kwartału 2018 r wynika, że największą liczbę wolnych stanowisk pracy posiadało województwo mazowieckie (22,6%), śląskie (13,0%) i wielkopolskie (10,7%), a najmniejsza liczba ofert wolnych miejsc pracy występowała w woje-

wództwie podlaskim (1,4%), warmińsko-mazurskim (1,7%) oraz opolskim (1,8%) (Główny Urząd Statystyczny, 2018a).

Można więc zauważyć, iż województwa o małej ilość wolnych miejsc pracy znacznie częściej szukają nowych pracowników przez urzędy pracy lub polecenia niżeli przez Internet. Poniżej zestawione zostały ze sobą wyniki modelu dla województwa podlaskiego (najmniejsza liczba wolnych miejsc pracy) z województwem mazowieckim (największa liczba ofert pracy).

2.5 Podsumowanie

Dane zebrane w Badaniu Kapitału Ludzkiego w latach 2010-2014 pozwoliły ocenić reprezentatywność poszczególnych źródeł danych wykorzystywanych w badaniach rynku pracy. Powszechnie uważa się, że Internet jest jednym z najczęstszych miejsc poszukiwania nowych pracowników lub pracy. Faktem jest, że liczba zamieszczanych w nim ogłoszeń jest ogromna. Duża liczba ofert pracy w Internecie nie pozwala jednak stwierdzić, czy faktycznie przedsiębiorcy z każdej branży dodają tam swoje ogłoszenia.

Wyniki przeprowadzonej analizy pokazują, iż Internet jest jednym z głównych miejsc zamieszczania ogłoszeń. Warto jednak zauważyć, iż nie wszystkie podmioty równie chętnie zamieszczają tam swoje oferty. Ogłoszenia dodawane w Internecie zazwyczaj kierowane są do dużej grupy osób, które niekoniecznie zamieszkują tereny faktycznego miejsca pracy. Przykładem takich ogłoszeń są sezonowe oferty pracy w kurortach nadmorskich. Dodatkowo przez Internet poszukiwani są pracownicy o dużych kwalifikacjach związanych z komunikacją lub umiejętnościami informatycznymi. Może to być związane z faktem, iż większość firm telekomunikacyjnych poszukujących takich pracowników znajduje się w dużych miastach, gdzie wykorzystanie Internetu jest największe.

Prawdopodobieństwo zamieszczenia ofert pracy w Internecie rośnie również wraz ze wzrostem popytu na prace w danym obszarze. Wyniki analizy potwierdzają, że w województwach z największą liczba wolnych miejsc pracy, wykorzystanie Internetu jest największe. Wskaźniki modeli IRT pokazują jednak, że Internet bardzo często wykorzystywany jest dopiero, jeżeli pracodawca nie znajdzie nowych pracowników poprzez polecenie lub urzędy pracy. Przykładami takich pracodawców są właściciele małych sklepów lub warsztatów. Analiza tych grup pokazuje jak ważne dla pracodawców jest zaufanie do nowych pracowników. Właściciele małych firm wolą korzystać z urzędów pracy lub poleceń, mając świadomość, że takie źródła zwiększają

ich szanse na zatrudnienie lojalnych pracowników. Internet jest wygodnym źródłem informacji o szerokim zasięgu, jednak osoby nieufne bardzo często boją się zbyt dużej anonimowości z nim związanej. Warto również zauważyć, iż oferty dotyczące zawodów na stałe uzależnionych od regionu pracy, rzadko umieszczane są w Internecie. Przykładami takich ofert mogą być stanowiska pracy dla górników lub marynarzy. Wspomniane oferty wymagają konkretnych umiejętności nabywanych w wyznaczonych miejscach Polski. Pracodawcy poszukujący takich pracowników znacznie łatwiej znajdują nowe osoby dzięki pomocy lokalnych instytucji lub społeczności. Dodatkowo w Internecie znacznie rzadziej umieszczane zostają oferty pracy związane z placówkami publicznymi. Przykładami taki stanowisk są ogłoszenia związane z edukacją lub służbą zdrowia.

Podsumowując zebrane wyniki można stwierdzić, iż wykorzystanie tylko i wyłącznie Internetu nie pozwala stworzyć reprezentatywnej próby badawczej do oceny sytuacji na polskim rynku pracy.

Mając na uwadze, że jednym ze źródeł wskazywanych przez pracodawców były Powiatowe i Wojewódzkie Urzędy pracy, w kolejnym rozdziale skupimy się na wykorzystaniu danych z Centralnej Bazy Ofert Pracy prowadzonej przez Ministerstwo Rodziny, Pracy i Polityki Społecznej na potrzeby popytu na pracę.

Rozdział 3

Estymacja popytu na pracę z wykorzystaniem danych Powiatowych Urzędów Pracy (Greta Białkowska)

3.1 Cel rozdziału

Głównym celem rozdziału jest wykorzystanie danych pochodzących z Powiatowych i Wojewódzkich Urzędów Pracy do opisu rynku pracy w Polsce. Aby to osiągnąć określono 3 cele szczegółowe. Pierwszym jest rozpoznanie danych z Centralnej Bazy Ofert Pracy (CBOP). Są to dane niestatystyczne, pochodzące z internetowej bazy urzędów pracy. Dane te nie były do tej pory stosowane w badaniach statystycznych, dlatego w tej pracy zostaną one porównane do źródła statystycznego jakim jest Badanie Popytu na Pracę GUS. Zostaną zbadane zmienne oraz błędy występujące w zbiorze CBOP, porównanie czy oba źródła zawierają te same zawody oraz rodzaje działalności itd. Porównanie tych źródeł jest drugim celem szczegółowym tego rozdziału. Trzecim celem będzie zweryfikowanie czy można poprawić błędy nielosowe zbiorze CBOP poprzez zastosowanie kalibracji z wykorzystaniem zmiennych pomocniczych, które znajdują się w danych z Badania Popytu na Pracę.

3.2 Metody korekcji braku reprezentatywności

W badaniach reprezentatywnych występują dwa rodzaje błędów: błędy losowe oraz nielosowe (systematyczne). Błędem losowym jest na przykład błąd statystyczny, ponieważ badanie re-

prezentatywne jest badaniem, w którym na podstawie próby wnioskuje się na całą populację. Działanie takie ze swej natury narażone jest na błąd. Wyeliminowanie błędu statystycznego możliwe jest tylko wtedy, gdy przeprowadzane są badania pełne. Jednak badania pełne często są zbyt kosztowne oraz czasochłonne, dlatego częściej przeprowadzane są badania reprezentacyjne (Szymkowiak, 2009).

Natomiast przykładem błędów nielosowych jest błąd pokrycia, błąd pomiaru czy braki danych. Błąd pokrycia występuje w sytuacji, gdy część docelowej populacji nie pokrywa się ze spisem jednostek podczas losowania próby lub gdy część jednostek losowana jest z prawdopodobieństwem innym niż to, które założone zostało przez osobę przeprowadzającą badanie (Lohr, 2010). Z błędem pokrycia mamy do czynienia, gdy wykorzystywany jest operat losowania, z którego nie można wylosować do próby jednostek faktycznie należących do badanej populacji bądź odwrotnie – istnieje możliwość wylosowania jednostki, która nie należy do populacji badania. Wykorzystanie operatu badania (na przykład pochodzącego od operatorów sieci komórkowych – ta sama osoba może znajdować się po kilkoma numerami telefonów), w którym informacja na temat danej jednostki jest umieszczona kilkukrotnie to również błąd pokrycia. Ponadto wyróżnia się również dwa rodzaje błędu pokrycia:

- *błąd niepokrycia* występuje w przypadku, gdy podczas losowania pominięte zostają jednostki należące do docelowej populacji, natomiast
- *błąd nadpokrycia* wynika z włączenia elementów, które nie należą do docelowej populacji. (Eurostat, 2019)

Błąd pomiaru występuje w przypadku, gdy odpowiedzi respondentów podczas badania nie są zgodne z prawdą, co prowadzi do uzyskania wyników niezgodnych z rzeczywistością. Można wymienić wiele przyczyn powstawania błędów losowych, między innymi są to niezrozumiałe pytania, pomyłki (spowodowane na przykład słabą pamięcią respondenta) lub celowe kłamstwa (Lohr, 2010).

Trzecim rodzajem, oraz jednym z głównych źródeł, błędów nielosowych są braki danych. Wśród braków danych wyróżnić możemy brak udziału lub brak odpowiedzi, czyli całkowity lub częściowy brak odpowiedzi. Brak udziału to sytuacja, w której znane są jedynie dane identyfikacyjne respondenta, jednak nie zostały uzyskane od niego żadne informacje – przyczyną jest odmowa bądź nieobecność respondenta podczas badania. Braki odpowiedzi z kolei dotyczą poszczególnych pytań, na które respondent nie udzielił odpowiedzi (Szymkowiak, 2009).

Braki odpowiedzi mogą mieć różne przyczyny, jednak bez względu na to z jakiego powodu powstały, są źródłem wielu problemów. Osoby odmawiające udziału w badaniu lub nieudzielające odpowiedzi zazwyczaj różnią się od tych, które biorą w nim udział oraz dostarczają niezbędnych informacji. Z tego powodu wyniki uzyskane podczas badania obarczone są zbyt dużymi błędami – ustalone oceny parametrów znacząco różnią się od rzeczywistych wartości, zmniejsza się efektywny rozmiar próby oraz zniekształcone zostają rozkłady wielu cech. Ponadto, gdy w badaniu występuje znaczna ilość braków odpowiedzi może ono być niekorzystnie postrzegane przez jego odbiorców lub okazać się dla nich kompletnie bezwartościowe. (Szymkowiak, 2009)

Według Lohr (2010) dobra próba powinna być, oczywiście w miarę możliwości, wolna od błędu pokrycia oraz posiadać dokładnie tyle odpowiedzi, ile jednostek jest badanych. Z tego względu zarówno błąd pokrycia, jak i pomiaru powinny zostać rozważone oraz zminimalizowane na etapie projektowania badania.

Jednak w praktyce badań statystycznych bardzo ciężko jest uzyskać kompletną próbę, która nie będzie zawierała braków danych. Z uwagi na ten fakt stosowanych jest wiele metod, które mają na celu zwiększenie frakcji odpowiedzi. Można podzielić je na trzy rodzaje:

- prewencyjne,
- redukujące frakcję braków odpowiedzi,
- korygujące.

Ogólnie rzecz ujmując podejście prewencyjne, które zapobiega występowaniu braków odpowiedzi stosuje się na etapie planowania badania, metody redukujące frakcję braków odpowiedzi na etapie zbierania danych do badania, natomiast techniki korygujące stosuje się w procesie estymacji, gdy zebrane zostały już niezbędne informacje (Szymkowiak, 2009).

Zbieranie danych do badań statystycznych wiąże się z kontaktem z respondentem, dlatego metody prewencyjne wywodzą się z nauk takich jak psychologia czy socjologia. Z tego względu dużą rolę odgrywają tutaj techniki mające na celu przełamanie niechęci respondentów do udzielania informacji czy promujące pozytywne nastawienie do badania. W ramach metod prewencyjnych dużą rolę odgrywają czynniki motywacyjne, które mają przekonać respondentów do udziału w badaniu, na przykład bodźce finansowe. Metody prewencyjne obejmują również kwestie takie jak przygotowanie kwestionariusza, właściwe przygotowanie operatu losowania czy odpowiednie przeszkolenie ankietera.

Techniki redukujące frakcję braków odpowiedzi polegają między innymi na ponownym kontakcie telefonicznym z respondentem, wysyłaniu monitów z prośbą o wzięcie udziału w badaniu czy zastępowaniu jednostek niewyrażających zgody na udział w badaniu. Metody te również w dużej mierze wywodzą się z nauk o zachowaniu się jednostek.

Metody korygujące to różnego rodzaju metody estymacji oraz ważenia danych, których celem jest zmniejszenie obciążenia, które jest konsekwencją braków danych. Ta grupa metod odgrywa coraz większa rolę z uwagi na fakt, iż braki danych występują w nawet najlepiej zaplanowanym badaniu. Szczególną rolę pełnią techniki oparte o system wag (Szymkowiak, 2009).

3.2.1 Imputacja

W przypadku wystąpienia braków odpowiedzi w badaniach statystycznych stosuje się dwie podstawowe metody: imputację oraz kalibrację. Imputacja polega na zastąpieniu braków danych konkretnymi wartościami w celu uzyskania kompletnego zbioru danych, natomiast kalibracja polega na opiera się na takim ustaleniu wag, by zminimalizować obciążenie wynikające z braków odpowiedzi (Szymkowiak, 2009).

Przed korektą powinno się dogłębnie przeanalizować inne źródła danych, które mogłyby zawierać brakujące informacje bądź służyć bezpośrednio do ich oszacowania (Balicki, 2014). Tu przykładem mogą być rejestry administracyjne czy źródła internetowe zawierające informacje, których brakuje w analizowanych danych.

Definicja 3.1. Według Carl-Erik Särndal (2005) imputacja jest to proces szacowania brakujących lub eliminowania niepoprawnych danych, oparty na wykrytych relacjach w zbiorze wartości tych samych lub innych zmiennych (lub obserwacji), dla których danych nie brakuje.

Jak wynika z powyższej definicji, w wyniku zastosowania imputacji w miejsce brakujących wartości przypisane zostają ich substytuty. Jednak, aby metoda ta właściwie odegrała swoją rolę w badaniu muszą zostać spełnione następujące warunki (Szymkowiak, 2009):

1. powinna w większym stopniu być uzależniona od danych, które pochodzą z próby, a nie odwoływać się do założeń, co do natury brakujących danych,
2. szacunki ważnych statystyk z próby nie powinny „zbyt mocno” opierać się na danych, które są imputowane,
3. nie powinna prowadzić do obciążeń lub zmian rozkładów cech w zbiorze danych, jak i do wzrostu wariancji stosowanych estymatorów.

W praktyce badań statystycznych nie łatwo jest dochować powyższych założeń. Oprócz tego podczas operacji na zbiorze danych zawierającym dane imputowane należy zachować szczególną uwagę. Nierozsądne użycie imputacji może prowadzić do dużego zniekształcenia uzyskanych wyników, co z kolei może być przyczyną źle wyciągniętych wniosków. W badaniach statystycznych wykorzystywanych jest wiele różnych metod imputacji, na przykład imputacja (Szymkowiak, 2009):

- **dedukcyjna** – metoda ta polega na szacowaniu brakujących danych drogą dedukcji na podstawie innych informacji uzyskanych w wyniku badania,
- **cold-deck** – metoda stosowana, gdy możliwe jest zastąpienie braków danych wartościami ze źródeł zewnętrznych (na przykład z rejestrów administracyjnych czy spisu) lub z poprzednich badań,
- **z wykorzystaniem średniej** – jest to metoda polegająca na zastąpieniu braków danych średnią wartością cechy obliczoną dla wszystkich jednostek, od których zostały uzyskane odpowiedzi,
- **z wykorzystaniem innej zmiennej** – dla pewnej zmiennej X, dla której brakuje odpowiedzi poszukuje się zmiennej Y, która jest blisko związana ze zmienną X oraz może być uznawana za „substytut" zmiennej X. Brakujące dane dla zmiennej X uzyskuje się w oparciu o wartości zmiennej Y,
- **najbliższego sąsiada** – metoda ta zakłada, że jeśli dwa obiekty posiadają zbliżone bądź te same wartości dla danej grupy cech (na przykład to samo wykształcenie, wiek czy płeć) to również powinny posiadać zbliżone wartości dla cechy *y*. Imputowaną wartością cechy *y* dla *k*-tego obiektu badania jest w tym przypadku $\widehat{y}_k = y_{l(k)}$, przy czym $l(k)$ jest dawcą dla tego obiektu. Dawca to obiekt należący do zbioru wszystkich respondentów *r* oraz dla którego wybrana funkcja odległości ma wartość minimalną,
- **hot-deck** – analogicznie jak w przypadku metody najbliższego sąsiada wartością imputowaną dla cechy *y* jest $\widehat{y}_k = y_{l(k)}$, przy czym $l(k)$ jest dawcą dla obiektu y_k losowo wybranym spośród wszystkich obiektów, które posiadają kompletny rekord danych lub spośród obiektów należących do tej samej klasy imputacyjnej,

- **ekspercka** – w tej metodzie wykorzystywana jest wiedza ekspertów na temat danej populacji, którzy po przestudiowaniu poszczególnych rekordów potrafią, w miejsce brakujących zmiennych, zaproponować realistyczne wartości.

W pakiecie statystycznym R (R Core Team, 2017) imputację można wykonać przy użyciu pakietu VIM (Kowarik & Templ, 2016) i funkcji takich jak na przykład *kNN* czy *hotdeck*.

3.2.2 Kalibracja w badaniach z brakami odpowiedzi

Oprócz imputacji drugą metodą, która wykorzystywana jest w badaniach statystycznych z brakami odpowiedzi jest wspomniana już wcześniej kalibracja. Jest to jedna z metod opartych na systemie wag.

Podstawy teoretyczne kalibracji sięgają lat 90-tych XX wieku i zostały sformułowane przez Särnadala i Deville'a, którzy w swojej pracy przedstawili sposób formułowania estymatora kalibracyjnego wartości globalnej, w którym tzw. wagi kalibracyjne zostały uzyskane z wag wynikających ze schematu losowania próby. Owe wagi kalibracyjne otrzymane były w oparciu o informacje zawarte w wektorze zmiennych pomocniczych. Wykorzystanie dodatkowych informacji w celu poprawy oszacowań parametrów znane było dużo wcześniej, niż teoria kalibracji. Jednak to, co wyróżnia tę metodę pod tym względem, to takie wykorzystanie zmiennych pomocniczych, aby spełnione było odpowiednie równanie kalibracyjne oraz jednocześnie zminimalizowane zostały odległości pomiędzy wartościami wag, które wynikają ze schematu losowania próby, a wagami kalibracyjnymi (Szymkowiak, 2009).

Warto przytoczyć dokładną definicję kalibracji, która według Szymkowiak (2009) kształtuje się następująco:

Definicja 3.2. *Kalibracja* to metoda polegająca na skorygowaniu wag wyjściowych wynikających ze schematu losowania próby, celem redukcji obciążenia wynikającego z istnienia braków odpowiedzi, tak aby spełnione było odpowiednie równanie kalibracyjne. Wagi te obliczane są w oparciu o wykorzystanie informacji dodatkowych lub spoza próby.

Znając teoretyczne podstawy tej metody, warto przedstawić jej formalne ujęcie – czyli proces ustalania wag kalibracyjnych, który można opisać w następujący sposób.

Zakładamy, że populacji składającej się z N elementów, losujemy próbę wielkości n. Ponadto zakładamy , że $\mathbf{w} = (w_1, ..., w_n)^T$ jest poszukiwanym wektorem wag kalibracyjnych, a **d**

= $(d_1,...,d_n)^T$ wektorem wag wynikających ze schematu losowania próby. Dodatkowo niech *G* będzie dowolną funkcją, która spełnia poniższe warunki:

- $G(1) = 0$,
- $G'(1) = 0$,
- $G''(1) = 1$,
- $G(\cdot) \geq 0$,
- $G(\cdot)$ jest ściśle wypukła oraz dwukrotnie różniczkowalna.

Niech głównym celem badania będzie oszacowanie wartości globalnej zmiennej *y*:

$$Y = \sum_{i=1}^{N} y_i, \tag{3.1}$$

gdzie y_i to wartość zmiennej *y* dla *i*-tej jednostki ($i = 1,...,N$). Oprócz tego niech $x_1,...,x_k$ oznaczają zmienne pomocnicze, które będą wykorzystywane podczas wyznaczania wag kalibracyjnych. Poprzez $\mathbf{X}_j$ oznaczona będzie wartość globalna zmiennej $x_j, (j = 1,...,k)$:

$$\mathbf{X}_j = \sum_{i=1}^{N} x_{ij}, \tag{3.2}$$

gdzie poprzez x_{ij} oznaczona jest wartość *j*-tej zmiennej pomocniczej dla *i*-tej jednostki badania. W ujęciu matematycznym proces poszukiwania wag kalibracyjnych można przedstawić w postaci trzech poniższych warunków (Szymkowiak, 2009):

1. minimalizacja funkcji odległości:

$$D(w,d) = \sum_{i=1}^{n} d_i G(\frac{w_i}{d_i}) \rightarrow \text{min}, \tag{3.3}$$

2. równania kalibracyjne:

$$\sum_{i=1}^{n} w_i x_{ij} = X_j, \quad j = 1,\ldots,k, \tag{3.4}$$

3. warunki ograniczające:

$$L \leq \frac{w_i}{d_i} \leq U, \quad \text{gdzie: } L \leq 1 \text{ i } U \geq 1, \quad i = 1,...,N \tag{3.5}$$

Warunek pierwszy (mówiący o minimalizacji funkcji odległości) zakłada, iż wagi kalibracyjne powinny być wyznaczone w taki sposób, by ich wartości były możliwie bliskie wartościom wag wynikającym ze schematu losowania. Funkcja *G* mierzy odległość pomiędzy ilorazem wag $\frac{w_i}{d_i}$ a 1. Warunek drugi, czyli równania kalibracyjne, to istota teorii kalibracji. Stanowi on, iż wagi powinny być dobrane w taki sposób, by po zastosowaniu ich do wszystkich zmiennych pomocniczych otrzymać ich wartości globalne. W sytuacji, gdy warunek ten będzie spełniony, to po wykorzystaniu wag kalibracyjnych do zmiennej *y*, powinno się otrzymać ocenę wartości globalnej, która jest bliska jej rzeczywistej wartości. Ostatni warunek to tak zwany warunek ograniczający, którego celem jest zapobieganie uzyskiwania ujemnych lub ekstremalnych wartości przez wagi kalibracyjne (Szymkowiak, 2009).

Ponadto istnieje dowolność w wyborze funkcji $G(\cdot)$. W literaturze najczęściej występują poszczególne jej postacie:

- liniowa

$$G_1(x) = \frac{1}{2}(x-1)^2, \tag{3.6}$$

- raking

$$G_3(x) = x(\log x - 1) + 1, \tag{3.7}$$

- sinus hiperboliczny

$$G_5(x) = \frac{1}{2\alpha} \quad \int_0^x sinh[\alpha(t - \frac{1}{t})]dt, \tag{3.8}$$

gdzie α to dodatni parametr, który pozwala sterować stopniem rozrzutu wag kalibracyjnych w stosunku do wag, które wynikają ze schematu losowania próby (domyślnie parametr ten przyjmuje wartość równą 1), natomiast *sinh* to funkcja sinusa hiperbolicznego zdefiniowanego jako $sinh(x) = \frac{e^x - e^- x}{2}$. (Szymkowiak, 2009)

Znając powyższe warto przedstawić postać estymatora kalibracyjnego. Zakłada się, że (Szymkowiak, 2009):

- dla każdego respondenta znana jest wartość dla każdej zmiennej pomocniczej – znana

jest macierz zmiennych pomocniczych:

$$X_r = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1k} \\ x_{21} & x_{22} & \ldots & x_{2k} \\ \ldots & \ldots & \ldots & \ldots \\ x_{m1} & x_{m2} & \ldots & x_{mk} \end{pmatrix} \tag{3.9}$$

- znany jest wektor wartości globalnych wszystkich zmiennych pomocniczych:

$$X = \left(\sum_{i=1}^{N} d_i x_{i1}, \sum_{i=1}^{N} d_i x_{i2}, \ldots, \sum_{i=1}^{N} d_i x_{ik}, \right)^T \tag{3.10}$$

Definicja 3.3. Według Szymkowiak (2009) estymator kalibracyjny wartości globalnej zmiennej Y przy znanym wektorze wartości globalnych zmiennych pomocniczych ma postać:

$$\widehat{Y}_X = \sum_{i=1}^{m} w_i y_i \tag{3.11}$$

Postać wektora wag kalibracyjnych w przypadku funkcji liniowej kształtuje się następująco:

$$w_i = d_i + d_i (X - \widehat{X})^T \left(\sum_{i=1}^{m} d_i x_i x_i^T \right)^{-1} x_i, \tag{3.12}$$

przy czym

$$\widehat{X} = \left(\sum_{i=1}^{m} d_i x_{i1}, \sum_{i=1}^{m} d_i x_{i2}, \ldots, \sum_{i=1}^{m} d_i x_{ik} \right)^T, \tag{3.13}$$

a

$$x_i = (x_{i1}, \ldots, x_{ik})^T \tag{3.14}$$

to wektor wartości wszystkich *k*-zmiennych pomocniczych dla *i*-tej jednostki badania ($i = 1, \ldots, m$).

3.2.3 Kalibracja w badaniach opartych na próbie nielosowej

Podczas losowania próby do badania reprezentacyjnego, do każdej wylosowanej jednostki w próbie przypisana jest waga d_i wynikająca ze schematu losowania. Posiadając taką informa-

cję możliwe jest zastosowanie procedury ustalania wag kalibracyjnych opisanej w poprzednim podrozdziale. Istnieją jednak sytuacje, gdzie niemożliwe jest uzyskanie wag wyjściowych – są to na przykład badania pełne (na przykład spisy ludności) czy dane niestatystyczne (rejestry administracyjne czy dane internetowe), w których również może występować problem braków danych, a zastosowanie metod korygujących błędy nielosowe będzie ważne z punktu widzenia końcowych wyników badania. Możliwość zastosowania podejścia kalibracyjnego na zbiorach danych innych niż próba reprezentacyjna polega na wykorzystaniu odpowiedniego rodzaju konstrukcji wyjściowych wag d_i, które podlegają kalibracji. (Klimanek & Szymkowiak, 2017)

W przypadku badań pełnych, ze względu na fakt, iż badana jest cała populacja, każdej jednostce przypisać można zatem wagę równą 1 ($d_i = 1$ dla $i = 1, .., N$). Dotyczy to zarówno przypadku, gdy wszystkie jednostki biorą udział w badaniu, jak i sytuacji, w której część jednostek odmawia lub z innych powodów nie bierze w nim udziału. Jednak, gdy w badaniu występują braki danych spowodowane nie wzięciem udziału w badaniu przez jednostki, można przypisać wagę równą 0 wszystkim jednostkom, które nie przystąpiły do badania oraz wagę równą 1 pozostałym. Tak przygotowany wektor wag może być poddany kalibracji w celu skorygowania braków danych (Klimanek & Szymkowiak, 2017).

W przypadku danych niestatystycznych wykorzystanie kalibracji kształtuje się identycznie, jak w przypadku badań pełnych. W tej sytuacji również odpowiednio ustalane są wagi d_i, jednak mogą przyjmować one różne postacie:

- waga równa 1,
- waga równa $\frac{N}{n}$,
- waga równa wartości zmiennej, do której odbywa się kalibracja.

W pakiecie statystycznym R (R Core Team, 2017) kalibrację można zastosować przy użyciu pakietu **survey** (Lumley, 2004) oraz funkcji *calibrate* lub wykorzystując funkcje z pakietu **leaken** (Alfons & Templ, 2013).

3.3 Wyniki analizy eksploracyjnej

3.3.1 Analiza zbioru danych z Centralnej Bazy Ofert Pracy

Przedmiotem analizy jest zbiór danych z CBOP, czyli opisywanej w pierwszym rozdziale Centralnej Bazy Ofert Pracy urzędów pracy. Zbiór ten obejmuje lata 2011-2017 oraz składa się ze 100

zmiennych, które zawierają informacje na temat zamieszczanych, przez pracodawców, w bazie ogłoszeń o pracę. Dane liczą 2 997 325 rekordów. Są to zarówno szczegółowe dane na temat samej oferty pracy (takie jak nazwa stanowiska pracy, jego adres, kod zawodu, ważność oferty, proponowane wynagrodzenie czy wymiar etatu itd.), jak dane systemowe – czyli informacje dotyczące daty wpłynięcia oferty do bazy czy nazwy placówki – urzędu pracy.

Przed rozpoczęciem każdego badania statystycznego należy szczegółowo zapoznać się z dostępnymi danymi – wybrać zmienne, które będą istotne podczas badania, określić ich jakość, a następnie zastosować metody, które tę jakość mają poprawić (kalibracja, imputacja itp.) oraz *oczyścić* zbiór danych. Dane pochodzące z Centralnej Bazy Ofert Pracy urzędów pracy są danymi internetowymi – nie zostały stworzone do celów statystycznych, dlatego nie wszystkie zmienne w tym zbiorze będą istotne dla badania rynku pracy. Najważniejsze zmienne zbioru CBOP kształtują się następująco:

- `prac_kod_pkd2007` – kod Polskiej Klasyfikacji Działalności, jest to pięcioznakowy kod określający rodzaj działalności prowadzonej przez przedsiębiorstwo,
- `kod_zawodu` – kod zawodu według Klasyfikacji Zawodów i Specjalności 2017,
- `prac_wielkosc_przed` – wielkość przedsiębiorstwa:
  - mikro (1-9 osób),
  - małe (10-49 osób),
  - średnie (50-249 osób),
  - duże (> 250 osób),
- `kod_woj` – województwo, w którym usytuowane jest miejsce pracy,
- `kod_powiatu` – powiat, w którym usytuowane jest miejsce pracy,
- `kod_rodz_zatr` – kod rodzaju zatrudnienia, czy jest to umowa o pracę na czas określony czy nieokreślony,
- `wymagany_staz` – staż pracy wymagany przez pracodawcę,
- `nazwa_stan` – nazwa oferowanego stanowiska,
- `wymiar_etatu` – proponowany wymiar etatu,

- liczba_miejsc_og – liczba wolnych miejsc pracy ogółem,
- data_waz_od – data, od której ważna jest oferta pracy,
- data_waz_do – data, z którą ważność oferty pracy kończy się.

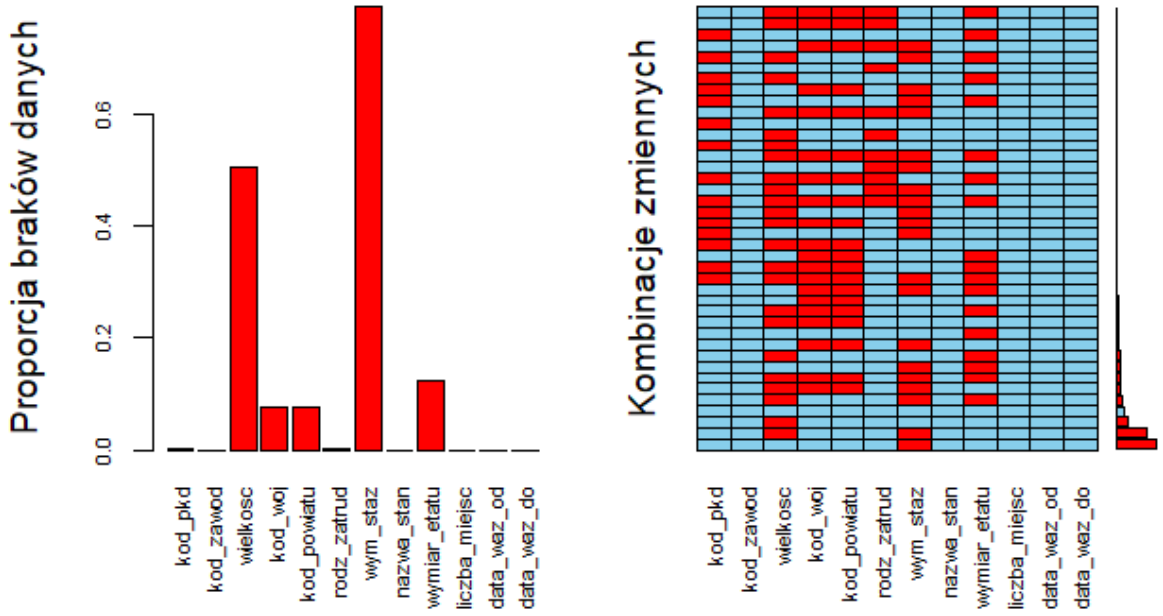
Na potrzeby dalszej analizy zbiór został ograniczony tylko do 2017 roku. Na podstawie zmiennej waznosc_do dodana została zmienna rok, a następnie wyfiltrowane zostały wiersze dla podanego roku. Ponadto wybrane zostały jedynie najważniejsze zmienne wymienione powyżej.

Zbiory danych, które obejmują kilka lat lub zbiory danych internetowych często nie są pozbawione błędów. Wynika to z faktu, iż celem powstawania danych niestatystycznych nie jest prowadzenie badań na ich podstawie. Ze względu na powyższe przed rozpoczęciem analizy należy znaleźć oraz wyeliminować wszystkie błędy występujące w dostępnym zbiorze danych.

W danych Centralnej Bazy Ofert Pracy również znajdują się błędy. Pierwszym z nich jest niejednolite kodowanie braków danych. W przypadku zmiennych prac_kod_pkd2007, wymagany_staz braki danych kodowane były jako NA oraz false. Zapis ten został ujednolicony. Ponadto kod PKD w zmiennej prac_kod_pkd2007 jest różnie kodowany, raz jest informacja pełna (pięcioznakowy kod), a raz kod dwu lub trzyznakowy. Pierwsza cyfra lub dwie pierwsze cyfry kodu działalności według klasyfikacji z 2007 roku oznaczają dział, dzięki temu można było znaleźć sekcje rodzaju działalności dla każdej oferty pracy. Do zbioru danych na podstawie zmiennej prac_kod_pkd2007 została dodana najpierw kolumna zawierająca numer działu, na podstawie której której powstała następnie zmienna sekcja_pkd informująca o sekcji PKD (litery od A do U). Po ujednoliceniu zmiennych można zbadać rozkład braków danych, w zbiorze CBOP dla 2017 kształtuje się on następująco:

Ze względu na długość nazwy niektórych zmiennych zostały zmienione dla poprawy czytelności wykresu: *wielkosc* to zmienna prac_wielkosc_przed, *rodz_zatrud* to zmienna kod_rodz_zatr, *wym_staz* to zmienna wymagany_staz, a *liczba_miejsc* to liczba_miejsc_og.

Jak wynika z wykresu 3.5, wskazującego proporcję braków danych, największym odsetkiem braków informacji charakteryzuje się zmienna wymagany_staz – 79,08% braków danych, na drugim miejscu pod względem braków danych jest zmienna informująca o wielkości przedsiębiorstwa (prac_wielkość_przed) – 50,34%, a dla zmiennej wymiar_etatu odsetek braków danych wynosi 12,48%. Zmienne określające kod zawodu, nazwę stanowiska, liczbę



Rysunek 3.1. Rozkład braków danych w zbiorze danych z CBOP dla 2017 roku

Źródło: Opracowanie własne.

miejsc pracy oraz datę ważności oferty pracy nie posiadają żadnych braków danych.

Drugi wykres przedstawia kombinacje braków odpowiedzi dla poszczególnych zmiennych. W danych z CBOP najczęściej występuje kombinacja, gdzie brakuje informacji dla zmiennej `wymagany_staz` oraz kombinacja, w której braki danych występują dla zmiennej `prac_wielkosc_przed` i `wymagany_staz`.

Tabela 3.1. Rozkład braków danych z zbiorze CBOP

Zmienna	Licba braków danych
prac_kod_pkd2007	1 939
kod_zawodu	0
prac_wielkosc_przed	303 966
kod_woj	46 756
kod_powiatu	46 756
kod_rodz_zatrud	367
wymagany_staz	477 495
nazwa_stan	0
wymiar_etatu	75 355
liczba_miejsc_og	0
data_waz_od	0
data_waz_do	0

Źródło: Opracowanie własne.

Badając popyt na pracę warto sprawdzić jakich zawodów poszukuje się w danej sekcji, województwie czy w jakiej wielkości przedsiębiorstwie. Do zbadania korelacji pomiędzy zawodem a innymi zmiennymi posłużył współczynnik V Cramera, który służy to badania zależności między zmiennymi nominalnymi, wśród których przynajmniej jedna przyjmuje więcej niż dwie wartości. Współczynnik ten interpretuje się następująco (Sobczyk, 2000):

- gdy V < 0,3 – korelacja jest słaba,
- gdy V < 0,5 – korelacja jest umiarkowana,
- gdy V > 0,5 – korelacja jest silna.

Zbadana została korelacja pomiędzy głównymi grupami zawodów, a kodem rodzaju działalności, wielkością przedsiębiorstwa, województwem i kodem rodzaju zatrudnienia (na czas określony lub nie). Główny Urząd Statystyczny (2019a) określa następujące grupy zawodów:

1. Przedstawiciele władz publicznych, wyżsi urzędnicy i kierownicy,
2. Specjaliści,
3. Technicy i inny średni personel,
4. Pracownicy biurowi,
5. Pracownicy usług i sprzedawcy,
6. Rolnicy, ogrodnicy, leśnicy i rybacy,
7. Robotnicy przemysłowi i rzemieślnicy,
8. Operatorzy i monterzy maszyn i urządzeń,
9. Pracownicy wykonujący prace proste.

Do zbioru danych dodana została zmienna *zawod_grupy1*, która informuje do której z powyższych grup zawodów należy dana oferta pracy. Tabela 3.2 zawiera rozkłady odsetków poszczególnych zawodów w badaniu Popytu na Pracę w 2017 roku. Oszacowania na podstawie tego badania zostaną wykorzystane do porównania z CBOP oraz wynikami kalibracji.

Jak wynika z tabeli 3.3, w ofertach pracy umieszczanych w bazie CBOP występuje bardzo słaba zależność pomiędzy zawodem a wielkością przedsiębiorstwa, województwem, w którym

Tabela 3.2. Odsetki dla poszczególnych zawodów w Badaniu Popytu na Pracę w 2017 roku

Grupa zawodu	kwartał I	kwartał II	kwartał III	kwartał IV
1. Władze	3,1	2,8	2,3	2,3
2. Specjaliści	15,3	15,0	16,0	15,9
3. Technicy	6,6	6,6	6,2	6,7
4. Biurowi	9,7	9,6	8,7	10,0
5. Usługi i sprzedaż	11,7	12,0	13,3	12,3
6. Rolnicy	0,3	0,2	0,3	0,2
7. Robotnicy	30,4	28,6	29,6	28,1
8. Operatorzy	14,7	18,0	15,8	17,1
9. Prace proste	8,1	7,3	7,8	7,3

Źródło: Opracowanie własne.

usytuowane jest miejsce pracy, oraz tym czy praca jest na czas określony czy nie. Dla sekcji PKD wartość statystyki jest znacznie wyższa niż dla pozostałych zmiennych, jednak pomiędzy tą zmienną a zawodem nadal występuje słaba korelacja.

Tabela 3.3. Wartość współczynnika V Cramera dla głównych grup zawodów i wybranych zmiennych w 2017 roku

Zmienna	sekcja_pkd	prac_wielkosc_przed	kod_woj	kod_rodz_zatrud
Statystyka	0,268	0,07	0,036	0,074

Źródło: Opracowanie własne.

W drugim kroku, dla porównania wyników, zbadana została bardziej szczegółowa grupa zawodów – kody dwucyfrowe. Dodana została zmienna *kod_zawodu2*. Wyniki kształtują się następująco:

Tabela 3.4. Wartość współczynnika V Cramera dla dwucyfrowych kodów zawodów i wybranych zmiennych w 2017 roku

Zmienna	sekcja_pkd	prac_wielkosc_przed	kod_woj	kod_rodz_zatrud
Statystyka	0,350	0,152	0,051	0,103

Źródło: Opracowanie własne.

Wartość współczynnika V Cramera w przypadku dwucyfrowych kodów zawodów spadła jedynie dla zmiennej *prac_kod_pkd*, wartości pozostałych współczynników wzrosły. W przypadku dwucyfrowych kodów zawodów występuje umiarkowana korelacja pomiędzy zawodem a sekcją PKD. Dla pozostałych zmiennych stopień korelacji z zawodem pozostaje taki sam jak w przy-

padku dużych grup zawodów.

Warto również zbadać czy zależności te będą stałe biorąc pod uwagę poszczególne kwartały 2017 roku. Poniżej znajduje się tabela z wynikami korelacji V Cramera w podziale na kwartały.

Tabela 3.5. Wartość współczynnika V Cramera dla głównych grup zawodów i wybranych zmiennych w podziale na kwartały 2017 roku

Zmienna	sekcja_pkd	prac_wielkosc_przed	kod_woj	kod_rodz_zatrud
kwartał I				
Statystyka	0,265	0,075	0,041	0,074
kwartał II				
Statystyka	0,271	0,073	0,039	0,075
kwartał III				
Statystyka	0,274	0,068	0,039	0,078
kwartał IV				
Statystyka	0,272	0,088	0,070	0,096

Źródło: Opracowanie własne.

Jak widać w tabeli 3.5 zależności są stałe. Wartości współczynnika V Cramera różnią się nieznacznie na przestrzeni kwartałów – oprócz kwartału ostatniego, gdzie wartość V jest znacznie większa w porównaniu do poprzednich. Jednak korelacja pomiędzy zawodem a czynnikami miejscem jego poszukiwania jest identyczna jak dla ogółu – zależność jest słaba.

3.3.2 Porównanie do Badania Popytu na Pracę

Na dane z badania popytu na pracę, które zostało opisane dokładnie w pierwszym rozdziale, składają się trzy zbiory danych obejmujące rok 2017. W przeciwieństwie do jednostkowych danych z bazy CBOP, dane z owego badania są danymi zagregowanymi. Przedstawiają one liczbę pracujących, wolnych oraz nowo utworzonych miejsc pracy w poszczególnych kwartałach danego roku, ze względu na województwo, zawód oraz sekcję PKD. W każdym ze zbiorów znajdują się następujące zmienne:

- `rok`,
- `kwartal` – I, II, III i IV kwartał,
- `zawod1` – grupa zawodu kodowana z użyciem jednej cyfry, gdzie poszczególne grupy są oznaczone

  - 1 = Przedstawiciele władz publicznych, wyżsi urzędnicy i kierownicy,
  - 2 = Specjaliści,
  - 3 = Technicy i inny średni personel,
  - 4 = Pracownicy biurowi,
  - 5 = Pracownicy usług i sprzedawcy,
  - 6 = Rolnicy, ogrodnicy, leśnicy i rybacy,
  - 7 = Robotnicy przemysłowi i rzemieślnicy,
  - 8 = Operatorzy i monterzy maszyn i urządzeń,
  - 9 = Pracownicy wykonujący prace proste.

- `zawod2` – grupa zawodu kodowana z użyciem dwóch cyfr. Szczegółowy opis można znaleźć na stronie Ministerstwa Rodziny, Pracy i Polityki Społecznej[1]

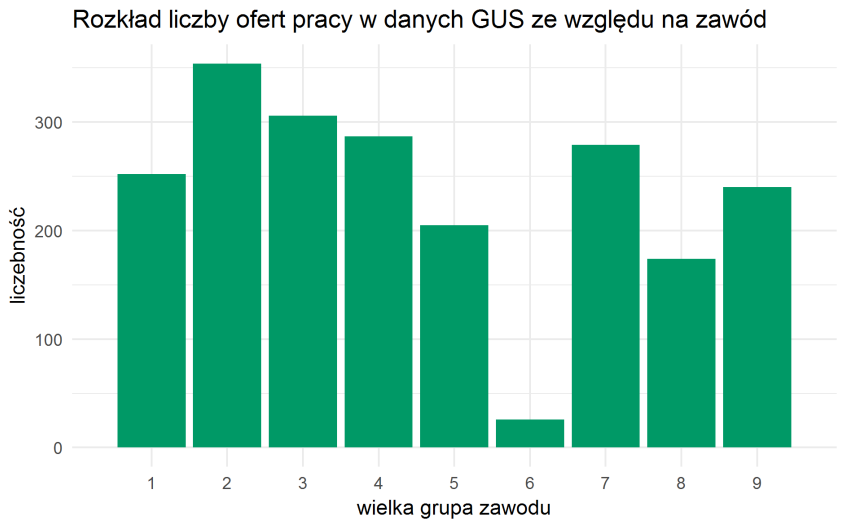
Ponadto, w zależności od zbioru, w danych znajdują się zmienne `woj` oraz `sekcja` oraz zawarte są informacje o liczebnościach wolnych miejsc pracy.

Jako, że dane z badania popytu na pracę nie są danymi jednostkowymi nie można ich szczegółowo porównać do danych z Centralnej Bazy Ofert Pracy pod względem jakości, ponieważ brakuje tutaj chociażby informacji o wielkości braków danych w tym zbiorze. Dane różnią się sposobem kodowania zmiennych – w badaniu popytu o pracę zawód oraz sekcja PKD są zmiennymi tekstowymi obejmującymi główne grupy zawodów i rodzajów działalności, a w bazie CBOP zostały zawarte kody zawodów i rodzaju działalności na szczegółowym poziomie.

Rozkład liczby ofert pracy ze względu na zawody w Badaniu Popytu o Pracę został pokazany na wykresie 3.2. Jak na nim widać, reprezentacja szóstej grupy zawodu w tych danych znacznie odbiega od pozostałych grup zawodów. Zważając na ten fakt, podczas dalszych analiz w obu zbiorach danych ta grupa zawodu nie będzie brana pod uwagę.

Analizując wykresy 3.3 i 3.4 można zauważyć, iż rozkłady liczby wolnych miejsc pracy ze względu na sekcję PKD różnią się od siebie. jednak te różnice nie są tak duże, jak można byłoby zakładać porównując dane statystyczne i niestatystyczne. W obu przypadkach bardzo dużą liczebnością charakteryzuje się sekcja C, czyli przetwórstwo przemysłowe. Zarówno dla danych pochodzących z Powiatowych Urzędów Pracy, jak i dla tych pochodzących z badania GUS sekcje

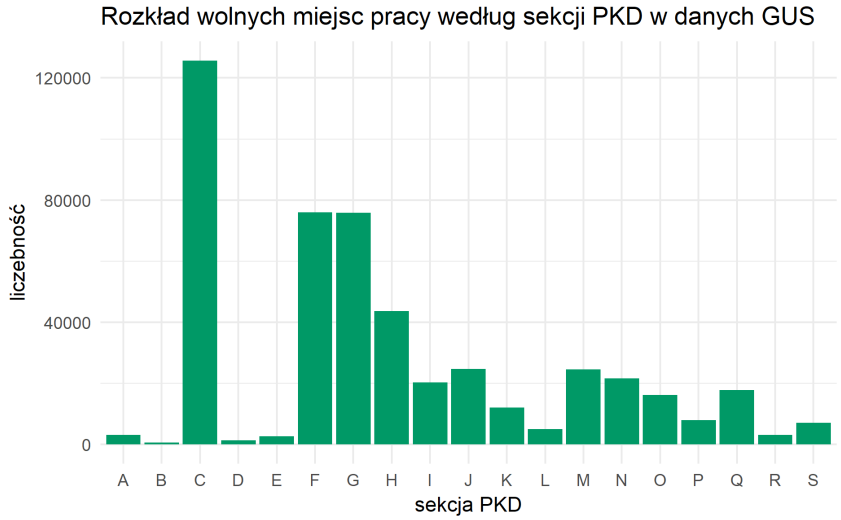
[1] Należy wejść na wyszukiwarkę zawodów http://psz.praca.gov.pl/rynek-pracy/bazy-danych/klasyfikacja-zawodow-i-specjalnosci/wyszukiwarka-opisow-zawodow



Rysunek 3.2. Rozkład liczby wolnych miejsc pracy (w tys.) ze względu na zawód dla danych pochodzących z Badania Popytu na Pracę dla całego 2017 roku

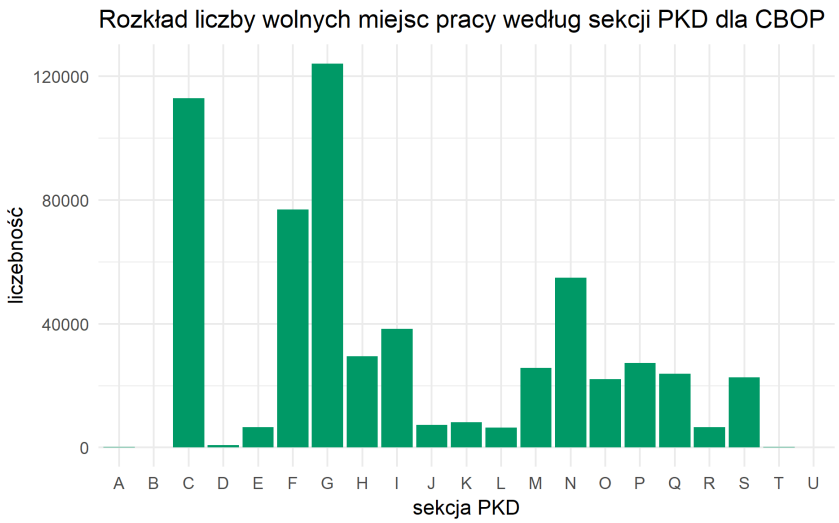
Źródło: Opracowanie własne. Oznaczenia: 1 = Przedstawiciele władz publicznych, wyżsi urzędnicy i kierownicy, 2 = Specjaliści, 3 = Technicy i inny średni personel, 4 = Pracownicy biurowi, 5 = Pracownicy usług i sprzedawcy, 6 = Rolnicy, ogrodnicy, leśnicy i rybacy, 7 = Robotnicy przemysłowi i rzemieślnicy, 8 = Operatorzy i monterzy maszyn i urządzeń, 9 = Pracownicy wykonujący prace proste.

A, B, D oraz E przyjmują minimalne wartości (w danych CBOP sekcje A i B oraz T i U wykazują bardzo małe liczebności, których nie widać na wykresie – nie są to braki danych). W obu zbiorach dużą liczbą wolnych miejsc pracy charakteryzują się sekcje F oraz G.



Rysunek 3.3. Rozkład liczby wolnych miejsc pracy ze względu na sekcję PKD dla danych pochodzących z Badania Popytu na Pracę

Źródło: Opracowanie własne.



Rysunek 3.4. Rozkład liczby wolnych miejsc pracy ze względu na sekcję PKD dla danych CBOP

Źródło: Opracowanie własne.

Ponadto jak widać na wykresie 3.3 w badaniu prowadzonym przez Główny Urząd Statystyczny wyłączone zostały dwie sekcje PKD:

- sekcja T: Gospodarstwa domowe zatrudniające pracowników; gospodarstwa domowe produkujące wyroby i świadczące usługi na własne potrzeby,
- sekcja U: Organizacje i zespoły eksterytorialne.

Z tego względu podczas analizy zbiór ofert z Centralnej Bazy Ofert Pracy został pomniejszony o oferty z powyższych sekcji Polskiej Klasyfikacji Działalności, które nie są uwzględniane w danych z Badania Popytu na Pracę.

Badanie to prowadzone jest kwartalnie, dlatego w zbiorze ofert pracy urzędów pracy została dodana zmienna `kwartal`, która zawiera informacje na koniec którego kwartału dana oferta była aktualna. Jeśli oferta była aktualna przez dłużej niż jeden kwartał została zduplikowana tak, by uzyskać informację o aktualnych ofertach na koniec każdego kwartału 2017 roku. Dodatkowo do zbioru została dodana zmienna *sekcja_pkd*, która informuje o sekcji rodzaju działalności.

3.4 Wyniki kalibracji

Kalibracja jest metodą polegającą na korygowaniu wag wynikających ze schematu losowania próby, tak by zredukować obciążenie wynikające z braków odpowiedzi. Ze względu na fakt, iż

dane CBOP są danymi internetowymi i nie posiadają wag, dodana została zmienna waga przyjmująca wartości równe wartościom zmiennej liczba_miejsc_og, ponieważ wagi kalibrowane były do liczby wolnych miejsc pracy według Głównego Urzędu Statystycznego. Ponadto na jedno ogłoszenie o pracę może przypadać jedno lub więcej wolnych miejsc pracy.

Zbiór został ograniczony jedynie do ofert pracy aktualnych na koniec danego kwartału 2017 roku. Według definicji Głównego Urzędu Statystycznego wybrane zostały jedynie te oferty, dla których dzień przypadający na zakończenie kwartału mieścił się w przedziale, który stanowiły daty ważności danej oferty pracy. Oferty, które były ważne na koniec więcej niż jednego kwartału zostały zduplikowane. W związku z tym zaktualizowany zbiór danych liczy 194 271 ofert pracy aktualnych na koniec danego kwartału.

Jako zmienne pomocnicze wybrane zostały zmienne kod_woj oraz sekcja_pkd. W początkowym zbiorze zmienna kod_woj posiadała 7,65 % braków danych. Ze względu na fakt, iż zmienne pomocnicze nie mogą posiadać braków danych, podczas kalibracji ze zbioru danych zostały wybrane tylko te wiersze, w których nie występowały braki danych w zmiennej kod_powiatu. W danych CBOP występuje zależność pomiędzy brakiem danych w zmiennej dotyczącej powiatu oraz województwa, z tego względu również niemożliwe było uzupełnienie braków w województwach na podstawie powiatów i zbiór został pomniejszony.

Przed kalibracją ze zbioru CBOP zostały wybrane tylko te sekcje działalności, które publikowane były przez GUS w Badaniu Popytu na Pracę. Ponadto wybrane zostały tylko te wiersze, dla których liczba miejsc ogółem w przedsiębiorstwie mieści się w przedziale od 0 do 100. Dodatkowo sekcje *A, B, D* oraz *R* zostały złączone oraz nazwane jako sekcje *inne* ze względu na niską oraz znacznie odbiegającą od pozostałych sekcji reprezentację w danych.

Kalibracja została zastosowana w trzech wariantach, gdzie zmiennymi pomocniczymi były kolejno:

1. województwo,
2. sekcja działalności,
3. województwo oraz sekcja działalności.

Kalibracja została wykonana w pakiecie *survey* za pomocą funkcji *calibrate*. Podczas wyznaczania wag kalibracyjnych, dla każdego z trzech wariantów, wykorzystana została liniowa funkcja odległości.

Tabela 3.6. Tabela dla sum wag kalibracyjnych i wartości globalnych dla populacji dla I kwartału 2017 roku (dla województw)

województwo	suma wag kalibracyjnych	wartości globalne
dolnośląskie	11 196	11 196
kujawsko-pomorskie	4 651	4 651
lubelskie	3 633	3 633
lubuskie	4 699	2 699
łódzkie	6 661	6 661
małopolskie	10 925	10 925
mazowieckie	25 875	25 875
opolskie	3 721	3 721
podkarpackie	3 333	3 333
podlaskie	1 221	1 221
pomorskie	6 236	6 236
śląskie	14 626	14 626
świętokrzyskie	2 994	2 994
warmińsko-mazurskie	2 252	2 252
wielkopolskie	13 254	13 254
zachodniopomorskie	5 226	5 226

Źródło: Opracowanie własne.

W pierwszym przypadku (gdzie zmienną pomocniczą są województwa) jak widać w tabeli A.11 wagi kalibracyjne w_i sumują się do wartości globalnych – do liczby wolnych miejsc pracy publikowanych przez Główny Urząd Statystyczny. Jeśli wagi w_i odtwarzają wartości znane dla całej zbiorowości, powinny, zgodnie z ideą kalibracji, zniwelować negatywny wpływ braków danych. W celu sprawdzenia jaki wpływ na braki danych w tym zbiorze miała kalibracja, obliczone zostały odsetki dla wag w_i oraz dla porównania dla wag d_i w podziale na zawody. Wyniki przedstawia tabela A.15.

Tabela 3.7. Odsetki dla wag d_i oraz w_i w zależności od zawodu dla I kwartału 2017 roku (dla województw)

główna grupa zawodu	suma wag d_i	suma wag w_i	odsetki dla wag d_i	odsetki dla wag w_i
1: *przedstawiciele władz publicznych, wyżsi urzędnicy i kierownicy*	874,00	929,91	0,75	0,78
2: *specjaliści*	6345,00	6616,53	5,47	5,58
3: *technicy i inny średni personel*	7892,00	8271,59	6,81	6,98
4: *urzędnicy biurowi*	8381,00	8905,90	7,23	7,52
5: *pracownicy usług i sprzedawcy*	22403,00	22898,61	19,33	19,32
7: *robotnicy przemysłowi i rzemieślnicy*	31147,00	31653,63	26,87	26,71
8: *operatorzy i monterzy maszyn i urządzeń*	15959,00	15796,05	13,77	13,33
9: *pracownicy wykonujący prace proste*	22909,00	23430,78	19,76	19,77

Źródło: Opracowanie własne.

Jak widać odsetki dla wag kalibracyjnych prawie wcale nie różnią się od tych dla wag wyjściowych – w tym przypadku działanie kalibracji nie przyniosło oczekiwanych rezultatów, dlatego w drugim kroku przeprowadzona została kalibracja, dla której zmienną pomocniczą była sekcja PKD. Na tym etapie wagi również zostały obliczone prawidłowo, ich suma odzwierciedla

wartości dla populacji.

Tabela 3.8. Tabela dla sum wag kalibracyjnych i wartości globalnych dla populacji dla I kwartału 2017 roku (dla sekcji PKD)

sekcja PKD	suma wag kalibracyjnych	wartości globalne
C	29 871	29 871
E	654	654
F	18 784	18 784
G	18 124	18 124
H	9 580	9 580
I	5 729	5 729
J	6 219	6 219
K	2 968	2 968
L	1 252	1 252
M	6 067	6 067
N	6 351	6 351
O	3 657	3 657
P	1 107	1 107
Q	4 308	4 308
S	1 975	1 975
inne	1 857	1 857

Źródło: Opracowanie własne.

Tabela 3.9. Odsetki dla wag d_i oraz w_i w zależności od zawodu dla I kwartału 2017 roku (dla sekcji PKD)

główna grupa zawodu	suma wag d_i	suma wag w_i	odsetki dla wag d_i	odsetki dla wag w_i
1: *przedstawiciele władz publicznych, wyżsi urzędnicy i kierownicy*	874,00	1007,87	0,75	0,85
2: *specjaliści*	6345,00	9131,90	5,47	7,71
3: *technicy i inny średni personel*	7892,00	9074,19	6,81	7,66
4: *urzędnicy biurowi*	8381,00	8747,70	7,23	7,38
5: *pracownicy usług i sprzedawcy*	22403,00	21339,12	19,33	18,01
7: *robotnicy przemysłowi i rzemieślnicy*	31147,00	32504,45	26,87	27,43
8: *operatorzy i monterzy maszyn i urządzeń*	15959,00	17370,70	13,77	14,66
9: *pracownicy wykonujący prace proste*	22909,00	19327,07	19,76	16,31

Źródło: Opracowanie własne.

Jak widać dla przypadku drugiego odsetki wag w_i kształtują się lepiej niż w poprzednim przypadku, gdzie zmienną pomocniczą było województwo.

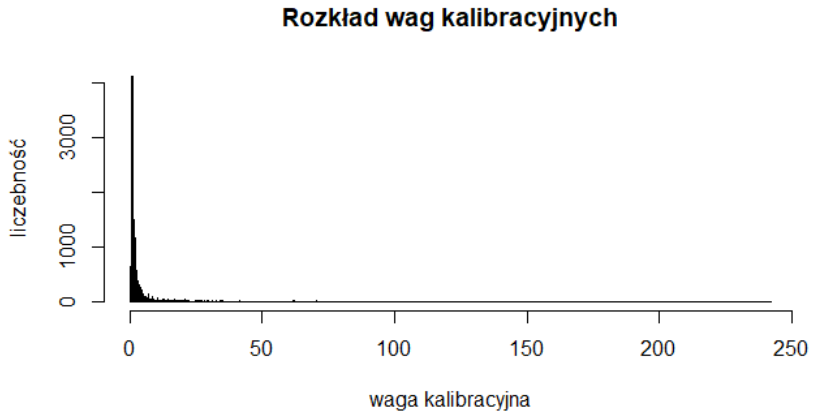
W ostatnim wariancie, w którym wykorzystywane były dwie zmienne jako zmienne pomocnicze wagi również odzwierciedlają liczbę wolnych miejsc pracy wynikających z Badania Popytu na Pracę. Analizując wyniki ostatniego wariantu w tabeli A.15 można zauważyć, iż nie różnią się one znacząco od odsetków z drugiego przypadku kalibracji, gdzie występowała jedna zmienna pomocnicza. Z tego względu wykorzystanie sekcji PKD i województwa razem jako zmiennych pomocniczych może być równorzędne z wykorzystaniem jedynie sekcji rodzaju działalności.

Podsumowując z wykorzystanych trzech wariantów najbardziej korzystnym przypadkiem jest wykorzystanie sekcji Polskiej Klasyfikacji Działalności oraz województwa jako zmienne po-

Tabela 3.10. Odsetki dla wag d_i oraz w_i w zależności od zawodu dla I kwartału 2017 roku (dla województwa oraz sekcji PKD)

główna grupa zawodu	suma wag d_i	suma wag w_i	odsetki dla wag d_i	odsetki dla wag w_i
1: *przedstawiciele władz publicznych, wyżsi urzędnicy i kierownicy*	874,00	1050,03	0,75	0,89
2: *specjaliści*	6345,00	9262,02	5,47	7,82
3: *technicy i inny średni personel*	7892,00	9323,84	6,81	7,87
4: *urzędnicy biurowi*	8381,00	8967,96	7,23	7,57
5: *pracownicy usług i sprzedawcy*	22403,00	21246,03	19,33	17,93
7: *robotnicy przemysłowi i rzemieślnicy*	31147,00	32298,06	26,87	27,26
8: *operatorzy i monterzy maszyn i urządzeń*	15959,00	17051,25	13,77	14,39
9: *pracownicy wykonujący prace proste*	22909,00	19303,81	19,76	16,29

Źródło: Opracowanie własne.



Rysunek 3.5. Rozkład wag kalibracyjnych w_i dla zmiennych pomocniczych: sekcja PKD oraz województwo

Źródło: Opracowanie własne.

mocnicze. Kalibrowanie jedynie według sekcji PKD daje identyczne rezultaty, jak w połączeniu jej ze zmienną informującą o województwie. Jednak odsetki wag kalibracyjnych nie nieznacznie różnią się od odsetków wag wyjściowych, dlatego kalibracja według podanych zmiennych pomocniczych nie przynosi oczekiwanych rezultatów.

3.5 Podsumowanie

Po przeanalizowaniu wszystkich aspektów dane z Powiatowych Urzędów Pracy mogą być dobrym źródłem do badania oraz opisu rynku pracy. Oczywiście nie jest to zbiór pozbawiony błędów nielosowych – występują błędy w kodowaniu zmiennych, kodowaniu braków danych oraz same braki danych. Z wybranych do analizy danych jedna zmienna (informująca o wyma-

ganym stażu pracy) charakteryzuje się 80% brakiem danych, a zmienna informująca o wielkości przedsiębiorstwa zawiera 50% danych, natomiast pozostałe zmienne nie posiadają lub posiadają, jednak mniej niż 20%.

Podczas analizy, za pomocą współczynnika V Cramera badano wpływ różnych zmiennych na publikowanie ofert z danego zawodu. Jedynie w przypadku dwucyfrowych kodów zawodów zmienna dotycząca sekcji PKD wykazała umiarkowaną korelację. Dla pozostałych zmiennych oraz głównych grup zawodów istnieje słaba zależność. Zbiór danych Centralnej Bazy Ofert Pracy zawiera wszystkie kody PKD oraz zawodu, w przeciwieństwie do danych z Badania Popytu na Pracę, gdzie nie uwzględnia się dwóch sekcji PKD (T oraz U).

Porównując rozkłady wolnych miejsc pracy ze względu na sekcję rodzaju działalności można zauważyć podobieństwo rozkładu dla niektórych sekcji, na przykład w obu przypadkach sekcje A, B, D oraz E charakteryzują się bardzo niskimi liczebnościami wolnych miejsc pracy, a sekcje C, F oraz G z kolei znacznie wyższymi niż pozostałe sekcje (w obu przypadkach).

Zastosowana kalibracja z wykorzystaniem zmiennej pomocniczej odnoszącej się do PKD nieznacznie zredukowała błąd pokrycia w CBOP. Wynika to głównie z faktu dość umiarkowanej zależności między zawodem, a sekcją PKD. Aby skorygować ten błąd należałoby poszukać innych zmiennych, silniej skorelowanych lub uzyskać dostęp do jednostkowych danych z badania Popytu na Pracę aby zintegrować zbiór z badania reprezentacyjnego oraz internetowego źródła danych.

Centralna Baza Ofert Pracy zawierała informacje dotyczące opisu zawodu, jak i przyporządkowanego kodu. Dodatkowo, pracownicy PUPy weryfikują poprawność opisów oraz kodów co może stanowić zbiór uczący i testowy na potrzeby uczenia maszynowego. Ten aspekt jest o tyle istotny, że internetowe źródła danych takie jak pracuj.pl czy OLX nie posiadają przypisanych zawodów zgodnie z KSiZ. W kolejnym rozdziale poruszony zostanie aspekt klasyfikacji ofert pracy do określonych zawodów z wykorzystaniem bazy CBOP oraz Bilansu Kapitału Ludzkiego (moduł ofert pracy).

Rozdział 4

Wykorzystanie uczenia maszynowego do klasyfikacji zawodów w portalach internetowych (Krzysztof Marcinkowski)

4.1 Cel rozdziału

Celem rozdziału jest zbudowanie modelu uczenia maszynowego zdolnego do klasyfikacji ofert pracy zgodnie z kategoriami Klasyfikacji Zawodów i Specjalności. Do przeprowadzenia takiej analizy użyty zostanie również *text mining*, czyli operacje związane z przetwarzaniem tekstu. W rozdziale przeprowadzone zostaną próby stworzenia modeli na podstawie danych z BKL i CBOP. Następnie zostanie on zaaplikowany do danych pochodzących z serwisu internetowego OLX. Sprawdzone zostaną dwa modele: wielomianowej regresji logistycznej LASSO oraz Naiwnego Bayesa. Modele zostaną ocenione na podstawie macierzy klasyfikacji oraz takich miar jak: wrażliwość, precyzja oraz miara F1.

4.1.1 Uczenie Maszynowe

Informatyk Tom Mitchell zaproponował formalną definicje uczenia maszynowego (ang. *machine learning* która stwierdzała że maszyna uczy się, gdy jest w stanie wykorzystać swoje doświadczenie, by polepszyć swoje wyniki w podobnych przypadkach w przyszłości. Podstawowy proces uczenia jest podobny, składa się on z następujących części:

- Przechowywania danych;

- Abstrakcji;
- Generalizacji;
- Ewaluacji.

Pierwsza z nich, czyli przechowywanie danych i obserwacji w taki sposób, by były one gotowe do przetwarzania. Abstrakcja zawiera w sobie tłumaczenie przechowywanych danych na reprezentacje pewnych zjawisk i koncepty. Proces generalizacji opisuje czynności jakie trzeba podjąć by móc zdobytą w poprzednim kroku wiedze zastosować w przyszłości, na przypadkach które są podobne, jednakże nie identyczne. Ostatnim krokiem jest ewaluacja modelu, czyli sprawdzenie wyników skuteczności stworzonego modelu. W praktyce wyróżnia się 5 kroków uczenia maszynowego:

- **Zbieranie danych**: w tym kroku dane są zbierane i magazynowane;
- **Eksploracyjna analiza danych i ich obróbka**: Jakość modeli zależy od tego na jakich danych modele będą szkolone. Ten krok zawiera w sobie sprawdzenie zależności występujących w danych, uporządkowaniu danych oraz dostosowaniu ich formy do wymagań modelu;
- **Trenowanie modelu**: w tym kroku wybrany model uczenia maszynowego trenowany jest na danych. Tworzony jest model który reprezentuje dane;
- **Testowanie modelu**: Zbudowany wcześniej model testowany jest na danych testowych;
- **Udoskonalanie modelu**: Jeżeli model nie osiągnął w poprzednim kroku wystarczających rezultatów to niezbędne jest zastosowanie bardziej zaawansowanych strategi doboru parametrów lub też zmiany samego modelu (Lantz, 2013);

Pojawiający się wyżej zbiór treningowy i testowy wiążę się z procesem sprawdzianu krzyżowego (ang *cross-validation*). Jest to symulowanie testu na przyszłych danych dzięki pomijaniu części danych historycznych podczas dopasowywania modelu. Model budowany jest na zbiorze treningowym, dlatego też jest on największy i stanowi w większości przypadków około 80% danych. Następnie model jest testowany na pozostałych 20% danych, czyli zbiorze testowym. Proces ten pozwala zapobiec takim sytuacją jak przeuczenie i niedouczenie modelu. Czyli kolejno sytuacji w której model opisuje dane zbyt szczegółowo oraz sytuacji w której model opisuje dane zbyt ogólnie (Conway & White, 2012).

4.1.2 Infrastruktura

Przeprowadzanie obróbki, transformacji i wreszcie modelowania danych było możliwe dzięki wykorzystaniu Laboratorium Interdyscyplinarnych Badań Naukowych UEP (InnoUEP). Obiekt ten powstał w ramach Programu Operacyjnego Innowacyjna Gospodarka. Celem tego przedsięwzięcia jest rozwój polskiej gospodarki bazując o innowacyjne przedsiębiorstwa. Jak można przeczytać na stronie Uniwersytetu Ekonomicznego w Poznaniu: *'Dla osiągnięcia celu zwiększenia roli nauki, konieczne jest zbudowanie infrastruktury pozwalającej na prowadzenie badań nieodstających zakresem i wnikliwością od badań prowadzonych w ośrodkach zagranicznych'*. Laboratorium to pozwoli na prowadzenie nowoczesnych badań z zastosowaniem zaawansowanych technologii i stworzy możliwość współpracy z międzynarodowymi naukowymi sieciami teleinformatycznymi. Zapewni ono stały i bezpieczny dostęp do zaawansowanej infrastruktury informatycznej. Na InnoUEP składa się 7 specjalistycznych pracowni:

- Pracownia Wirtualnej i Wzbogaconej Rzeczywistości;
- Pracownia Internetu Rzeczy;
- Pracownia Internetu Następnej Generacji;
- Pracownia Energetyki Odnawialnej;
- Pracownia Platform i Usług Korporacyjnych;
- Pracownia Oceny Oddziaływania na Środowisko;
- Pracownia Badań Konsumenckich;

Ponadto pracownie połączone są infrastrukturą sieciową. Współdzielą również segment obliczeniowy i przechowywania danych: Common Operational Reasearch Enviroment(CORE).

4.2 Zbiory danych wykorzystane do uczenia maszynowego

4.2.1 Proces przetwarzania danych

Dane na podstawie których zbudowane zostały modele pochodzą z Badania Kapitału Ludności z 2014 roku oraz Centralnej Bazy Ofert Pracy z 2017 roku. Zbiór CBOP z 2017 roku to 12 312 unikalnych nazw zawodów (obserwacji), natomiast rok 2014 reprezentowany jest przez 7 820

unikalnych nazw zawodów (obserwacji). Większy zbiór posiadał zmienną opisującą kategorie pracę według metodologii BKL, która potem została przydzielona do odpowiedniej kategorii Klasyfikacji Zawodów i Specjalności. Również do kategorii Klasyfikacji Zawodów i Specjalności przyporządkowane zostały oferty z roku 2014, tym razem na podstawie nagłówków, metod przetwarzania tekstu oraz indywidualnie dla każdej oferty.

Sam tekst występujący w ofercie został obrobiony za pomocą algorytmów przetwarzania tekstu. W pierwszej kolejności dokonano tokenizacji. Krok też zmienia strukturę obiektu, tak że jego rezultatem jest: unikalny numer identyfikujący dokument oraz każde pojedyncze słowo występujące w tym dokumencie, przedstawione w osobnym wierszu. Tak więc każda oferta została przetworzona na tyle wierszy w tabeli ile słów w niej występowało. Kolejnym krokiem było usunięcie słów nie wnoszących nic do dalszej analizy (ang. *stopwords*) (Silge & Robinson, 2019). Są to słowa występujące często w tekście, jednakże nie definiujące go, dla przykładu: 'albo' czy też 'aczkolwiek'. Usunięte zostały również słowa których długość jest mniejsza niż 3.

Ostatecznie dokonano *stematyzacji* oraz *lematyzacji*. Stematyzacja polega na wyodrębnieniu z wybranego wyrazu tak zwanego rdzenia. Jest to ta cześć, która nie ulega zmianie podczas odmiany przez rodzaje czy przyimki. Lematyzacja natomiast, oznacza proces sprowadzenie kilku wyrazów będących odmianą danego zwrotu do wspólnej postaci. Oba te procesy są niezbędne by móc wychwycić poszczególne wyrazy, mimo tego że występują w innej formie (Jivani i in., 2011).

Dzięki tym zabiegom analiza oparta będzie na bezokolicznikach i mianownikach, ponadto nie będzie zawierała słów o małym znaczeniu, jak na przykład spójniki. Zabieg ten był oparty na dane pochodzące z repozytorium Marcina Kosińskiego (https://github.com/MarcinKosinski/trigeR5/blob/master/dicts/polimorfologik-2.1.zip). Ostatecznie usunięte zostały polskie znaki oraz w miarę możliwości poprawione błędy gramatyczne. Wszystkie duże litery zostały sprowadzone do małych.

Podział na część uczącą i testową będzie oparty o proporcje 8:2, to znaczy że 80% danych posłuży do budowy modelu, a 20% do jego weryfikacji. Ponadto podczas procesu podziału zbiór zostanie wymieszany. Dane zostały przygotowane do modelowania poprzez stworzenie obiektu *Document Term Matrix*. Opiera się on na macierzy w której w wierszach występuje oznaczenie dokumentu oraz jego kategorii, natomiast kolumny reprezentują konkretne słowa, w komórkach znajduję się ich częstość w konkretnym dokumencie. Tabela 4.1 przedstawia przykład takiego obiektu.

Tabela 4.1. Przykładowy obiekt typu *Document Term Matrix*

ID	oferta	praca	fryzjer	monter	koscian
1	1	1	1	0	0
2	1	1	0	1	1

Źródło: Opracowanie własne

Kolumna ID oznacza unikalny identyfikator oferty. Kolejne kolumny są wyrazami które pojawiają się w wszystkich ofertach. Oferta pierwsza dotyczy oferty pracy jako fryzjer. Natomiast drugi wierszy odwołuje się do oferty pracy jako monter w Kościanie. Opisane wyżej operacje zostały wykonane w dwóch pakietach języka *R*: `tm` (Feinerer, Hornik & Meyer, 2008) oraz `tidytext` (Silge & Robinson, 2016). Ponadto przetwarzanie danych wykonane było za pomocą pakietu `tidyverse` (Wickham, 2017), wizualizacje powstały za pomocą pakietu `ggplot2` (Wickham, 2016).

4.2.2 Oferty pracy z portalu OLX

Dane na których dokonana będzie klasyfikacja pochodzą z serwisu internetowego OLX. Dane były pobierane za pomocą *webscrapingu*. Proces ten polega na automatycznym pobieraniu danych ze stron internetowych przy użyciu odpowiednich skryptów. Modele będą bazować na tytułach ofert pracy. Okres ważności oferty definiowany jest dwoma zmiennymi: data utworzenia oferty oraz data automatycznej dezaktualizacji oferty. Domyślnie okres ten trwa miesiąc,oczywiście możliwe jest przedłużenie okresu ważności. Dane poddane klasyfikacji zostały przefiltrowane tak, by ich okres ważności przypadał na ostatni miesiąc każdego kwartału. Pozwoliło to na wyodrębnienie 411 787 wpisów.

W celu wyeliminowania duplikatów sprawdzone zostały unikalne identyfikatory ofert oraz ich status. Ponadto każda oferta posiadała zdefiniowaną kategorie. W danych znajdowało się 30 unikalnych identyfikatorów kategorii pracy. Każdy z nich określa pewny dział rynku pracy którego oferta dotyczy. Identyfikatory te nie pokrywają się z kategoriami Klasyfikacji Zawodów i Specjalności. Kategoriami występującymi w danych z portalu OLX były między innymi:

- Edukacja
- IT / telekomunikacja
- Przedstawiciel handlowy

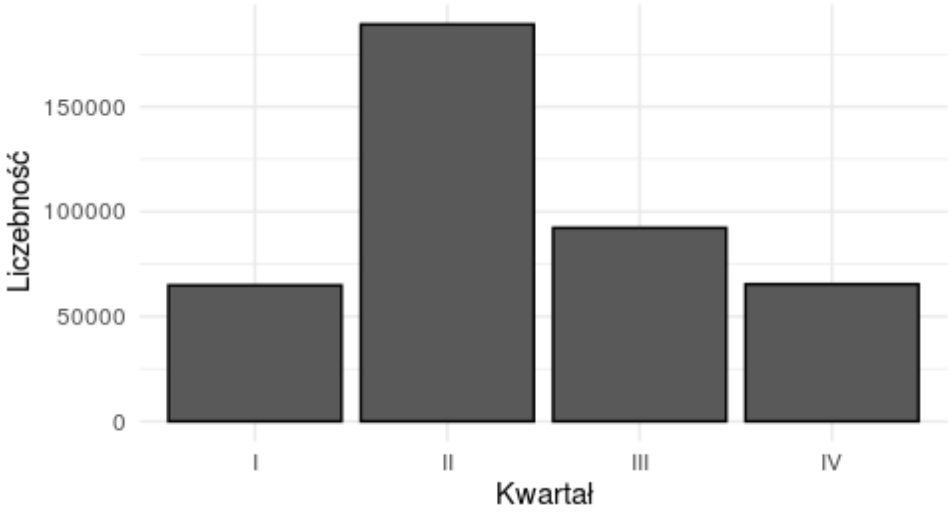
- Kierowca / kurier
- Pozostałe oferty pracy
- Opiekunki dla dzieci
- Zdrowie i opieka
- Obsługa klienta / call Center
- Ochrona

Każda oferta serwisu internetowego przyporządkowana była do konkretnego województwa. Kodowanie województw wymagało jednak dodatkowego przekształcenia, z wewnętrznego systemu OLX, na system kodowania Głównego Urzędu Statystycznego. Oferty zostały poddane tokenizacji, następnie stematyzacji i lematyzacji. Usunięte zostały słowa nieistotne dla analizy(*stopwords*). Utworzony został *Document Term Matrix*, i to właśnie na tym obiekcie przeprowadzona zostanie klasyfikacji. Ponadto, kolumny oznaczające słowa występujące w danych z portalu internetowego, zostały przefiltrowane tak, by pozostawić tylko te które występowały w danych zbioru treningowego. Następnie, kolumny które występował w zbiorze budującym model, zostały dodane do danych OLX. Kroki te były niezbędne w celu zachowania wymaganych przez modele wymiarów macierzy. Tak więc ostatecznym rezultatem były dwa obiekty typu: *Document Term Matrix* posiadające taką samą ilość kolumn.

4.2.3 Eksploracyjna analiza danych

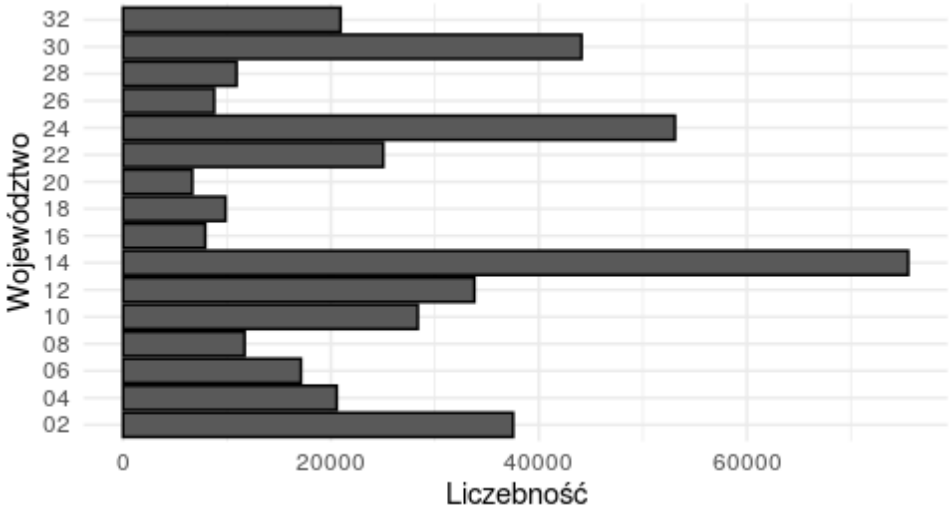
Przefiltrowane dane z serwisu internetowego to 411 787 oferty pracy. Zdecydowana większość ofert była aktualna podczas końca drugiego kwartału (rysunek: 4.1). Najmniejszą liczbę ofert można zaobserwować na początku i końcu roku.

Rysunek 4.2 przedstawia rozkład liczebności ofert pracy według województw. Oś X reprezentuje łączną liczbę ofert w danych województwie, natomiast oś Y przedstawia numer województwa. Zdecydowana większość ofert pacy pochodzi z województwa Mazowieckiego. Ponadto duża liczba ofert charakteryzuje takie województwa jak: Wielkopolskie, Śląskie czy Dolnośląskie. Regiony w których liczba ofert jest najmniejsza to: województwo Podlaskie czy też Opolskie.



Rysunek 4.1. Rozkład liczebności danych OLX według kwartałów

Źródło: Opracowanie własne na podstawie danych z portalu internetowego.



Rysunek 4.2. Rozkład liczebności danych OLX według województw

Źródło: Opracowanie własne na podstawie danych z portalu internetowego.

4.3 Wybrane algorytmy uczenia maszynowego

4.3.1 Regresja logistyczna

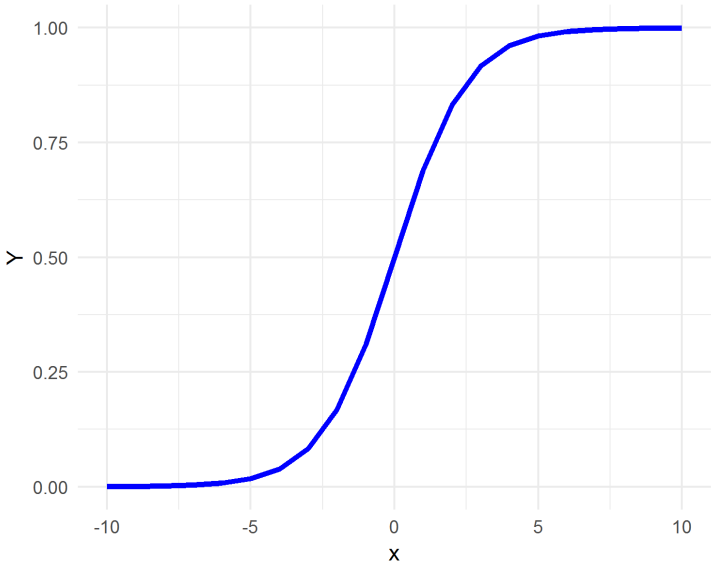
Regresja logistyczna charakteryzuje się zmienną zależną która jest dyskretna oraz zmiennymi niezależnymi które mogą mieć charakter zarówno ciągły jak I skokowy. W najbardziej podstawowej formie metoda ta pozwala na prognozowanie zmiennej binarnej, jest to wtedy binarna

regresja logistyczna. Jednakże jeżeli prognozowana jest zmienna przybierająca więcej niż dwie różne wartości, wtedy jest to wielomianowa regresja logistyczna.

Binarna regresja logistyczna określana jest wzorem:

$$P(Y) = \frac{1}{1 + e^{-(b_0 + b_1 x_{1i})}} \tag{4.1}$$

gdzie $P(Y)$ jest prawdopodobieństwem zajścia zdarzenia Y. Zmienna predykcyjna jest oznaczana przez x_1. Natomiast b_1 oznacza współczynniki regresji przypisane do zmiennej x_1. Należy zauważyć że wartość $P(Y)$ przyjmuje wartości z przedziału od (0,1) (Field, Miles & Field, 2012).



Rysunek 4.3. Wykres funkcji regresji logistycznej

Źródło: Opracowanie własne.

Model regresyjny może być zapisany wzorem:

$$ln\frac{\hat{p}(2|x)}{1 - \hat{p}(2|x)} = \alpha + \beta^T x \tag{4.2}$$

Rozwiązanie modelu (4.2) pozwala na wyodrębnienie dwóch estymatorów:

- $$\hat{p}(2|x) = \frac{exp(\alpha + \beta^T x)}{1 + exp(\alpha + \beta^T x)} \tag{4.3}$$
- $$\hat{p}(1|x) = \frac{1}{1 + exp(\alpha + \beta^T x)} \tag{4.4}$$

Estymatory (4.3) oraz (4.4) należą do przedziału (0,1). Estymatory pokazują kierunek zgodnie z którym jedno z prawdopodobieństw rośnie do 1, natomiast drugie spada do 0. Obserwacja x zostanie przydzielona do klasy 2 jeżeli wzór (4.2) będzie dodatni, oraz do klasy 1 kiedy wyrażenie to będzie ujemne.

Na rysunku 4.3 przedstawiony jest charakterystyczny wykres regresji logistycznej. Estymatory wskazują kierunek wzdłuż którego jedno z prawdopodobieństw wzrasta do 1, a drugie spada do 0. Parametry modelu estymowane są na podstawie próby, m.in. za pomocą metody największej wiarygodności. Metoda ta polega na maksymalizacji funkcji wiarygodności, w przypadku modelu binarnego stosuje się model dwumianowym, w przypadku większej ilości zmiennych: wielomianowym (Koronacki & Ćwik, 2008). Służy do tego wzór: (4.5).

$$\prod_{i=1}^{n} \hat{p}(2|x_i)^{y_i} \hat{p}(1|x_i)^{1-y_i}, \tag{4.5}$$

Maksymalizacja ze względu na parametry α oraz β przebiega iteracyjnie. Reguła dyskryminacyjna wybiera większą z wartości (4.3) i (4.4) i na podstawie tego wyboru przydziela obserwacje x. Model (4.2) można uogólnić na przypadek wielomianowy gdzie g to liczba zmiennych objaśniających:

$$ln\frac{\hat{p}(g-1|x)}{\hat{p}(g|x)} = \beta_{(g-1)0} + \beta_{g-1}^{T} x \tag{4.6}$$

Model (4.6) daje:

$$\hat{p}(g|x) = \frac{1}{1+\sum_{l=1}^{g-1} exp(\beta_{l0} + \beta_l^T x)} \tag{4.7}$$

Logarytm funkcji wiarygodności definiowany jest następującym wzorem:

$$\sum_{i=1}^{n} ln\hat{p}_{k_i}(x_i;\theta) \tag{4.8}$$

gdzie k_i to klasa i-tego elementu próby. Maksymalizacja funkcji względem parametru θ skutkuje znalezieniem rozwiązania zadania estymacji parametru oraz prawdopodobieństw (Koronacki & Ćwik, 2008).

Jakość modelu można oceniać bazując na mierze dopasowania jaką jest pseudo-R^2. Ta znana przede wszystkim z regresji liniowej miara, znajduje swoje zastosowanie również w regresji logistycznej. W tym przypadku definiowane jest jako proporcja odchylenia wyjaśniona

przez model:

$$R^2_{pseudo} = 1 - \frac{dev_\omega}{dev_{\omega_0}}, \tag{4.9}$$

gdzie dev_{ω_0} jest odchyleniem modelu ω_0: y $\sim$ 1, w którym za zmienną niezależną przyjmowana jest stała wartość (Koronacki & Ćwik, 2008).

W celu określenia skuteczności modelu regresji logistycznej wykorzystywane są dane zaobserwowane oraz za prognozowane. W tym celu wykorzystywana jest funkcja wiarygodności. Miara ta jest oparta na sumie prawdopodobieństw prognoz i faktycznych obserwacji. Miara ta wskazuje ile zmienności zostało niewytłumaczone po dopasowaniu modelu. Stad, jest to miara którą należy minimalizować. Ponieważ im większa wartość współczynnika prawdopodobieństwa, tym więcej niewyjaśnionych przez model obserwacji znajduje się w danych.

Regresja logistyczna powinna spełniać następujące założenia (Field i in., 2012):

- Istnieje linowa zależność między ciągłymi zmiennymi objaśniającymi a logarytmem szans;
- Obserwacje muszą być niezależne;
- W danych nie występuje silna współliniowość.

4.3.1.1 Regresja LASSO

Zakładając że N to zbiór par zmiennych zależnych i niezależnych: $(x_i, y_i)_{i=1}^N$, regresja LASSO szuka rozwiązania $(\hat{\beta}_0, \hat{\beta})$ do zadania optymalizacji:

$$\min \left(\frac{1}{2N} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 \right) \tag{4.10}$$

ze względu na:

$$\sum_{j=1}^{p} |\beta_j| \leq t \tag{4.11}$$

Warunek (4.11) może być zapisany jako $||\beta||_1 \leq t$. Granica t ogranicza sumę wartości bezwzględnych oszacowanych parametrów. Pozwala to na określenie jak dobrze model będzie dopasowany do danych. Wartość ta powinna być oszacowana poprzez procedurę sprawdzianu krzyżowego. Zmienne niezależne powinny zostać podane standaryzacji. Krok ten jednakże

może zostać pominięty jeżeli zmienne posiadają takie same jednostki (Tibshirani, Wainwright & Hastie, 2015).

Zakładając że kodowanie zmiennej binarnej będzie odbywało się za pomocą oznaczenia: (-1, +1), w oparciu o model binarnej regresji logistycznej (4.2), negatywna logarytmiczna funkcja wiarygodności LASSO będzie miała postać:

$$\frac{1}{N}\sum_{i=1}^{N}\log\left(1+e^{-y_i f(x_i;\beta_0,\beta)}\right)+\lambda||\beta||_1 \tag{4.12}$$

gdzie $f(x_i;\beta_0,\beta) := \beta_0+\beta^T x_i$. Dla każdej pary zmiennej objaśnianej i objaśniającej (x,y), iloczyn $yf(x)$ determinuje poprawność klasyfikacji. Dodatnia wielkość oznacza poprawną klasyfikacje, podczas gdy ujemna: niepoprawną klasyfikacje. Dla problemu klasyfikacji większej ilości klas, dla obserwacji danych przez: $(x_i,y_i)_{i=1}^N$ negatywna logarytmiczna funkcja wiarygodności poddana regularyzacji przybiera postać (Tibshirani i in., 2015):

$$-\frac{1}{N}\sum_{i=1}^{N}\log P\left(Y=y_i|x_i;\{\beta_0,\beta_k\}_{k=1}^{K}\right)+\lambda\sum_{k=1}^{K}||\beta_k||_1. \tag{4.13}$$

Zgodnie z dokumentacją pakietu H2O (LeDell i in., 2019), w wielomianowej regresji logistycznej LASSO prawdopodobieństwa są definiowane poprzez:

$$P(y=c|x)=\frac{e^{x^T\beta_c+\beta_{c0}}}{\sum_{k=1}^{K}\left(e^{x^T\beta_k+\beta_{k0}}\right)}. \tag{4.14}$$

Natomiast funkcja wiarygodności z karą jest definiowana poprzez wzór (4.15).

$$-\left[\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}\left(y_{i,k}\left(x_i^T\beta_k+\beta_{k0}\right)\right)-\log\left(\sum_{k=1}^{K}e^{x_i^T\beta_k+\beta_{k0}}\right)\right]+\lambda\left[\frac{(1-\alpha)}{2}||\beta||_F^2+\alpha\sum_{j=1}^{P}||\beta_j||_1\right] \tag{4.15}$$

gdzie β_c jest wektorem współczynników dla klasy c, oraz $y_{i,k}$ to k-ty element wektora binarnego będącego efektem przekształcenia zmiennej zależnej za pomocą przekształcenia *'one-hot encoding'*. Przekształcenie to, wymagane przez niektóre modele, polega na przekształceniu zmiennej zależnej, dla przykładu: $y_{i,k}$ jest równe 1 jeżeli i-ta obserwacja to 'k', w przeciwnym wypadku wynosi 0 (Nykodym, Hussami, Kraljevic, Rao & Wang, 2015).

Mając estymator $\hat{e}$, jednym ze sposobów redukcji jego obciążenia jest zastosowanie próby uczącej jako równocześnie00 próby testowej. Następuje to poprzez podział na dwa podzbiory: próbę uczącą oraz próbę testową. Wykorzystanie części informacji do stworzenia re-

guły klasyfikacyjnej może skutkować zawyżeniem wartości estymatora błędu. Problem może zostać rozwiązany poprzez wspomnianą wcześniej metodę sprawdzania krzyżowego(ang. cross-validation). Zakładając że τ_n^{-j} to próba trenująca τ_n z której usunięto obserwację $Z_j = (X_j, Y_j)$. Klasyfikator konstruowany jest na podstawie próby τ_n^{-j}. Następnym krokiem jest test na pojedynczej obserwacji Z_j. Procedura ta powtarzana jest n razy. Estymator kształtuje się następująco:

$$\hat{e}_{CV} = \frac{1}{n} \sum_{j=1}^{n} I(\hat{d}(X_j; \tau_n^{-j}) \neq Y_j) \tag{4.16}$$

Każda możliwa obserwacja próby jest w procedurze (4.16) wykorzystana jest do stworzenia klasyfikatora $\hat{d}$. (Krzyśko, Wołyński, Górecki & Skorzybut, 2008)

4.3.2 Naiwny Bayes

Należy założyć że każdej hipotezie przypisano prawdopodobieństwo a priori P(h). Określenie *a prori* oznacza, że przy obliczaniu tego prawdopodobieństwa nie były brane pod uwagę żadne, mogące wpływać na jego wartość obserwacje odnoszące się do danej dziedziny, której to dotyczą hipotezy. Branie pod uwagę takowych danych prowadzi do obliczenia prawdopodobieństwa hipotezy h jako prawdopodobieństwa warunkowego $P(h|D)$, gdzie D jest pewnym zestawem danych mogących wpływać na ocenę prawdopodobieństwa hipotez. W przypadku którym dane *D* nie wpływają na ocenę prawdopodobieństwa hipotezy *h*, czyli prawdopodobieństwo to jest od nich niezależne, to zachodzi równość: P(h|D) = P(h). Przedmiotem badania jest jednak sytuacja, gdy zaobserwowanie danych skutkuje rewizją wcześniej przyjętych prawdopodobieństw hipotez. Twierdzenie Bayesa określa jak taka rewizja powinna się odbyć (Cichosz, 2000).

Klasyfikator ten zakłada niezależność zmiennych objaśniających. Samo prawdopodobieństwo zajścia zdarzenia oblicza się poprzez podzielenie ilości zdarzeń kiedy miało ono miejsce przez całkowitą liczbę prób. Obliczenie prawdopodobieństwa dwóch niezależnych od siebie zdarzeń będzie wynikiem ich iloczynu. Jednakże jeżeli dwa wydarzenia są zależne od siebie, wtedy niezbędne jest wykorzystanie teorii Bayes'a. Uzyskania prawdopodobieństwa $P(A|B)$ czyli prawdopodobieństwa zajścia zdarzenia A, wiedząc że zdarzenie B już miało miejsce, jest możliwe przy wykorzystaniu następującego wzoru:

$$P(h|D) = \frac{P(h)P(D|h)}{P(D)}. \tag{4.17}$$

Równanie (4.17) opisuje twierdzenie Bayesa dla dowolnej hipotezy h oraz zbioru danych

D (Dangeti, 2017). Wzór ten do wyrażenia związku między prawdopodobieństwami hipotezy wykorzystuje dwa różne prawdopodobieństwa. Prawdopodobieństwo danych P(D) oznacza prawdopodobieństwo odnotowania danych D, które będą podstawą wnioskowania, bez dokonywania wcześniejszych założeń odnośnie poprawności żadnej z hipotez. Prawdopodobieństwo P(D|h) jest prawdopodobieństwem odnotowania tychże danych zakładając, że hipoteza h jest poprawna. Najczęściej celem obliczanie prawdopodobieństw jest stworzenie hipotezy najbardziej prawdopodobnej, bazując na wybranych danych. Znajdujące się w mianowniku we wzorze (4.17) prawdopodobieństwo danych P(D), nie jest zależne od hipotez, nie ma wpływu na wynik i stąd może być pominięte. Wystarczające jest uwzględnienie iloczynu występującego w liczniku: $P(h)P(D|h)$. Bezwzględna wartość $P(h|D)$ wymaga do jej obliczenia prawdopodobieństwa $P(D)$. Może być ono obliczone czyniąc pewne założenia i bazując na wyznaczeniu prawdopodobieństw $P(D|h)$ dla różnych hipotez. Zakładając że przestrzeń hipotez składa się ze skończonej liczby wykluczających się parami hipotez, które wyczerpują wszystkie możliwości, stąd:

$$(\forall h_1, h_2 \in H) P(h_1 \wedge h_2) = 0 \tag{4.18}$$

$$\sum_{h \in H} P(h) = 1 \tag{4.19}$$

Warunek (4.18) informuje, że żadne z dwóch hipotez nie są jednocześnie poprawne. Warunek (4.19) natomiast że poprawna jest przynajmniej jedna z nich. Po przyjęciu tych założeń można obliczyć prawdopodobieństwo całkowite:

$$P(D) = \sum_{h \in HH} P(h)P(D|h) \tag{4.20}$$

Nie tworząc założeń odnośnie przestrzeni H i sposobu reprezentowania jej elementów, zakładając jedynie że jest ich pewna skończona i ograniczona liczba, to zadanie modelowania można sformułować jako zadanie wyboru hipotezy z przestrzeni najlepszej w świetle zbioru przykładów T. Wyróżnia się dwa standardowe metody precyzowania, jakie warunki musi spełniać hipoteza, w celu uznania za najbardziej odpowiednią w sensie probabilistycznym:

- Zasada maksymalnej zgodności: trzeba wybrać hipotezę $h_{ML} \in H$, maksymalizującą warunkowe prawdopodobieństwo danych uczących,

- Zasada maksymalnego prawdopodobieństwa, trzeba wybrać hipotezę $h_{MAP} \in H$ o maksymalnym prawdopodobieństwie.

Do wykorzystania obu opisanych wyżej metod klasyfikacji probabilistycznej nie jest wymagane obliczenie prawdopodobieństwa zbioru uczącego $P(T)$. Jednakże klasyfikacja zgodna z zasadą maksymalnej zgodności, jak i maksymalnego prawdopodobieństwa wymaga wyznaczenia warunkowych prawdopodobieństw $P(T|h)$. Prawdopodobieństwo danych trenujących do trenowania pojęć jest prawdopodobieństwem znajdujących się w nich w postaci etykiet przykładów informacji determinującej ocenę hipotez. Prawdopodobieństwo P(T|h) informuje, jak bardzo prawdopodobne jest, aby elementy ze zbioru T miały takie etykiety kategorii, jakie rzeczywiście pojawiają się w zbiorze T, jeśli poprawna jest hipoteza h. Ponadto jeżeli zbiór uczący jest poprawny, czyli pojawiające się w nim etykiety wybranych danych są etykietami kategorii, jakie przydziela im pojecie c, to dla dowolnej hipotezy można stwierdzić że jeżeli hipoteza h zakwalifikuje wybranym danym ze zbioru T takie same etykiety, jakie pojawiają się w zbiorze treningowym, i jest to zbiór poprawny, to jego warunkowe prawdopodobieństwo równe jest 1. Etykiety występujące w zbiorze uczącym są pewne, niemożliwe są inne etykiety, co należy rozumieć, że jeżeli hipoteza h przypisuje choćby jednemu przykładowi ze zbioru T etykietę inną niż pojawiającą się w zbiorze treningowym, to prawdopodobieństwo tego zbioru, zakładając poprawność hipotezy, wynosi 0 (Cichosz, 2000).

Klasyfikatora bayesowskiego nie dotyczą problemy związane z obliczeniem prawdopodobieństw a posteriori. Skupia się on na reprezentowaniu hipotezy dzięki tworzonym na podstawie zbioru treningowego oszacowań wybranych prawdopodobieństw i klasyfikuje wybrane dane, wybierając kategorie najbardziej prawdopodobne, bazując na tych oszacowaniach. Naiwny klasyfikator bayesowski zakłada że przykłady scharakteryzowane są pewnym zestawem atrybutów $a_1, a_2, ..., a_n$. Bazując na zbiorze trenującym T, są obliczane prawdopodobieństwa konkretnych kategorii pojęcia docelowego c oraz prawdopodobieństwa wybranych wartości wszystkich atrybutów dla wybranych danych różnych kategorii. Zakładając że zbiór uczący stworzony jest z przykładów wybranych z dziedziny zgodnie z pewnym rozkładem prawdopodobieństwa i skutkiem tego jest fakt że: prawdopodobieństwa określane w trakcie uczenia się będą prawdopodobieństwami przy założeniu wyboru przykładów zgodnie z tymże rozkładem. Oszacowania dotyczą prawdopodobieństw $P_\Omega(c(x) = d)$ dla wszystkich kategorii $d \in C$ pojęcia docelowego c oraz $P_\Omega(a_i(x) = v|c(x) = d)$ dla wszystkich kategorii $d \in C$ oraz wartości $v \in A_i$ dla i = 1,2 ..,m gdzie A_i jest przeciwdziedziną atrybutu a_i. W celu obliczenia $P_{x\in\Omega}(c(x) = d)$ dla

wszystkich d liczone jest, ile znajduje się w zbiorze T przykładów konkretnych kategorii, następnie wartości te dzielone są przez liczbę wszystkich przykładów, można to zapisać w następujący sposób:

$$P_{x\in\Omega}(c(x)=d)=\frac{|\,T^d\,|}{|\,T\,|}. \tag{4.21}$$

Analogicznie, w celu ustalenia $P_\Omega(a_i(x)=v|c(x)=d)$ dla kategorii i wartości wszystkich cech można policzyć, jaka jest liczba przykładów każdej kategorii dla wybranych wartości każdego atrybutu i podzielić te liczbę przez ogólną liczbę przykładów tejże kategorii (Cichosz, 2000):

$$P_{x\in\Omega}(c(x)=d)=\frac{|\,T^d\,|}{|\,T\,|}. \tag{4.22}$$

Ostatecznie warto zaznaczyć że zdarzenia przeciwne definiowane są jako:

$$P(A)+P(A')=1. \tag{4.23}$$

Podczas obliczeń może wydarzyć się sytuacja w której wybrane słowa nie pojawiają się w niektórych kategoriach, a występują tylko w wybranych przypadkach. Powoduje to problemy przy obliczeniu iloczynów, dlatego też, by uniknąć takich sytuacji, dodawane są niewielkie liczby do tabeli występowania słów. Pozwala to na uniknięcie zerowych prawdopodobieństw,zabieg ten nazywany jest estymatorem Laplaca (ang. *Laplace estimator*) i najczęściej wynosi on 1. Dzięki temu zabiegowi pewne jest że każde słowo wystąpi przynajmniej raz i nie zakłóci obliczeń (Dangeti, 2017).

4.3.3 Metody oceny modeli klasyfikacji

Jednym z elementów modelowania jest weryfikacja jego poprawności. By to osiągnąć, niezbędne jest określenie jak często model dokonywał błędnej klasyfikacji. Zestawieniem przywidywań stanu zmiennej oraz jej faktycznego, zaobserwowanego stanu jest macierz klasyfikacji. W tabeli 4.2 można zaobserwować przykład takowej tabeli dla problemu klasyfikacji binarnej. Zakładając że zadaniem modelu było określenie czy obserwacja przyjmuje stan: 1 czy 0. W tabeli prawidłowe określenie stanu zmiennej określone jest jako TP (ang. *true positive*), prawidłowe wskazanie przeciwnego stanu: TN (ang. *true negative*). Obserwacje wliczające się w te pola zostały poprawnie zaklasyfikowane. Ich przeciwieństwem są: błędnie wskazania stanu FN (ang. *false negative*) oraz FP (ang. *false positive*) czyli błędne wskazanie stanu przeciwnego

do 1. Tabela klasyfikacji pozwala w przejrzysty sposób porównać stan faktyczny ze wskazaniami modelu. Dobry model to ten który minimalizuje liczbę FN oraz FP, maksymalizując liczbę TP oraz TN.

Tabela 4.2. Przykładowa macierz klasyfikacji wykorzystywana do oceny algorytmów uczenia maszynowego

	Zaobserwowano stan 1	Nie zaobserwowano stanu 1
Przewidywano stan 1	TP	FP
Nie przywidywano stanu 1	FN	TN

Źródło: Opracowanie własne na podstawie Cook (2016)

Na tabeli 4.2 opierają się również reguły decyzyjne. Ich wyliczenie pozwala na porównywanie modeli między sobą . Najważniejsze z nich to:

- Wrażliwość (ang. *Recall*), definiowana wzorem:

$$\frac{TP}{TP+FN}, \tag{4.24}$$

- Specyficzość (ang. *Specificity*), definiowana przez:

$$\frac{TN}{FP+TN}, \tag{4.25}$$

- Precyzja (ang: *Precision*):

$$\frac{TP}{TP+FP}, \tag{4.26}$$

- Celność (ang. *Accuracy*):

$$ACC = \frac{TN+TP}{TN+TN+FN+FP}. \tag{4.27}$$

- Miara f1: Miara ta, jeżeli osiąga wartość bliższą 1, oznacza że zarówno miara precyzji jak i wrażliwości osiągają wysokie wartości. Oznacza to że model poprawnie klasyfikuje wystąpienia zmiennej i nie przypisuje dużej ilości fałszywie pozytywnych klasyfikacji. Niskie wartości precyzji i wrażliwości miały by swoje odbicie w wartości F1 bardziej zbliżającej

się do zera (Dangeti, 2017).

$$\frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{4.28}$$

- Miara Hit Ratio jest miarą opisującą skuteczność modelu. Opiera się ono na liczbie poprawnych klasyfikacji podzielonej przez całkowitą liczbę klasyfikacji. Jest to miara używana między innymi przez modele z pakietu H2O (LeDell i in., 2019). W pakiecie tym podawana jest również informacja o skuteczności modelu przy kolejnych podejściach klasyfikacji (Cook, 2016).

4.4 Wyniki klasyfikacji zawodów w ofertach pracy

4.4.1 Wykorzystane narzędzia

Praca z danymi odbywała się w całości w języku R. Ponadto zostały użyte dodatkowe pakiety. W przypadku modelowanie były to:

- pakiet `glmnet` (Friedman i in., 2010);
- pakiet H2O (LeDell i in., 2019).

Powyższe pakiety pozwalają na kompleksowe modelowanie danych. Pakiet H2O (LeDell i in., 2019) został wykorzystany podczas przetwarzania dużych ilości danych, ponieważ pozwala on na równoległą prace algorytmów. Język R działa na jednym wątku, dlatego też wykorzystanie odpowiednich pakietów jest ważne kiedy obiektem zainteresowania są duże zbiory danych. Pakiet ten zawiera interfejsy dla takich języków jak: R, Python, Jave czy Scala. Zaimplementowano w nim wiele popularnych algorytmów uczenia maszynowego, takich jak: modele liniowe, Naiwny Bayes, algorytmy PCA czy grupowanie metodą k-średnich. Całość jest flagowym produktem firmy *H2O.ai*, ta platforma opierająca się o open source wykorzystywana jest przez ponad 9 000 organizacji i 80 000 analityków, włączając w to największe banki i towarzystwa ubezpieczeniowe (Landry, 2018).

Ostatecznie porównanie modeli będzie możliwe dzięki sprawdzeniu ich macierzy klasyfikacji. Pierwszy z nich to wielomianowa regresja logistyczna LASSO, stworzona za pomocą pakietu `glmnet` (Friedman i in., 2010). Kolejnym modelem jest również wielomianowa regresja logistyczna LASSO lecz tym razem stworzona już w H2O (LeDell i in., 2019). Trzecim modelem

jest Naiwny bayes również stworzony w H2O (LeDell i in., 2019). Dane na których zbudowano model zostały podzielone na zbiory uczące się i testowe(oraz walidacyjne w przypadku pakietu H2O (LeDell i in., 2019)). W celu sprawdzenia jakości modeli zostaną przedstawione macierze klasyfikacji oraz miary błędów i poprawności. Zostały jednakże one w pierwszej kolejności policzone na tym samym rodzaju danych, na których zostały zbudowane modele. Tak wytrenowane i ocenione modele dopiero następnie zostaną użyte do klasyfikacji danych z portalu OLX. Ostatnim krokiem będzie zbudowanie macierzy klasyfikacji i sprawdzenie jakości predykcji kategorii dla danych z portalu internetowego.

4.4.2 Wielomianowa regresja logistyczna w pakiecie `glmnet`

Wielomianowa regresja logistyczna LASSO w pakiecie `glmnet` (Friedman i in., 2010) została wygenerowana bazując na poniższym kodzie:

```
library(glmnet)
cv.glmnet(x, y, family = "multinomial", alpha = 1, nfolds = 10, type.
    measure="deviance")
```

Program 4.1. Kod w języku R modelujący metodą wielomianowej regresji logistycznej LASSO z wykorzystaniem pakietu glmnet

gdzie:

- *x* oznacza zbiór treningowy, stanowiący 80% wszystkich danych;
- *y* to zmienna objaśniana z zbioru danych treningowych;
- *family* określa jakiego rodzaju równanie powinno zostać zaimplementowane, *multinomial* oznacza że będzie to klasyfikacja z większą liczbą zmiennych opisywanych niż w przypadku modelu binarnego;
- *alpha* określa czy model powinien być regresją grzbietową czy regresją LASSO, dopuszczalne są również warianty pomiędzy 0 a 1, wartość 1 oznacza że będzie to regresja LASSO;
- *nfolds* określa na ile grup będzie dzielony zbiór podczas sprawdzania krzyżowego, liczba 10 została wybrana arbitralnie jednakże jest to również wartość domyślna dla tego modelu;
- *type.measure* to miejsce na zdefiniowanie miary dopasowania modelu. Parametr ten jest używany podczas sprawdzianu krzyżowego. W tym przypadku jest on ustawiony na

'deviace'. Miara ta określa różnice między dopasowanym modelem a abstrakcyjnym modelem który idealnie by pasował do wszystkich danych(ang: saturated model). Warto zaznaczyć że jest to w tym pakiecie miara domyślna dla problemów klasyfikacji.

Macierz 4.3 przedstawia macierz klasyfikacji dla modelu wielomianowej regresji logistycznej LASSO.

Tabela 4.3. Macierz klasyfikacji wykonana w pakiecie `glmnnet` dla modelu wielomianowej regresji logistycznej LASSO

	2:Spec	3:Technicy	4:Biuro	5:Usługi	7:Przemysł	8:Operatorzy	9:Proste
2:Spec	571	153	131	54	13	7	7
3:Technicy	427	696	223	236	246	154	223
4:Biuro	33	65	597	35	11	28	47
5:Usługi	61	51	47	681	8	5	24
7:Przemysł	18	52	3	1	608	82	48
8:Operatorzy	8	36	36	13	125	777	41
9:Proste	47	112	128	145	154	112	775

Źródło: Opracowanie własne.

Macierz 4.3 została wygenerowana w pakiecie `glment`. W kolumnach znajdują się wartości referencyjne, natomiast wiersze oznaczają predykcje. Dla przykładu, w tabeli 4.3 pierwsza komórka o wartości 571 oznacza że model ten zakwalifikował 571 obserwacji poprawnie, ponieważ ofertą z drugiej kategorii właśnie takową przydzielił. Pozostałe wartości w tym wierszu oznaczają wartości które zostały zakwalifikowane przez model do drugiej kategorii, jednakże faktycznie znajdowały się w innej. Wartości po przekątnej tabeli są ilością poprawnie zaklasyfikowanych ofert pracy. Na podstawie macierzy 4.3 można również policzyć opisane w rozdziale wcześniej miary oceny modelu. W tym modelu **trafność wynosiła: 0.58.**. Tabela 4.4 pokazuje kształtowanie się wartości: wrażliwości, precyzji oraz miary F1 w zależności od kategorii:

Bazując na wartości F1 można stwierdzić że model dobrze klasyfikuje oferty zwłaszcza z ósmej kategorii, nazwa tej kategorii to: 'Operatorzy i monterzy maszyn i urządzeń'. Najgorzej w kategorii trzeciej, czyli: 'Technicy i inny średni personel'.

4.4.3 Wielomianowa regresja lasso z wykorzystaniem oprogramowania H2O

Model Wielomianowej logistycznej regresji LASSO zaimplementowany w H2O (LeDell i in., 2019) został stworzony poniższym kodem:

Tabela 4.4. Tabela z weryfikacją jakości klasyfikacji wielomianowej regresji logistycznej LASSO w pakiecie `glmnet`

	Precyzja	Wrażliwość	F1
2	0.61	0.49	0.54
3	0.32	0.60	0.41
4	0.73	0.51	0.60
5	0.78	0.58	0.67
7	0.75	0.52	0.62
8	0.75	0.67	0.71
9	0.53	0.67	0.59

Źródło: Opracowanie własne.

```
 library(h2o)
h2o.glm(x = x,
        y = y,
        training_frame = train,
        model_id = "glm",
        family = 'multinomial',
        validation_frame = valid,
        nfolds = 10,
        alpha = 1)
```

Program 4.2. Kod w języku R modelujący metodą Wielomianowej regresji logistycznej LASSO z wykorzystaniem pakietu H2O

Gdzie:

- *x* zawiera wektor nazw zbioru danych, z wykluczeniem zmiennej objaśnianej;
- *y* nazwa zmiennej objaśnianej;
- *training_frame* zbiór danych treningowych stanowiący 70% wszystkich danych;
- *model_id* jest unikalnym identyfikatorem modelu;
- *family* określa jakiego rodzaju równanie powinno zostać zaimplementowane, *multinomial* oznacza że będzie to klasyfikacja z większą liczbą zmiennych opisywanych;
- *validation_frame* określa zbiór będący odniesieniem dla modelu, stanowi on 15% całości danych;
- *nfolds* określa na ile grup będzie dzielony zbiór podczas sprawdzania krzyżowego, liczba 10 została wybrana arbitralnie jednakże jest to również wartość domyślna dla tego modelu;

- *alpha* określa czy powinna być to regresja grzbietowa czy regresja LASSO czy też wariant między tymi modelami, wartość 1 oznacza że będzie to regresja LASSO

Tabela 4.5. Macierz klasyfikacji dla modelu wielomianowej regresji logistycznej LASSO w pakiecie H2O

	2:Spec	3:Technicy	4:Biuro	5:Usługi	7:Przemysł	8:Operatorzy	9:Proste
2:Spec	368	401	19	50	9	5	42
3:Technicy	101	531	53	39	33	26	79
4:Biuro	76	185	433	49	0	27	111
5:Usługi	51	188	36	512	0	10	97
7:Przemysł	6	272	5	3	442	66	97
8:Operatorzy	7	142	15	4	62	608	93
9:Proste	4	211	39	27	27	36	512

Źródło: Opracowanie własne.

Tabela 4.5 zawiera macierz klasyfikacji. Wiersze oznaczają wartości prawdziwe, kolumny zaś: predykcje modelu. Prognozy poprawne(True Positive) znajdują się więc po przekątnej. Liczba obserwacji w macierzy jest mniejsza niż w przypadku 4.3 ze względu na inny podział zbioru danych. W przypadku modelowania w pakiecie H2O dane są podzielone na podzbiory: uczący, walidacyjny i testowy. Proporcja wynosi kolejno: 70%, 15%, 15%. Dotyczy to obu modeli zbudowanych w tym pakiecie. Tabela 4.6 przedstawia wyliczone na podstawie macierzy klasyfikacji miary:

Tabela 4.6. Tabela z miarami weryfikującymi jakość klasyfikacji modelu wielomianowej regresji LASSO

	Precyzja	Wrażliwość	F1
2	0.41	0.60	0.41
3	0.61	0.27	0.62
4	0.49	0.72	0.49
5	0.57	0.74	0.57
7	0.49	0.77	0.49
8	0.65	0.78	0.65
9	0.59	0.49	0.59

Źródło: Opracowanie własne.

Bazując na tabelce 4.6 można stwierdzić że model osiąga porównywalne wyniki co model stworzony w pakiecie `glmnet` (Friedman i in., 2010). Najwyższe wartości osiągane są dla kategorii 3 i 8. Gorsze wyniki pojawiają się w kategorii 2.

Model Naiwnego Bayesa powstał natomiast za pomocą kodu: 4.3. Warto zauważyć że w tym przypadku model osiąga gorsze rezultaty. Model wielomianowej regresji liniowej LASSO sprawdzał się w tym problemie zdecydowanie lepiej. Model ten nie przydzielił żadnej oferty do kategorii numer 3. Ponadto zdecydowaną większość danych zaklasyfikował do kategorii numer 4. Przedstawione to zostało w tabeli 4.7. Poza kategorią 4, jedyną inną kategorią jaką model wydaje się dostrzegać jest kategoria 7. Pozostałe zawody zostały przez model zignorowane całkowicie(jak w przypadku 3 kategorii) lub przynajmniej w dużej części.

```
h2o.naiveBayes(x = x, y = y, training_frame = train,
               model_id = "nb",validation_frame = valid, nfolds = 10)
```

Program 4.3. Kod w języku R modelujący metodą Naiwnego Bayesa z wykorzystaniem pakietu H2O

Wszystkie parametry mają taką samą interpretacje jak w modelu wielomianowej regresji logitycznej LASSO. Macierz klasyfikacji dla tego modelu zaprezentowana została w tabeli 4.7.

Tabela 4.7. Macierz klasyfikacji na podstawie modelu Naiwnego Bayesa

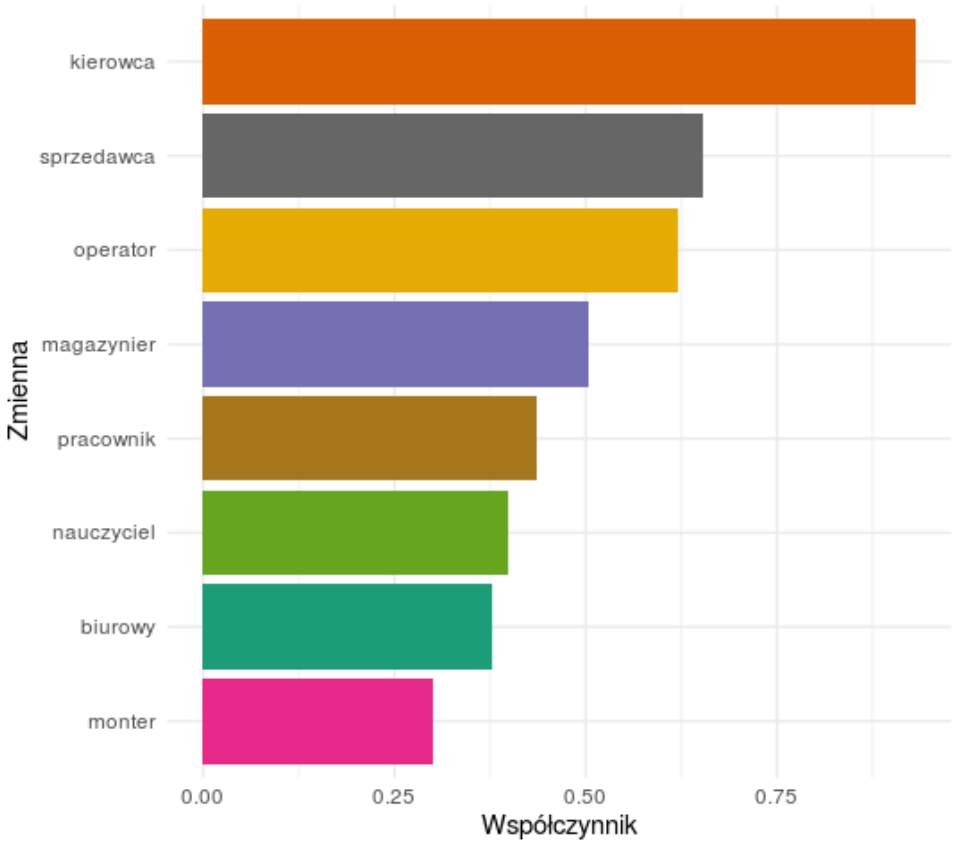
	2:Spec	3:Technicy	4:Biuro	5:Usługi	7:Przemysł	8:Operatorzy	9:Proste
2:Spec	0	0	686	15	188	2	3
3:Technicy	3	0	609	12	229	2	7
4:Biuro	2	0	710	9	155	4	1
5:Usługi	2	0	716	17	144	7	8
7:Przemysł	8	0	628	14	233	4	4
8:Operatorzy	13	0	755	9	142	5	7
9:Proste	10	0	577	63	172	8	26

Źródło: Opracowanie własne.

Na rysunku 4.4 przedstawione zostało 10 najważniejszych zmiennych determinujących kategorie w modelu wielomianowej regresji logistycznej LASSO zbudowanego w pakiecie H2O (LeDell i in., 2019). Są to zmienne które według modelu mają największy wpływ na wybór kategorii w danych treningowych. Oś X tego wykresu opisuje ważność zmiennej (ang: *variable importance*), wartość ta liczona jest za pomocą wielkości współczynników w modelu.

4.4.4 Wyniki klasyfikacji ofert z serwisu internetowego

Wyżej opisane modele zostały użyte do klasyfikacji danych pochodzących z serwisu internetowego. Wszystkich danych jest 411 787. W celu weryfikacji klasyfikacji wylosowano po 1% wszystkich ofert dla każdego kwartału. Tabela 4.8 przedstawia rozkład ilości kategorii pracy



Rysunek 4.4. Słowa mające największe znaczenie przy klasyfikacji według modelu wielomianowej regresji logistycznej LASSO z pakietu H2O.

Źródło: Opracowanie własne na podstawie danych z portalu internetowego.

w zależności od użytego modelu.

Tabela 4.8. Rozkład prognozowanych ofert pracy w próbie według kategorii i modeli

Model	2:Spec	3:Technicy	4:Biuro	5:Usługi	7:Przemysł	8:Operatorzy	9:Proste
`glmnet` glm	1121	13990	359	323	130	54	495
H2O nb	103	0	12514	866	2454	111	424
H2O glm	613	7523	1000	2471	981	2013	1871

Źródło: Opracowanie własne na podstawie. Uwagi: glm = regresja logistyczna wielomianowa; nb = Naiwny Bayes.

Oba modele wielomianowej regresji logistycznej LASSO najczęściej prognozowały kategorie numer 3 (tabela: 4.8). W przypadku naiwnego bayesa klasyfikacja ta nie występuje w ogóle. Jednymi z najpopularniejszych słów w modelu z pakietu `glmnet` (Friedman i in., 2010) z kategorii 3 były takie słowa jak: 'doradca'czy 'kierowca'.

W pakiecie H2O (LeDell i in., 2019) dla tego samego modelu i kategorii były to między in-

nymi: ’opiekunka’, ’niania’ czy ’przedstawiciel’. W tej kategorii znajdują się takie oferty jak: ’Operator CNC’, ’Fryzjerka’, ’Operator koparki’ czy też kierowcy.

Modele wielomianowej regresji logistycznej wspólnie przydzieliły takie oferty jak ’Specjalista ds. organizacji ruchu’ czy tez ’Specjalista ds.Wsparcia Klienta’ do kategorii numer 2. Do tej kategorii zaklasyfikowało się również kilka ofert dla pracy dla nauczycieli.

Kategoria ’Pracownicy biurowi’ jest za to reprezentowana przez takie oferty pracy jak: recepcjonistki, sekretarki czy też równego rodzaju oferty mające ’Praca biurowa’ w swej nazwie.

Do kategorii piątej wielomianowe modele regresji przydzielają oferty z zakresu gastronomii czy też handlu detalicznego. Kategoria ’robotnicy przemysłowi i rzemieślnicy’ to przede wszystkim oferty pracy dla monterów instalacji i mechaników.

W kategorii 9 znajdują się oferty dla pracowników zajmujących się utrzymaniem czystości, w tym przypadku liczba ofert o tej tematyce jest zdecydowanie największa. Kwartalnie oba modele wielomianowej regresji logistycznej przydzielają najwięcej ofert do trzeciej kategorii we wszystkich kwartałach. Sytuacja wygląda analogicznie dla modelu Naiwnego Bayesa z tą różnicą że jest to kategoria czwarta. Na podstawie wylosowanej próby kreuje się następujący obraz:

- Modele wielomianowe regresji logistycznej LASSO, w obu pakietach, poradziły sobie lepiej z problemem klasyfikacji niż model Naiwnego Bayesa;
- Obiecujące wyniki można zaobserwować na grupach posiadających mocno wyróżniające się słownictwo, przykładem takiej grupy może być kategoria: ’Specjaliści’. Trafiają do niej, i słusznie, wszystkie oferty adresowane do ludzi z specjalistycznym wykształceniem, duży udział w tej grupie mają również oferty adresowane do nauczycieli;
- Najpopularniejszą klasą jest kategoria 3, zasób słownictwa w tych ofertach jest najbardziej ogólny, a tematyka prac najszersza;
- Algorytm wielomianowej regresji LASSO z pakietu `H2O` (LeDell i in., 2019) lepiej poradził sobie z wychwyceniem słów charakterystycznych między ofertami z kategorii 3, 7 a 8. Ofert pracy które zostały uznane za należące do 8 w obu modelach wielomianowej regresji nie było w ogóle. Model z pakietu `glmnet` (Friedman i in., 2010) przydzielił zdecydowanie za dużo ofert do szerokiej kategorii: trzeciej;
- O ogólnym charakterze trzeciej kategorii mogą również świadczyć same kategorie serwisu

internetowego. W przypadkach obu regresji najpopularniejszą kategorią w tej klasyfikacji była kategoria: 'Pozostałe oferty pracy';

- Model wielomianowej regresji logistycznej LASSO z pakietu H2O (LeDell i in., 2019), bazując na wylosowanej próbie, poradził sobie najlepiej z problemem klasyfikacji. Potrafił on najbardziej równomiernie rozłożyć kategorie i odnotować mniej widoczne różnice między ofertami. W przypadku kategorii 7 najwięcej(83% wszystkich przydzielonych ofert) było ofert oznaczonych do kategorii: 'produkcja', 'budowa' oraz 'mechanik/blacharz/lakiernik'. W kategorii 8 praktycznie wszystkie oferty należą do kategorii 'kierowca/kurier', co prawda z samej definicji 'operatorzy i monterzy maszyn i urządzeń' może to jasno nie wynikać, jednakże zgodnie z przyjętymi definicjami zbioru treningowego, kierowcy powinni znaleźć się właśnie w kategorii 8. Warto odnotować, że kierowcy w zdecydowanej większości trafili do grupy trzeciej w algorytmie tworzonym w pakiecie `glmnet` (Friedman i in., 2010).

4.4.5 Manualna ocena uzyskanych wyników

Ostatnim krokiem weryfikacji jakości algorytmu klasyfikacji jest jego użycie do już sklasyfikowanych danych. Wylosowane zostało 4 000 obserwacji, po 1 000 z każdego kwartału. Następnie oferty te zostały zostały manualnie zweryfikowane i przydzielone do odpowiedniej kategorii. Wyniki tej klasyfikacji zostały skonfrontowane z wynikami klasyfikacji modelu wielomianowej regresji logistycznej LASSO zbudowanym w pakiecie H2O (`LeDell i in., 2019`). Macierz 4.9 przedstawia wyniki klasyfikacji.

Tabela 4.9. Macierz klasyfikacji dla modelu wielomianowej regresji logistycznej LASSO

	2:Spec	3:Technicy	4:Biuro	5:Usługi	7:Przemysł	8:Operatorzy	9:Proste
2:Spec	121	122	3	20	2	0	0
3:Technicy	7	443	9	107	8	1	39
4:Biuro	4	79	224	15	0	2	27
5:Usługi	24	393	20	529	0	7	68
7:Przemysł	7	262	2	6	249	16	165
8:Operatorzy	1	44	3	1	6	509	2
9:Proste	4	211	19	4	18	8	189

Źródło: Opracowanie własne.

W tabeli 4.9 kolumny oznaczają przydział dokonany przez model, podczas gdy wiersze są

rezultatem ręcznej weryfikacji. Na podstawie macierzy klasyfikacji możliwe jest policzenie miar jakości modelu co zostało przedstawione w tabeli: 4.10.

Tabela 4.10. Tabela weryfikująca jakość klasyfikacji modelu wielomianowej regresji logistycznej LASSO

	Precyzja	Wrażliwość	F1
2	0.45	0.72	0.55
3	0.72	0.28	0.41
4	0.64	0.8	0.71
5	0.51	0.77	0.61
7	0.35	0.88	0.50
8	0.90	0.94	0.92
9	0.42	0.38	0.40

Źródło: Opracowanie własne.

Na podstawie tabeli 4.10 można stwierdzić że klasyfikacja odnosi lepsze wyniki w kategoriach takich jak: 'operatorzy i monterzy maszyn i urządzeń' gdzie miara F1 jest najwyższa. Jednakże warto zwrócić uwagę na niższą wartość w kategorii 'pracownicy przy pracach prostych'. Oferty z tej kategorii często były nieprawidłowo klasyfikowane do grupy: 'technicy i inny średni personel'. Kategoria 3 i 9 osiągają najsłabsze wyniki.

4.5 Podsumowanie

Na podstawie przeprowadzonej analizy można stwierdzić że modele wytrenowane na BKL osiągają dobre wyniki na danych z serwisu Internetowego OLX. Przetestowane zostały dwa różne modele:

- Wielomianowej regresji logistycznej LASSO, model ten stworzony został w dwóch wersjach: w modelu `H2O` (LeDell i in., 2019) oraz `glmnet` (Friedman i in., 2010);
- Model Naiwnego Bayesa stworzony w pakiecie `H2O` (LeDell i in., 2019).

Na podstawie oceny modeli na zbiorze testowym można było stwierdzić że modele wielomianowej regresji logistycznej LASSO osiągały lepsze rezultaty. Pomiędzy tymi dwoma modelami bardziej satysfakcjonujące wyniki osiągał model stworzony w pakiecie `H2O` (LeDell i in., 2019). Dane z serwisu internetowego OLX musiały pierw zostać poddane obróbce, zarówno w celu sprawdzenia terminu ich obowiązywania, ale także by móc wyeliminować powtarzające się

oferty. Dane z serwisu OLX dostosowane zostały do kodowania województw według standardu Głównego Urzędu Statystycznego. Oferty pracy OLX jednakże posiadały inne oznaczenia zawodów niż ustalone w Klasyfikacji Zawodów i Specjalności. Wszystkich danych które były brane pod uwagę było 411 787. W pierwszym kroku wszystkie trzy modele zostały użyte do modelowania na próbie składającej się z 4% całości danych, po 1% na kwartał. Sprawdzone zostały rozkłady klasyfikacji oraz w sposób ogólny i manualny sprawdzone zostały oferty tam się znajdujące. Następnie wydzielony został 1% całości danych, rozłożony na równo między 4 kwartały i następnie został manualnie zaklasyfikowany do kategorii Klasyfikacji Zawodów i Specjalności. Następnie model wielomianowej regresji logistycznej LASSO użyty został do wcześniej przydzielonych danych. Na tak przygotowanym zbiorze powstało ostateczne podsumowanie skuteczności modelu klasyfikującego.

Model wielomianowy regresji logistycznej osiągnął mocno zróżnicowane rezultaty, w zależności od kategorii oferty pracy. Najskuteczniejszy był podczas przydzielania ofert z kategorii 9, czyli: 'Pracownicy przy pracach prostych'. Miara *f1* wynosiła tam 0.92 co jest bardzo dobrym wynikiem. Następną kategorią którą model skutecznie opisywał była kategoria 4, czyli 'Pracownicy biurowi'. Miara *f1* wynosiła w tym przypadku 0.71. Kategorie 2, 5, oraz 7 uzyskały przeciętny wynik, mieszczący się między 0.5 a 0.6. Najgorzej wypadły kategorie 3 i 9, czyli 'Technicy i inny średni personel' oraz 'Pracownicy przy pracach prostych'. Miara *f1* w przypadku tych kategorii wynosiła 0.4. Oferty z 9 kategorii były często przydzielane do kategorii numer 7, ponadto algorytm błędnie przydzielał dużo ofert z 3 kategorii, do kategorii dziewiątej. Oferty z kategorii trzeciej były często mylone z kategorią 5. Zdecydowana większość oferty z drugiej kategorii została poprawnie zaklasyfikowana, jednakże model zaklasyfikował do tej kategorii dużą ilość ofert z kategorii 3.

Rozdział 5

Porównanie danych z Badania Popytu na Pracę, Centralnej Bazy Ofert Pracy oraz portalu OLX (Greta, Magdalena, Krzysztof)

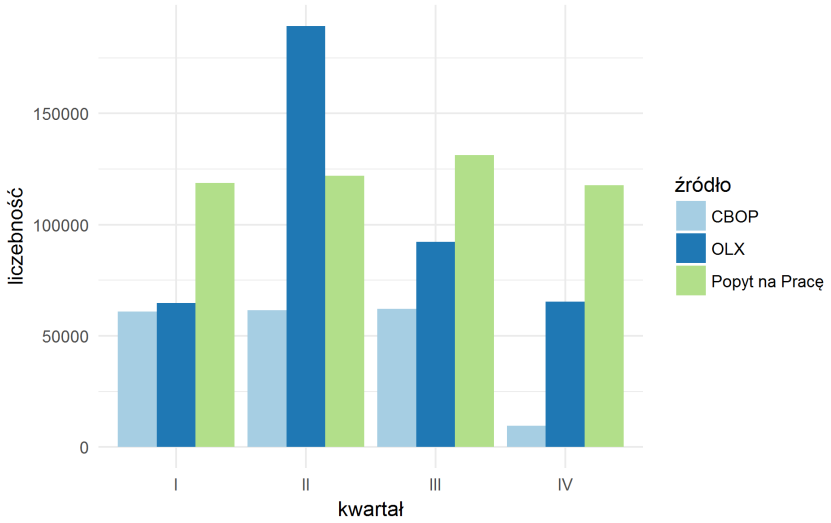
5.1 Cel rozdziału

Celem tego rozdziału jest porównanie rozkładów odsetków oraz liczebności wolnych miejsc pracy dla danych pochodzących z Badania Popytu na Pracę Głównego Urzędu Statystycznego, Centralnej Bazy Ofert Pracy urzędów pracy oraz portalu OLX. Ponadto w rozdziale tym zostaną również zbadanie korelacje pomiędzy źródłem danych a umieszczaniem ofert ze względu na województwo oraz zawód, korelacje pomiędzy odsetkami oraz obciążenie.

5.2 Rozkład liczby ofert o pracę na koniec kwartału

Na wykresie 5.1 przedstawiony został rozkład liczebności wolnych miejsc dla danych pochodzących z Badania Popytu na Pracę, Centralnej Bazy Ofert Pracy oraz z portalu OLX w podziale na kwartały. Rozkłady liczebności ofert pracy według kwartałów w poszczególnych zbiorach są bardziej zróżnicowane, niż w przypadku rozkładu ich odsetków ze względu na województwa. W Badaniu Popytu na Pracę liczebności różnią się od siebie minimalnie w poszczególnych kwartałach, co wynika z faktu, iż jest to badanie kwartalne i w każdym z badanych okresów próba powinna posiadać taką samą liczebność. Na portalu OLX najwięcej ofert pracy umieszczanych było w drugim kwartale, następnie w trzecim a najmniej było ich w kwartale pierwszym oraz

ostatnim. W bazie CBOP natomiast pierwszy trzy kwartały charakteryzowały się jednakowymi liczebnościami, a kwartale ostatnim umieszczonych było o wiele mniej ofert w porównaniu do pozostałych. Podsumowując rozkład liczby ofert umieszczanych w bazie urzędów pracy oraz na portalu internetowym różni się od rozkładu liczby ofert publikowanych w badaniu GUS ze względu na fakt, iż liczba ofert w badaniu dla każdego kwartału jest planowana. Natomiast częścią wspólną dla obu źródeł internetowych jest niska liczba ofert pracy publikowanych w ostatnim kwartale.

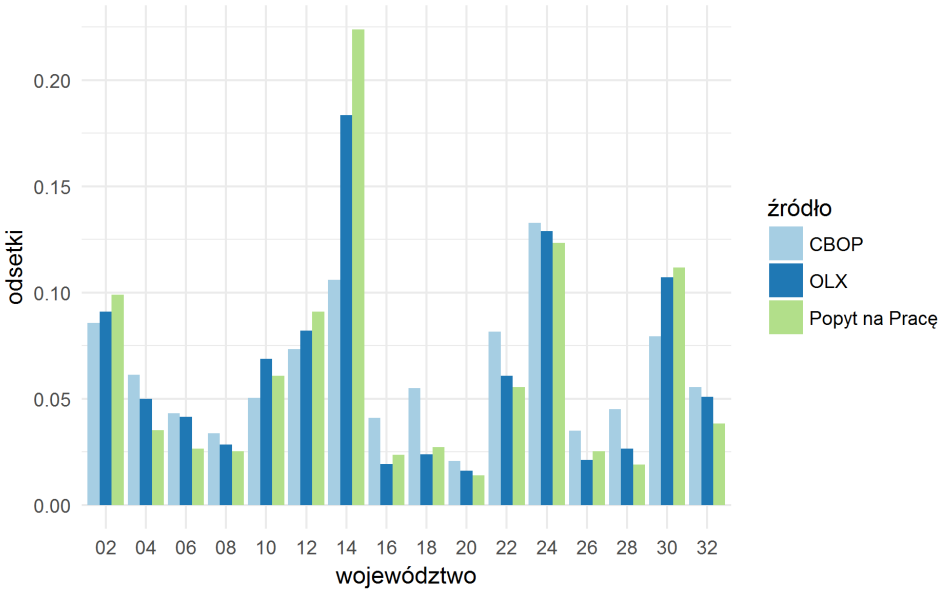


Rysunek 5.1. Rozkład liczebności wolnych miejsc pracy w danym kwartale dla danych pochodzących z Badania Popytu na Pracę, Centralnej Bazy Ofert Pracy oraz portalu OLX

Źródło: Opracowanie własne na podstawie danych GUS, CBOP i OLX

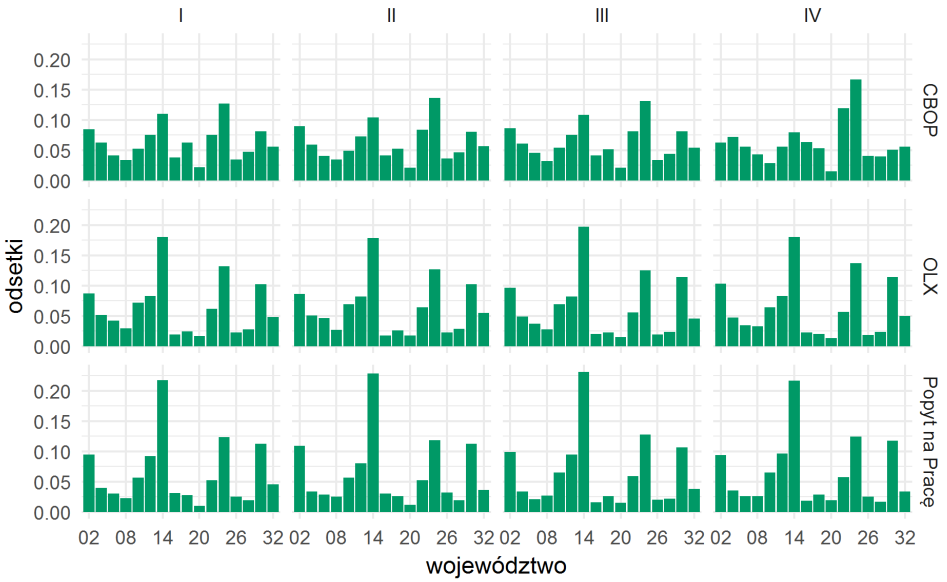
Wykres 5.2, przedstawia rozkład odsetków wolnych miejsc pracy ze względu na województwa. Porównując rozkłady wolnych miejsc pracy dla CBOP, OLX oraz Badania Popytu o Pracę widać, że zarówno w danych statystycznych, jak i niestatystycznych najwięcej wolnych miejsc pracy przypada na województwa: mazowieckie, śląskie, wielkopolskie, dolnośląskie (dla CBOP jest to jeszcze pomorskie) oraz małopolskie. Najmniej ofert z Badania Popytu na Pracę przypada na województwo podlaskie, w bazie CBOP oraz OLX również z tego województwa umieszczane jest najmniej ofert o pracę. Podsumowując, rozkłady wolnych miejsc pracy ze względu na województwo z zbiorze CBOP, OLX oraz GUS są do siebie bardzo podobne, te same województwa w obu zbiorach wykazują największe oraz najmniejsze odsetki dla liczby ofert pracy. Wartość odsetków jest większa dla danych pochodzących z Badania Popytu na Pracę.

Na wykresie 5.3 widać rozkład odsetków ofert pracy w porównywanych zbiorach w przekroju województw i kwartałów. Jak można zauważyć, tendencja przedstawiona na wykresie 5.2



Rysunek 5.2. Rozkład odsetków wolnych miejsc pracy ze względu na województwo dla danych pochodzących z Badania Popytu na Pracę, Centralnej Bazy Ofert Pracy oraz portalu OLX

Źródło: Opracowanie własne.



Rysunek 5.3. Rozkład odsetków wolnych miejsc pracy ze względu na województwo oraz kwartał dla danych pochodzących z Badania Popytu na Pracę, Centralnej Bazy Ofert Pracy oraz portalu OLX

Źródło: Opracowanie własne.

(przedstawiającym odsetki dla samych poszczególnych województw), utrzymuje się również w podziale na kwartały.

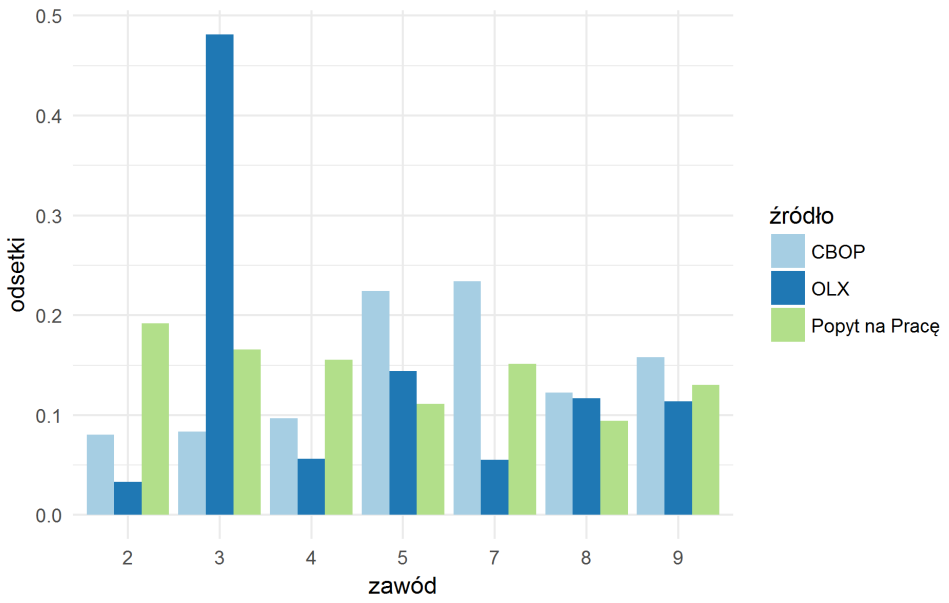
Wykres 5.4 przedstawia rozkład udziałów wolnych miejsc pracy w poszczególnych zawodach dla danych pochodzących z Badania Popytu na Pracę, Centralnej Bazy Ofert Pracy oraz z portalu OLX. Z każdego zbioru, we wcześniejszych etapach, z powodu małej reprezentacji wyeliminowana została 6 grupa zawodu, dlatego nie jest ona również uwzględniona z porównaniu. Ponadto, w danych OLX podczas analiz nie była brana pod uwagę również 1 grupa zawodu Wszystkie porównywane zbiory danych składają się z ofert aktualnych na koniec danego kwartału 2017 roku.

Przed porównaniem warto przypomnieć jak kształtują się wielkie grupy zawodów według Klasyfikacji Zawodów i Specjalności:

1. Przedstawiciele władz publicznych, wyżsi urzędnicy i kierownicy,
2. Specjaliści,
3. Technicy i inny średni personel,
4. Pracownicy biurowi,
5. Pracownicy usług i sprzedawcy,
6. Rolnicy, ogrodnicy, leśnicy i rybacy,
7. Robotnicy przemysłowi i rzemieślnicy,
8. Operatorzy i monterzy maszyn i urządzeń,
9. Pracownicy wykonujący prace proste.

Jak widać rozkład odsetków wolnych miejsc pracy w źródle statystycznym znacznie różni się od źródeł niestatystycznych. W danych pochodzących z Głównego Urzędu Statystycznego największą grupę zawodów stanowią specjaliści (2), a zaraz po niej grupa trzecia – technicy i inny średni personel. Natomiast najmniejszą jest grupa ósma oraz grupa piąta.

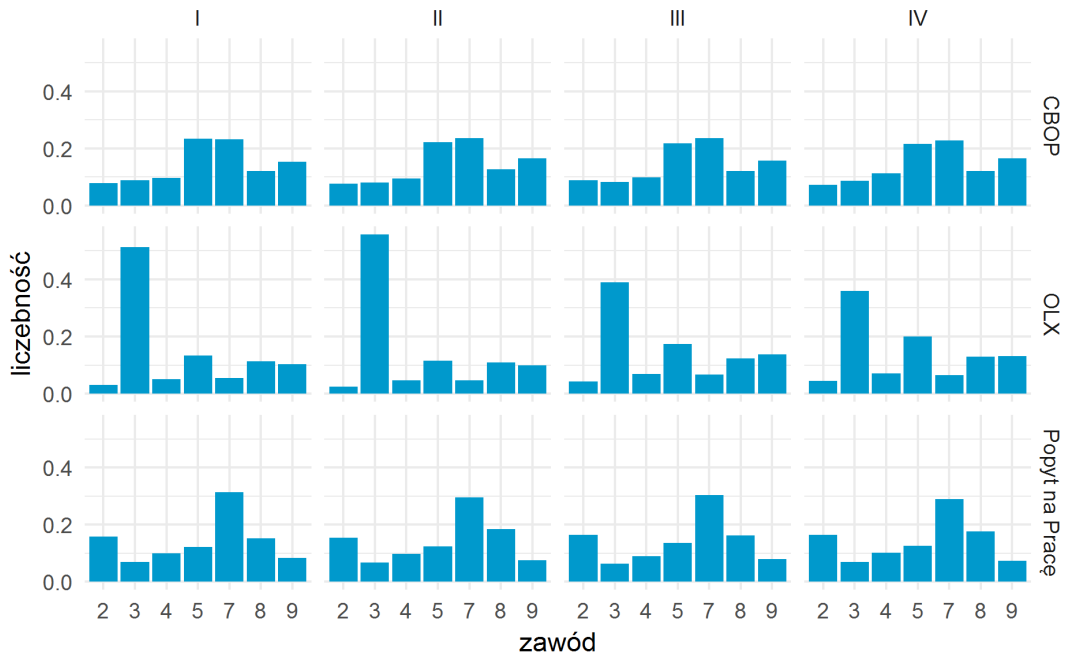
W Centralnej Bazie Ofert Pracy najczęściej umieszczane są ogłoszenia o pracę dla robotników przemysłowych i rzemieślników (7) oraz dla pracowników usług i sprzedawców (5). Natomiast najmniejszy udział przypada na grupę drugą, a następnie trzecią oraz czwartą – pierwsze trzy grupy zawodów, które charakteryzują się największymi odsetkami w badaniu GUS, w bazie Urzędów Pracy stanowią największą grupę. W zbiorze OLX natomiast największy odsetek wolnych miejsc pracy przypada na trzecią grupę zawodu, dla której wysoki wynik jest również



Rysunek 5.4. Rozkład odsetków wolnych miejsc pracy ze względu na zawód dla danych pochodzących z Badania Popytu na Pracę, Centralnej Bazy Ofert Pracy oraz portalu OLX

Źródło: Opracowanie własne.

w danych GUS, a w danych CBOP jest to jedna z mniejszych wartości. Najmniej ofert na portalu OLX umieszczanych jest dla przedstawicieli władz publicznych, wyższych urzędników i kierowników co pokrywa się w tym przypadku z bazą CBOP. Podsumowując rozkłady wolnych miejsc pracy według zawodów różnią się dla wszystkich porównywanych źródeł danych. Na portalu OLX najwięcej umieszczanych ofert o pracę pochodzi jest dla techników oraz średniego personelu – ta grupa odstaje od pozostałych grup zawodów. Na drugim miejscu pod względem odsetków umieszczanych ofert pracy jest grupa piąta, czyli pracownicy usług i sprzedawcy. Natomiast najmniej ofert pojawia się dla grupy drugiej. W bazie urzędów pracy również najrzadziej umieszczane są ofert dla urzędników państwowych i kierowników. Wysokim udziałem wolnych miejsc pracy charakteryzują się w danych CBOP grupa siódma oraz piąta, natomiast najmniejszy udział wolnych miejsc pracy przypada dla grupy drugiej, co pokrywa się z danymi z OLX. Udziały w danych GUS przyjmują mniej zróżnicowane wartości niż dla wymienionych danych niestatystycznych. Największa wartość kształtuje się dla grupy urzędników państwowych, wyższych urzędników i kierowników, a jedna z mniejszych dla pracowników usług i sprzedawców, co jest odwrotnością danych OLX i CBOP. Dla ósmej oraz dziewiątej grupy zawodu wartość odsetków jest identyczna we wszystkich trzech zbiorach. Grupa czwarta oraz piąta wypada podobnie, pod względem odsetków, dla obu zbiorów internetowych.



Rysunek 5.5. Rozkład odsetków wolnych miejsc pracy ze względu na zawód oraz kwartał dla danych pochodzących z Badania Popytu na Pracę, Centralnej Bazy Ofert Pracy oraz portalu OLX

Źródło: Opracowanie własne.

Wykres 5.5 przedstawia rozkład odsetków wolnych miejsc pracy względem źródła danych oraz zawodu i kwartału. Tendencja dla zawodów w każdym ze źródeł utrzymuje się również w kwartałach.

5.2.1 Korelacje i inne miary

Po przeanalizowaniu rozkładów wolnych miejsc pracy względem poszczególnych zmiennych w każdym ze zbiorów zbadana została zależność pomiędzy źródłem umieszczania ogłoszeń (portal OLX, baza CBOP lub oferty pochodzące z badania Głównego Urzędu Statystycznego) a województwem oraz zawodem. Do obliczenia tych zależności wykorzystany został współczynnik V Cramera, który opisywany był już w rozdziale trzecim. Wyniki prezentuje tabela 5.1. Jak widać istnieje umiarkowana zależność pomiędzy źródłem ogłoszenia o pracę a województwem i zawodem.

Następnie zbadana została korelacja Pearsona między odsetkiem ofert pracy aktualnych w danym kwartale a źródłem oferty pracy. Wykresy został wygenerowane za pomocą pakietu `corrplot` (Wei & Simko, 2017). Wyniki prezentuje rysunek: 5.6. Istnieje znacząca, pozytywna

Tabela 5.1. Wartość współczynnika V Cramera dla poszczególnych źródeł umieszczania danych oraz województwa i zawodu

Zmienna	województwo	zawód
Statystyka	0,352	0,322

Źródło: Opracowanie własne.

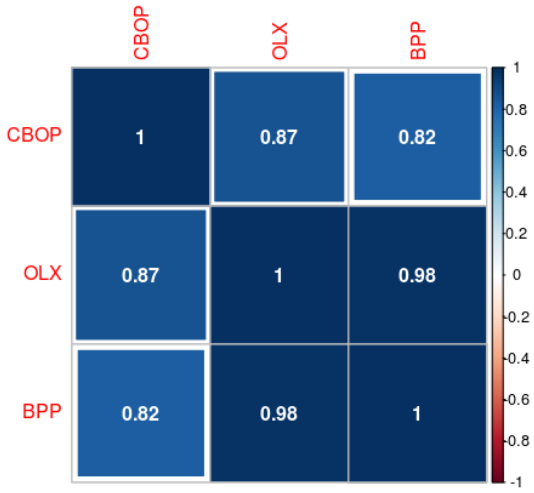
korelacja między liczbą ofert pracy w konkretnych kwartałach między źródłami takimi jak CBOP i Badaniami Popytu o Prace. Ponadto odsetek ofert w danym kwartale w CBOP skorelowana również jest z liczbą ofert w portalu internetowym OLX. Między OLX a Badaniami Popytu o pracy korelacja nie występuje.

Rysunek 5.6. Rozkład miar korelacji Pearsona między źródłami danych a odsetkami w poszczególnych kwartałach

Źródło: Opracowanie własne.

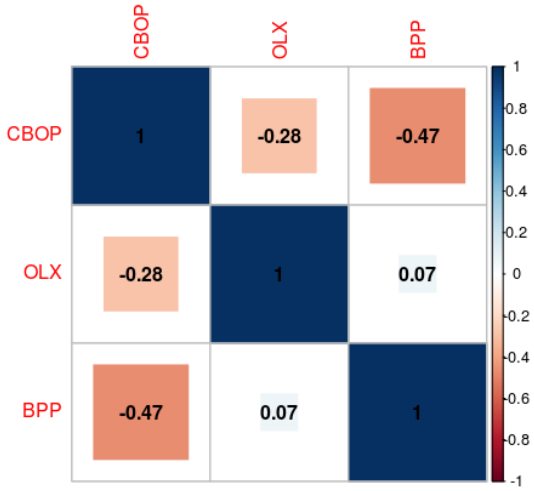
Rysunek 5.7 przedstawia korelacje między odsetkami ofert pracy w poszczególnych województwach a źródłami z których one pochodzą. Najsilniejsza korelacja występuje między liczbą ofert z portalu OLX a Badaniami Popytu o Prace. Mniejsza korelacja występuje między danymi z CBOP a pozostałymi źródłami. Jednakże we wszystkich przypadkach jest ona bardzo silna.

Rysunek 5.8 przedstawia rozkład korelacji Persona między liczbą ofert pracy w poszczególnych zawodach między źródłami. Pomiędzy Badaniami Popytu o Prace a CBOP istnieje umiarkowana, negatywna korelacja. Korelacje związane z portalem internetowym OLX należy uznać że nieistniejące w przypadku korelacji z Badaniami Popytu o Prace (0.07) oraz umiarkowanie negatywne w przypadku korelacji z CBOP (-0.47).



Rysunek 5.7. Rozkład miary korelacji Pearsona między rozkładem odsetek na województwa w każdym ze źródeł

Źródło: Opracowanie własne.



Rysunek 5.8. Rozkład miar korelacji Pearsona między rozkładem odsetek w podziale na zawody w każdym ze źródeł

Źródło: Opracowanie własne.

W analizie obciażenia wykorzystane zostały następujące miary

- Obciążenie

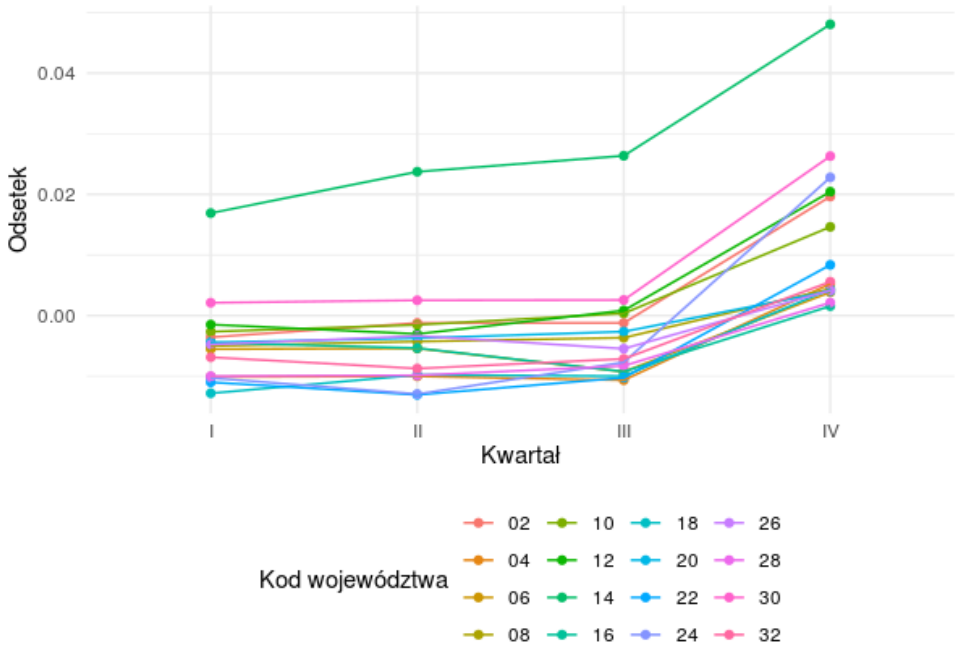
$$\text{Bias} = \hat{\theta}_{\text{olx / cbop}} - \hat{\theta}_{\text{popyt}}, \tag{5.1}$$

- Relatywne absolutne obciążenie

$$\text{ARB} = |\hat{\theta}_{\text{olx / cbop}} - \hat{\theta}_{\text{popyt}}|/\hat{\theta}_{\text{popyt}}, \tag{5.2}$$

gdzie $\hat{\theta}_{\text{olx / cbop}}$ oznaczają oszacowania ze źródła OLX oraz CBOP oraz przyjęto założenie, że $\hat{\theta}_{\text{popyt}}$ oszacowanie jest nieobciążone. Obydwie miary nie uwzględniają również wariancji oszacowań z badania Popytu na pracę.

Rysunek 5.9 przedstawia rozkład obciążenia w czasie między Badaniem Popytu na Pracę a CBOP według województw. Widać znacząco duży odsetek dla województwa Mazowieckiego. Ponadto widoczna jest, opisana wcześniej, korelacja między kształtowaniem się liczby ofert pracy w województwach w poszczególnych źródłach danych. Największe obciążenie we wszystkich województwach występuje w ostatnim kwartale. Przez pierwsze trzy kwartały obciążenie utrzymuje się na podobnym poziomie.

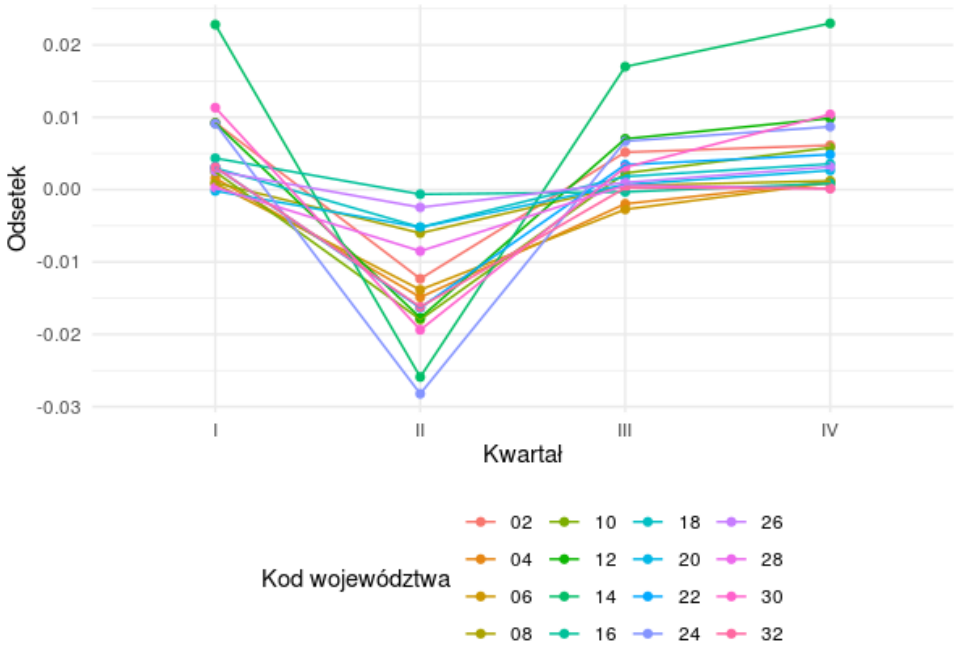


Rysunek 5.9. Rozkład obciążenia między odsetkiem ofert pracy w Badaniu Popytu na Prace a CBOP według województw

Źródło: Opracowanie własne.

Rozkład obciążeń (5.10) dla Badania Popytu na Prace a portalem OLX charakteryzuje się stosunkowo wysokim obciążeniem dla województwa Mazowieckiego. Ponadto, w drugim kwartale występuje znaczący spadek obciążenia, szczególnie duży dla województw takich jak: Ma-

zowieckie czy też Śląskie. Jest to spowodowane dużym odsetkiem ofert pracy które pojawiły się drugim kwartale w portalu internetowym OLX. W ostatnich dwóch kwartałach poziom obciążenia ponownie wraca do wielkości podobnych z pierwszego kwartału.



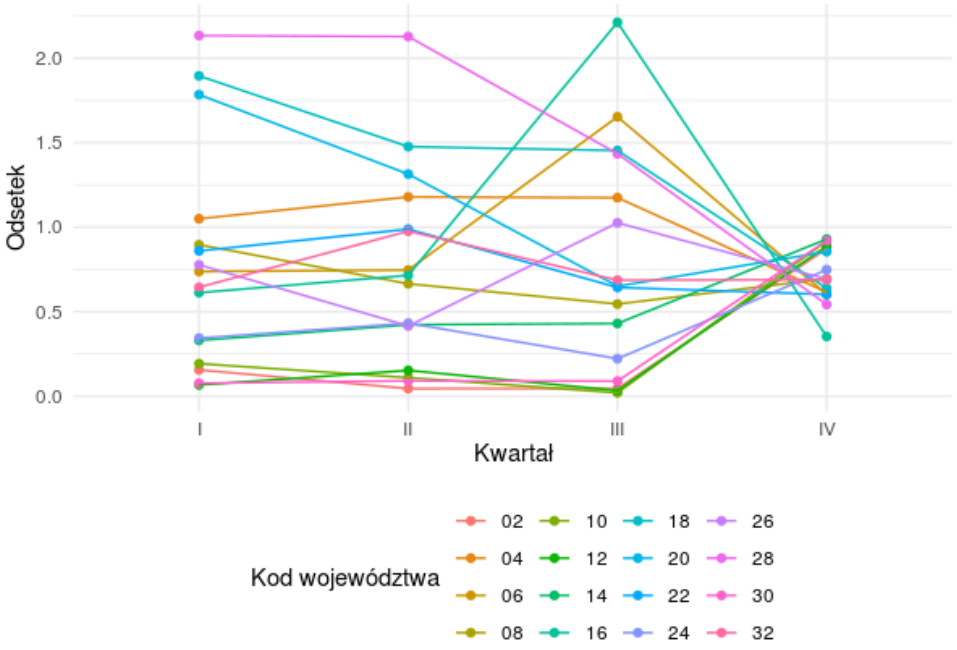
Rysunek 5.10. Rozkład obciążenia między odsetkiem ofert pracy w Badaniu Popytu na Prace a OLX według województw

Źródło: Opracowanie własne.

Rysunek 5.11 przedstawia rozkład relatywnego absolutnego obciążenia między odsetkami z Badania Popytu na Pracę a CBOP. Pierwsze dwa kwartały charakteryzują się stabilizacją i małymi zmianami w relatywnym absolutnym obciążeniu. Duże zmiany zachodzą w kwartale trzecim, współczynnik rośnie dla takich województw jak Mazowieckie czy Kujawsko-Pomorskie. Ponadto maleje on dla takich województw jak: Podlaskie czy Warmińsko-Mazurskie. W ostatnim kwartale współczynnik osiąga bardzo podobne wielkości dla wszystkich województw.

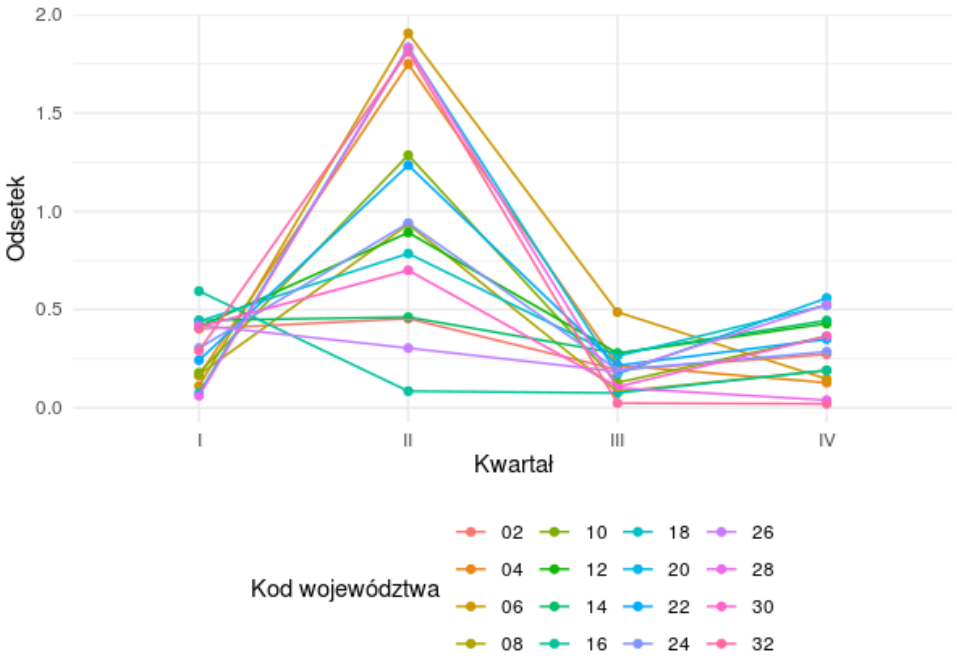
Rysunek 5.12 przedstawia rozkład relatywnego absolutnego obciążenia między odsetkiem ofert pracy w Badaniu Popytu na Prace a OLX według województw. Ponownie, duże zmiany współczynnika zachodzą w kwartale drugim. W kwartałach 1, 3 i 4 współczynnik charakteryzuje się podobnymi wartościami.

Duże wielkości relatywnego absolutnego obciążenia między Badaniami Popytu na Prace a CBOP można zaobserwować zwłaszcza w kategorii ’Pracownicy przy pracach prostych’ jak i ’Pracownicy usług i sprzedawcy’(rysunek: 5.15). Różnica ta jest szczególnie widoczna podczas



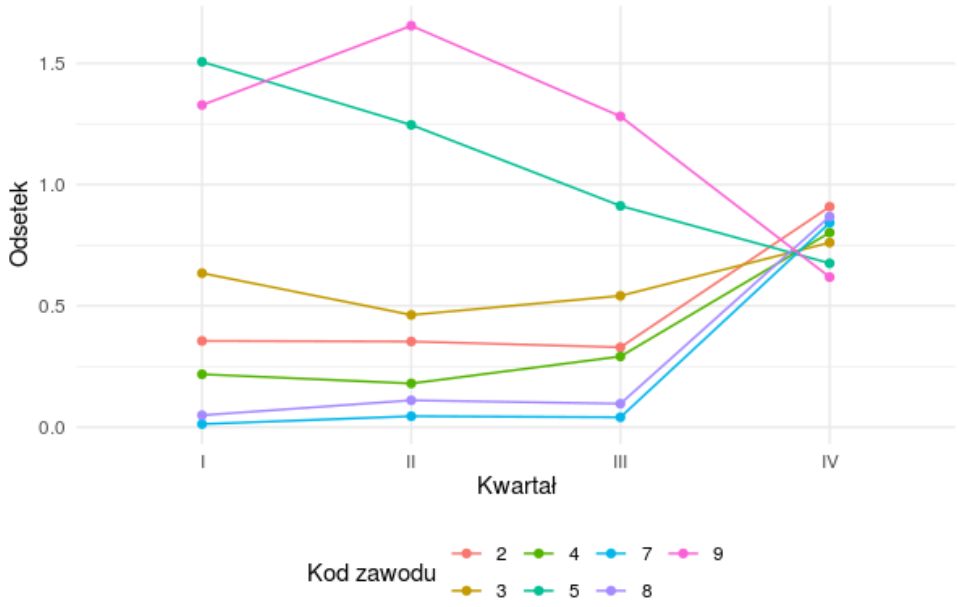
Rysunek 5.11. Rozkład relatywnego absolutnego obciążenia między odsetkiem ofert pracy w Badaniu Popytu na Prace a CBOP według województw

Źródło: Opracowanie własne.



Rysunek 5.12. Rozkład relatywnego absolutnego obciążenia między odsetkiem ofert pracy w Badaniu Popytu na Prace a OLX według województw

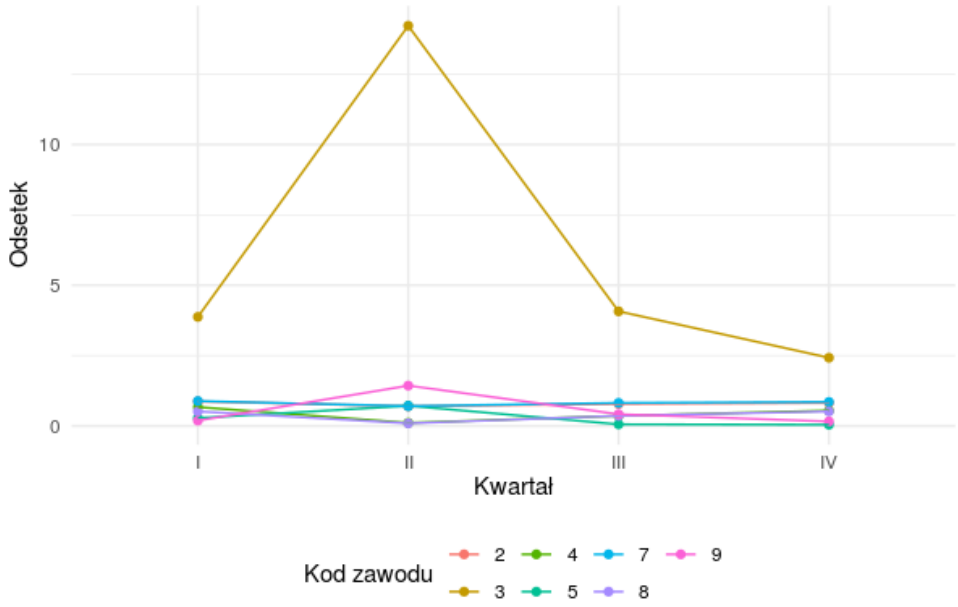
Źródło: Opracowanie własne.



Rysunek 5.13. Rozkład relatywnego absolutnego obciążenia między odsetkiem ofert pracy w Badaniu Popytu na Prace a CBOP według zawodów

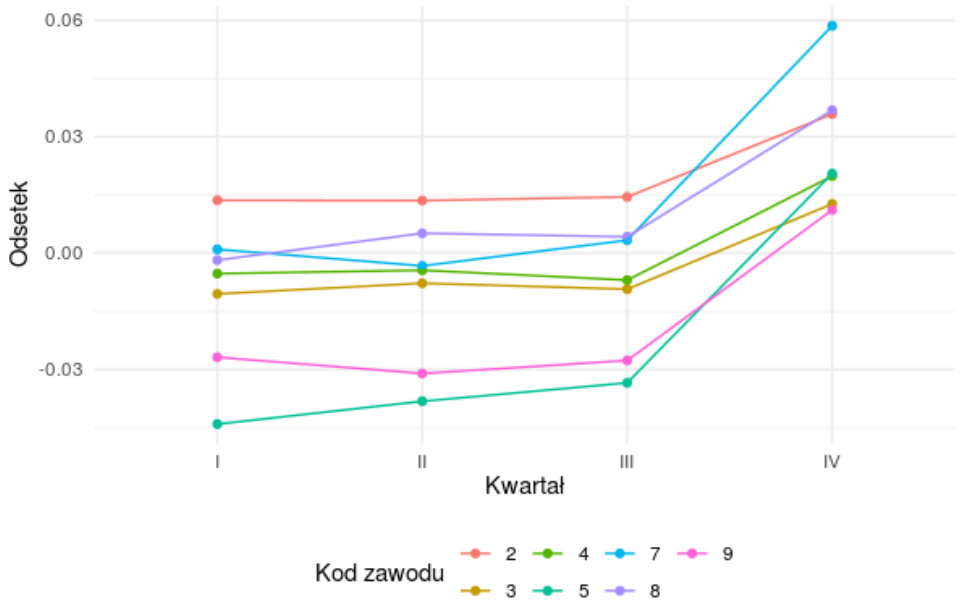
Źródło: Opracowanie własne.

pierwszych trzech kwartałów. Jednakże maleje ona, tak by w ostatnim kwartale zrównać się z obciążeniem dla pozostałych zawodów, dla których współczynnik ten rósł.



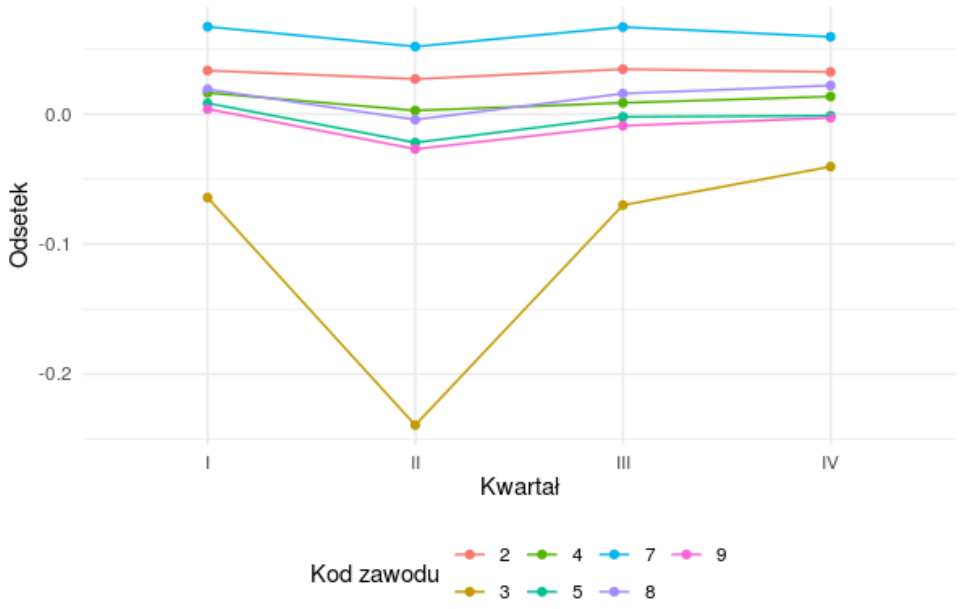
Rysunek 5.14. Rozkład relatywnego absolutnego obciążenia między odsetkiem ofert pracy w Badaniu Popytu na Prace a OLX według zawodów

Źródło: Opracowanie własne.



Rysunek 5.15. Rozkład obciążenia między odsetkiem ofert pracy w Badaniu Popytu na Prace a CBOP według zawodów

Źródło: Opracowanie własne.



Rysunek 5.16. Rozkład obciążenia między odsetkiem ofert pracy w Badaniu Popytu na Prace a OLX według zawodów

Źródło: Opracowanie własne.

Na rysunku 5.16 widać rozkład relatywnego absolutnego obciążenia między odsetkiem

ofert pracy w Badaniu Popytu na Prace a OLX według zawodów. Szczególnie duże wartości obciążenia można zaobserwować w zawodzie trzecim: 'Technicy i inny średni personel'. Dotyczy to zwłaszcza kwartału drugiego. Jednakże nie jest to obserwacja która powinna dziwić, ponieważ zgadza się ona z poprzednim obserwacjami dotyczącymi kształtowania się ofert pracy w portalu OLX. Znaczące obciążenie dla tego zawodu spowodowane jest słabą jakością klasyfikacji modelu dla grupy trzeciej. Model ten przydzielał do tej grupy za dużo ofert. Ponadto szczególny wzrost w drugim kwartale spowodowany jest ogólnie dużą ilością ofert na portalu OLX w tym właśnie czasie. Warte odnotowanie są również nieduże wartości obciążenia dla pozostałych grup zawodów we wszystkich kwartałach.

Rysunek 5.15 przedstawia rozkład obciążenia między odsetkiem ofert pracy w Badaniu Popytu na Prace a CBOP według zawodów. Widoczna jest, przytoczona wcześniej, korelacja negatywna między tymi źródłami danych. Dla większości zawodów można zaobserwować trend rosnący. Najmocniejszy wzrost obciążenia można zaobserwować w czwartym kwartale.

Rysunek 5.16 przedstawia rozkład obciążenia między odsetkiem ofert pracy w Badaniu Popytu na Prace a OLX według zawodów. Ponownie, widać duże obciążenie trzeciej kategorii, największe w drugim kwartale. Pozostałe ofert utrzymują się na stały, stosunkowo niskim poziomie obciążenia.

5.3 Podsumowanie

Porównanie trzech źródeł informacji o rynku pracy: CBOP, Badania Popytu na Prace oraz portalu internetowego OLX pozwoliło na wyodrębnienie kilku cech łączących i różniących te źródła. Wszystkie trzy źródła cechują się dużą korelacją rozkładu odsetek ofert pracy dla poszczególnych województw. Jednakże portal internetowy OLX charakteryzował się dużą ilością ofert pracy w drugim kwartale, pozostałe trzy kwartały posiadały równomiernie rozłożony rozkład odsetek. W przypadku CBOP kwartał czwarty osiągał najmniejsze liczebności ofert pracy. Badania Popytu na Prace charakteryzowały się równomiernym rozłożeniem odsetek ofert pracy według kwartałów. W OLX oraz BPP szczególnie dużo ofert przypadało na województwo Mazowieckie. Rozłożenie odsetek ofert pracy według zawodów wykazało negatywną korelacje między CBOP a Badaniem Popytu na Prace. W przypadku OLX duża część ofert została przydzielona do trzeciej kategorii. Pozwala to na zaobserwowanie potencjalnych wad modelu klasyfikującego. Analiza ze względu na te trzy czynniki: województwo zawód oraz czas pokazała że: CBOP po-

siada korelacje przynajmniej średnią z pozostałymi źródłami ofert pracy. Podczas gdy OLX oraz BPP w przypadku zawodów i kwartałów nie wykazują korelacji. Największa korelacja między źródłami występuje w aspekcie rozdziału ofert na województwa.

Podsumowanie

Celem niniejszej pracy była ocena niestatystycznych źródeł danych na potrzeby opisu polskiego rynku pracy, w szczególności skupiono się na popycie na pracę. Aby ten cel osiągnąć wykorzystano następujące źródła danych:

- badanie Popyt na Pracę GUS,
- Badanie Kapitału Ludzkiego moduł poświęcony pracodawcom i ofertom pracy,
- Centralną Bazy Ofert Pracy (CBOP) Ministerstwa Rodziny, Pracy i Polityki Społecznej,
- Oferty pracy z portalu OLX.

W pierwszym rozdziale skupiono się na przeglądzie literatury oraz opisano wykorzystywane źródła statystyczne i niestatystyczne. Następnie porównano wyżej wymienione źródła uwzględniając takie charakterystyki jak cel powstania, definicję populacji i jednostki, sposób powstania, koszt czy terminowość.

W rozdziale drugim przeanalizowano wyniki Badania Kapitału Ludzkiego w zakresie przedsiębiorców poszukujących nowych pracowników. Wykorzystanie modeli Item Response Theory (IRT) umożliwiło zbadanie wpływu skłonności pracodawców, do zamieszczania ogłoszeń w wielu miejscach, na wybór konkretnego kanału komunikacji. Na podstawie otrzymanych wyników można stwierdzić, iż wykorzystanie Internetu podczas rekrutacji nowych pracowników zależy od wielkości firmy, sekcji PKD, której dotyczy ogłoszenie oraz województwa, w którym poszukiwani są nowi pracownicy. Oferty pracy zamieszczane w Internecie bardzo często pochodzą z wielkich firm. Duża liczba pracowników zmusza pracodawcę do wykorzystania źródła o szerokim zasięgu dotarcia. Priorytetem takich firm jest znalezienie pracownika, który będzie przede wszystkim efektywnie wykonywał swoją prace. Małe przedsiębiorstwa znacznie rzadziej poszukują nowych ludzi przez Internet. Jest to spowodowane inną motywacją przedsiębiorcy. Pracodawca takiego rodzaju firm również szuka osoby o dużych kompetencjach, jednak

woli zawęzić obszar swoich poszukiwań do lokalnych form zamieszczania ogłoszeń. Świadczy to potrzebie pozyskania przede wszystkich zaufanych i lojalnych pracowników. Anonimowość i różnorodność osób odpowiadających na ogłoszenia zamieszczane w Internecie może budzić niepokój wśród pracodawców. Dodatkowo oferty pracy z branż silnie związanych z lokalizacją miejsca pracy zdecydowanie częściej zamieszczane są w urzędach pracy niżeli w Internecie. Przykładami takich ofert są wolne stanowiska pracy dla górników lub marynarzy. Warto również wspomnieć, iż portale Internetowe bardzo często nie uwzględniają stanowisk państwowych dla lekarzy, pielęgniarek, nauczycieli czy też policjantów. Na podstawie wyników modeli można również zauważyć, iż wykorzystanie Internetu rośnie wraz ze wzrostem popytu na pracę. Wyniki potwierdzają, iż w województwach z największym popytem na pracę wstępuje duża popularność Internetu jako miejsca dodawania ogłoszeń. Na podstawie rozdziału drugiego podważone została więc reprezentatywność wykorzystania internetowych danych dotyczących polskiego rynku pracy. Rzetelna i efektywna analiza rynku pracy możliwa będzie jedynie przy użyciu danych pochodzących w wielu źródeł. Zgodnie z wiedzą autorów jest to pierwsze w Polsce zastosowanie modelu IRT do oceny reprezentatywności kanałów komunikacji pracodawców z potencjalnymi pracownikami.

Jednym z celów szczegółowych trzeciego rozdziału było zbadanie danych pochodzących z Centralnej Bazy Ofert Pracy Urzędów Pracy. Po analizie CBOP, można stwierdzić, że dane te mogą być dobrym źródłem do badania rynku pracy w Polsce w zakresie popytu na pracę. Zbiór danych z tego źródła nie był pozbawiony błędów, jednak były to błędy w dużej mierze wynikającego z nieprawidłowego kodowania zmiennych oraz braków danych. Ponadto, w przeciwieństwie do danych pochodzących z badania GUS, zbiór ten zawierał informacje o wszystkich sekcjach PKD. Badając rozkłady wolnych miejsc pracy względem owych sekcji zauważono podobieństwo do danych z Badania Popytu na Pracę dla poszczególnych sekcji: sekcje A (Rolnictwo, leśnictwo, łowiectwo i rybactwo), B (Górnictwo i wydobywanie), D (Wytwarzanie i zaopatrywanie w energię elektryczną, gaz, parę wodną, gorącą wodę i powietrze do układów klimatyzacyjnych) i E (Dostawa wody; gospodarowanie ściekami i odpadami oraz działalność związana z rekultywacją) wykazały niskie wartości odsetków wolnych miejsc pracy w obu zbiorach, a sekcje C (Przetwórstwo przemysłowe), F (Budownictwo) oraz G (Handel hurtowy i detaliczny; naprawa pojazdów samochodowych, włączając motocykle) znacznie wyższymi niż pozostałe sekcje. Dodatkowo, za pomocą współczynnika V Cramera, zbadano wpływ poszczególnych zmiennych na publikowanie ofert z danego zawodu w bazie urzędów pracy. Okazało się, że istnieje słaba zależ-

ność pomiędzy zmiennymi. Porównanie dostępnego źródła niestatystycznego ze statystycznym było drugim celem szczegółowym trzeciego rozdziału. Natomiast ostatni cel stanowiła korekcja błędów nielosowych w zbiorze CBOP z wykorzystaniem zmiennych z danych GUS. W tym celu zastosowana została kalibracja. Jako zmienne pomocnicze wykorzystano zmienne informujące o województwie oraz zawodzie, najpierw każdą z osobna, a następnie obie zmienne razem. Kalibracja z wykorzystaniem zmiennej pomocniczej odnoszącej się do PKD jedynie nieznacznie zredukowała błąd pokrycia w bazie CBOP. Według wiedzy autorów jest to pierwsze wykorzystanie danych pochodzących z Centralnej Bazy Ofert Pracy do pomiaru oraz opisu rynku pracy w Polsce w kontekście metody reprezentacyjnej.

Klasyfikacyjne modele uczenia maszynowego były obiektem zainteresowania rozdziału czwartego. Celem modeli była klasyfikacja ofert z portalu internetowego OLX. Zastosowane zostały modele: wielomianowej regresji logistycznej LASSO oraz Naiwnego Bayesa. W tym celu wykorzystano dwa pakiety: `H2O` (LeDell i in., 2019) oraz `glmnet` (Friedman i in., 2010). Dane, zarówno na których podstawie budowano, jak i same dane z OLX, podane zostały obróbce związanej z *text mining.* Test ofert pracy został w pierwszej kolejności poddany tokenizacji, następnie stematyzacji oraz lematyzacji. Usunięte zostały słowa wliczające się do *stop words*, słowa mające mniej niż trzy znaki, ostatecznie usunięte polskie znaki. Dla obu zbiorów zbudowane zostały obiekty *Document Term Matrix*. Wszystkie te kroki były niezbędne do zbudowania modeli mogących klasyfikować oferty pracy. Proces modelowania odbywał się z wykorzystaniem takich metod uczenia maszynowego jak sprawdzanie krzyżowe czy też repolaryzacji. Modele stworzone za pomocą pakietu `H2O` zostały obliczane za pomocą przetwarzania równoległego na wielu rdzeniach. Było to możliwe dzięki wykorzystaniu infrastruktury w ramach *InnoUEP*. Oferty pracy pochodzące z portalu internetowego OLX zostały przefiltrowane tak, by zostały tylko te które był aktualne podczas ostatniego miesiąca każdego z kwartałów. Ponadto usunięte zostały powtarzające się oferty oraz zweryfikowany został ich status. Dane te zostały uzyskane za pomocą *webscrapingu*. Ponadto, oferty pracy z OLX charakteryzowały się innym kodowaniem województw niż ten z którego korzysta Główny Urząd Statystyczny(GUS). Ważnym elementem było przekodowanie zmiennej określającej województwo do standardów kodowania GUS. Zostało to wykonane za pomocą manualnego sprawdzania wewnętrznego kodowania serwisu OLX, weryfikacji jakie miasta obejmuje i przydzielania go do kategorii GUS. Oferty portalu internetowego posiadają również inny sposób kodowania niż ten zgodny z Klasyfikacja Zawodów i Specjalności. Całość danych z OLX wynosiła 411 787 ofert pracy. Wylosowano 1%

z całości danych, tak by każdy kwartał posiadał równą reprezentacje. Następnie wylosowana próbka została poddana manualnej klasyfikacji do standardów kategorii z Klasyfikacji Zawodów i Specjalności. Tak oznaczone dane zostały skonfrontowane z klasyfikacjami wcześniej wyszkolonych modeli. Jakość klasyfikacja została zweryfikowana za pomocą macierzy klasyfikacji oraz miary F1. Modelem osiągającym najlepsze wyniki był model wielomianowej regresji logistycznej LASSO zbudowany w pakiecie H2O. Model ten charakteryzował się wysokimi wynikami w kategorii takiej jak: 'operatorzy i monterzy maszyn i urządzeń'. Gorsze wyniki natomiast osiągał w kategoriach: 'technicy i inny średni personel'(kategoria 3) oraz 'personel przy pracach prostych'(kategoria 9). Kategoria 3 była najczęściej wybierana przez model. Ogólny charakter tych kategorii nie pozwolił modelowi na wyodrębnienie unikalnych słów które mógłby w sposób jednoznaczny klasyfikować oferty. Świadczą o tym chodzi by słowa, które według modelu miały największy wpływ na klasyfikacje. Słowa te to między innymi: kierowca, sprzedawca, operator, magazynier czy nauczyciel. W pierwszej 10 najbardziej istotnych słów nie ma takich które mogłyby determinować kategorie 3 lub 9. Najważniejsze słowa pochodzą z kategorii 8, 5 czy też 2. Ostatecznie można stwierdzić że oferty z portalu internetowego mogą być z sukcesem poddawane klasyfikacji przez zbudowane na innych danych modele. Jednakże warto stworzyć rozgraniczenie między kategoriami. Do przyszłych rozważań warto wpisać zastosowanie bardziej szczegółowego poziomu Klasyfikacji Zawodów i Specjalności, zwłaszcza w kategorii 3 i 9. Ponadto warto by było sprawdzić czy różne modele nie sprawdzają się lepiej w konkretnych kategoriach. Proces ten mógłby polegać na sprawdzaniu z jakim prawdopodobieństwem pierwszy model klasyfikuje ofertę, jeżeli prawdopodobieństwo jest za małe, to oferta mogła by być analizowana przez kolejne modele. Warto również byłoby zweryfikować dane pozyskane na innych portalach internetowych. Zgodnie z wiedzą autora jest to pierwsze wykorzystanie portalu OLX na potrzeby opisu popytu na pracę ze szczególnym uwzględnieniem aspektu reprezentatywności.

Pracę kończy rozdział porównujący badanie Popytu na Pracę GUS z CBOP oraz OLX w zakresie liczby ofert w 2017 roku według województw oraz zawodów. W Centralnej Bazie Ofert Pracy najczęściej umieszczane są ogłoszenia o pracę dla robotników przemysłowych i rzemieślników (7) oraz dla pracowników usług i sprzedawców (5). Najmniej ofert na portalu OLX umieszczanych jest dla przedstawicieli władz publicznych, wyższych urzędników i kierowników co pokrywa się w tym przypadku z bazą CBOP. Na portalu OLX najwięcej umieszczanych ofert o pracę pochodzi jest dla techników oraz średniego personelu – ta grupa odstaje od pozostałych grup

zawodów. Na drugim miejscu pod względem odsetków umieszczanych ofert pracy jest grupa piąta, czyli pracownicy usług i sprzedawcy.

W dalszych możliwych krokach badań należałoby skupić się na połączeniu bazy CBOP oraz OLX aby stworzyć jedno niestatystyczne źródło, które można następnie połączyć z badaniem Popytu na Pracę. Kolejnym ciekawym zagadnieniem mogłaby być analiza trendów oraz indeksów popularności poszczególnych zawodów.

Bibliografia

Alfons, A. & Templ, M. (2013). Estimation of Social Exclusion Indicators from Complex Surveys: The R Package laeken. *Journal of Statistical Software*, *54*(15), 1–25.

Antosz, P. (2014). *Raport metodologiczny z badań realizowanych w 2014 roku w ramach V edycji Balansu Kapitału Ludzkiego*.

Baker, F. & Kim, S.-H. (2017). *The Basics of Item Response Theroy Using R*.

Balicki, A. (2014). Metody imputacji brakujących danych w badaniach statystycznych.

Beręsewicz, M. (2016). *Internet data sources for real estate market statistics* (Rozprawa doktorska).

Beręsewicz, M. (2017). A Two-Step Procedure to Measure Representativeness of Internet Data Sources. *International Statistical Review*, *85*(3), 473–493.

Beręsewicz, M. & Szymkowiak, M. (2015). Big Data w Statystyce Publicznej – nadzieje, osiągnięcia, wyzwania i zagrożenia. *Ekonometria*, *2*, 9–22.

Biuro Inwestycji i Cykli Ekonomicznych. (2019). Dostęp z https://http://biec.org/produkty/

Brzezińska, J. (2016). Modele IRT i modele Rascha w Badaniach testowych.

Carl-Erik Särndal, S. L. (2005). *Estimation in Surveys with Nonresponse*. John Wiley i Sons, Ltd.

Chalmers, R. P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software*, *48*(6), 1–29. doi:10.18637/jss.v048.i06

Cichosz, P. (2000). *Systemy uczące się*. Wydawnictwa Naukowo-Techniczne.

Conway, D. & White, J. M. (2012). *Machine learning for hackers*. O'Reilly Media.

Cook, D. (2016). *Practical machine learning with H2O: powerful, scalable techniques for deep learning and AI*. O'Reilly Media, Inc.

Dangeti, P. (2017). *Statistics for machine learning*. Packt Publishing Ltd.

Eurostat. (2019). Statistics Explained – glossary. Dostęp z https://ec.europa.eu/eurostat/statistics-explained/index.php/Glossary:Coverage

Feinerer, I., Hornik, K. & Meyer, D. (2008). Text Mining Infrastructure in R. *Journal of Statistical Software*, *25*(5), 1–54.

Field, A., Miles, J. & Field, Z. (2012). *Discovering Statistics Using R*. SAGE Publications.

Forbes. (2017). Ile waży praca? Dostęp z https://www.forbes.pl/technologie/jak-wiele-danych-produkujemy-kazdego-dnia/4mn4w69

Friedman, J., Hastie, T. & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, *33*(1), 1–22.

Główny Urząd Statystyczny. (2015). Popyt na pracę w 2015r.

Główny Urząd Statystyczny. (2017). *Społeczeństwo informacyjne w Polsce w 2017 roku*. GUS, Warszawa.

Główny Urząd Statystyczny. (2018a). *Popyt na pracę w kwartale II w 2018 roku*.

Główny Urząd Statystyczny. (2018b). *Program Badań Statystycznych, Statystyki Publicznej*.

Główny Urząd Statystyczny. (2019a). Klasyfikacja Zawodów i Specjalności 2014. Dostęp z https://stat.gov.pl/Klasyfikacje/doc/kzs/kzs_pp.htm

Główny Urząd Statystyczny. (2019b). *Pojęcia stosowane w statystyce publicznej*. GUS, Warszawa.

Grupa OLX. (2019). Regulamin serwisu. Dostęp z https://pomoc.olx.pl/hc/pl/articles/360000828525#r5

Jivani, A. G. i in. (2011). A comparative study of stemming algorithms. *Int. J. Comp. Tech. Appl*, *2*(6), 1930–1938.

Kleka, P. (2012). *Zastosowanie teorii odpowiadania na pozycje testowe (IRT) do tworzenia skróconej wersji testów i kwestionariuszy psychologicznych* (Rozprawa doktorska).

Klimanek, T. & Szymkowiak, M. (2017). Podejście kalibracyjne w badaniu losów absolwentów na przykładzie projektu „Kadry dla gospodarki". *Wiadomości statystyczne*.

Konarski, R. (2004a). MODEL CECHY LATENTNEJ W ANALIZIE PSYCHOMETRYCZNEJ TESTÓW I POZYCJI TESTOWYCH.

Konarski, R. (2004b). Model cechy latentnej w analizie psychometrycznej testów i pozycji testowych. *W: B. Niemierko, H. Szaleniec.(red.) Standadry wymagań i normy testowe w diagnostyce edukacyjnej. Kraków: Polskie Towarzystwo Diagnostyki Edukacyjnej*.

Kondratek, B. & Pokropek, A. (2013). Teoria odpowiedzi na pozycje testowe: jednowymiarowe modele dla cech ukrytych o charakterze ciągłym.

Koronacki, J. & Ćwik, J. (2008). *Statystyczne systemy uczące się*. Akademicka Oficyna Wydawnicza EXIT.

Kowarik, A. & Templ, M. (2016). Imputation with the R Package VIM. *Journal of Statistical Software*, *74*(7), 1–16. doi:10.18637/jss.v074.i07

Krzyśko, M., Wołyński, W., Górecki, T. & Skorzybut, M. (2008). Systemy uczące się. *Rozpoznawanie wzorców, analiza skupień i redukcja wymiarowości. WNT, Warszawa*.

Landry, M. (2018). *Machine Learning with R and H2O*. H2O.ai, Inc.

Lantz, B. (2013). *Machine learning with R*. Packt Publishing Ltd.

LeDell, E., Gill, N., Aiello, S., Fu, A., Candel, A., Click, C., … Malohlava, M. (2019). *h2o: R Interface for 'H2O'*. R package version 3.24.0.5.

Lohr, S. (2010). *Sampling: Design and Analysis*. Brooks/Cole, Cengage Learning.

Lumley, T. (2004). Analysis of Complex Survey Samples. *Journal of Statistical Software*, *9*(1), 1–19. R package verson 2.2.

Ministerstwo Rodziny, Pracy i Polityki Społecznej. (2019). Dostęp z http://psz.praca.gov.pl/-/926922-centralna-baza-ofert-pracy-cbop-

Nykodym, T., Hussami, N., Kraljevic, T., Rao, A. & Wang, A. (2015). *Generalized Linear Modeling with H2O*.

OECD. (2018). Strengthening SMEs and Entrepreneurship for Productivity and Inclusive Growth.

Pater, R. (2017). Internetowe oferty pracy jako źródło informacji o zapotrzebowaniu na kompetencji.

Publiczne Służby Zatrudnienia. (2019). Dostęp z https://www.praca.gov.pl/eurzad/index.eup#/inneSprawy/wyborUrzedu?dest=EURZAD

R Core Team. (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.

Rizopoulos, D. (2006). ltm: An R package for Latent Variable Modelling and Item Response Theory Analyses. *Journal of Statistical Software*, *17*(5), 1–25.

Silge, J. & Robinson, D. (2016). tidytext: Text Mining and Analysis Using Tidy Data Principles in R. *JOSS*, *1*(3). doi:10.21105/joss.00037

Silge, J. & Robinson, D. (2019). *Text Mining with R: A Tidy Approach*. O'Reilly Media, Inc.

Sobczyk, M. (2000). *Statystyka*. Wydawnictwo Naukowe PWN.

Szymkowiak, M. (2009). *Estymatory kalibracyjne w badaniu budżetów gospodarstw domowych* (Rozprawa doktorska).

Tibshirani, R., Wainwright, M. & Hastie, T. (2015). *Statistical learning with sparsity: the lasso and generalizations*. Chapman i Hall/CRC.

Wei, T. & Simko, V. (2017). *R package ćorrplot": Visualization of a Correlation Matrix*. (Version 0.84).

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wickham, H. (2017). *tidyverse: Easily Install and Load the 'Tidyverse'*. R package version 1.2.1.

Wikipedia. (2019). Urząd Pracy. Dostęp z https://pl.wikipedia.org/wiki/Urz%C4%85d_pracy

Spis tabel

Spis rysunków

Spis Programów

Dodatek A

Wyniki kalibracji dla wszystkich kwartałów (Rozdział 3)

Tabela A.1. Tabela dla sum wag kalibracyjnych i wartości globalnych dla populacji dla IV kwartału (dla województw)

województwo	suma wag kalibracyjnych	wartości globalne
dolnośląskie	10 997	10 997
kujawsko-pomorskie	4 127	4 127
lubelskie	3 097	3 097
lubuskie	3 067	3 067
łódzkie	7 669	7 669
małopolskie	11 345	11 345
mazowieckie	25 506	25 506
opolskie	2 141	2 141
podkarpackie	3 383	3 383
podlaskie	2 272	2 272
pomorskie	6 808	6 808
śląskie	14 611	14 611
świętokrzyskie	2 946	2 946
warmińsko-mazurskie	1 896	1 896
wielkopolskie	13 818	13 818
136 Uncompiled Changes zachodniopomorskie	3 914	3 914

Źródło: Opracowanie własne.

Tabela A.2. Odsetki dla wag d_i oraz w_i w zależności od zawodu dla IV kwartału (dla województw)

główna grupa zawodu	suma wag d_i	suma wag w_i	odsetki dla wag d_i	odsetki dla wag w_i
1: *przedstawiciele władz publicznych, wyżsi urzędnicy i kierownicy* 113,00	884,61	0,64	0,75	
2: *specjaliści* 777,00	4 933,83	4,40	4,20	
3: *technicy i inny średni personel*	1 165,00	7 094,60	6,59	6,03
4: *urzędnicy biurowi*	1 426,00	10 999,93	8,07	9,35
5: *pracownicy usług i sprzedawcy*	3 078,00	20 386,81	17,42	17,34
7: *robotnicy przemysłowi i rzemieślnicy*	4 345,00	26 055,51	24,59	22,1
8: *operatorzy i monterzy maszyn i urządzeń*	2 949,00	19 252,57	16,69	16,37
9: *pracownicy wykonujący prace proste*	3 816,00	27 989,14	21,60	23,80

Źródło: Opracowanie własne.

Tabela A.3. Tabela dla sum wag kalibracyjnych i wartości globalnych dla populacji dla IV kwartału (dla sekcji PKD)

sekcja PKD	suma wag kalibracyjnych	wartości globalne
C	30 135	30 135
E	738	738
F	17 933	17 933
G	18 546	18 546
H	11 885	11 885
I	4 131	4 131
J	5 702	5 702
K	2 996	2 996
L	1 146	1 146
M	6 001	6 001
N	3 747	3 747
O	4 635	4 635
P	2 041	2 041
Q	4 960	4 960
S	1 228	1 228
inne	1 773	1 773

Źródło: Opracowanie własne.

Tabela A.4. Odsetki dla wag d_i oraz w_i w zależności od zawodu dla IV kwartału (dla sekcji PKD)

główna grupa zawodu	suma wag d_i	suma wag w_i	odsetki dla wag d_i	odsetki dla wag w_i
1: *przedstawiciele władz publicznych, wyżsi urzędnicy i kierownicy*	113,00	698,13	0,64	0,59
2: *specjaliści*	777,00	6 963,64	4,40	5,92
3: *technicy i inny średni personel*	1 165,00	9 449,12	6,59	8,04
4: *urzędnicy biurowi*	1 426,00	11 103,56	8,07	9,44
5: *pracownicy usług i sprzedawcy*	3 078,00	20 077,00	17,42	17,07
7: *robotnicy przemysłowi i rzemieślnicy*	4 345,00	30 940,64	24,59	26,31
8: *operatorzy i monterzy maszyn i urządzeń*	2 949,00	19 027,43	16,69	16,18
9: *pracownicy wykonujący prace proste*	3 816,00	19 337,47	21,60	16,44

Źródło: Opracowanie własne.

Tabela A.5. Odsetki dla wag d_i oraz w_i w zależności od zawodu dla IV kwartału (dla województwa oraz sekcji PKD)

główna grupa zawodu	suma wag d_i	suma wag w_i	odsetki dla wag d_i	odsetki dla wag w_i
1: *przedstawiciele władz publicznych, wyżsi urzędnicy i kierownicy*				
2: *specjaliści*				
3: *technicy i inny średni personel*				
4: *urzędnicy biurowi*				
5: *pracownicy usług i sprzedawcy*				
7: *robotnicy przemysłowi i rzemieślnicy*				
8: *operatorzy i monterzy maszyn i urządzeń*				
9: *pracownicy wykonujący prace proste*				

Źródło: Opracowanie własne.

Tabela A.6. Tabela dla sum wag kalibracyjnych i wartości globalnych dla populacji dla III kwartału (dla województw)

województwo	suma wag kalibracyjnych	wartości globalne
dolnośląskie	13 248	13 248
kujawsko-pomorskie	4 089	4 089
lubelskie	3 485	3 485
lubuskie	3 106	3 106
łódzkie	6 918	6 918
małopolskie	9 782	9 782
mazowieckie	27 889	27 889
opolskie	3 597	3 597
podkarpackie	3 209	3 209
podlaskie	1 388	1 388
pomorskie	6 394	6 394
śląskie	14 436	14 436
świętokrzyskie	3 877	3 877
warmińsko-mazurskie	2 262	2 262
wielkopolskie	13 702	13 702
zachodniopomorskie	4 442	4 442

Źródło: Opracowanie własne.

Tabela A.7. Odsetki dla wag d_i oraz w_i w zależności od zawodu dla III kwartału (dla województw)

główna grupa zawodu	suma wag d_i	suma wag w_i	odsetki dla wag d_i	odsetki dla wag w_i
1: *przedstawiciele władz publicznych, wyżsi urzędnicy i kierownicy*	874,00	885,11	0,69	0,73
2: *specjaliści*	6 777,00	6 583,26	5,36	5,40
3: *technicy i inny średni personel*	8 112,00	7 875,58	6,41	6,46
4: *urzędnicy biurowi*	9 706,00	9 780,84	7,67	8,03
5: *pracownicy usług i sprzedawcy*	20 863,00	20 139,03	16,49	16,53
7: *robotnicy przemysłowi i rzemieślnicy*	34 646,00	33 118,47	27,38	27,19
8: *operatorzy i monterzy maszyn i urządzeń*	18 743,00	17 236,93	14,81	14,15
9: *pracownicy wykonujący prace proste*	26 796,00	26 204,79	21,18	21,51

Źródło: Opracowanie własne,

Tabela A.8. Tabela dla sum wag kalibracyjnych i wartości globalnych dla populacji dla III kwartału (dla sekcji PKD)

sekcja PKD	suma wag kalibracyjnych	wartości globalne
C	33 534	33 534
E	674	674
F	20 323	20 323
G	20 811	20 811
H	10 510	10 510
I	5 409	5 409
J	6 320	6 320
K	2 935	2 935
L	1 382	1 382
M	7 277	7 277
N	5 822	5 822
O	4 254	4 254
P	3 117	3 117
Q	4 421	4 421
S	1 931	1 931
inne	5 409	5 409

Źródło: Opracowanie własne.

Tabela A.9. Odsetki dla wag d_i oraz w_i w zależności od zawodu dla III kwartału (dla sekcji PKD)

główna grupa zawodu	suma wag d_i	suma wag w_i	odsetki dla wag d_i	odsetki dla wag w_i
1: *przedstawiciele władz publicznych, wyżsi urzędnicy i kierownicy*	874,00	1 007,01	0,69	0,77
2: *specjaliści*	6 777,00	9 218,64	5,36	7,05
3: *technicy i inny średni personel*	8 112,00	9 514,66	6,41	7,27
4: *urzędnicy biurowi*	9706,00	10136,86	7,67	7,75
5: *pracownicy usług i sprzedawcy*	20 863,00	22 739,58	16,49	17,39
7: *robotnicy przemysłowi i rzemieślnicy*	34 646,00	36 119,50	27,38	27,62
8: *operatorzy i monterzy maszyn i urządzeń*	18 743,00	19 721,46	14,81	15,08
9: *pracownicy wykonujący prace proste*	26 796,00	22 337,29	21,18	17,08

Źródło: Opracowanie własne,

Tabela A.10. Odsetki dla wag d_i oraz w_i w zależności od zawodu dla III kwartału (dla województwa oraz sekcji PKD)

główna grupa zawodu	suma wag d_i	suma wag w_i	odsetki dla wag d_i	odsetki dla wag w_i
1: *przedstawiciele władz publicznych, wyżsi urzędnicy i kierownicy*	874,00	1 056,88	0,69	0,81
2: *specjaliści*	6 777,00	9 648,39	5,36	7,38
3: *technicy i inny średni personel*	8 112,00	9 452,40	6,41	7,23
4: *urzędnicy biurowi*	9 706,00	10 358,58	7,67	7,92
5: *pracownicy usług i sprzedawcy*	20 863,00	22 232,34	16,49	17,00
7: *robotnicy przemysłowi i rzemieślnicy*	34 646,00	35 574,56	27,38	27,20
8: *operatorzy i monterzy maszyn i urządzeń*	18 743,00	19 512,63	14,81	14,92
9: *pracownicy wykonujący prace proste*	26 796,00	22 959,21	21,18	17,55

Źródło: Opracowanie własne.

Tabela A.11. Tabela dla sum wag kalibracyjnych i wartości globalnych dla populacji dla II kwartału (dla województw)

województwo	suma wag kalibracyjnych	wartości globalne
dolnośląskie	13 248	13 248
kujawsko-pomorskie	4 089	4 089
lubelskie	3 485	3 485
lubuskie	3 106	3 106
łódzkie	6 918	6 918
małopolskie	9 782	9 782
mazowieckie	27 889	27 889
opolskie	3 597	3 597
podkarpackie	3 209	3 209
podlaskie	1 388	1 388
pomorskie	6 394	6 394
śląskie	14 436	14 436
świętokrzyskie	3 877	3 877
warmińsko-mazurskie	2 262	2 262
wielkopolskie	13 702	13 702
zachodniopomorskie	4 442	4 442

Źródło: Opracowanie własne.

Tabela A.12. Odsetki dla wag d_i oraz w_i w zależności od zawodu dla II kwartału (dla województw)

główna grupa zawodu	suma wag d_i	suma wag w_i	odsetki dla wag d_i	odsetki dla wag w_i
1: *przedstawiciele władz publicznych, wyżsi urzędnicy i kierownicy*	775,00	785,65	0,61	0,64
2: *specjaliści*	5 778,00	5 718,69	4,58	4,69
3: *technicy i inny średni personel*	8 043,00	7 613,51	6,38	6,25
4: *urzędnicy biurowi*	9549,00	9773,42	7,58	8,02
5: *pracownicy usług i sprzedawcy*	21 103,00	20 600,98	16,74	16,91
7: *robotnicy przemysłowi i rzemieślnicy*	33 912,00	31 951,19	26,90	26,23
8: *operatorzy i monterzy maszyn i urządzeń*	19 215,00	18 190,21	15,24	14,93
9: *pracownicy wykonujący prace proste*	27 671,00	27 190,34	21,95	22,32

Źródło: Opracowanie własne,

Tabela A.13. Tabela dla sum wag kalibracyjnych i wartości globalnych dla populacji dla II kwartału (dla sekcji PKD)

sekcja PKD	suma wag kalibracyjnych	wartości globalne
C	32 153	32 153
E	681	681
F	18 921	18 921
G	18 374	18 374
H	11 707	11 707
I	5 078	5 078
J	6 475	6 475
K	3 202	3 202
L	1 333	1 333
M	5 229	5 229
N	5 443	5 443
O	3 735	3 735
P	1 689	1 689
Q	4 092	4 092
S	1 924	1 924
inne	1 788	1 788

Źródło: Opracowanie własne.

Tabela A.14. Odsetki dla wag d_i oraz w_i w zależności od zawodu dla II kwartału (dla sekcji PKD)

główna grupa zawodu	**suma wag d_i**	**suma wag w_i**	**odsetki dla wag d_i**	**odsetki dla wag w_i**
1: *przedstawiciele władz publicznych, wyżsi urzędnicy i kierownicy*	775,00	865,80	0,61	0,71
2: *specjaliści*	5 778,00	7 384,62	4,58	6,06
3: *technicy i inny średni personel*	8 043,00	8 796,96	6,38	7,22
4: *urzędnicy biurowi*	9549,00	10523,00	7,58	8,64
5: *pracownicy usług i sprzedawcy*	21 103,00	20 455,34	16,74	16,79
7: *robotnicy przemysłowi i rzemieślnicy*	33 912,00	33 047,90	26,90	27,13
8: *operatorzy i monterzy maszyn i urządzeń*	19 215,00	19 175,61	15,24	15,74
9: *pracownicy wykonujący prace proste*	27 671,00	21 574,77	21,95	17,71

Źródło: Opracowanie własne,

Tabela A.15. Odsetki dla wag d_i oraz w_i w zależności od zawodu dla II kwartału (dla województwa oraz sekcji PKD)

główna grupa zawodu	**suma wag d_i**	**suma wag w_i**	**odsetki dla wag d_i**	**odsetki dla wag w_i**
1: *przedstawiciele władz publicznych, wyżsi urzędnicy i kierownicy*	775,00	904,00	0,61	0,74
2: *specjaliści*	5 778,00	7 622,74	4,58	6,26
3: *technicy i inny średni personel*	8 043,00	8 688,39	6,38	7,13
4: *urzędnicy biurowi*	9 549,00	11 006,77	7,58	9,03
5: *pracownicy usług i sprzedawcy*	21 103,00	2 0359,53	16,74	16,71
7: *robotnicy przemysłowi i rzemieślnicy*	33 912,00	32 592,26	26,90	26,75
8: *operatorzy i monterzy maszyn i urządzeń*	19 215,00	18 918,23	15,24	15,53
9: *pracownicy wykonujący prace proste*	27 671,00	21 732,09	21,95	17,84

Źródło: Opracowanie własne.