UNIWERSYTET EKONOMICZNY W POZNANIU

Instytut Informatyki
i Ekonomii Ilościowej

Mateusz Marszałkowski

Rozszerzenie badania Popyt na Pracę o informacje z Centralnej Bazy Ofert Pracy z wykorzystaniem metod integracji danych

An extension of the the Demand for Labour survey using information from the Central Job Offers Database using data integration methods

Praca licencjacka

Promotor: dr Maciej Beręsewicz

Pracę przyjęto dnia:

Podpis promotora

Kierunek: Informatyka i Ekonometria

Poznań 2021

Spis treści

Wstęp **1**

1 Źródła danych w statystyce **3**

1.1 Znaczenie badań statystycznych 3

1.2 Źródła danych w statystyce 4

1.2.1 Klasyfikacja badań statystycznych 4

1.2.2 Etapy przeprowadzania badań statystycznych 5

1.2.3 Źródła statystyczne i niestatystyczne 6

1.2.4 Sposoby doboru próby 8

1.2.5 Klasyfikacja błędów losowych i nielosowych 11

1.3 Próby nielosowe w badaniach statystycznych 13

1.3.1 Badania internetowe i błędy nielosowe 13

1.3.2 Charakterystyka Big Data (”the 4 Vs”) jako próby nielosowej 15

1.4 Podsumowanie 16

2 Metody estymacji na podstawie prób nielosowych **18**

2.1 Główne założenia i oznaczenia 18

2.2 Estymacja z wykorzystaniem Propensity Score estimator 21

2.3 Estymacja z wykorzystaniem Doubly Robust estimator 24

2.4 Implementacja w języku python 25

2.5 Podsumowanie 27

3 Wyniki integracji badania Popyt na Pracę z Centralną Bazą Ofert Pracy **28**

3.1 Badanie symulacyjne 28

3.1.1 Założenia 28

3.1.2 Założenia badania symulacyjnego 29

3.1.3 Wyniki badania symulacyjnego 30
3.1.4 Wnioski 31
3.2 Badanie empiryczne 32
3.2.1 Założenia 32
3.2.2 Źródła danych 33
3.2.3 Wyniki badania empirycznego 36
3.2.4 Wnioski 40
3.3 Podsumowanie wyników 40

Podsumowanie **42**

Literatura **46**

Spis tabel **47**

Spis rysunków **48**

Spis programów w języku Python **49**

A Załączniki **50**
A.1 Wniosek do zgłoszenia krajowej oferty pracy 50
A.2 Opis zmiennych wykorzystanych w badaniu 53

Wstęp

Badanie cech danej populacji to jedno z ważniejszych zadań w procesie podejmowania decyzji opartych na danych (ang. *data-driven policy making*). Szczegółowe poznanie danej zbiorowości pozwala na podejmowanie efektywnych działań, począwszy od badania sentymentu wśród pewnej społeczności, poprzez prowadzenia kampanii marketingowej czy na sprawnym zarządzaniem państwem kończąc.

Najpełniejszy obraz rzeczywistości dostarcza spis powszechny (badanie pełne), który zapewnia otrzymanie informacji od każdego obiektu z określonej zbiorowości. Niemniej, przeprowadzanie spisu powszechnego to przedsięwzięcie bardzo kosztowne i czasochłonne, dlatego w większości współczesnych państw przeprowadzany jest co 10 lat.

Alternatywą do badania pełnego jest przeprowadzenie badań częściowych. Polega ona na wylosowaniu pewnego podzbioru populacji do przebadania. Dzięki temu, że mechanizm doboru próby jest losowy możliwe jest wnioskowanie na temat całej populacji tylko przy pomocy próby. Pierwsze podejścia do próby oszacowania danej populacji przy pomocy pewnej tylko części populacji sięgają połowy XVII wieku, natomiast w 1934 roku polski uczony Jerzy Spława-Neyman ugruntował metodę reprezentacyjną jako standardowe narzędzie statystyczne do wnioskowania o populacji dzięki wprowadzeniu pojęcia przedziałów ufności.

Niemniej, współcześnie coraz częściej do badań wykorzystywane są źródła niestatystyczne. Szczególnie atrakcyjnymi źródłami danych mogą być dane znane jako *Big Data* czy rejestry administracyjne. Ich atrakcyjność wynika z faktu, że pokrywają one znaczną część populacji. Jednakże trzeba pamiętać, że niestatystyczne źródła nie mogą być bezpośrednio wykorzystywane do badania cech populacji ze względu na błędy pokrycia. selekcji czy pomiaru.

Stąd pojawił się problem zachowania balansu między wykorzystaniem wygodnych i łatwo dostępnych źródeł Big Data (z których wnioskowanie statystyczne cechuje się obciążeniem), a wykorzystaniem kosztownych i czasochłonnych w wytworzeniu źródeł pochodzących z badań reprezentacyjnych (które pozwalają na wnioskowanie statystyczne). Rozwiązanie problemu

tego balansu zostało zaproponowane przez J. K. Kim i Wang (2019) w artykule *Sampling Techniques for Big Data Analysis*.

Autorzy proponują zastosowanie techniki zwanej Integracją Danych (*ang. Data Integration*). Polega ona na wykorzystaniu obu zbiorów – zarówno niestatystycznego zbioru (np. Big Data) jak i statystycznego badania reprezentacyjnego. Oba zbiory są następnie łączone i na podstawie połączonych zbiorów można przeprowadzić estymację poprzez zastosowanie podejść takich jak ważenie przez prawdopodobieństwo przynależności do źródła nielosowego (*ang. propensity score weighting*) czy estymatory podwójnie odporne (*ang. doubly robust estimators*), które umożliwiają korektę błędów zbioru Big Data oraz są oparte na solidnych podstawach teoretycznych.

Niniejsza praca skupi się na przedstawieniu powyższej metody. W pierwszym rozdziale zaprezentowane zostaną zagadnienia teoretyczne związane ze źródłami danych w statystyce oraz sposobami pozyskiwania danych. Drugi rozdział został poświęcony przedstawieniu obu najważniejszych estymatorów z metody Integracji Danych, czyli *propensity score estimator* oraz *doubly robust estimator*. Pokazana została również implementacja tychże estymatorów w języku `python`. Trzeci rozdział został podzielony na dwie sekcje. Pierwsza z nich to odtworzenie badania symulacyjnego z artykułu J. K. Kim i Wang (2019). Ma to na celu przedstawienie efektywności obu estymatorów. Druga sekcja trzeciego rozdziału to badanie empiryczne mające na celu praktyczne wykorzystanie estymatorów w estymacji pewnych cech populacji. Zbiorami wykorzystanymi w badaniu są dane pochodzące z Centralnej Bazy Ofert Pracy (CBOP), tutaj traktowane jako zbiór Big Data. Zbiorem statystycznym w badaniu empirycznym będzie badanie z popytu na pracę przeprowadzone przez Główny Urząd Statystyczny[1].

W badaniu empirycznym rozważone zostaną dwie zmienne. Pierwsza określa czy wakat oferowany jest w wymiarze 40 godzin tygodniowo (*pelen*). Druga wskazuje czy wakat dotyczy zatrudnienia tylko w systemie jednozmianowym (*jedna_zmiana*). Obie zmienne występują jedynie w zbiorze danych z CBOP. Dzięki metodzie Integracji Danych i połączeniu z danymi pochodzącymi z badania popytu na pracę możliwe będzie dostarczenie bardziej rzetelnych szacunków odsetka ofert pracy spełniających te warunki przez zmniejszenie obciążenia wynikającego z nielosowego charakteru zbioru CBOP.

[1]Dane z Popytu na Pracę zakupiono ze środków dotacji projakościowej dla Wydziału Informatyki i Gospodarki Elektronicznej (teraz Instytutu Informatyki i Ekonomii Ilościowej) Uniwersytetu Ekonomicznego w Poznaniu dla kierunku Informatyka i Ekonometria przez Ministerstwo Nauki i Szkolnictwa Wyższego (teraz Ministerstwo Edukacji i Nauki) na lata 2015-2017.

Rozdział 1

Źródła danych w statystyce

1.1 Znaczenie badań statystycznych

Zapotrzebowanie na informacje dotyczące pewnej dużej zbiorowości nie jest zjawiskiem nowym. Już w VI w. p. n. e. król rzymski *Servius Tullius* kazał przeprowadzić spis powszechny swojego królestwa. Pierwotne metody sprowadzały się do zbierania danych o stanie i funkcjonowaniu państwa. Były to badania o tyle istotne, że pozwalały ówczesnym władcom na zdobycie wiedzy o podlegających im zasobach (Ostasiewicz 2011, s. 13). Stan wojska, osoby gotowe do służby wojskowej czy liczba ludności do opodatkowania – znajomość tych danych była strategicznym elementem sprawowania władzy (Ostasiewicz 2011, s. 15).

Współcześnie, badania statystyczne pomagają zarówno organom państwowym w efektywnym zarządzaniu państwem jak i pozwalają na szczegółowy wgląd w opinie publiczne, poglądy oraz zmiany kulturowe i dynamiczne kształtujące się trendy społeczno-ekonomiczne. Badania statystyczne dają możliwość przeanalizowania przyczyn zachodzących zjawisk i zrozumienia fundamentalnych, jak i złożonych, procesów. Dzięki temu procesy decyzyjne mogą zostać zoptymalizowane, źródła wielu problemów zidentyfikowane, zaś sama materia społeczeństwa w którym żyjemy dogłębnie zrozumiana.

Jedną z ważniejszych organizacji statystycznych o zasięgu światowym jest *United Nations Statistics Division* (UNSD) będąca częścią *United Nations Department of Economic and Social Affairs* (UNDESA) (*United Nation's offical webpage* 2020). Główną rolą UNSD jest promowanie standaryzacji metodologii, terminologii i definicji używanych przez lokalne narodowe urzędy statystyczne. UNSD zapewnia również scentralizowane źródło danych na temat większości państw (wydając roczniki statystyczne obejmujące prawie wszystkie państwa świata). Orga-

nizacja ta pomaga również w utrzymaniu odpowiedniego standardu w przeprowadzaniu badań w celu zapewnienia rzetelnych i reprezentatywnych danych z całego świata(Ostasiewicz 2011), (*United Nation's offical webpage* 2020).

W Polsce najważniejszą instytucją odpowiedzialną za zbieranie danych statystycznych jest Główny Urząd Statystyczny (GUS) (*Ustawa z dnia 29 czerwca 1995 r. o statystyce publicznej.* 1995). Od ponad 100 lat jest on „oficjalnym organem administracyjnym państwowej statystyki administracyjnej" (*Ustawa z dnia 21 października 1919 r. o organizacji statystyki administracyjnej.* 1919). GUS poddaje badaniu większość sektorów państwa. Zbierane są informacje w dziedzinach takich jak: „Stan i ochrona środowiska", „Organizacja państwa", „samorząd terytorialny", „Gospodarka społeczna", „Rodzina", „Ludność, procesy demograficzne", „Wyznania religijne, grupy etniczne", „Rynek pracy" i wiele innych. Na 2021r. urząd planuje 249 badań statystycznych (GUS 2020).

Pierwotny sposób zbierania informacji (spis ludności) jest bardzo kosztowny, czasochłonny i wymagający logistycznie. Wymaga on dużej kadry przeszkolonych rachmistrzów, wydrukowania formularzy i przygotowania wszelkich materiałów pomocniczych. W związku z tym przeprowadzany jest w większości współczesnych państw średnio co 10 lat (Michalski 2004, s. 18–19). W zaplanowanym w Polsce na 2021r. „Narodowym Spisie Powszechnym Ludności i Mieszkań" maksymalny budżet na wszystkie prace spisowe wynosi 386 mln złotych (*Ustawa z dnia 9 sierpnia 2019 r. o narodowym spisie powszechnym ludności i mieszkań w 2021 r.* 2019).

1.2 Źródła danych w statystyce

1.2.1 Klasyfikacja badań statystycznych

W celu znalezienia odpowiedzi na pytania dotyczące pewnej populacji konieczne jest zdecydowanie się na dobór odpowiedniej metody badania statystycznego. Każda z metod będzie dyktować sposób przeprowadzania badania oraz to jakie wnioski możemy wyciągać. Sobczyk (2007, s. 16–20) zidentyfikował trzy główne metody przeprowadzania badań statystycznych:

1. **Badania pełne (całkowite, wyczerpujące)** – obejmują wszystkie jednostki danej zbiorowości statystycznej. Wyróżnić można dwa rodzaje badań pełnych:
    a. Spisy statystyczne – doraźne lub okresowe badanie statystyczne obejmujące wszystkie jednostki danej zbiorowości statystycznej

b. Rejestracja bieżąca – systematyczne notowanie faktów dotyczącej badanej zbiorowości statystycznej

2. **Badania niepełne (częściowe)** – badana jest tylko jakaś część całej populacji (część ta zwana jest próbką). Badania częściowe dzielą się na:

   a. Badania ankietowe – informacje zbierane są przy pomocy ankiet rozsyłanych do ściśle wybranych jednostek z danej populacji
   b. Badania monograficzne – analizowana jest tylko jedna starannie wybrana jednostka lub mały zespół jednostek, które są typowe, powszechnie występujące lub wskazujące kierunek rozwoju
   c. Badania metodą reprezentacyjną – badaniu podlega mały podzbiór badanej populacji wybrany w sposób losowy.

3. **Szacunki interpolacyjne i ekstrapolacyjne** – jest to szacowanie danej wartości na podstawie posiadanych już informacji. Możemy szacować nieznane wartości cechy na podstawie jej wartości sąsiednich (interpolacja) lub możemy szacować wartości wykraczające poza przedział wartości znanych.

Co więcej, zarówno badania pełne jak i niepełne mogą różnić się swoim charakterem (Sobczyk 2007, s. 17). Wyróżniamy trzy główne charaktery badań statystycznych:

1. **Badania ciągłe** – są to badania prowadzone nieprzerwanie, gdzie dane zjawisko jest rejestrowane i analizowane sukcesywnie (np. rejestracja samochodów, zawieranie związków małżeńskich)
2. **Badania okresowe** – są to badania podejmowane w ściśle określonych odstępach czasu np. co 10 lat (np. polski powszechny spis ludności)
3. **Badania doraźne** – są to badania przeprowadzane w celu jednorazowego zbadania jakiegoś zjawiska

1.2.2 Etapy przeprowadzania badań statystycznych

Oprócz wybrania odpowiadającej nam metody badania należy również starannie przygotować wszystkie etapy przeprowadzania badania statystycznego. (Sobczyk 2007, s. 20–32) opisał szczegółowo cztery główne etapy każdego badania statystycznego :

1. **Przygotowanie (programowanie) badania** – polega na obraniu jasnego celu dla danego badania, wybrania metody badawczej i ustaleniu podstawowych informacji takich jak zdefiniowanie jednostki statystycznej i badanej zbiorowości.
2. **Obserwacja statystyczna** – polega ona na *„ustaleniu wartości cech ilościowych lub odmian cech jakościowych u wszystkich jednostek tworzących zbiorowość statystyczną lub u prawidłowo dobranej ich reprezentacji”* (Sobczyk 2007, s. 20).
3. **Opracowanie i prezentacja materiału statystycznego** – na tym etapie weryfikujemy zebrane obserwacje i przeprowadzamy grupowanie, a także prezentujemy zebrany materiał w postaci graficznej (histogramy, diagramy itp.) tak, aby można było wyciągnąć z nich odpowiednie wnioski w etapie kolejnym
4. **Opis lub wnioskowanie statystyczne** – dzięki wcześniej przygotowanym w etapie trzecim danym możemy opisać daną zbiorowość (opisywać możemy tylko na podstawie badań całkowitych) lub spróbować wyciągnąć pewne wnioski na podstawie losowej, reprezentacyjnej próby z populacji.

1.2.3 Źródła statystyczne i niestatystyczne

Zdecydowanie się na przeprowadzenie badania pełnego jest najprostszym metodologicznie podejściem do próby zbadania danej populacji. Dzięki takiemu badaniu posiadamy całkowity obraz danego zjawiska. Jednakże, ze względu na charakter badań całkowitych (koszt, długi czas przeprowadzania badania) często wykonuje się badania znacznie tańszą i szybszą metodą niepełną. W przypadku badań częściowych kluczowym elementem jest wybór sposobu wygenerowania próby losowej oraz wyznaczenia wielkości tejże próby. Jest to o tyle ważne, że błędne wylosowanie próby lub pewne założenia dotyczące populacji mogą obciążyć wynik badania w takim stopniu, że wyniki staną się niereprezentatywne.

Możemy wyodrębnić dwa sposoby generowania prób oraz wyznaczania ich wielkości, mianowicie sposoby statystyczne i niestatystyczne:

- **Sposoby statystyczne** opierają się na mechanizmach rachunku prawdopodobieństwa. Ideą przewodnią sposobów statystycznych jest to, że dany element populacji ma znane i różne od zera prawdopodobieństwo bycia wylosowanym (Showkat i Parveen 2017). Co więcej, sposoby te pozwalają na precyzyjne wyznaczenie prawdopodobieństwa wystąpienia błędu szacunku. Warto zaznaczyć, że wysokość popełnianego błędu można wy-

znaczyć tylko dla metody reprezentacyjnej. Badania ankietowe i monograficzne nie dają takiej możliwości (Sobczyk 2007) .

- **Sposoby niestatystyczne** opierają się w głównej mierze na wiedzy eksperckiej i ocenie własnej badacza. Ich główną charakterystyką jest brak użycia mechanizmów rachunku prawdopodobieństwa. Nie wszystkie jednostki będące przedmiotem badania mają takie samo prawdopodobieństwo bycia wylosowanym. W związku z tym w większości przypadków wyniki badań mogą w małym stopniu nadawać się do wyciągania wniosków na temat całej populacji (Showkat i Parveen 2017). Niemniej, sposoby niestatystyczne są tańszą i szybszą alternatywą w porównaniu do sposobów statystycznych (Wu i Thompson 2020).

Ważnym elementem przeprowadzania badań jest również określenie operatu losowania. Jest to lista wszystkich badanych obiektów, które mogą zostać wylosowane. Wu i Thompson (2020) podają dwa najczęściej występujące operaty losowania :

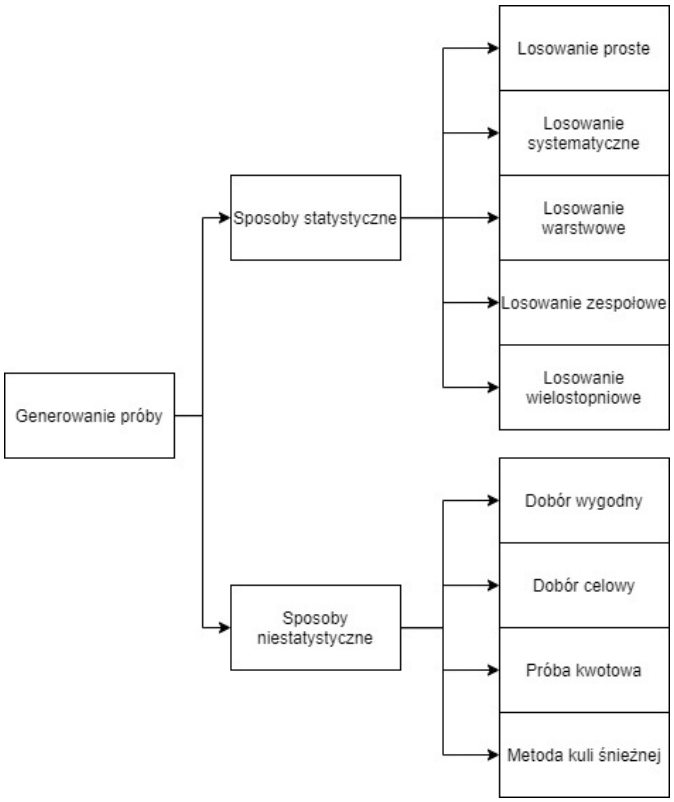
- Kompletna lista wszystkich jednostek. Może to być na przykład spis wszystkich studentów danej uczelni.
- Lista niepokrywających się klastrów, które wspólnie reprezentują całą populację. Na przykład danym klastrem może być konkretna uczelnia wyższa w Polsce. Wtedy populację generalną będzie tworzyć zbiór wszystkich uczelni (klastrów) w Polsce.

Zadbanie o aktualność i kompletność operatu losowania jest wymagane, aby dana próba mogła być reprezentatywna. W Polsce istnieją dwa operaty zawierające pełne informacje ogólnopolskie w zakresie ludności i mieszkań. Są to PESEL (Powszechny Elektroniczny System Ewidencji Ludności) zawierający spis wszystkich Polaków oraz dane na temat miejsca ich zameldowania oraz TERYT, który zawiera informacje o wszystkich mieszkaniach w Polsce (TERYT wykorzystywany jest przede wszystkim przez GUS). W kontekście przedsiębiorstw najczęściej wykorzystywane są trzy operaty, mianowicie REGON (Krajowy Rejestr Urzędowy Podmiotów Gospodarki Narodowej) prowadzony przez Prezesa Głównego Urzędu Statystycznego, NIP (Numer Identyfikacji Podatkowej) służący do identyfikacji podatników w Polsce oraz KRS (Krajowy Rejestr Sądowy) prowadzony przez wybrane sądy rejonowe i Ministerstwo Sprawiedliwości. Gdy posiadamy określony operat losowania możemy wybrać odpowiedni sposób generowania próby z operatu.

1.2.4 Sposoby doboru próby

W celu dobrania próby do przeprowadzania danego badania statystycznego istnieją dwa główne sposoby generowania prób: losowy i nielosowy. Zasadniczą różnicą między tymi dwoma sposobami jest to, że w przypadku sposobów losowych nasza próba znana jest przed rozpoczęciem zbierania informacji od wylosowanych jednostek. Najpierw generujemy próbę przy pomocy mechanizmów rachunku prawdopodobieństwa, a następnie badamy wylosowane jednostki. W sposobach nielosowych nasza próba powstaje w trakcie zbierania informacji od respondentów. Respondenci ci nie są wybrani wcześniej w sposób losowy tylko często są to jednostki, do których najprościej dotrzeć, same wyrażają chęć brania udziału w badaniu lub które akurat zostaną spotkane przez ankietera w danym miejscu.

Co więcej, wykorzystywanie prób losowych pozwala na ocenę otrzymanych danych przy pomocy przedziałów ufności. Dzięki temu nie potrzeba przyjmować żadnych założeń dotyczących populacji przed podejściem do próbkowania. Zostało to udowodnione przez polskiego uczonego Jerzego Neymana w 1934 roku (Bethlehem i Biffignandi (2012, s. 6), Neyman (1934)).



Rysunek 1.1. Sposoby doboru próby

Źródło: Opracowanie własne

Pierwszą rodziną generowania próby z operatu losowania są próby losowe. Są to sposoby statystyczne opierające się na regułach Rachunku Prawdopodobieństwa. Można wyodrębnić parę głównych sposobów generowania prób losowych:

- **losowanie proste** – dzieli się na losowanie proste ze zwracaniem i bez zwracania. W losowaniu prostym ze zwracaniem każdy element ma dokładnie takie samo prawdopodobieństwo bycia wylosowanym. W przypadku wylosowania dwóch tych samych elementów populacji generalnej losowanie powtarza się aż otrzymana zostanie lista n unikalnych elementów próby, będąca podzbiorem całej populacji. W losowaniu prostym bez zwracania różnicą jest to, że wylosowany element zostaje usunięty z puli elementów do wylosowania. Nie ma więc możliwości wylosowania tej samej jednostki dwukrotnie (Showkat i Parveen 2017).
- **losowanie systematyczne** –- w tym sposobie losowania wybierany jest co któryś element populacji w równych interwałach. Na przykład, możemy wylosować liczbę z zakresu od 1 do 10. Jeżeli po losowaniu otrzymamy liczbę 7 to z operatu losowania wybierzemy co siódmy element, tj. element siódmy, czternasty, dwudziesty pierwszy itd. (Showkat i Parveen 2017) Jest to szczególnie przydatne jeżeli badania przeprowadzane są na sporym obszarze geograficznym, na przykład przy badaniu rolnictwa, gdyż badaczom łatwiej jest odwiedzać gospodarstwa oddalone od siebie w równej odległości (Wu i Thompson 2020, s. 23).
- **losowanie warstwowe** –- w tym przypadku operat losowania zostaje podzielony na niepokrywające się podzbiory, zwane warstwami, których suma to wszystkie elementy populacji generalnej (Wu i Thompson 2020, s. 8). Ważną cechą losowania warstwowego jest homogeniczność jednostek będących w danej warstwie i heterogeniczność jednostek między warstwami. Gdy dany operat losowania zostanie podzielony na warstwy to z każdej warstwy zostaje wylosowany podzbiór. Losować można proporcjonalnie lub nieproporcjonalnie do wielkości danej warstwy. W przypadku losowania proporcjonalnego, z warstw o większej ilości elementów zostanie wylosowana większa ilość jednostek niż z warstw o mniejszej ilości elementów. W losowaniu nieproporcjonalnym z każdej warstwy wybierana jest ta sama ilość obiektów do wylosowania, niezależnie od wielkości danej warstwy. Suma podzbiorów ze wszystkich warstw stanowi próbę do badania (Showkat i Parveen 2017). Ważne jest to, że w losowaniu warstwowym elementy z każdej warstwy znajdą się w próbie ostatecznej.

- **losowanie zespołowe** – ten typ losowania opiera się na zasadzie grupowania jednostek w zespoły (zwane również klastrami) o podobnych cechach. Zespołami mogą być na przykład miasta, bloki mieszalne czy uczelnie wyższe. Budowanie klastrów przebiega podobnie do budowania warstw w przypadku losowania warstwowego. Po wyznaczeniu klastrów następuje losowanie zespołów, które zostaną całkowicie przebadane. W związku z tym tylko niektóre klastry są poddawanie badaniu (Wu i Thompson 2020, s. 9). Na przykład można podzielić każdą uczelnię ekonomiczną w Polsce jako osobny klaster, następnie zaś wylosować parę klastrów (uczelni wyższych), które zostaną przebadane całkowicie. Niewylosowane klastry nie będą badane.
- **losowanie wielostopniowe** –- jest to zastosowanie paru technik losowania na paru różnych poziomach. Próba losowana jest z najniższego poziomu podziału danych grup (Showkat i Parveen 2017). Przykładem losowania zespołowego może być połączenie losowania zespołowego z losowaniem prostym. Najpierw wydzielamy klastry z danego operatu losowania, losujemy klastry do badania, a na końcu przeprowadzamy losowanie proste aby wybrać próbę spośród wybranych wcześniej klastrów.

Kolejną rodziną wybierania prób są próby nielosowe. Są to metody niestatystyczne, nieopierające się na mechanizmach Rachunku Prawdopodobieństwa, które bardzo często są obciążone i wnioskowanie z takich prób na temat całej populacji może być mylące (Wu i Thompson 2020, s. 6). Wyróżniamy między innymi:

- **dobór wygodny** – metoda ta zakłada, że dobór jednostek do próby nie jest ustalony odgórnie przy pomocy mechanizmu losowego. Zamiast tego badacz zbiera dane w sposób dla niego wygodny. Może zbierać informacje na przykład tylko od ludzi, których zna (rodzina, znajomi) lub w miejscach gdzie szybko i w prosty sposób może liczyć na dużą ilość zebranych odpowiedzi, w miejscach takich jak centra handlowe w weekend (Showkat i Parveen 2017).
- **dobór celowy** –- w tym przypadku badacz skupia się na jednostkach, które jego zdaniem będą dobrze pasować do postawionego przez siebie celu badania. Taki rodzaj doboru jest o wiele szybszy, prostszy i tańszy niż prawdziwie losowa próba i często nadaje się do przeprowadzania badań wstępnych. Niemniej, ciężko zadbać o reprezentatywność próby, gdy jej dobór opiera się na ocenie własnej badacza (Showkat i Parveen 2017).
- **próba kwotowa** – w swoich założeniach przypomina trochę losowanie warstwowe. Mianowicie, dana populacja dzielona jest na rozłączne podzbiory, a następnie zaś ankieter

ma za zadanie przebadać wyznaczoną część (kwotę) każdego z podzbiorów (na przykład 200 mężczyzn i 250 kobiet). Jednakże, w odróżnieniu od losowania warstwowego, jednostki które zostaną przebadane nie są losowanie odgórnie. Zamiast tego leży to w kwestii ankietera, które osoby z danej warstwy zostaną przebadane (Showkat i Parveen 2017). Tutaj również występuje problem z reprezentatywnością, gdyż ankieter (świadomie lub nie) może wybierać osoby do badania, które są na przykład uprzejmiejsze lub może wybierać „pierwsze lepsze" osoby tak, aby jak najszybciej przebadać wyznaczoną kwotę jednostek i szybciej skończyć swoją pracę.

- **metoda kuli śnieżnej** – metoda ta wykorzystywana jest głównie w przypadku, gdy badana populacja jest ciężka do zidentyfikowania w całości. Polega ona na zachęcaniu przebadanych ludzi do przekazywania informacji kontaktowych o podobnych jednostkach. Nowe, nieznane wcześniej kontakty, poddaje się badaniu i prosi znów o przekazanie informacji o nowych możliwych respondentach. Schemat powtarza się tak długo aż zbierze się wystarczająco dużą próbę . Metoda ta sprawuje się całkiem dobrze w przypadku badania ludzi bezdomnych czy imigrantów bez ważnych dokumentów tożsamości (Showkat i Parveen 2017).

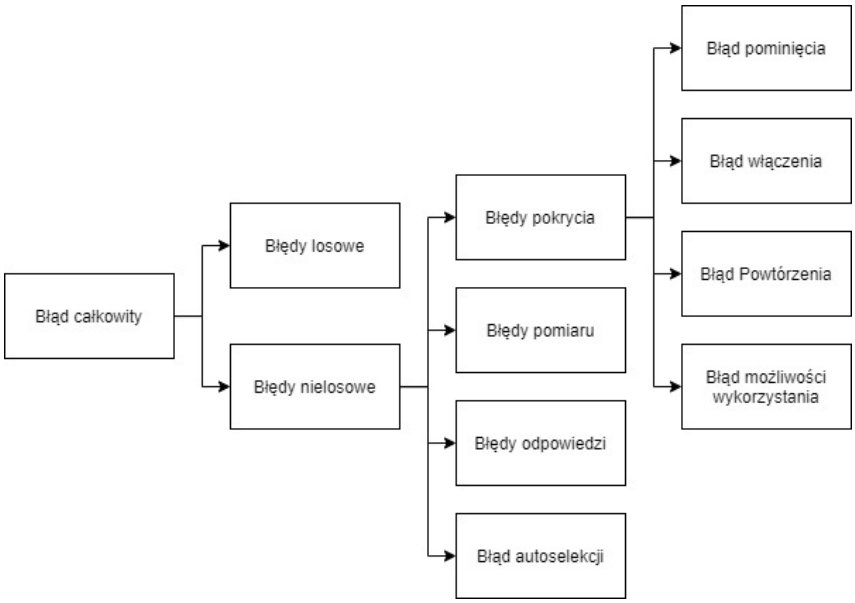
1.2.5 Klasyfikacja błędów losowych i nielosowych

Każde badanie statystyczne obarczone jest w mniejszym lub większym stopniu błędami, które mogą zniekształcać ostateczny wynik. W każdym badaniu powinno dążyć się do minimalizacji błędów, tak aby wyniki były jak najbardziej reprezentatywne. Błędy możemy podzielić na dwie główne kategorie, błędy losowe i błędy nielosowe.

Błędy losowe wynikają z samego charakteru badań częściowych i są nieuniknione przy wnioskowaniu o populacji. Wraz ze wzrostem liczebności próby błędy losowe maleją. Jedynym sposobem na wyeliminowanie błędów losowych jest przebadanie całej próby (Bethlehem i Biffignandi 2012, s. 99). Są to błędy nieobciążone, co oznacza, że „wartość oczekiwana estymatora jest równa szacowanemu parametrowi"(Sobczyk 2007, s. 141).

Błędami nielosowymi (zwanymi również systematycznymi) są wszystkie błędy powstałe w wyniku nieprawidłowego przygotowania operatu losowania, dokonania pomiarów lub występujących braków danych. Możemy wyróżnić cztery główne kategorie błędów nielosowych:

- **błędy pokrycia (błędy operatu)** – są to wszystkie błędy związane z tym, że dobrany operat losowania różni się znacząco od rzeczywistej populacji generalnej (Wu i Thompson 2020,



Rysunek 1.2. Kategorie błędów statystycznych w badaniach reprezentacyjnych

Źródło: opracowanie własne na podstawie diagramu z Bethlehem i Biffignandi (2012, s. 100)

s. 5). Błędy pokrycia są szczegółowiej podzielone przez (Bethlehem i Biffignandi 2012) na:

- **błąd pominięcia** – jest to obranie operatu losowania, który nie zawiera w sobie wszystkich jednostek populacji badanej
- **błąd włączenia** – jest to obranie operatu losowania, który zawiera w sobie jednostki nienależące do populacji badanej
- **błąd powtórzenia** – jest to sytuacja, w której obrany operat losowania zawiera w sobie więcej niż jedno wystąpienie danej jednostki co może skutkować kilkukrotnym przebadaniem tego samego obiektu
- **błąd możliwości wykorzystania** – błąd ten występuje, gdy obrany operat losowania zawiera w sobie jednostki, z którymi w żaden badacz nie jest w stanie się skontaktować lub do nich dotrzeć

- **błędy pomiaru** – są to błędy związane z błędnym zbieraniem i zapisywaniem informacji. Mogą wynikać z błędnego zapisu odpowiedzi przez ankietera lub niezrozumienia pytania przez osobę odpowiadającą. Występują również w ankietach prowadzonych drogą elektroniczną. Łatwo jest przypadkowo kliknąć i wybrać złą odpowiedź (Bethlehem i Biffignandi 2012, s. 102).

- **brak odpowiedzi** – jest to przypadek, gdy wylosowana jednostka nie udziela odpowiedzi na przedstawione pytania lub gdy przedstawiona odpowiedź jest bezużyteczna. Niektóre osoby chętniej odpowiadają na pytania niż inne co może skutkować nadmierną reprezentacją jednej grupy ludzi w porównaniu do osób, które niechętnie odpowiadają na pytania ankieterów (Bethlehem i Biffignandi 2012, s. 103).
- **błąd autoselekcji** – występuje, gdy dana ankieta jest ogólnodostępna i każdy ma do niej dostęp. Jednostki badane nie są więc wylosowanymi obiektami z operatu losowania, a same decydują o tym czy podejmą się wzięcia udziału w badaniu czy nie (Bethlehem i Biffignandi 2012, s. 303–304).

1.3 Próby nielosowe w badaniach statystycznych

1.3.1 Badania internetowe i błędy nielosowe

Wraz z rozwojem Internetu i coraz większą bazą ludzi korzystających z Sieci na co dzień, badania elektroniczne zaczęły stawać się coraz popularniejsze. Aktualnie, w większości państw, dostęp do Internetu ma między 60% a 90% obywateli (Bethlehem i Biffignandi 2012, s. 41). Można łatwo zauważyć, że badania internetowe są kolejnym narzędziem o ogromnym potencjale do masowego przeprowadzania analiz, podobnie do badań telefonicznych czy badań wysyłanych pocztą.

Tak samo jak „klasyczne" metody przeprowadzania badań (osobiście przez ankietera, telefonicznie czy drogą pocztową) tak i badania internetowe mogą być źródłem danych zebranych w sposób statystyczny. Należy zadbać o dobór odpowiedniego operatu losowania nieobarczonego błędami pokrycia, zaprojektować ankietę w taki sposób, żeby była prosta i zrozumiała w obsłudze (w celu wyeliminowania błędu pomiaru po stronie użytkownika odpowiadającego na ankietę) oraz wyznaczyć próbkę w sposób losowy (Bethlehem i Biffignandi 2012, s. 40–43).

W przypadku badań internetowych operatem losowania może być na przykład lista wszystkich pracowników w danym przedsiębiorstwie. Każdy pracownik może posiadać swój adres e-mail, dzięki któremu można rozesłać ankietę do ówcześnie wylosowanych osób.

Jednakże sprawa operatu losowania komplikuje się, gdy chcemy przeprowadzić badanie na szerszą skalę. Nie wszystkie osoby posiadające dostęp do Internetu posiadają adres e-mail. Niektóre osoby posiadające taki adres nie korzystają z niego zbyt często. Co więcej, ciężkie może okazać się uzyskanie kompletnej listy adresów e-mail danej populacji. Fakt, że ktoś dany adres

posiada nie oznacza, że jest od publicznie dostępny. W ten sposób prosto może dojść do błędów pominięcia. Na przykład, w Stanach Zjednoczonych ludność latynoska jest niedostatecznie reprezentowana w związku z problemami z dostępem do Internetu czy posiadaniem adresu e-mail (Bethlehem i Biffignandi 2012, s. 422). Istnieje wiele przedsięwzięć opierających się na tak zwanych „panelach internetowych". Zrzeszają one internautów do cyklicznego lub jednorazowego wypełniania ankiet internetowych na dane tematy. Część z tych paneli internetowych opera się na zasadach badań przeprowadzanych w sposób statystyczny. Próbka osób zapraszanych do zapisania się do danego panelu jest losowana i zapraszana poprzez wiadomość e-mail, pocztę tradycyjną lub odwiedzana przez badacza i zapraszana werbalnie . Jednym z przykładów takiego panelu w Europie jest holenderski panel internetowy „The CentERpanel" prowadzony przez instytut badawczy CentERdata zlokalizowany na kampusie Uniwersytetu w Tilburgu. CentERdata rekrutuje wylosowanych wcześniej uczestników poprzez kontakt telefoniczny. Jeżeli dany uczestnik badania nie posiada dostępu do Internetu to CentERdata zapewnia łącze z Internetem poprzez „Net.Box", które podłącza się do telewizora, co pozwala na rozwiązywanie ankiet on-line (Bethlehem i Biffignandi 2012, s. 425).

Aktualnie, bardzo często odbiega się przeprowadzania badań przy pomocy ankiet internetowych w sposób statystyczny na rzecz rozwiązań niestatystycznych. Spowodowane jest to znacznie wyższym kosztem przeprowadzenia badania w sposób statystyczny. Najczęstszym sposobem przeprowadzania badań w sposób niestatystyczny z wykorzystaniem Internetu jest zastosowanie tak zwanych „paneli dobrowolnych" (Ang. Opt-In Panels). Rekrutacja do takich paneli przebiega w sposób autoselekcyjny. Dany użytkownik Internetu może dobrowolnie zapisać się i stać się członkiem danego panelu dobrowolnego. Będąc członkiem danego panelu może on regularnie wypełniać ankiety na różne tematy. Z takich rozwiązań korzystają głównie firmy badające rynek oraz sprawdzające opinie na temat danych produktów (Bethlehem i Biffignandi 2012, s. 420). Oznacza to oczywiście, że wyniki badań uzyskane w sposób niestatystyczny mogą być silnie obciążone, głównie poprzez błędy pokrycia i błędy autoselekcji. Może to prowadzić do niereprezentatywności próby i trudnościami w wykorzystywaniu mechanizmów Rachunku Prawdopodobieństwa do wnioskowania o całej populacji.

Niemniej taki sposób zbierania informacji powoduje nagromadzenie się dużej ilości odpowiedzi od respondentów. W połączeniu z informacjami dodatkowymi z innych źródeł danych powstaje ogromna i obfita baza danych. Ogrom informacji pozyskiwanych przy pomocy dobrowolnych paneli internetowych może być traktowany jako Big Data. Pomimo problemów z błę-

dami pokrycia i błędami autoselekcji takie ogromne bazy danych mają spory potencjał przy badaniu zbiorowości. (J. K. Kim i Wang 2019) w swoim artykule „Sampling Techniques for Big Data Analysis" przedstawili metody pomagające w rozwiązaniu problemu niereprezentatywności próbek pozyskanych z ogromnych zbiorów danych pochodzących z badań internetowych.

1.3.2 Charakterystyka Big Data ("the 4 Vs") jako próby nielosowej

Duże zbiory danych (ang. "Big Data") charakteryzują się wielkością, zmiennością i różnorodnością zapisanych w nich informacji. Zbiory te z roku na rok stają się coraz bardziej atrakcyjne dla badaczy, analityków oraz osób decyzyjnych. Efektywne wykorzystanie informacji zwartych w Big Data jest niezwykle cenne, gdyż często jest kluczem do zdobycia przewagi konkurencyjnej na rynku (McAfee i in. 2012).

Niemniej, trzeba bardzo ostrożnie podchodzić do takich zbiorów danych. W Big Data najważniejsza jest ilość, dostępność, a także aktualność danych. W porównaniu do badań reprezentacyjnych, które są przeprowadzane na starannie wyselekcjonowanej, reprezentatywnej próbie, Big Data pada ofiarą nie reprezentatywności ze względu na swoje niestatystyczne pochodzenie. Pomimo problemów z nie reprezentatywnością takie ogromne zbiory wciąż mają spory potencjał przy badaniu danej populacji. (J. K. Kim i Wang 2019) w swoim artykule „Sampling Techniques for Big Data Analysis" przedstawili metody pomagające w rozwiązaniu problemu nie reprezentatywności ogromnych zbiorów danych. Jedna z tych metod, tj. integracja próby losowej i nielosowej (w oryginale "Data Integration"), zostanie zaprezentowana w późniejszej części tej pracy.

Ogólna charakterystyka Big Data i jej najważniejsze założenia zostały opisane m. in. przez (Franke i in. 2016), którzy wyszczególnili cztery główne cechy dużych zbiorów danych (tzw. "Four V's"):

- **Volume (ilość)** – jak sama nazwa wskazuje, zbiór Big Data musi być zbiorem dużym, zawierającym ogromną liczbę zapisanych rekordów. Jednakże z roku na rok cena przechowywania danych cyfrowych stale maleje przez co możliwe jest przechowywanie coraz to większej ilości informacji w zbliżonej cenie. W związku z tym nie istnieje jedna prawidłowa odpowiedź na pytanie "jak obszerny powinien być zbiór danych, aby traktować go jako zbiór Big Data?", gdyż z roku na rok granica ta jest stale przesuwana. Głównym wyznacznikiem jest to, że przetwarzanie i dostęp do całego zbioru jednocześnie jest niemożliwe (lub bardzo niepraktyczne) ze względu na ograniczenia w mocy obliczeniowej do-

stępnych komputerów. Z tego powodu często wykorzystuje się specjalną infrastrukturę do przechowywania i odczytywania danych np. Apache Hadoop (Shvachko i in. 2010).

- **Variety (różnorodność)** – informacje zapisane w dużych zbiorach danych znacznie różnią się od informacji, które możemy znaleźć w relacyjnych bazach danych. W zbiorach Big Data dane często mają postać częściowo ustrukturyzowaną lub niestrukturyzowaną. Taki typ danych pochodzi najczęściej z postów internetowych, wpisów na Twitterze, informacji z portali społecznościowych, nagrań dźwiękowych, odczytów z różnych sensorów czy wszelakich nagrań. Różnorodność ta wprost wpływa na złożoność we wnioskowaniu z takich niestrukturyzowanych danych.
- **Veracity (wiarygodność)** – ze względu na swój masowy charakter, Big Data często miewa problemy z niejednorodnością. Często duże zbiory danych powstają w wyniku połączenia paru mniejszych zbiorów. Mowa wtedy o ”danych znalezionychlub ”danych wygodnych”. Często te dane nie zostały zebrane z pewnym pytaniem badawczym w głowie tylko korzysta się z tego co jest akurat dostępne. Big Data również prezentuje duży problem w kwestii oczyszczania zbioru z błędnych danych. Źle oczyszczony zbiór może być powodem do złego wnioskowania w przyszłości. W zastosowaniu administracyjnym również powstaje wiele błędów na samym etapie zbierania danych, gdyż urzędnicy mogą różnie interpretować przepisy lub błędnie wprowadzać dane. Wszystkie te czynniki mocno nadszarpują wiarygodność zbiorów Big Data dlatego zawsze należy podchodzić do nich z pewną rezerwą. Obszerny zbiór danych nie gwarantuje reprezentatywności na całą populację.
- **Velocity (prędkość)** – powodem, dla którego zbiory Big Data są tak obszerne, jest ilość informacji, które są zapisywane nieustannie przez różnego rodzaju urządzenia. Dane spływają ciągłym strumieniem przez co zbiory Big Data są cały czas rozbudowywane o nowe rekordy. Dane te mogą pochodzić z różnych źródeł, np. z sensorów samochodów autonomicznych czy z sesji użytkowników korzystających z platform streamingowych. Czasami prędkość spływających nowych informacji jest tak duża, że niemożliwe jest zapisanie wszystkiego, chociażby w przypadku badania fizyki cząstek elementarnych.

1.4 Podsumowanie

Znaczenie badań statystycznych utrzymuje się na wysokim poziomie od tysiącleci. Istnieje wiele rodzajów badań statystycznych, źródeł danych, a także sposobów na generowanie prób

potrzebnych do badań. Spisy powszechne i badania częściowe są dobrze znane i wciąż efektywnie wykorzystywane. Niemniej, wraz z rozwojem technologii jesteśmy w stanie otrzymać dostęp do zupełnie nowego rodzaju danych, mianowicie Big Data. Ze względu na rosnącą dostępność i obszerność zbiory Big Data mogą być chętnie wykorzystywane przez badaczy do estymacji pewnych cech populacji. Trzeba pamiętać jednak, że mają one pewne ograniczenia i aby móc z nich w pełni korzystać należy posłużyć się odpowiednimi metodami estymacji.

W następnym rozdziale zostaną przedstawione główne założenia estymacji przy pomocy źródeł Big Data. Zaprezentowana zostanie metoda integracji prób losowych i nielosowych (ang. *Data Integration*), która ma na celu otrzymanie estymatorów bazujących na próbie Big Data. Pokazana zostanie również implementacja metody Data Integration w języku Python wraz ze wszystkimi wykorzystanymi modułami.

Rozdział 2

Metody estymacji na podstawie prób nielosowych

2.1 Główne założenia i oznaczenia

Wraz ze wzrostem dostępności prób nielosowych (np. Big Data, rejestry administracyjne) rośnie chęć wykorzystywania ich w celu opisywania cech pewnych populacji. Jednak, jak zostało wspomniane w sekcji 1.3.2, należy ostrożnie podchodzić do takich zbiorów danych ze względu na ich charakter. Najczęściej zbiory te są niereprezentatywne względem całej populacji będące wynikiem błędów pokrycia (ang. coverage error) i autoselekcji (ang. *selection error*). Ich niestatystyczne pochodzenie podważa prawdziwość wyciąganych wniosków o populacji generalnej. Z drugiej strony są mogą być cennym źródłem informacji, które warto wykorzystać w badaniach ze względu na swoją obszerność.

W ostatnim czasie, w literaturze przedmiotu pojawiło się wiele artykułów poświęconych metodom estymacji w przypadku prób nielosowych. Wśród najważniejszych ostatnich prac podejmujących problematykę estymacji i proponujące konkretne rozwiązania należy wymienić:

- Chen, J. K. T., Valliant, R. L. & Elliott, M. R. (2018). Model-assisted calibration of non-probability sample survey data using adaptive LASSO. *Survey Methodology*, *44*(1), 117–144
- Chen, J. K. T., Valliant, R. L. & Elliott, M. R. (2019). Calibrating non-probability surveys to estimated control totals using LASSO, with an application to political polling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *68*(3), 657–681

- Tam, S.-M. & Kim, J.-K. (2018). Big Data ethics and selection-bias: An official statistician's perspective. *Statistical Journal of the IAOS*, *34*(4), 577–588
- Kim, J. K. & Wang, Z. (2019). Sampling techniques for big data analysis. *International Statistical Review*, *87*, S177–S191
- Yang, S., Kim, J. K. & Song, R. (2020). Doubly robust inference when combining probability and non-probability samples with high dimensional data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *82*(2), 445–465
- Chen, Y., Li, P. & Wu, C. (2020). Doubly robust inference with nonprobability survey samples. *Journal of the American Statistical Association*, *115*(532), 2011–2021
- Chen, S., Yang, S. & Kim, J. K. (2021). Nonparametric Mass Imputation for Data Integration. *Journal of Survey Statistics and Methodology*

W niniejszej pracy skupiono się na jednej pracy:

Kim, J. K., i Wang, Z. (2019). **Sampling techniques for big data analysis**. *International Statistical Review*, 87, S177-S191,

w której, J. K. Kim i Wang proponują połączenie źródła statystycznego i źródła niestatystycznego w technice zwanej *integracją próby losowej i nielosowej (ang. Data Integration)*. Ma ona na celu minimalizację błędu pokrycia zbioru Big Data poprzez integrację z badaniem statystycznym. Dzięki temu możliwe będzie wnioskowanie na temat całej populacji na podstawie obu połączonych źródeł.

J. K. Kim i Wang (2019) przedstawili dwa estymatory: *Propensity Score Estimator*[1] (PS) oraz podwójnie odporny estymator (ang. *Doubly Robust Estimator*; DR). W tabeli 2.1 przedstawiono podstawowe oznaczenia, które będą wykorzystywane do zdefiniowana powyższych estymatorów.

Zarówno PS jak i DR bazują na zbliżonych założeniach. Do ich wyznaczenia potrzebne są dwa zbiory danych. Pierwszym zbiorem jest reprezentatywny zbiór powstały w wyniku przeprowadzenia badania opartego na próbie losowej (oznaczony symbolem "A"). Zakładamy, że badanie to nie zawiera w sobie informacji o zmiennej celu Y. Drugim zbiorem jest zbiór Big Data (oznaczony symbolem "B"), który zawiera informacje o zmiennej celu Y. Ze względu na swój charakter, zbiór B nacechowany jest błędami pokrycia implikującymi brak reprezentatywności.

[1] W pracy używane będzie pojęcie *Propensity Score Estimator* ponieważ pod taką nazwą pojawia się w literaturze. Polskie tłumaczenie *estymacji przez dopasowanie* używane przez m.in. prof. Marka Gruszczyńskiego dotyczy badania przyczynowości.

Tablica 2.1. Zestawienie oznaczeń

Symbol	Wyjaśnienie
U	populacja celu
N	wielkość populacji
A	próba losowa, podzbiór populacji U
B	próba nielosowa, podzbiór populacji U
$i = 1, .., .N$	indeksy jednostek z populacji
$I_i \in \{0, 1\}$	indeks przynależności do próby losowej A
$\delta_i \in \{0, 1\}$	indeks przynależności do próby nielosowej B
n	wielkość próby A
m	wielkość próby B
Y	zmienna celu
y_i	wartość zmiennej celu dla i-tej jednostki
$\boldsymbol{X}$	macierz zmiennych pomocniczych (np. demograficznych)
$\boldsymbol{x}_i$	wektor zmiennych pomocniczych dla i-tej jednostki
π_i	prawdopodobieństwo inkluzji do próby A
$d_i = 1/\pi_i$	wagi wynikające z losowania (dla próby A)
w_i	wagi d_i skorygowane o braki odpowiedzi i błąd pokrycia (dla próby A)
p_i	prawdopodobieństwo przynależności do próby B
$Pr(\delta_i = 1\|\boldsymbol{x}_i)$	prawdopodobieństwo warunkowe przynależności do próby B
θ	charakterystyka celu (np. średnia, mediana)
$\boldsymbol{\lambda}$	wektor parametrów związany z modelem $Pr(\delta_i = 1\|\boldsymbol{x}_i)$
$\boldsymbol{\beta}$	wektor parametrów związany z modelem warunkowym $E(y_i\|\boldsymbol{x}_i)$
$\hat{\theta}_{\text{PS}}$	estymator parametru celu oparty na metodzie PS
$\hat{\theta}_{\text{DR}}$	estymator parametru celu oparty na metodzie DR
$\hat{\theta}_{\text{naïve}}$	estymator parametru celu oparty na średniej arytmetycznej

Tabela 2.2 przedstawia poglądowy układ danych wykorzystywanych w na potrzeby estymacji. Symbol ✓ oznacza, że dana zmienna(e) występują w określonym zbiorze. Natomiast kolumna reprezentatywność wskazuje czy dany zbiór umożliwia wnioskowanie na daną populację. Celem zainteresowania obydwu metod jest estymacja średniej z populacji tj. $\theta = \sum_{i=1}^{N} y_i/N$.

Tablica 2.2. Idea łączenia danych – przykład dwóch zbiorów

Zbiór	Reprezentatywność	$\boldsymbol{X}$	Y
A	Tak	✓	–
B	Nie	✓	✓

Źródło: Opracowanie własne na podstawie J. K. Kim i Wang (2019)

W celu poprawnego wykorzystania obydwu metod ważne jest przyjęcie następujących założeń:

1. zbiór danych B (np. big data) musi powstać w sposób nielosowy i każda jednostka z populacji U ma określone prawdopodobieństwo przynależności p_i,
2. musimy posiadać badanie reprezentatywne (zbiór A) dotyczące tej samej populacji U, przy czym próba ta stanowi niewielki odsetek badanej populacji,
3. potrafimy zidentyfikować jednostki występujące w obydwu zbiorach danych i możemy określić δ_i czyli indykatorem przynależności do źródła Big Data dany

$$\delta_i = \begin{cases} 1 & \text{jeżeli } i \in B, \\ 0 & \text{w przeciwnym wypadku}, \end{cases} \tag{2.1}$$

4. zakładamy, że mechanizm wyboru próbki Big Data można zignorować (ang. *Missing at Random*, zob. Rubin (1976)):

$$\forall_{i \in U} \quad P(\delta_i = 1 \mid \boldsymbol{x}_i, y_i) = P(\delta_i = 1 \mid \boldsymbol{x}_i),$$

i jest zgodny z modelem parametrycznym:

$$\forall_{i \in U} \quad P(\delta_i = 1 \mid \boldsymbol{x}_i,) = p_i(\boldsymbol{\lambda}; \boldsymbol{x}) \in (0, 1>,$$

gdzie $p_i(\boldsymbol{\lambda}; \boldsymbol{x}) = p(x_i^T \boldsymbol{\lambda})$ dla danej znanej funkcji $p(\cdot)$, która ma drugą ciągłą pochodną w odniesieniu do nieznanego parametru $\boldsymbol{\lambda}$. W tym celu zwykle stosuje się regresję logistyczną (logitową):

$$p_i(\boldsymbol{\lambda}) = \frac{\exp\left(\boldsymbol{x}_i^T \boldsymbol{\lambda}\right)}{1 + \exp(\boldsymbol{x}_i^T \boldsymbol{\lambda})}, \tag{2.2}$$

natomiast część liniowa zostanie zapisana następująco logit $p_i(\boldsymbol{\lambda}) = \boldsymbol{x}_i^T \boldsymbol{\lambda}$.

W następnych dwóch sekcjach 2.2 i 2.3 zaprezentowana zostanie idea obu estymatorów przedstawionych przez (J. K. Kim i Wang 2019), które bazują na powyższych założeniach.

2.2 Estymacja z wykorzystaniem Propensity Score estimator

Metoda PS wymaga estymacji wektora parametrów $\boldsymbol{\lambda}$ przy założeniu parametrycznej postaci funkcji $p()$. J. K. Kim i Wang (2019) założyli, że w próbie losowej A możemy zidentyfikować jednostki z próby B i tę informację wykorzystujemy do wyznaczenia $\boldsymbol{\lambda}$. J. K. Kim i Wang (2019) rozważyli estymację parametrów metodą największej wiarygodności, przy czym w związku z wykorzystaniem próby losowej oraz wag d_i (lub w_i chociaż w oryginalnej pracy nie wspominano

o wagach finalnych) mówimy o tzw. pseudo-funkcji największej wiarygodności. J. K. Kim i Wang (2019) dokonali estymacji parametrów wykorzystując logarytm pseudo-funkcji wiarygodności dany wzorem (2.3)

$$\log L(\boldsymbol{\lambda}; \boldsymbol{x}) = \sum_{i \in A} d_i [\delta_i \log\{p_i(\boldsymbol{\lambda}; \boldsymbol{x}_i)\} + (1 - \delta_i) \log\{1 - p_i(\boldsymbol{\lambda}; \boldsymbol{x}_i)\}]. \tag{2.3}$$

Celem jest znalezienie wektora $\hat{\boldsymbol{\lambda}}$ maksymalizując funkcję $\log L(\boldsymbol{\lambda}; \boldsymbol{x})$

$$\hat{\boldsymbol{\lambda}} = \underset{\boldsymbol{\lambda} \in \Omega}{\operatorname{argmax}} \log L(\boldsymbol{\lambda}; \boldsymbol{x}_i), \tag{2.4}$$

gdzie Ω jest przestrzenią możliwych parametrów.

W praktyce (2.4) dokonuje się stosując metodę Newtona-Raphson'a, która wymaga wyznaczenia pochodnych pierwszego oraz drugiego rzędu. J. K. Kim i Wang (2019) przyjęli następujące oznaczenia wektora pierwszych pochodnych

$$S(\boldsymbol{\lambda}; \boldsymbol{x}) = \frac{\partial \log L(\boldsymbol{\lambda}; \boldsymbol{x})}{\partial \boldsymbol{\lambda}}, \tag{2.5}$$

gdzie nazwa $S(\boldsymbol{\lambda}; \boldsymbol{x})$ jest skrótem od angielskiego słowa *score* oraz macierzy drugich pochodnych (hesjanu)

$$H(\boldsymbol{\lambda}; \boldsymbol{x}) = \frac{\partial^2 \log L(\boldsymbol{\lambda}; \boldsymbol{x})}{\partial \boldsymbol{\lambda}^2}, \tag{2.6}$$

które następnie wykorzystywane są aby w sposób iteracyjny wyznaczyć wektor parametrów $\boldsymbol{\lambda}$

$$\boldsymbol{\lambda}_{k+1} = \boldsymbol{\lambda}_k - H(\boldsymbol{\lambda}; \boldsymbol{x})^T S(\boldsymbol{\lambda}; \boldsymbol{x}). \tag{2.7}$$

Poniżej przedstawiono wyprowadzenie pochodnej pierwszego oraz drugiego rzędu z funkcji (2.3):

Pierwsza pochodna z funkcji największej wiarygodności:

$$\frac{\partial \log L}{\partial \lambda} = \frac{\partial}{\partial \lambda}\Big[\sum d_i[\delta_i \log(p_i(\lambda; x_i)) + (1-\delta_i)\log(1-p_i(\lambda; x_i))]\Big] =$$
$$\sum_i \frac{\partial}{\partial \lambda}\Big[d_i\delta_i \log\Big(\frac{\exp\big(x_i^T\lambda\big)}{1+\exp(x_i^T\lambda)}\Big)\Big] + \frac{\partial}{\partial \lambda}\Big[d_i(1-\delta_i)\log\Big(1-\frac{\exp\big(x_i^T\lambda\big)}{1+\exp(x_i^T\lambda)}\Big)\Big] = \tag{2.8}$$
$$\sum_i \frac{d_i\delta_i x}{1+\exp(x_i^T\lambda)} + \frac{d_i(\delta_i-1)x \times \exp\big(x_i^T\lambda\big)}{1+\exp(x_i^T\lambda)}$$

Druga pochodna z funkcji największej wiarygodności:

$$\frac{\partial^2 \log L}{\partial \lambda^2} = \frac{\partial^2}{\partial \lambda^2}\Big[\sum_i \frac{d_i\delta_i x}{1+\exp(x_i^T\lambda)} + \frac{d_i(\delta_i-1)x \times \exp\big(x_i^T\lambda\big)}{1+\exp(x_i^T\lambda)}\Big] =$$
$$\sum \frac{\partial^2}{\partial \lambda^2}\Big[\frac{d_i\delta_i x}{1+\exp(x_i^T\lambda)}\Big] + \frac{\partial^2}{\partial \lambda^2}\Big[\frac{d_i(\delta_i-1)x \times \exp\big(x_i^T\lambda\big)}{1+\exp(x_i^T\lambda)}\Big] = \tag{2.9}$$
$$\sum_i -\frac{d_i\delta_i x^2 \exp\big(x_i^T\lambda\big)}{(1+\exp(x_i^T\lambda))^2} + \frac{d_i(\delta_i-1)x^2 \exp\big(x_i^T\lambda\big)}{(1+\exp(x_i^T\lambda))^2}$$

Zwykle, za kryterium stopu można przyjąć warunki:

- $|\boldsymbol{\lambda}_{k+1} - \boldsymbol{\lambda}_k| < \epsilon$, lub
- $\sum_i |S(\boldsymbol{\lambda}_k; \boldsymbol{x}_i)| < \epsilon$,

gdzie ϵ to pewna mała wartość (np. 1×10^{-6}).

Po wyznaczeniu wektora parametrów $\hat{\boldsymbol{\lambda}}$ estymator wartości średniej $\hat{\theta}$ dany jest wzorem:

$$\hat{\theta}_{\mathsf{PS}} = \frac{\sum_{i\in B} p_i(\hat{\boldsymbol{\lambda}})^{-1} y_i}{\sum_{i\in B} p_i(\hat{\boldsymbol{\lambda}})^{-1}}. \tag{2.10}$$

Należy zwrócić uwagę, że do wyznaczenia $\hat{\theta}_{\mathsf{PS}}$ wykorzystujemy próbę B, a nie A. Oznacza to, że musimy wyznaczyć dla wszystkich jednostek $i \in B$ wartość $p_i(\hat{\boldsymbol{\lambda}})^{-1}$.

J. K. Kim i Wang (2019) zaproponowali aby wyznaczyć wariację estymatora (2.10) stosując następujący układ równań, który można rozwiązać zarówno metodą największej wiarygodnosci, jak i uogólnioną metodą momentów:

$$\begin{aligned} U(\theta, \boldsymbol{\lambda}; \boldsymbol{x}) &\equiv \sum_{i\in B} p_i(\boldsymbol{\lambda}; \boldsymbol{x})^{-1}\left(y_i - \boldsymbol{x}_i^T\boldsymbol{\lambda}\right) = 0 \\ S(\boldsymbol{\lambda}; \boldsymbol{x}) &\equiv \sum_{i\in A} d_i\left\{\delta_i - p_i(\boldsymbol{\lambda}; \boldsymbol{x})\right\} g_i(\boldsymbol{\lambda}; \boldsymbol{x}) = \mathbf{0}, \end{aligned} \tag{2.11}$$

którą można równoważnie zapisać wykorzystując sumowanie po całej populacji

$$
\begin{aligned}
U(\theta, \boldsymbol{\lambda}; \boldsymbol{x}) &\equiv \sum_{i=1}^{N} \delta_i p_i(\boldsymbol{\lambda}; \boldsymbol{x}_i)^{-1} \left(y_i - \boldsymbol{x}_i^T \boldsymbol{\lambda}\right) = 0 \\
S(\boldsymbol{\lambda}; \boldsymbol{x}) &\equiv \sum_{i=1}^{N} I_i d_i \left\{\delta_i - p_i(\boldsymbol{\lambda}; \boldsymbol{x}_i)\right\} g_i(\boldsymbol{\lambda}; \boldsymbol{x}_i) = \mathbf{0},
\end{aligned}
\tag{2.12}
$$

gdzie $g_i(\boldsymbol{\lambda}; \boldsymbol{x}) = \partial \,\text{logit}(p_i(\boldsymbol{\lambda}))/\partial \boldsymbol{\lambda}$. W pracy jednak nie wyznaczono wariancji tego estymatora ponieważ nie autor nie dysponował pełną informacją o schemacie losowania Badania Popyt na Pracę.

2.3 Estymacja z wykorzystaniem Doubly Robust estimator

Drugim estymatorem zaprezentowanym przez J. K. Kim i Wang (2019) jest *Doubly Robust Estimator*. Jest on rozwinięciem *Propensity Score Estimator* i również wymaga estymacji parametru $\hat{\lambda}$.

Głównym założeniem *Doubly Robust Estimator* jest to, że musimy być w stanie skonstruować model regresji liniowej dla $E(Y \mid \boldsymbol{x}) = x^T\beta$ lub regresji logistycznej $E(Y \mid \boldsymbol{x}) = \frac{\exp\left(x^T\beta\right)}{1+\exp(x^T\beta)}$. W pracy wykorzystano regresję logistyczną

Po przeprowadzeniu regresji liniowej na próbie Big Data możemy dokonać estymacji przy pomocy poniższego wzoru:

$$
\hat{\theta}_{DR} = \frac{1}{N}\left\{\sum_{i\in B} \frac{1}{p_i(\hat{\boldsymbol{\lambda}})}\left(y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}\right) + \sum_{i\in A} d_i \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}\right\} \tag{2.13}
$$

gdzie $\hat{\beta}$ jest estymowanym parametrem regresji liniowej z próby Big Data. Natomiast w przypadku regresji logistcznej dany jest wzorem

$$
\hat{\theta}_{DR} = \frac{1}{N}\left\{\sum_{i\in B} \frac{1}{p_i(\hat{\boldsymbol{\lambda}})}\left(y_i - \frac{\exp\left(\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}\right)}{1+\exp\left(\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}\right)}\right) + \sum_{i\in A} d_i \frac{\exp\left(\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}\right)}{1+\exp\left(\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}\right)}\right\} \tag{2.14}
$$

Wariancja $\hat{\theta}_{B,DR}$ może być estymowana poprzez przeprowadzenie regresji liniowej na próbie losowej A i wyznaczenie wariancji z estymatora próby losowej $\hat{\theta}_{A,reg}$. Wymagane jest założenie, że $n/N_B = o(1)$. Dowód na nieobciążoność estymatora DR podany jest w pracy J. K. Kim i Wang (2019).

2.4 Implementacja w języku python

Implementacja estymatorów z sekcji 2.2 i 2.3 wymagała użycia paru różnych pakietów języka Python oraz samodzielnego utworzenia niektórych funkcji. Poniżej znajduje się opis najważniejszych pakietów oraz funkcji niezbędnych do prawidłowego przeprowadzenia badania.

Najważniejsze pakiety, które zostały wykorzystane w badaniu:

- **Pakiet NumPy** – NumPy jest fundamentalnym pakietem wykorzystywanym do wykonywania obliczeń naukowych w języku Python. Najważniejszym obiektem dostarczanym przez pakiet NumPy jest obiekt '*ndarry*'. Jest to n-wymiarowa tablica o stałej wielkości zawierająca homogeniczne typy danych. Wykorzystywanie obiektu '*ndarry*' zapewnia o wiele szybsze wykonywanie kodu. Spowodowane jest to faktem, że obiekty '*ndarray*' traktowane są jako wektory. Dzięki temu, w pisanym kodzie nie występują wprost pętle ani indeksowania. Operacje te mają miejsce w zoptymalizowanym, wstępnie skompilowanym kodzie w języku C. Takie wykorzystanie wektorów pozwala również na wykorzystanie zapisu zbliżonego do zapisu matematycznego, który powoduje zmniejszenie wymaganych linijek kodu oraz zwiększa jego czytelność (w porównaniu do klasycznego zapisywania operacji w postaci pętli) (*NumPy Documentation* 2021).
- **Pakiet Pandas** – Pandas dostarcza szybkie, elastyczne i ekspresyjne stryktury danych. Są one stworzone do pracy z oznaczonymi danymi (podobnymi do danych, które możemy znaleźć w relacyjnych bazach danych). Głównym celem pakietu Pandas jest dostarczenie narzędzi do przeprowadzania wysokopoziomowych i praktycznych analiz na różnych zbiorach danych. Z tego pakietu korzysta się bardzo często do przeprowadzania analiz finansowych, statystycznych, naukach społecznych czy w wielu obszarach nauk technicznych. (*Pandas Documentation* 2021)
- **Pakiet SciPy** – Pakiet SciPy wykorzystywany jest głównie do obliczeń naukowych i technicznych. Zawiera on w sobie moduły, dzięki którym możliwe jest poruszenie takich zagadnień jak rozwiązywanie problemów optymalizacyjnych, algebra liniowa, całkowanie czy równania różniczkowe. Do przeprowadzenia badania szczególnie ważne były dwie funkcje z modułu *Optimize*. Moduł ten poświęcony jest funkcjom, które wykorzystywane są do minimalizowania (lub maksymalizowania) określonych funkcji celu. Obejmuje on rozwiązywanie problemów nieliniowych (z obsługą zarówno lokalnych, jak i globalnych algorytmów optymalizacji), programowanie liniowe, ograniczone i nieliniowe metody

najmniejszych kwadratów, znajdowanie pierwiastków i dopasowywanie krzywych (*Scipy Documentation* 2021):

- **scipy.optimize.minimize** – Funkcja ta zapewnia wspólny interfejs dla nieograniczonych i ograniczonych algorytmów minimalizacji dla wielowymiarowych funkcji skalarnych w scipy.optimize.

- **Pakiet scikit-learn** – Pakiet ten zawiera różne algorytmy klasyfikacji, regresji i grupowania, w tym: maszyny wektorów nośnych, lasy losowe, wzmocnienie gradientu, algorytm centroidów i DBSCAN (ang. *Density-based spatial clustering of applications with noise*). Jest zaprojektowany do współpracy z numerycznymi i naukowymi bibliotekami Pythona NumPy i SciPy. Do badania wykorzystana została jedna z funkcji pakietu sklearn (Pedregosa i in. 2011):

  - **sklearn.linear_model.LogisticRegression** – Jest to implementacja regresji logistycznej.

Najważniejsze funkcje, które zostały wykorzystane w badaniu:

```
def estim_ps_ll(lamb, data, ind, xs):
  eta = np.matmul(np.array(data[xs]), lamb)
  pi = rho(eta)
  delta = data[ind] == 1
  ll = delta*np.log(pi) + (1-delta)*np.log(1-pi)
  return -sum(ll)
```

Program 2.1. Implementacja funkcji największej wiarygodności

```
def ps_ll(big_data, sample, membership, lamb, xs, y):
  res = minimize(estim_ps_ll, lamb,
        method = 'Nelder-Mead',
        args=(sample, membership, xs))
  Rho = rho(np.matmul(np.array(big_data[xs]), res.x))
  weight = 1/Rho
  ps_ll = np.average(np.array(big_data[y]), weights= weight)
  return np.array([res.x[0], res.x[1], ps_ll]), res.message
```

Program 2.2. Funkcja wyznaczająca estymator PS przy pomocy maksymalizacji funkcji największej wiarygodności

```
def dr_ll(population, sample_name, sample_data, membership, lamb, xs, y,
    lin_model):
  res = minimize(estim_ps_ll, lamb,
        method = 'Nelder-Mead',
        args=(sample_data, membership, xs))
  Rho = rho(np.matmul(np.array(population[xs]), res.x))
  lin_estimation = np.transpose(np.array(lin_model.predict(population[xs]))
      )
  error = np.subtract( np.array(population[y])  ,  lin_estimation )
  delta_b = np.array(population[membership])
  sample_i = np.array(population[sample_name])
  N = len(population.index)
```

```
    n = np.sum(sample_i)                                                    11
    first_sum_500 = np.sum(delta_b * error / Rho)                           12
    second_sum_500 =  np.sum(sample_i * N / n * lin_estimation)             13
    DR = (first_sum_500 + second_sum_500) / N                               14
    return np.array([res.x[0], res.x[1], DR]), res.message                  15
```

Program 2.3. Funkcja wyznaczająca estymator DR przy pomocy maksymalizacji funkcji największej wiarygodności

2.5 Podsumowanie

Przedstawione założenia, opisane pakiety oraz funkcje w języku Python stanowią podstawę do przeprowadzania dwóch badań. Pierwszym badaniem będzie badanie symulacyjne przeprowadzone w sekcji 3.1. Ma ono na celu pokazać dokładność otrzymanych estymatorów *Propensity Score Estimator* oraz *Doubly Robust Estimator*. Drugim badaniem będzie badanie empiryczne przeprowadzone w sekcji 3.2. Wykorzystane zostaną wcześniej wspomniane estymatory podczas implementacji metody *Data Integration*. W tym przypadku próbą losową A będzie badanie ankietowe przeprowadzone przez GUS zaś próbą Big Data B będą dane z Centralnej Bazy Ofert Pracy.

Rozdział 3

Wyniki integracji badania Popyt na Pracę z Centralną Bazą Ofert Pracy

3.1 Badanie symulacyjne

3.1.1 Założenia

Poniższe badanie symulacyjne jest odtworzeniem badania symulacyjnego przeprowadzonego przez (J. K. Kim i Wang 2019). Jego celem jest zaprezentowane i skontrastowane estymatorów opisanych w sekcjach 2.2 i 2.3 z estymatorem naiwnym (opartym wyłącznie na zbiorze Big Data). Punktem odniesienia do porównania efektywności wszystkich trzech estymatorów będzie ich obciążenie i odchylenie standardowe. Dzięki temu zaprezentowana zostanie efektywność estymatorów *Propensity Score Estimator* (PS) oraz *Doubly Robust Estimator* (DR) pozwalającymi na estymację przez integracje badania częściowego ze zbiorem Big Data.

Pierwszym krokiem badania symulacyjnego będzie wygenerowanie populacji. Następnie wygenerowany zostanie indykator przynależności do zbioru Big Data, dzięki któremu i-ta obserwacja z populacji zostanie zaklasyfikowana jako element zbioru Big Data lub nie. Kolejnym krokiem będzie przeprowadzenie 500 symulacji metodą Monte Carlo. Każda kolejna iteracja symulacji będzie pociągać za sobą wygenerowanie nowego podzbioru populacji generalnej, który będzie traktowany jako badanie ankietowe (próba losowa). Symulacje zostały przeprowadzone dla dwóch wariantów. Wariant pierwszy zakłada wielkość próby losowej na poziomie n = 500. Wariant drugi przewiduje wielkość próby losowej wynoszącą n = 1000. Próba Big Data pozostanie niezmieniona przez wszystkie iteracje symulacji. Na podstawie 500 symulacji Monte Carlo

zestawione ze sobą zostaną wszystkie trzy estymatory dla obu wariantów obszerności prób losowych.

3.1.2 Założenia badania symulacyjnego

Do przeprowadzenia badania symulacyjnego niezbędne jest wygenerowanie populacji. Do tego celu został przyjęty model liniowy (3.1) z dwoma zmiennymi losowymi. Rozkłady poszczególnych zmiennych losowych modelu przedstawione zostały poniżej. Wielkość populacji N została przyjęta na poziomie 1 000 000.

$$y_i = 1 + x_{1,i} + x_{2,i} + \epsilon_i, \tag{3.1}$$

gdzie: $N = 1000000$, $i = 1, ..., N$, $x_{1,i} \sim N(1,1)$, $x_{2,i} \sim Exp(1)$, $\epsilon_i \sim N(0,1)$ i $(x_{1,i}, x_{2,i}, \epsilon_i)$ są niezależne parami

Z powyższej populacji generalnej wygenerowanie zostanie zbiór Big Data. Zbiór ten postanie na podstawie indykatora przynależności do zbioru Big Data δ_i, który przyjmuje wartość 0, gdy i-ta obserwacja z populacji nie należy do zbioru Big Data i 1, gdy i-ta obserwacja należy do tego zbioru. Indykator przynależności do próby Big Data δ_i ma rozkład Bernoulliego z prawdopodobieństwem równym p_i ($\delta_i \sim Ber(p_i)$), gdzie $logit(p_i) = x_{2,i}$. Przyjęcie takiej postaci funkcji logitowej sprawia, że $\sim 60\%$ elementów z populacji generalnej zostanie zakwalifikowana jako element zbioru Big Data. Poniżej znajduje się fragment kodu odpowiedzialny za generowanie wyżej opisanej populacji oraz zbioru Big Data:

```
# 0. Wygenerowana populacja (1 000 000 obserwacji)
N = 1000000
ones = np.ones(N)
x1 = np.random.normal(loc=1, scale=1, size=N)
x2 = np.random.exponential(scale=1, size=N)
e = np.random.normal(loc=0, scale=1, size=N)
y = 1 + x1 + x2 + e

data = {'x0': ones,
        'x1': x1,
        'x2': x2,
        'e': e,
        'y': y}

populacja = pd.DataFrame.from_dict(data)

Lambda = 1
p1 = np.exp( x2 *  Lambda) / (1 + np.exp(x2 *  Lambda))
register = np.random.binomial(n=1, p = p1, size = N)
populacja["big_data"] = register == 1
```

Program 3.1. Generowanie danych w badaniu symulacyjnym

Poniżej widoczne jest pięć pierwszych wierszy zbioru populacji generalnej. Zmienna x_0 to wektor składający się z samych jedynek, zaś zmienna big_data to wektor wskazujący przynależność poszczególnej obserwacji do zbioru Big Data (zgodnie z założeniami w 2.1 zbiór Big Data to podzbiór populacji generalnej).

Tablica 3.1. Fragment zbioru populacji generalnej

	x0	x1	x2	e	y	big_data
0	1.0	-0.085631	0.554665	-0.568493	0.900541	True
1	1.0	1.997345	1.197582	0.260294	4.455222	True
2	1.0	1.282978	0.843429	-0.100796	3.025612	True
3	1.0	-0.506295	4.283358	-0.683621	4.093442	True
4	1.0	0.421400	1.567052	-0.774417	2.214034	False

Źródło: Opracowanie własne na podstawie J. K. Kim i Wang (2019)

Zbiór reprezentujący badanie ankietowe również jest podzbiorem populacji generalnej. Niemniej, jest on generowany przy pomocy losowania prostego przy każdej następującej iteracji symulacji Monte Carlo.

3.1.3 Wyniki badania symulacyjnego

Wyniki przeprowadzonych symulacji Monte Carlo znajdują się w poniższej tabeli 3.2. Punktem odniesienia do porównania efektywności wszystkich trzech estymatorów będzie ich obciążenie oraz odchylenie standardowe. Im dana wartość obciążenia i odchylenia standardowego jest bliższa zeru tym efektywniejszy jest dany estymator. Dla poniższego badania średnia wartość zmiennej objaśnianej y dla populacji wynosi $3,0016$.

Tablica 3.2. Obciążenie i odchylenie standardowe różnych metod Integracji Danych po wykonaniu 500 symulacji metodą Monte Carlo

Estymator	n = 500			n = 1000		
	Wartość	Obciążenie	Odchylenie Standardowe	Wartość	Obciążenie	Odchylenie Standardowe
Naiwny	3,1873	0,1857	1,7876	3,1873	0,1857	1,7876
Propensity Score	3,0003	-0,0013	0,0214	3,0011	-0,0005	0,0152
Doubly Robust	3,0003	-0,0013	0,0451	3,0015	-0,0001	0,0318

Powyższe wyniki obrazują problemy wynikające z wyciągania wniosków jedynie na podstawie niestatystycznego zbioru Big Data. Estymator naiwny, który bazuje wyłączenie na zbiorze Big Data, charakteryzuje się największym obciążeniem oraz odchyleniem standardowym spośród wszystkich zaprezentowanych estymatorów.

Zdecydowanie najlepsze wyniki osiągają estymatory przedstawione przez J. K. Kim i Wang (2019), czyli Propensity Score Estimator oraz Doubly Robust Estimator. Zarówno wartości obciążenia oraz odchylenia standardowego dla obu tych estymatorów są najbardziej zbliżone do zera ze wszystkich zaprezentowanych estymatorów.

3.1.4 Wnioski

Przeprowadzone badanie symulacyjne wskazuje na efektywność estymatorów zaprezentowanych przez (J. K. Kim i Wang 2019). Spośród wszystkich trzech estymatorów cechują się one najmniejszym obciążeniem, a co za tym idzie są najdokładniejszym oszacowaniem wartości średniej z populacji.

Można zatem wyciągnąć wniosek, że metoda Integracji Danych pozwala na znacznie dokładniejszą estymację. Połączenie obu źródeł danych, statystycznego badania częściowego z niestatystycznym zbiorem Big Data, owocuje uzyskanych wyników znacznie lepiej opisujących rzeczywistość niż wykorzystanie samego zbioru Big Data.

W związku z tym podobne badanie zostanie przeprowadzone na rzeczywistych danych w sekcji 3.2. Do empirycznego sprawdzenia funkcjonowania estymatorów jako badanie zo-

stanie wykorzystane badanie przeprowadzone przez GUS w ramach badania popytu na pracę. Źródłem Big Data natomiast będą dane uzyskane z Centralnej Bazy Ofert Pracy (CBOP).

3.2 Badanie empiryczne

3.2.1 Założenia

Poniższe badanie będzie przestawiać praktyczne wykorzystanie estymatorów *Propensity Score Estimator* oraz *Doubly Robust Estimator*. Oba estymatory zostaną skontrastowane z estymatorem naiwnym (czyli średnią ze zmiennej celu znajdującej się w zbiorze Big Data).

Prawidłowa estymacja przy pomocy PS i DR wymaga użycia dwóch różnych zbiorów danych. Ich szczegółowy opis wraz z objaśnieniem pochodzenia znajduje się w następnej podsekcji. Zgodnie z założeniami przyjętymi przez J. K. Kim i Wang (2019) jeden ze zbiorów to zbiór powstały w sposób statystyczny. Drugi zaś powstały w sposób niestatystyczny. Zbiory te powinny powstać w sposób niezależny od siebie, niemniej powinniśmy być w stanie zidentyfikować, które elementy z pierwszego zbioru (o pochodzeniu statystycznym) mogą należeć do drugiego zbioru (o pochodzeniu niestatystycznym). W przypadku danych użytych w poniższym badaniu indeksem przynależności do próby nielosowej (δ_i) będzie informacja, czy w badaniu ankietowym (statystycznym) dany pracodawca zgłosił swoją ofertę pracy do Powiatowego urzędu pracy czy nie, tj:

$$\delta_i = \begin{cases} 1 & \text{jeżeli pracodawca zgłosił daną ofertę do PUP}, \\ 0 & \text{w przeciwnym wypadku}, \end{cases} \tag{3.2}$$

Informacja o przynależności obserwacji ze zbioru statystycznego do zbioru niestatystycznego jest kluczowa. Zakładamy, że zmienne celu, których wartości dla danej populacji będziemy starać się estymować, znajdują się tylko w zbiorze Big Data. Źródło statystyczne nie może posiadać bezpośredniej informacji o zmiennej celu. To właśnie dzięki połączeniu obu zbiorów przy wykorzystaniu indykatora przynależności δ_i otrzymamy dokładniejsze informacje o zmiennej celu niż tylko przy wykorzystaniu zbioru niestatystycznego (Big Data).

Estymacja została przeprowadzana dla dwóch okresów, dla pierwszego kwartału 2018r oraz dla pierwszego kwartału 2019r. Tak jak zostało wspomniane to wcześniej, celem badania jest estymacja wartości dwóch zmiennych, mianowicie zmiennych '*pelen*' oraz '*jedna_zmiana*'.

Zgodnie z założeniami, obie zmienne występują tylko w zbiorze CBOP. Zarówno zmienna '*pelen*' jak i '*jedna_zmiana*' przyjmują wartości *true* lub *false* zgodnie z poniższym schematem:

$$pelen = \begin{cases} true & \text{jeżeli oferta dotyczy pracy w wymiarze 40 godzin tygodniowo,} \\ false & \text{w przeciwnym wypadku,} \end{cases} \tag{3.3}$$

$$jedna\_zmiana = \begin{cases} true & \text{jeżeli oferta pracy dotyczy zatrudnienia w systemie jednozmianowym,} \\ false & \text{w przeciwnym wypadku,} \end{cases} \tag{3.4}$$

3.2.2 Źródła danych

Zbiorem statystycznym (zbiorem A, zgodnie z oznaczeniami z tabeli 2.2) są dane pozyskane w wyniku przeprowadzenia badania częściowego przez Główny Urząd Statystyczny. Badanie, przeprowadzane kwartalnie od 2019r, to nosi nazwę "Badania popytu na pracę". Dane obejmują "*liczbę pracujących osób oraz liczbę i strukturę wolnych miejsc pracy, w tym nowo utworzonych oraz wolnych miejsc pracy zgłoszonych do urzędów pracy. Informacje o nowo utworzonych i zlikwidowanych miejscach pracy.*"(Główny Urząd Statystyczny 2020). Badanie przeprowadzana jest metodą reprezentacyjną. Próba do badania losowana jest oddzielnie dla jednostek o liczbie pracujących powyżej 9 osób oraz dla jednostek zatrudniających do 9 osób. Operat losowania dla tego badania to Baza Jednostek Statystycznych (BJS). Udział w badaniu popytu na pracę jest obowiązkowy zgodnie z art. 30 i 30a Ustawy o statystyce publicznej. Szczegóły metodologii przyjętej przez GUS dostępne są w Zeszycie Metodologicznym (Główny Urząd Statystyczny 2019).

Zbiorem niestatystycznym (zbiorem B, zgodnie z oznaczeniami z tabeli 2.2) są dane pozyskane z Centralnej Bazy Ofert Pracy (CBOP). Można odnaleźć w niej oferty pracy składane przez przedsiębiorców szukających nowych pracowników. Jeżeli dany pracodawca szuka nowych pracowników do pracy w swoim przedsiębiorstwie może zgłosić się do Powiatowego Urzędu Pracy (PUP) i złożyć odpowiedni wniosek (wniosek można zobaczyć w załączniku A.1). Baza ofert pracy jest uzupełniania przez urzędników państwowych pracujących w Powiatowych Urzędach Pracy. Urzędnicy prowadzą również cały proces rejestrowania ofert w systemie CBOP, począw-

szy od przyjmowania wniosków, przez ich weryfikację aż po umieszczanie ofert w systemie. Szczegóły zasad obowiązujących urzędników państwowych podczas procesowania zgłoszeń od pracodawców zostały określone w Ministerstwo Pracy i Polityki Społecznej (2014).

Niemniej, na potrzeby badania została wylosowana grupa Powiatowych Urzędów Pracy, z którymi przeprowadzono wywiad telefoniczny. Miał on na celu sprawdzenie poziomu weryfikacji ofert pracy przez urzędników zanim trafią one do Centralnej Bazy Ofert Pracy. Ważne jest zrozumienie w jaki sposób urzędnicy interpretują Rozporządzenie Ministra oraz jak weryfikują oferty pracy. To właśnie dane wprowadzane przez urzędników zasilają rekordy w Centralnej Bazie Ofert Pracy. Powiatowe urzędy pracy, które zostały wylosowane metodą losowania prostego to:

- PUP w Kościanie
- PUP w Będzinie
- PUP w Bochni
- PUP w Lubaczowie
- PUP w Ostrowcu Świętokrzyskim
- PUP w Sejnach
- PUP w Wolsztynie

We wszystkich Powiatowych Urzędach Pracy, z którymi się skontaktowano, podstawowym źródłem informacji dotyczących ofert pracy jest wniosek składany przeze pracodawcę. Za każdym razem gdy pracodawca składa dany wniosek to jego dane kontaktowe są uaktualniane w systemie urzędu. Sama weryfikacja pracodawcy odbywa się poprzez sprawdzenie KRS, REGON lub CDEIG.

W większości przypadków przepytani urzędnicy zgodnie odpowiadali, że kwestionowanie informacji zamieszczonych przez pracodawcę na wniosku zdarza się jedynie, gdy zachodzą jakieś znacznie niespójności, Niektóre urzędy wychodziły nawet z inicjatywą i organizowały spotkania z pracodawcami, aby samemu pozyskiwać nowe oferty pracy. Kwestią różniącą większość urzędów to daty ważności danej oferty pracy. Na przykład w Powiatowym Urzędzie Pracy w Lubaczowie pracodawca samemu ustala jak długo dana oferta pracy ma być ważna, może ją wycofać w każdej chwili a urząd dodatkowo kontaktuje się z pracodawcą raz w tygodniu w celu potwierdzenia czy oferta jest wciąż ważna. Natomiast w Powiatowym Urzędzie Pracy w Sejnach każda nowa oferta ważna jest domyślnie przez miesiąc i to interesem pracodawcy jest to, aby raz w miesiącu powiadomić urząd o chęci przedłużenia oferty. Kolejną ważną kwestią jest fakt,

że wynagrodzenia wpisywane na ofertach pracy nie są kwotami wiążącymi. Wszystkie urzędy potwierdziły, że gdy dana osoba szukająca pracy skontaktuje się z pracodawcą to mogą oni negocjować warunki niezależnie od kwoty wpisanej na ofercie pracy. Niemniej, jeżeli pracodawca złoży propozycję wynagrodzenia niższą niż na ofercie pracy to osoba szukająca pracy ma prawo zrezygnować z oferty i złożyć zażalenie w Urzędzie Pracy.

W związku z powyższym jakość danych z Centralnej Bazy Ofert Pracy może nie być idealna. Najważniejszym wyznacznikiem jakości danych jest skrupulatność urzędników i dokładność pracodawcy przy wypełnianiu wniosku z ofertą. Zaniedbania z któreś ze stron mogą powodować to, że obraz ofert pracy proponowanych przez polskich pracodawców może zostać zniekształcony.

Oba zbiory danych (badanie popytu na pracę oraz dane z CBOP) opisują oferty pracy zgłoszone przez pracodawców. Każdy wiersz danego zbioru to jedna obserwacja, a dokładniej poszczególna oferta pracy zgłoszona przez pracodawcę. Poniżej w tablicy 3.3 przedstawiono zmienne, które zostaną użyte w badaniu. Szczegółowy opis zmiennych wraz z wartościami, które przyjmują dostępny jest w załączniku A.2.

Jako zmienne objaśniające przyjęte zostały: *woj*, *klasa_pr*, *sek*, *sekc_pkd*, *occup*. Zmiennymi objaśnianymi będą: *pelen* oraz *jedna_zmiana*. Zmienna *delta* zostanie wykorzystana do znalezienia początkowych wartości wektora λ, który będzie niezbędny w procesie maksymalizowania funkcji największej wiarygodności, zgodnie ze wzorem (2.3). Poniżej przedstawiono fragmenty zarówno zbioru częściowego jak i zbioru CBOP, które zostały użyte przy estymacji przy pomocy estymatorów DR i PS .

Tablica 3.4. Fragment zbioru badania częściowego

woj	sek	klasa_pr	sekc_pkd	occup	delta	waga_final
14	1	D	O	1	0	3
14	1	D	O	2	0	9
14	1	D	O	4	0	1
12	1	D	R + S	4	0	5
12	1	D	O	2	0	5

Źródło: Opracowanie własne na podstawie danych ankietowych z GUSu

Tablica 3.3. Zestawienie oznaczeń zmiennych

Zmienna	Wyjaśnienie
woj	województwo, w którym dana oferta pracy została złożona
klasa_pr	wielkość przedsiębiorstwa, którego dotyczy oferta
sek	sektor, którego dotyczy ogłoszenie
sekc_pkd	sekcja pkd
occup	kod zawodu
delta	czy dana oferta została złożona przez przedsiębiorcę do Powiatowego Urzędu Pracy
waga_final	waga obserwacji wynikająca z losowania
liczba_miejsc_og	waga obserwacji wynikająca z ilości wakatów zgłoszonych na dane stanowisko
pelen	czy dana oferta dotyczy pełnego wymiaru pracy (pełen etat)
jedna_zmiana	czy dana oferta dotyczy pracy tylko na jedną zmianę

Tablica 3.5. Fragment zbioru Centralnej Bazy Ofert Pracy

woj	sek	sekc_pkd	klasa_pr	liczba_miejsc_og	pelen	jedna_zmiana	occup
2	2	C	D	2	True	False	7
2	2	N	D	12	True	False	4
2	2	N	D	6	True	False	4
2	2	C	D	7	True	False	8
2	2	C	D	7	True	True	8

Źródło: Opracowanie własne na podstawie danych z Centralnej Bazy Ofert Pracy

3.2.3 Wyniki badania empirycznego

Przed przystąpieniem do estymacji zbadane zostały związki między zmiennymi przyjętymi do badania. W tym celu obliczono współczynniki V Cramera dla każdej ze zmiennych zarówno

dla badania częściowego jak i danych ze zbioru CBOP. Współczynnik V Cramera bazuje na statystyce χ^2. Wzór na wartość statystyki V Cramera podany jest poniżej:

$$V = \sqrt{\frac{\chi^2}{n \cdot \sqrt{(k-1)(r-1)}}} \tag{3.5}$$

gdzie: χ^2 – wyznaczona wartość statystyki χ^2, n – liczba wszystkich obserwacji, k – liczba kolumn tabeli kontyngencji bez sumy (liczba wariantów pierwszej cechy), r – liczba wierszy tabeli kontyngencji bez sumy (liczba wariantów drugiej cechy)

Otrzymane wartości współczynnika V Cramera można zinterpretować następująco: 1) od 0,00 do 0,29 – słaby związek pomiędzy zmiennymi; 2) od 0,30 do 0,49 – umiarkowany związek pomiędzy zmiennymi; 3) od 0,50 do 0,10 – silny związek pomiędzy zmiennymi.

Po wyznaczeniu relacji między zmiennymi okazało się, że większość zmiennych przejawia dosyć słaby lub umiarkowanie silny związek ze sobą nawzajem. Większość wartości oscyluje w granicy 0,2-0,4. Zmienne o najsilniejszych związkach w obu zbiorach zostały przedstawione poniżej, w tablicach 3.6 i 3.7.

Tablica 3.6. Zmienne o najsilniejszych związkach ze zbioru badania popyty na pracę

	Kwartał	
Para zmiennych	2018 kw1	2019 kw1
occup i sek	0,35	0,35
sekc_pkd i occup	0,42	0,39
sekc_pkd i sek	0,75	0,76
klasa_pr i sekc_pkd	0,31	0,34
klasa_pr i sek	0,31	0,32

Tablica 3.7. Zmienne o najsilniejszych związkach ze zbioru CBOP

	Kwartał	
Para zmiennych	2018 kw1	2019 kw1
sekc_pkd i jedna_zmiana	0,49	0,47
klasa_pr i jedna_zmiana	0,45	0,46
sekc_pkd i sek	0,71	0,75
sekc_pkd i occup	0,33	0,34
klasa_pr i sekc_pkd	0,33	0,33

Mimo, że większość zmiennych wskazuje na raczej słaby związek ze sobą, przykłady par z tablic 3.6 i 3.7 pokazują, że istnieją pewne zależności między zmiennymi. Siła związków wystę-

pująca miedzy zmiennymi może rzutować na mniejszą dokładność przeprowadzonej estymacji przy pomocy PS i DR.

Poniżej przedstawiono dwie kolejne tablice pokazujące relacje między zmiennymi. Tym razem przedstawiają one relacje między zmiennymi objaśniającymi ze zbioru popyt na pracę, a indykatorem przynależności δ do zbioru CBOP (tablica 3.8), oraz związki między zmiennymi objaśniającymi, a zmiennymi objaśnianymi ze zbioru CBOP (tablica 3.9).

Tablica 3.8. Związki zmiennych objaśniających ze zbioru popytu na pracę z indykatorem przynależności δ do zbioru CBOP dla 1 kwartału 2018 i 2019

	Kwartał	
Zmienna	2018 kw1	2019 kw1
sek	0,076	0,022
occup	0,22	0,17
sekc_pkd	0,2	0,18
klasa_pr	0,19	0,11
woj	0,22	0,26

Tablica 3.9. Związki zmiennych objaśnianych ze zmiennymi objaśniającymi ze zbioru CBOP

	Zmienna objaśniana			
	pelen		jedna_zmiana	
	Kwartał		Rok	
Zmienna objaśniająca	2018 kw1	2019 kw1	2018 kw1	2019 kw1
sek	0,059	0,058	0,09	0,098
occup	0,089	0,12	0,29	0,31
sekc_pkd	0,17	0,18	0,49	0,47
klasa_pr	0,02	0,085	0,45	0,46
woj	0,12	0,092	0,16	0,24

Poniższa tablica 3.10 pokazuje wyniki przeprowadzonego badania empirycznego. Zestawiono w niej wartości trzech estymatorów (naiwny, PS oraz DR) zmiennych *'pelen'* i *'jedna_zmiana'* dla dwóch okresów (pierwszego kwartału 2018r i pierwszego kwartału 2019r):

Tablica 3.10. Wyniki przeprowadzonej estymacji przy pomocy PS i DR dla połączonych zbiorów badania popytu na prace i zbioru CBOP

Kwartał	Zmienna celu	Estymator	Wartość estymatora	Kwartał	Zmienna celu	Estymator	Wartość estymatora
2018 k1	pelen	Naiwny	0.9665	2019 k1	pelen	Naiwny	0.9675
		PS	0.9569			PS	0.9587
		DR	1.0563			DR	0.9725
2018 k1	jedna_zmiana	Naiwny	0.4846	2019 k1	jedna_zmiana	Naiwny	0.4936
		PS	0.6690			PS	0.7315
		DR	0.7989			DR	0.7556

W badaniu empirycznym estymator naiwny to średnia ważona ze zmiennej celu. Zastosowaną wagą jest zmienna *'liczba_miejsc_og'*, która oznacza ile wolnych miejsc pracy jest oferowanych przez pracodawcę dla konkretnej oferty pracy. Wykorzystanie sumy ważonej pozwala na pokazanie rzeczywistej częstości występowania danej zmiennej celu. W tym przypadku estymator naiwny wyrażony jest wzorem:

$$\hat{\theta}_{\text{naïve}} = \frac{\sum_{i \in B} y_i n_i}{\sum_{i \in B} n_i} \tag{3.6}$$

gdzie: y_i – wartość i-tej obserwacji zmiennej celu ze zbioru CBOP (Big Data), n_i – waga i-tej obserwacji, wagą obserwacji jest wartość zmiennej *'liczba_miejsc_og'*.

Dla obu okresów wartość estymatora naiwnego dla zmiennej *'pelen'* wynosiła ponad 96%. Estymator PS wskazuje wynik bliższy 95% dla obu okresów. Wynik dla DR w pierwszym kwartale 2018 wynosi 105% co jest wynikiem błędnym (niemożliwe jest, aby więcej niż 100% ogłoszeń o pracę oferowało pracę w pełnym wymiarze godzin pracy, maksymalnie może być to 100%, nie więcej). Tak obciążony wynik może być spowodowany dwoma czynnikami. Po pierwsze, tabela 3.9 wskazuje, że zmienna *'pelen'* posiada bardzo słabą relację z innymi zmiennymi. Może to powodować niedokładną estymację. Po drugie, taki wynik może być spowodowany zaokrągleniami występującymi w trakcie obliczeń co również może zaburzać poprawność wyniku. Niemniej, dla pierwszego kwartału 2019r estymacja przy pomocy DR wykazała wynik na poziomie 97%. Jest to wynik większy zarówno od estymatora naiwnego jak i PS.

W przypadku zmiennej *'jedna_zmiana'* wartości estymatora naiwnego oscylowały w przedziale 48-49% dla obu okresów. Dla tej zmiennej widoczne są też większe różnice po przeprowadzonej estymacji przy pomocy PS i DR. Dla PS jest to odpowiednio 67% ogłoszeń oferujących pracę na jedną zmianę w pierwszym kwartale 2018r i 73% w pierwszym kwartale 2019. DR również wykazuje inne wyniki w porównaniu z estymatorem naiwnym, niemniej są one zbli-

żone do estymatora PS. Dla obu okresów estymator DR wynosi odpowiednio 80% i 76%. Te estymowane wyniki są dokładniejsze niż dla zmiennej *'pelen'*, ponieważ istnieją silniejsze relacje między zmienną *'jedna_zmiana'*, a pozostałymi zmiennymi objaśniającymi (co widoczne jest w tabeli 3.9). Ta różnica między estymatorem naiwnym a estymatorami PS i DR będąca rzędu 30 punków procentowych pokazuje dokładność omawianych estymatorów, gdy warunki (relacje między zmiennymi) są wystarczająco dobre.

3.2.4 Wnioski

Przeprowadzone badanie empiryczne pokazuje, że estymatory przedstawione przez J. K. Kim i Wang (2019) dają wiarygodne wyniki. Jeżeli występują istotne związki między zmiennymi to estymatory PS i DR prezentują dokładniejsze wyniki niż estymator naiwny.

Widać to szczególnie dobrze w przypadku estymacji zmiennej *'jedna_zmiana'*, gdzie różnica między estymatorem naiwnym, a estymatorem PS wynosi 0,1844 w 2018r i 0,2379 w 2019r. Estymacja zmiennej *'jedna_zmiana'* przy pomocy estymatora DR również daje dokładniejsze wyniki. Różnica między estymatorem naiwnym, a estymatorem DR wynosi 0,3143 w 2018r i 0,262 w 2019r. Podtrzymują to wyniki badania symulacyjnego, które wskazały efektywność estymatorów PS i DR.

Jeżeli jednak związki między zmiennymi nie są dostatecznie silne to estymatory PS i DR mogą nie dawać wystarczająco dokładnych wyników. Widać to szczególnie w przypadku wartości estymatora DR dla zmiennej *'pelen'* gdzie estymowana wartość zmiennej celu przekracza dopuszczalną wartość 1. Niemniej, mimo obciążoności wynikającej ze słabych związków między zmiennymi, estymatory PS i DR wciąż można wykorzystywać do opisu danej zmiennej celu. Trzeba tylko pamiętać, że wyniki mogę nie być tak dokładne jak w przypadku danych, w których występują silniejsze związki między zmiennymi.

3.3 Podsumowanie wyników

Estymacja z wykorzystaniem estymatorów PS i DR przynosi lepsze wyniki w porównaniu do zastosowania estymatora naiwnego. Zostało to potwierdzone poprzez przeprowadzenie badania symulacyjnego, które, dzięki kontrolowanym warunkom, pokazało zwiększoną dokładność obu estymatorów. Potwierdzone jest to poprzez obliczone wartości obciążenia i odchylenia

standardowego, które zgodnie z tabelą 3.2 wskazują na większą dokładność zarówno PS jak i DR.

Dzięki potwierdzeniu dokładności otrzymanych wyników poprzez przeprowadzenia badania symulacyjnego oba estymatory mogły zostać użyte w praktyce. Dzięki temu, po użyciu dwóch zbiorów danych (zbioru statystycznego z badania popytu na pracę i zbioru niestatystycznego z Centralnej Bazy Ofert Pracy), otrzymane zostały również wyniki estymowanych zmiennych *'jedna_zmiana'* oraz *'pelen'*. W tym przypadku dużą rolę w precyzji estymacji odgrywały związki zmiennych celu ze zmiennymi objaśniającymi. Dla zmiennej *'pelen'* związki były dosyć słabe, przez co wyniki nie były aż tak dokładnie. Niemniej, dla zmiennej *'jedna_zmiana'*, gdzie relacje ze zmiennymi objaśniającymi były znacznie silniejsze, wyniki estymacji okazały się znacznie dokładniejsze w porównaniu do estymatora naiwnego.

Zarówno badanie symulacyjne jak i badanie empiryczne pokazały, że zastosowanie estymatorów zaprezentowanych przez J. K. Kim i Wang (2019) pozwala na praktyczne wykorzystanie metody integracji danych w celu estymacji pewnych cech danej zbiorowości. Oba badania pokazały, że estymatory te (przy spełnieniu odpowiednich założeń) są dobrym sposobem na poszerzenie wiedzy na temat danej populacji, gdy jesteśmy w stanie zdobyć dane ankietowe i rozszerzyć je zbiorem Big Data.

Podsumowanie

Celem niniejszej pracy było zaprezentowanie oraz praktyczne wykorzystanie estymatorów propenisty score estimator (PS) oraz doubly robust estimator (DR) przedstawionych przez J. K. Kim i Wang (2019) w artykule *Sampling Techniques for Big Data Analysis*.

Badanie symulacyjne, które było odtworzeniem badania symulacyjnego przedstawionego w J. K. Kim i Wang (2019), wykazało efektywność obu estymatorów. Samo badanie przeprowadzone zostało metodą Monte Carlo. Ze wcześniej wygenerowanej w sposób losowy populacji wygenerowano podzbiór będący zbiorem Big Data (ok 60% liczebności populacji). Następnie przeprowadzono 500 prób Monte Carlo. Po uzyskaniu 500 wyników dla obu estymatorów policzono średnią ze wszystkich wartości. Z tej ostatecznej wartości obliczono obciążenie oraz odchylenie standardowe. W porównaniu do estymatora naiwnego (będącego średnią ze zmiennej celu), w obu przypadkach obciążenie estymatorów PS i DR wynosiło mniej niż 3% wartości obciążenia estymatora naiwnego dla próby o wielkości 500 i mniej niż 1% wartości obciążenia estymatora naiwnego dla próby o wielkości 1000. Oznacza to, że oba estymatory był średnio o 15 do 30 razy bardziej dokładne niż estymator naiwny. Odchylenie standardowe również wahało się na poziomie 2-3% wielkości odchylenia standardowego estymatora naiwnego.

Estymacja w badaniu empirycznym została przeprowadzona przy pomocy wykorzystania dwóch odrębnych zbiorów danych. Pierwszym zbiorem był statystyczny zbiór z badania popytu na pracę przeprowadzonego przez Główny Urząd Statystyczny. Drugim zbiorem był niestatystyczny zbiór pochodzący z Centralnej Bazy Ofert Pracy. W obu zbiorach ograniczono się do danych z pierwszego kwartału 2018 i 2019 roku. Oba zbiory pokazują strukturę kształtowania się popytu na pracę. Przedstawiają one oferty pracy składane przez pracodawców poszukujących nowych pracowników.

W przypadku danych z CBOP dokonano dodatkowej weryfikacji jakości danych poprzez przeprowadzenie wywiadu telefonicznego z wybranymi Powiatowymi Urzędami Pracy. Próba urzędów została wylosowana metodą losowania prostego i wynosiła ona 7 jednostek z terenu całej

Polski. Przeprowadzone rozmowy z urzędnikami wskazały, że jakość danych zależy w głównej mierze od skrupulatności pracowników urzędów oraz dokładności pracodawców podczas wypełniania wniosku na złożenie nowej oferty pracy. W związku z tym, że zależnie od urzędu procesy weryfikacyjne oraz wytyczne odnośnie kontaktowania się z pracodawcami różniły się nieznacznie od siebie to należy pamiętać, że istnieje pewna doza niepewności w otrzymanych danych. Może to powodować pewne zakrzywienie obrazu ofert pracy składanych przez polskich pracodawców.

W badaniu empirycznym zmiennymi, które zostały poddane estymacji były *'pelen'* (czy dana oferta pracy jest na pełen etat?) oraz *'jedna_zmiana'* (czy dana oferta pracy jest na jedną zmianę?). W przypadku zmiennej *'jedna_zmiana'* wyniki okazały się dokładniejsze niż w przypadku zmiennej *'pelen'*. Estymatory PS i DR zmiennej *'jedna_zmiana'* miały zbliżone wartości ok 65-75% dla obu okresów podczas gdy estymator naiwny wyniósł ok 48-49%. Zmienna *'pelen'* okazała się bardziej problematyczna. Wszystkie trzy estymatory dla obu okresów miały wartości w przedziale 95-97%. Niemniej dla pierwszego kwartału 2018 roku wartość estymatora DR wyniosła 105% co znacząco przekracza dopuszczalny wynik (nie może być więcej niż 100% ofert pracy gwarantujących pracę na jedną zmianę). Wynikać to może przede wszystkim z dosyć słabej relacji między zmienną *'pelen'* a zmiennymi objaśniającymi oraz zaokrągleniami podczas obliczeń. Siły tych relacji zostały policzone przy pomocy statystyki V Cramera. Dla porównania, zmienna *'jedna_zmiana')* cechuje się silniejszymi relacjami ze zmiennymi objaśniającymi co przełożyło się na o wiele dokładniejszą estymację.

Zgodnie z wiedzą autora i promotora jest to pierwsze badanie w Polsce, którego celem jest rozszerzenie badania Popyt na Pracę o informacje ze źródeł niestatystycznych z wykorzystaniem metod integracji danych. Dotychczas nie zostały dostarczone podobne szacunki wykorzystujące zaprezentowaną metodę.

Bibliografia

Bethlehem, J. & Biffignandi, S. (2012). *Handbook of web surveys* (Vol. 567). John Wiley & Sons.

Chen, J. K. T., Valliant, R. L. & Elliott, M. R. (2018). Model-assisted calibration of non-probability sample survey data using adaptive LASSO. *Survey Methodology*, *44*(1), 117–144.

Chen, J. K. T., Valliant, R. L. & Elliott, M. R. (2019). Calibrating non-probability surveys to estimated control totals using LASSO, with an application to political polling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *68*(3), 657–681.

Chen, S., Yang, S. & Kim, J. K. (2021). Nonparametric Mass Imputation for Data Integration. *Journal of Survey Statistics and Methodology*.

Chen, Y., Li, P. & Wu, C. (2020). Doubly robust inference with nonprobability survey samples. *Journal of the American Statistical Association*, *115*(532), 2011–2021.

Franke, B., Plante, J.-F., Roscher, R., Lee, E.-s. A., Smyth, C., Hatefi, A., Chen, F., Gil, E., Schwing, A., Selvitella, A. i in. (2016). Statistical inference, learning and models in big data. *International Statistical Review*, *84*(3), 371–389.

Główny Urząd Statystyczny. (2019). *Zeszyt metodologiczny popytu na pracę*. Pobrane czerwiec 10, 2021 z https://stat.gov.pl/obszary-tematyczne/rynek-pracy/popyt-na-prace/zeszyt-metodologiczny-popyt-na-prace,3,1.html

Główny Urząd Statystyczny. (2020). *Popyt na pracę*. Pobrane czerwiec 10, 2021 z https://stat.gov.pl/obszary-tematyczne/rynek-pracy/popyt-na-prace/popyt-na-prace-w-2019-roku,1,15.html

GUS. (2020). Program badań statystycznych statystyki publicznej na rok 2021.

Kim, J. K. & Wang, Z. (2019). Sampling techniques for big data analysis. *International Statistical Review*, *87*, S177–S191.

McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D. & Barton, D. (2012). Big data: the management revolution. *Harvard business review*, *90*(10), 60–68.

Michalski, T. (2004). *Statystyka*. Sklep WSiP 10% rabatu.

Ministerstwo Pracy i Polityki Społecznej. (2014). ROZPORZĄDZENIE MINISTRA PRACY I POLITYKI SPOŁECZNEJ w sprawie szczegółowych warunków realizacji oraz trybu i sposobów prowadzenia usług rynku pracy.

Neyman, J. (1934). On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection. *Journal of the Royal Statistical Society*, *97*(4), 558–625.

NumPy Documentation. (2021). Pobrane kwiecień 21, 2021 z https://numpy.org/doc/stable/user/whatisnumpy.html

Ostasiewicz, W. (2011). *Badania statystyczne*. Wolters Kluwer.

Pandas Documentation. (2021). Pobrane kwiecień 21, 2021 z https://pandas.pydata.org/docs/getting_started/overview.html

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Rubin, D. B. (1976). Inference and Missing Data. *Biometrika*, *63*(3), 581–592. http://www.jstor.org/stable/2335739

Scipy Documentation. (2021). Pobrane kwiecień 21, 2021 z https://www.scipy.org/about.html

Showkat, N. & Parveen, H. (2017). Non-Probability and Probability Sampling.

Shvachko, K., Kuang, H., Radia, S. & Chansler, R. (2010). The Hadoop Distributed File System. *2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST)*, 1–10. https://doi.org/10.1109/MSST.2010.5496972

Sobczyk, M. (2007). *Statystyka*. Wydawnictwo Naukowe PWN.

Tam, S.-M. & Kim, J.-K. (2018). Big Data ethics and selection-bias: An official statistician's perspective. *Statistical Journal of the IAOS*, *34*(4), 577–588.

United Nation's offical webpage: Department of Economic and Social Affairs. (2020). Pobrane listopad 14, 2020 z https://www.un.org/en/desa/about-us/organigramme

United Nation's offical webpage: About Us. (2020). Pobrane listopad 14, 2020 z https://unstats.un.org/home/about/

Ustawa z dnia 21 października 1919 r. o organizacji statystyki administracyjnej. (1919).

Ustawa z dnia 29 czerwca 1995 r. o statystyce publicznej. (1995).

Ustawa z dnia 9 sierpnia 2019 r. o narodowym spisie powszechnym ludności i mieszkań w 2021 r. (2019).

Wu, C. & Thompson, M. E. (2020). Sampling Theory and Practice.

Yang, S., Kim, J. K. & Song, R. (2020). Doubly robust inference when combining probability and non-probability samples with high dimensional data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *82*(2), 445–465.

Spis tablic

2.1 Zestawienie oznaczeń 20
2.2 Idea łączenia danych – przykład dwóch zbiorów 20

3.1 Fragment zbioru populacji generalnej 30
3.2 Obciążenie i odchylenie standardowe różnych metod Integracji Danych po wykonaniu 500 symulacji metodą Monte Carlo 31
3.4 Fragment zbioru badania częściowego 35
3.3 Zestawienie oznaczeń zmiennych 36
3.5 Fragment zbioru Centralnej Bazy Ofert Pracy 36
3.6 Zmienne o najsilniejszych związkach ze zbioru badania popyty na pracę 37
3.7 Zmienne o najsilniejszych związkach ze zbioru CBOP 37
3.8 Związki zmiennych objaśniających ze zbioru popytu na pracę z indykatorem przynależności δ do zbioru CBOP dla 1 kwartału 2018 i 2019 38
3.9 Związki zmiennych objaśnianych ze zmiennymi objaśniającymi ze zbioru CBOP . 38
3.10 Wyniki przeprowadzonej estymacji przy pomocy PS i DR dla połączonych zbiorów badania popytu na prace i zbioru CBOP 39

Spis rysunków

1.1 Sposoby doboru próby 8
1.2 Kategorie błędów statystycznych w badaniach reprezentacyjnych 12

Spis programów

2.1 Implementacja funkcji największej wiarygodności 26
2.2 Funkcja wyznaczająca estymator PS przy pomocy maksymalizacji funkcji największej wiarygodności 26
2.3 Funkcja wyznaczająca estymator DR przy pomocy maksymalizacji funkcji największej wiarygodności 26
3.1 Generowanie danych w badaniu symulacyjnym 29

Dodatek A

Załączniki

A.1 Wniosek do zgłoszenia krajowej oferty pracy

POWIATOWY URZĄD PRACY W POZNANIU
Ul. Czarnieckiego 9, 61-538 Poznań
fax: 61-8330-252
www. poznan.praca.gov.pl
ofertypracy@poznan.praca.gov.pl, tel. 61 8345 672

Test rynku: testrynku@poznan.praca.gov.pl, tel. 61 8345 673

Forma upowszechnienia oferty:

☐ - **otwarta- zawierająca dane umożliwiające identyfikację pracodawcy krajowego**

☐ - **zamknięta –nie zawierająca danych pracodawcy krajowego**

Oferta zgłaszana w celu uzyskania opinii Starosty związanej z zatrudnieniem cudzoziemca TAK / NIE *

ZGŁOSZENIE KRAJOWEJ OFERTY PRACY

1. Nr zgłoszenia:	**2. Data przyjęcia zgłoszenia**

I DANE DOTYCZĄCE PRACODAWCY KRAJOWEGO

3. Nazwa pracodawcy

.....
.....
……

Imię, Nazwisko i stanowisko osoby reprezentującej pracodawcę (numer telefonu do kontaktu)
.....
…….

5.Numer statystyczny pracodawcy - (REGON)

☐☐☐☐☐☐☐☐☐

6. Numer identyfikacji Podatkowej - NIP

☐☐☐-☐☐☐-☐☐-☐☐

7. Podstawowy rodzaj działalności wg PKD ☐☐☐☐ ☐

Charakterystyka prowadzonej działalności……
……
……

4.Adres pracodawcy

☐☐-☐☐☐ .....
kod pocztowy miejscowość

ulica .....

gmina .....
e – mail .....
strona internetowa .....

Forma prawna prowadzonej działalności:
- osoba fizyczna prowadząca działalność gospodarczą
- spółka……..
- przedsiębiorstwo państwowe
- inna……..

Nie jestem / jestem* agencją zatrudnienia
Zgłaszam ofertę pracy tymczasowej **TAK / NIE***

8.Liczba zatrudnionych pracowników	**9.Data rozpoczęcia działalności**
…….	…….

10.Pouczony o odpowiedzialności karnej przewidzianej w art. 233 § 1 K.K. **Oświadczam co następuje:**

Oferta jest w tym samym czasie zgłoszona do innego urzędu pracy **TAK / NIE***

W okresie 365 dni przed zgłoszeniem oferty pracy zostałem ukarany lub skazany prawomocnym wyrokiem za naruszenie przepisów prawa pracy albo jestem objęty postępowaniem dotyczącym naruszenia przepisów prawa pracy **TAK / NIE***

Złożona oferta pracy nie może naruszać zasad równego traktowania w zatrudnieniu w rozumieniu przepisów prawa pracy i nie może zawierać wymagań dyskryminujących ze względu na płeć, wiek, niepełnosprawność, rasę, religię, narodowość, przekonania polityczne, przynależność związkową, pochodzenie etniczne, wyznanie lub orientację seksualną.

Ponadto pouczamy, iż w przypadku braku w zgłoszeniu krajowej oferty pracy danych wymaganych, powiatowy urząd pracy powiadamia pracodawcę krajowego, w formie ustalonej dla wspólnych kontaktów, o konieczności uzupełnienia zgłoszenia. Nieuzupełnienie przez pracodawcę krajowego zgłoszenia w terminie do 7 dni od dnia powiadomienia, spowoduje, że oferta pracy nie zostanie przyjęta do realizacji przez powiatowy urząd pracy.
Podstawa prawna: Ustawa z dnia 20 kwietnia 2004r. o promocji zatrudnienia i instytucjach rynku pracy.

…… ……
Miejscowość, data Imię i Nazwisko /Podpis pracodawcy/ osoby upoważnionej

***niewłaściwe skreślić**

II DANE DOTYCZĄCE ZGŁOSZONEGO MIEJSCA PRACY

11.Nazwa zawodu
.....
.....

12.Nazwa stanowiska
.....
.....
.....
.....

14.Liczba wolnych miejsc zatrudnienia, ☐☐
w tym dla osób niepełnosprawnych ☐☐

13.Kod zawodu
☐☐☐☐☐☐☐☐

15. Wnioskowana liczba kandydatów ☐☐

16.Adres miejsca pracy
.....
.....
.....

17. Dodatkowe informacje

18.Rodzaj umowy
- ☐ o pracę na czas nieokreślony
- ☐ o pracę na czas określony
- ☐ o pracę na okres próbny
- ☐ umowa zlecenie
- ☐ umowa o dzieło
- ☐ umowa na zastępstwo
- ☐ umowa o pracę tymczasową

19. Wymiar czasu pracy
- ☐ pełny etat
- ☐ ¾ etatu
- ☐ ½ etatu
- ☐ inne (ilość godzin pracy)……

20. Rozkład pracy w godz.
- ☐ jedna zmiana – godz.……..…
- ☐ dwie zmiany – godz.……...
- ☐ trzy zmiany – godz.……
- ☐ ruch ciągły
- ☐ inne…….

21. System wynagrodzenia
- ☐ Miesięczny
- ☐ Godzinowy
- ☐ Akordowy
- ☐ Prowizyjny

22 Wysokość wynagrodzenia (brutto)
…..... **brutto**
W przypadku umowy cywilno-prawnej – stawka godzinowa.

23. Data rozpoczęcia zatrudnienia
Od ……
Do ……
Okres zatrudnienia w przypadku umowy o pracę albo okres wykonywania umowy w przypadku umowy cywilno-prawnej

24.Data ważności oferty (nie dłużej niż 30 dni)
.....

III DANE DOTYCZĄCE OCZEKIWAŃ WOBEC KANDYDATA

25. Wykształcenie (poziom/kierunek) ……

26. Doświadczenie zawodowe…….

27. Umiejętności……..

28. Uprawnienia……..……

29. Znajomość j. obcych (stopień znajomości)……..

30. Oczekiwania mile widziane.…….……

31. Inne……...

32. Charakterystyka lub rodzaj wykonywania pracy……..
…….
…….

33. Preferowana forma kontaktu z pracodawcą:
1) kontakt osobisty w godz. ……
2) telefoniczne umówienie spotkania……..
3) inne…….

34. Jakie dokumenty ma złożyć potencjalny kandydat:
☐ CV ☐ świadectwo pracy
☐ list motywacyjny ☐ inne……..

IV DANE DOTYCZĄCE POSTEPOWANIA Z OFERTĄ PRACY

35. Okres aktualności oferty pracy: od……. do …….
36. Sposób i częstotliwość kontaktów: co najmniej raz na 3 dni / w wyznaczonym * terminie……
37. Informacja o działaniach urzędu podjętych na rzecz pracodawcy/ dodatkowe informacje dot. realizacji oferty pracy:……...
38. Jestem / nie jestem *zainteresowany przekazaniem oferty pracy do innych powiatowych urzędów pracy, w celu jej upowszechnienia (proszę podać jakich)……
39. Jestem/ nie jestem zainteresowany upowszechnieniem oferty prac w wybranych państwach EOG – proszę podać w jakich……..
40. Jestem / nie jestem* zainteresowany zatrudnieniem kandydatów z państw EOG (w przypadku zainteresowania zatrudnieniem obywatela EOG proszę wypełnić załącznik nr 1)
41. W przypadku ubiegania się o zezwolenie na pracę cudzoziemca wyrażam zgodę/nie wyrażam zgody* na rekrutację kandydatów zgodnie z § 6 ust. 1 pkt.7 Rozporządzenia Ministra Rodziny, Pracy i Polityki Społecznej z dnia 7 grudnia 2017r. w sprawie wydawania zezwolenia na pracę cudzoziemca oraz wpisu oświadczenia o powierzeniu wykonywania pracy cudzoziemcowi do ewidencji oświadczeń (Dz. U. 2017 poz. 2345).

* niewłaściwe skreślić

A.2 Opis zmiennych wykorzystanych w badaniu

Nazwa zmiennej	Wartość zmiennej	Etykieta
Delta	0	Oferta zgłoszona poza Powiatowym Urzędem Pracy
	1	Oferta zgłoszona do Powiatowego Urzędu Pracy
pelen	True	Oferta dotyczy pełnego wymiaru pracy
	False	Oferta nie dotyczy pełnego wymiaru pracy
jedna_zmiana	True	Oferta dotyczy pracy tylko na jedną zmianę
	False	Oferta nie dotyczy pracy tylko na jedną zmianę
woj	2	Dolnośląskie
	4	Kujawsko-pomorskie
	6	Lubelskie
	8	Lubuskie
	10	Łódzkie
	12	Małopolskie
	14	Mazowieckie
	16	Opolskie
	18	Podkarpackie
	20	Podlaskie
	22	Pomorskie
	24	Śląskie
	26	Świętokrzyskie
	28	Warmińsko-mazurskie
	30	Wielkopolskie
	32	Zachodniopomorskie
klasa_pr	M	Małe przedsiębiorstwo - mniej niż 50 pracowników, przychody netto mniejsze lub równe 10 mln euro;
	S	Średnie przedsiębiorstwo - mniej niż 250 pracowników, przychody netto mniejsze lub równe 43 mln euro;
	D	Duże przedsiębiorstwo - wszystkie, które nie mieszczą się w powyższych kategoriach
sek	1	Sektor publiczny
	2	Sektor prywatny
sekc_pkd	A	Rolnictwo, leśnictwo, łowiectwo i rybactwo
	B	Górnictwo i wydobywanie
	C	Przetwórstwo przemysłowe
	D	wytwarzanie i zaopatrywanie w energię elektryczną, gaz, parę wodną, gorącą wodę i powietrze do układów klimatyzacyjnych
	E	dostawa wody; gospodarowanie ściekami i odpadami oraz działalność związana z rekultywacją
	F	Budownictwo
	G	Handel hurtowy i detaliczny; naprawa pojazdów samochodowych, włączając motocykle
	H	Transport i gospodarka magazynowa
	I	Działalność związana z zakwaterowaniem i usługami gastronomicznymi
	J	Informacja i komunikacja
	K	Działalność finansowa i ubezpieczeniowa
	L	Działalność związana z obsługą rynku nieruchomości
	M	Działalność profesjonalna, naukowa i techniczna
	N	Działalność w zakresie usług administrowania i działalność wspierająca
	O	Administracja publiczna i obrona narodowa; obowiązkowe zabezpieczenia społeczne
	P	Edukacja
	Q	Opieka zdrowotna i pomoc społeczna
	R	Działalność związana z kulturą, rozrywką i rekreacją
	S	Pozostała działalność usługowa
occup	1	Przedstawiciele władz publicznych, wyżsi urzędnicy i kierownicy
	2	Specjaliści
	3	Technicy i inny średni personel
	4	Pracownicy biurowi
	5	Pracownicy usług osobistych i sprzedawcy
	6	Rolnicy, ogrodnicy, leśnicy i rybacy
	7	Robotnicy przemysłowi i rzemieślnicy
	8	Operatorzy i monterzy maszyn i urządzeń
	9	Pracownicy przy pracach prostych