UNIWERSYTET
EKONOMICZNY
W POZNANIU

Wydział Informatyki
i Gospodarki Elektronicznej

Jakub Skotarek, Krzysztof Sukiennicki

Wielowymiarowa analiza uniwersytetów i szkół wyższych o profilu ekonomicznym na podstawie badania Ekonomicznych Losów Absolwentów w 2016 roku

Multivariate analysis of universities and higher education institutions with a focus on economic sciences based on the Polish Graduate Tracking System in 2016

Praca licencjacka

Promotor: dr Maciej Beręsewicz

Pracę przyjęto dnia:

Podpis promotora

Kierunek: Informatyka i ekonometria

Specjalność: Analityka gospodarcza

Poznań 2019

Spis treści

Wstęp (Jakub Skotarek, Krzysztof Sukiennicki)

Od wielu lat w Polsce jak i na świecie prowadzony jest monitoring losów absolwentów. Mimo stałego rozwoju tego badania, obarczone jest ono dużym ryzykiem błędu, wynikającym z ograniczonej grupy respondentów biorących w nim udział. Monitoring ten może mieć charakter ankietowy. W konsekwencji otrzymywane są wyniki różnej jakości, których interpretacja może być dość utrudniona.

Celem poniższej pracy jest wielowymiarowa analiza uniwrsytetów i szkół wyższych o profilu ekonomicznym, z wykorzystaniem analizy głównych składowych oraz analizy skupień. Do projektu wykorzystano bazę danych badania ekonomicznych losów absolwentów (ELA) przygotowaną przez Ośrodek Przetwarzania Informacji – Państwowy Instytut Badawczy (OPI PIB) na zlecenie Ministerstwa Nauki i Szkolnictwa Wyższego. Automatyczna analiza, jaką posługuje się ELA, dokonywana jest przy pomocy oprogramowania oraz metodologii, które stworzyli naukowcy z Uniwersytetu Warszawskiego pod kierownictwem prof. Mikołaja Jasińskiego.

Praca dyplomowa składa się z czterech rozdziałów. Pierwszy z nich skupia się na pomiarze losów absolwentów uczelni wyższych. Opisana w nim jest problematyka badania, jaki jest jego cel, doświadczenia polskie oraz zagraniczne w prowadzeniu monitoringu, a także jak tworzone są rankingi uniwersyteckie. Rozdział ten został przygotowany przez Krzysztofa Sukiennickiego.

Drugi z rozdziałów dotyczy ogólnopolskiego systemu monitorowania ekonomicznych losów absolwentów. Szczegółowo opisuje on metodykę prowadzenia badania, a także wskaźniki jakie zostały w nim wykorzystane, m.in..: zatrudnienie absolwentów, kontynuację edukacji, poziom bezrobocia, doświadczenie zawodowe, a także jak kształtują się ich zarobki po ukończeniu studiów. Rozdział podsumowuje eksploracyjna analiza danych. Tę część opracował Jakub Skotarek.

Trzeci rozdział obejmuje analizę głównych składowych w badaniu losów absolwentów. Pierwsza część przedstawia teoretyczne podstawy analizy, następnie przeprowadzona jest im-

plementacja jej metod w języku R, a następnie prezentowane są otrzymane wyniki. Rozdział został przygotowany przez Krzysztofa Sukiennickiego.

Ostatni z rozdziałów jest nawiązaniem do wcześniejszego i dotyczy analizy skupień dla badanej grupy absolwentów. W części teoretycznej opisana została idea prowadzenia analizy skupień, miary niepodobieństwa, metody hierarchiczne i optymalizacyjne oraz miary oceny jakości. W dalszej części rozdziału przeprowadzona została implementacja metod analizy skupień w języku R, a na jej podstawie opracowano wyniki końcowe podsumowane ich interpretacją. Rozdział został przygotowany przez Jakuba Skotarka.

Głównym źródłem informacji była literatura z zakresu metodyki badania losów absolwentów, wykorzystania danych administracyjnych, a także analizy głównych składowych oraz analizy skupień. Praca dodatkowo została uzupełniona o dane z projektu ELA, a także informacje publikowane przez szkoły wyższe w Polsce.

Rozdział 1

Pomiar losów absolwentów uczelni wyższych - Krzysztof Sukiennicki

1.1 Problematyka pomiaru losów absolwentów

Pomiary losów zawodowych absolwentów szkół wyższych stanowią jeden z czynników umożliwiających podnoszenie jakości kształcenia zarówno w Polsce, jak i na świecie. Regularny monitoring pozwala na dopasowanie materiałów przekazywanych studentom, mając na celu ich lepsze przygotowanie do warunków panujących na rynku pracy. Badania te stanowią podstawę do zmian w systemie edukacji w wielu krajach Unii Europejskiej, m.in. Hiszpanii, Austrii czy na Węgrzech.

Przed 2011 rokiem zainteresowanie losami zawodowymi absolwentów w Polsce było znikome. Badania te miały charakter dobrowolny i prowadzone były nieregularnie oraz wybiórczo, najczęściej w ramach indywidualnych inicjatyw instytucji edukacyjnych lub projektów badawczych (Bożykowski i in., 2014). W ostatnich latach obserwuje się jednak znaczy wzrost zainteresowania wynikami badań, które mogą przyczynić się do dalszego rozwoju szkolnictwa oraz podnoszenia kwalifikacji zawodowych absolwentów uczelni wyższych. W 2011 roku badania te zostały uregulowane poprzez zmiany w ustawie o szkolnictwie wyższym, która nakazuje uczelniom monitorowanie kariery zawodowej absolwentów, zwłaszcza w pierwszych pięciu latach po ukończeniu studiów.

Mimo iż coraz więcej polskich uczelni przeprowadza badania mające na celu zbadanie losu absolwentów, sposób ich realizacji nie jest w żaden sposób ustandaryzowany. Brak regulacji dotyczący sposobu monitorowania losów zawodowych skutkuje otrzymywaniem wyników

różnej jakości. Monitoring często prowadzony jest w formie nieobowiązkowych ankiet internetowych, co uniemożliwia otrzymanie pełnego przekroju danych, ponadto specyfika badania uniemożliwia uczelniom kontrolowanie jego zasięgu. Dodatkowo nieobowiązkowy charakter badań powoduje, że analizowana próba jest zaniżona, a uzyskane wyniki mogą okazać się niewystarczające i niemiarodajne.

Badania losów absolwentów szkół wyższych pozwalają na monitorowanie ich sytuacji na rynku pracy, a także informują uczelnię o tym, czy zostali oni odpowiednio przygotowani do podjęcia pracy w zawodzie. Poziom kształcenia najłatwiej jest zmierzyć poprzez analizę sytuacji zawodowej absolwentów w pierwszych miesiącach po ukończeniu uczelni. Ponieważ wielokrotnie nie posiadają oni wtedy doświadczenia zawodowego, ich zatrudnienie opiera się przede wszystkim na rodzaju ukończonych studiów, a także wiedzy zdobytej w trakcie ich trwania.

Od 1990 roku obserwuje się znaczny wzrost zainteresowania szkolnictwem wyższym, a także rosnącą liczbę studentów. Wzrost ten spowodował wiele zmian w polskim systemie edukacji, jednak wpłynął również na zwiększenie poziomu bezrobocia wśród osób kończących uczelnie wyższe (Jasiński, Bożykowski, Zając, Chłoń-Domińczak & Żołtak, 2017).

Dzięki badaniom losów absolwentów można zaobserwować trendy panujące na aktualnym rynku pracy analizując liczbę bezrobotnych absolwentów, tych którzy podjęli pracę w zawodzie, a także tych, którzy podjęli pracę w innym obszarze.

Należy pamiętać, że badania te są źródłem informacji nie tylko dla uczelni, ale również samych studentów oraz maturzystów. Otrzymane wyniki odzwierciedlają sytuację na rynku pracy i mogą stanowić podstawę wyboru kierunku kształcenia. Dzięki otrzymanym rezultatom przyszli i obecni studenci mogą zorientować się, czy ukończenie danego kierunku studiów będzie wiązało się w przyszłości z mniejszym ryzykiem bezrobocia lub koniecznością podjęcia pracy w innym obszarze zawodowym, a także na jakie zarobki będą mogli liczyć otrzymując zatrudnienie w dziedzinie związanej z ich wykształceniem.

Od rozpoczęcia obligatoryjnego monitoringu losów ekonomicznych absolwentów minęło już 8 lat, co pozwoliło na znaczne podniesienie jakości badania, jednak aby wyeliminować występujące problemy badawcze, konieczne są kolejne zmiany. Zwiększanie próby badanych absolwentów, uzupełnione o wnioski wyciągane przez uczelnię z poprzednich edycji badania, mogą znacznie poprawić otrzymywane w przyszłości raporty dzięki bardziej rzeczywistych danych o mniejszych odchyleniach.

1.2 Badania losów absolwentów na świecie i w Polsce

1.2.1 Doświadczenia międzynarodowe

Mimo iż pomiary losów absolwentów przeprowadzane są w wielu krajach Unii Europejskiej i na świecie, sposoby przeprowadzania pomiarów nie są jednakowe, co utrudnia porównywanie otrzymanych wyników ze względu na brak tejże standaryzacji. Badania te prowadzone są z wykorzystaniem różnych metod, dodatkowo na różnice wpływają (Bożykowski i in., 2014) m.in.:

- różni organizatorzy badań, którymi mogą być instytucje państwowe lub też komercyjne,
- częstotliwość, w jakiej prowadzone są pomiary,
- procedura badawcza wybrana przez organizatora,
- dobór narzędzi badawczych, w których odpowiedzi mogą być uzyskiwane poprzez ankiety papierowe, telefoniczne, czy też wywiady bezpośrednie,
- czy liczby szkół biorącej udział w badaniu.

Podobnie jak w przypadku badań prowadzonych w Polsce, wyniki zbierane przez zagraniczne uczelnie wyższe służą dopasowaniu poziomu kształcenia studentów do rzeczywistych warunków panujących na rynku pracy danego kraju oraz właściwemu przygotowaniu studenta do podjęcia pracy po ukończeniu studiów.

Jednym z badań prowadzonych w Europie był projekt TRACKIT koordynowany przez European University Association[1] Celem projektu było dostarczenie informacji na temat ścieżek kariery absolwentów w 31 krajach Europy, poprzez analizę poziomu szkolnictwa wyższego w kontekście poszczególnych państw (Gaebel, Hauschildt, Muehleck & Smidt, 2012). TRACKIT skupiał się nie tylko na sposobie i częstotliwości przeprowadzania badań, ale również w jaki sposób otrzymana wiedza była wykorzystywana poprzez szkoły wyższe w Europie.

Raporty przygotowywane w ramach projektu TRACKIT dostarczyły informacji na temat sposobu przeprowadzania badania, instytucji odpowiedzialnych oraz jakości otrzymywanych danych. Jednym z krajów wymienionych w raporcie, w którym proces monitorowania losów absolwentów wyższych był najbardziej rozwinięty, jest Dania. Badanie te przeprowadzane są zarówno przez rządowe agencje, jak i bezpośrednio przez szkoły wyższe. Kiedy pracowano nad

[1] Organizacja skupiająca uczelnie z Europy, której głównym celem jest ujednolicenie procesu kształcenia studentów

projektem TRACKIT, system monitorowania losów ekonomicznych studentów w Polsce nie był jeszcze wystarczająco rozwinięty. Pod względem jakości otrzymywanych wyników, Polska plasowała się daleko za Danią, co wynikało z braku systematyczności w prowadzeniu danych i jego dobrowolnym charakterze. Uległo to jednak znacznej zmianie w kolejnych latach. Aktualnie dane GUS nie stanowią jedynego źródła danych, a zmiany w sposobie prowadzenia badania umożliwiają zebranie informacji na temat absolwentów w różnych okresach od momentu ukończenia studiów. Prestiż badania oraz chęć podnoszenia poziomu oferty edukacyjnej powoduje, iż coraz więcej uczelni nie tylko w Polsce, ale również w Europie jest zainteresowana aktywnym braniem udziału w badaniu.

Jednym z państw o najdłuższej historii prowadzenia badań losów absolwentów są Niemcy. Główną metodą wykorzystywaną w celu pozyskania danych są ankiety prowadzone w formie panelowej, głównie przez same uczelnie. Charakter ich jest dobrowolny, jednak z biegiem lat cieszy się coraz to większą popularnością. Indywidualne sposoby opracowywania danych powodują, że nie istnieje ustandaryzowany system monitorowania kariery absolwentów. Jednym z najbardziej rozwiniętych i popularnych badań jest INCHER koordynowany przez Uniwersytet w Kassel (INCHER-Kassel, n.d.). W swoich badaniach Uniwersytet w Kassel analizuje nie tylko sytuację ekonomiczną studentów i absolwentów, ale również przyczyny nierówności społecznych, które mogą być konsekwencją przerwania studiów i rezygnacji z nauki. Wielobszarowa analiza przeprowadzana w ramach projektu INCHER pozwala na wprowadzenie zmian zarówno w samym programie nauczania, ale również wyeliminowania ryzyka związanego z indywidualnym niedopasowaniem programu lub trybu studiów do potrzeb studentów.

Monitorowanie losów studentów i absolwentów w Wielkiej Brytanii, ma swój początek jeszcze w trakcie studiów. Dane zbierane są zarówno na poziomie państwowym, jak też poprzez poszczególne instytucje. Jedną z największych instytucji monitorującej losy absolwentów jest HESA,czyli Higher Education Statistics Agency, która specjalizuje się w analizach dotyczących szkolnictwa wyższego na terenie Zjednoczonego Królestwa (Higher Education Statistics Agency, 2019). W swoich analizach HESA opiera się nie tylko na danych prezentowanych przez uczelnie wyższe, ale również na informacjach dostarczonych przez absolwenta jeszcze w trakcie studiów. W swoich badaniach HESA wykorzystuje program Data Future, który pozwala na bardziej wydajne gromadzenie danych. Data Future umożliwia zebranie danych w dłuższych okresie, a ich uzupełnienie możliwe jest w dowolnym momencie. Zebrane informacje dzielone są następnie na trzy grupy, zwane okresami referencyjnymi. Bardzo ważnym elementem badania HESA jest

to, że dane są aktualizowane na bieżąco, przez co odzwierciedlają dynamikę zmian badanej grupy. Dzięki regularnej aktualizacji danych zebranych w modelu, raporty generowane przez HESA zaspokajają potrzeby ich odbiorców oraz umożliwiają szybkie reagowanie i wprowadzanie koniecznych zmian w edukacji. Dodatkowo dane te prezentowane są z wykorzystaniem business intelligence i narzędzia Heidi Plus, które pozwala na wizualizację danych dopasowanych do użytkownika. Wyniki prezentowane są w formie pulpitu BI, który ze względu na swoją intuicyjną i przyjazną użytkownikowi formę pozwala na indywidualną pracę z danymi, a także ich dalszą analizę.

1.2.2 Polskie doświadczenia w zakresie monitorowania losów absolwetów

Monitorowanie kariery absolwentów polskich uczelni przed rokiem 2011 było znikome, niewiele uniwersytetów prowadziło badania oraz projekty w tej tematyce. Jedną z uczelni, która już w latach 2007/08 zaczęła generować pierwsze raporty o karierze swoich absolwentów był Uniwersytet Śląski (Uniwersytet Śląski w Katowicach, 2019). Wyniki z badań poruszały tematykę m.in. sytuacji zawodowej absolwentów, ich statusu na rynku pracy, zgodności pracy zawodowej z wybranym kierunkiem studiów. Projekt ten sukcesywnie jest kontynuowany przez biuro karier uniwersytetu (Uniwersytet Śląski w Katowicach, n.d.). Na przestrzeni lat UŚ modyfikuje pytania do respondentów, co daje lepsze odniesienie w kwestii dopasowania do aktualnych warunków na rynku pracy. Z aktualnego badania, które dotyczy absolwentów lat 2016-2017 zostały skonstruowane wnioski mówiące o chęci podejmowania pierwszej pracy już podczas studiów, pierwszych praktykach zawodowych w kraju i zagranicą, wymiarze zatrudnienia oraz czynników poza zarobków mających największy wpływ na wybór miejsca pracy. Pomimo wielu akcji marketingowych, mających na celu udział w przeprowadzonych badaniach liczba respondentów nie zwiększa się i ciągle oscyluje w granicach 11 procent wszystkich absolwentów uczelni. Przez to nie jest możliwe matematyczne określenie parametrów takich jak błąd statystyczny. Przyjęta metoda doboru umożliwia konstruowanie wniosków o dominujących bądź sporadycznie występujących cechach i opiniach w odniesieniu do badanych grup. Wpływ na tak niewielką liczbę absolwentów ma charakter badania, który jest dobrowolny i nieprzystąpienie do niego, nie wiąże się z żadnymi konsekwencjami.

1.2.2.1 Ekonomiczne Losy Absolwentów

W ciągu ostatnich lat w Polsce obserwuje się wzrost zainteresowania wynikami pomiarów losów zawodowych absolwentów uczelni wyższych. Monitoring ten umożliwia opisany w kolejnych rozdziałach projekt ELA, czyli analiza Ekonomicznych Losów Absolwentów stworzona oraz stale rozwijana przez Ministerstwo Nauki i Szkolnictwa Wyższego. Wyniki pomiarów są nie tylko podstawą do podnoszenia poziomu kształcenia w ramach konkretnych uczelni, stanowią również informację dla kandydatów na studia, którzy na ich podstawie mogą podjąć decyzję dotyczącą wyboru kierunku studiów. Od 2014 roku system ELA poddał analizie sytuację na rynku pracy ponad 1,1 mln absolwentów polskich uczelni.

Powszechny i bezpłatny dostęp do systemu ELA pozwala na uzyskanie informacji dotyczących:

- rankingów kierunków, dziedzin i obszarów,
- zestawień dla poszczególnych uczelni, jak i całego kraju,
- raporty o poziomie bezrobocia, czasie poszukiwania pracy, wynagrodzeniach czy geograficznym położeniu absolwentów.

Wyniki publikowane po przeprowadzeniu badań ELA wykorzystywane są przez wiele instytucji, a następnie prezentowane w poszczególnych rankingach. Jednym z nich wspomniany jest wcześniej „Ranking Perspektywy", jednak wyniki publikowane są nie tylko przez opiniotwórcze portale czy gazety, ale również przez uczelnie, które w ten sposób mogą zaprezentować jakość oferowanego przez siebie kształcenia, które stanowi kluczową informację dla przyszłych studentów.

1.2.2.2 Kadry dla Gospodarki

Jedną z uczelni prezentujących powyższe wyniki był Uniwersytet Ekonomiczny w Poznaniu. Dane udostępniane były na stronie internetowej uczelni, która wykorzystywała je do oceny przekazywanych treści i ich dopasowania do warunków panujących na rynku pracy. W razie potrzeby treści te mogły być również modyfikowane, aby zapewnić lepszą jakość kształcenia (Uniwersytet Ekonomiczny w Poznaniu, 2019). Ze względu na zróżnicowanie ścieżek kariery absolwentów, badania te były realizowane dla poszczególnych wydziałów, aby móc lepiej zobrazować sytuację ekonomiczną studentów poszczególnych kierunków.

Zakres badania obejmował wszystkich absolwentów studiów stacjonarnych i miał charakter panelowy. W badaniu losów absolwentów UEP studenci ankietowani byli czterokrotnie (Ławrynowicz & Michoń, 2011):

- *badanie wstępne* – kwestionariusz papierowy wypełniany w chwili ukończenia studiów, w którym absolwenci deklarowali, czy chcą uczestniczyć w kolejnych badaniach prowadzonych przez uczelnię. Udział w badaniu wstępnym był obowiązkowy, jednak studenci mieli prawo odmówienia udziału poprzez pisemne oświadczenie odmowy. Powyższy sposób rozpoczęcia badań ekonomicznych losów absolwentów pozwolił na uzyskanie większej grupy potencjalnych respondentów i w związku z tym bardziej miarodajnych wyników w porównaniu do uczelni, w których badania wstępne nie były obowiązkowe.

- *badanie pełne* – w formie kwestionariusza elektronicznego, które przeprowadzane było po roku, trzech oraz pięciu latach od ukończenia studiów. W badaniu tym wykorzystano specjalistyczne narzędzie IBM SPSS Data Collection, które pozwoliło na indywidualną analizę ścieżki kariery studenta przy każdorazowym wypełnieniu kwestionariusza. Wyniki otrzymane w ramach kwestionariusza były spersonalizowane i dodatkowo uzupełnione o informacje na temat studiów z okresu studiów, co pozwoliło na uzyskanie pełnych i wiarygodnych wyników stanowiących podstawę w kontekście zmian w sposobie nauczania na UEP.

Badania ekonomicznych losów absolwentów ze względu na ich nieustandaryzowany charakter prowadzone są z wykorzystaniem różnych metodologii. Jednym z przykładów może być monitoring Szkoły Głównej Handlowej, w której badania te od początku mają charakter dobrowolny. Badania te prowadzone są od 2010 z wykorzystaniem środków własnych uczelni. (Macioł, Miniewicz & Moskalewicz-Ziółkowska, 2013) Wyniki badania prowadzonego przez SGH publikowane są w formie raportu z dominującą charakterystyką opisową otrzymanych wyników, uzupełnione o wykresy. Szczegółowo opisane są obszary badania, takie jak ocena programu studiów i warunków studiowania, czy status zawodowy absolwentów bezpośrednio po ukończeniu studiów. Ważnym aspektem badania jest również samoocena sytuacji studentów, która umożliwia uczelni zidentyfikowanie luki edukacyjnej oraz wyeliminowanie jej w kolejnych latach poprzez wprowadzenie kluczowych zmian w sposobie nauczania. Kolejnym przykładem wykorzystania badania jest monitoring prowadzony przez Uniwersytet Ekonomiczny we Wrocławiu, który od 2011 zbiera informacje na temat losów absolwentów w formie dobrowolnych

kwestionariuszy. W badaniach UEW wyróżnia grupy interesariuszy w celu otrzymania pełnych i wysokiej jakości danych, które mogą być analizowane dla poszczególnych zbiorowości (Pałys, Pałys, Prokopowicz & Sykuła, 2013). Wyniki badania przeprowadzonego przez UEW podzielone są ze względu na stopień oraz tryb ukończonych studiów, a także na wydział, który ukończyli respondenci. Otrzymane wyniki prezentowane są w formie graficznej oraz opisowej i obrazują liczbę absolwentów poszczególnych kierunków, którzy wzięli udział w badaniu. Aby ułatwić analizę rezultatów dane te ponadto są podzielone na poszczególne grupy wiekowe. Wszelkiego rodzaju wyniki z badań oraz analiz statystycznych służą temu, aby móc zaprezentować je w odpowiedniej formie, która umożliwiałaby ich interpretacje na podstawie, której można formułować wnioski. Jedną z takich form są rankingi, które zostaną omówione w następnym rozdziale.

1.3 Rankingi szkół wyższych

Aby uzyskać właściwe pomiary potrzebne do oceny szkół wyższych konieczne jest uwzględnienie ich specyfiki oraz charakteru. Ze względu na poziom studiów rankingi uczelni dzieli się na (Prawo o szkolnictwie wyższym: Dz. U. 2005 Nr 164 poz. 1365, 2005):

- I stopnia - formę kształcenia, na którą są przyjmowani kandydaci posiadający świadectwo dojrzałości, kończącą się uzyskaniem kwalifikacji pierwszego stopnia,
- II stopnia - formę kształcenia, na którą są przyjmowani kandydaci posiadający co najmniej kwalifikacje pierwszego stopnia, kończącą się uzyskaniem kwalifikacji drugiego stopnia,
- jednolite studia magisterskie - formę kształcenia, na którą są przyjmowani kandydaci posiadający świadectwo dojrzałości, kończącą się uzyskaniem kwalifikacji drugiego stopnia,

W rankingach uwzględniania jest również forma studiów, do których zalicza się studia (Prawo o szkolnictwie wyższym: Dz. U. 2005 Nr 164 poz. 1365, 2005):

- stacjonarne - dla grup absolwentów studiów stacjonarnych rozpoczętych po wprowadzeniu w Polsce Krajowych Ram Kwalifikacji,
- niestacjonarne - formę studiów wyższych, inną niż studia stacjonarne, wskazaną przez senat uczelni,
- inne.

Do stworzenia rankingów (Edusfera, 2019), opisanych w dalszej części pracy, bierze się pod uwagę wiele obszarów, takich jak:

- prestiż, czyli reputację danej uczelni, w której analizowana jest opinia na temat jej absolwentów, a także pracowników naukowych. Stanowi to 30 procent oceny.
- siła naukowa, będąca pomiarem zdolności uczelni do przeprowadzania badań naukowych i jej zasobów intelektualnych. Uzyskane wyniki stanowią 40 procent oceny rankingu.
- warunki studiowania, wyceniane na 20 procent w ogólnej ocenie, które często stanowią kluczowy czynnik przy wyborze danej szkoły wyższej. Do warunków studiowania zalicza się: stopień wykwalifikowania kadry naukowej, zasoby biblioteczne, czy też wspieranie studentów przez uniwersyteckie biura karier.
- umiędzynarodowienie studiów jest ostatnim obszarem, stanowiącym 10 procent całej oceny. Brana jest w nim pod uwaga liczba przyjeżdżających oraz wyjeżdżających studentów, liczba kierunków i przedmiotów prowadzona w językach obcych, a także wielokulturowość środowiska uniwersyteckiego.

Podział wyników badań ze względu na poziom oraz formę studiów pozwala uzyskać miarodajne wyniki, dzięki którym szkoły wyższe w precyzyjny sposób są w stanie określić, czy aktualne formy nauczania są dopasowane do aktualnej sytuacji ekonomicznej na rynku, a także czy odpowiadają one na potrzeby absolwentów przygotowując ich do wybranych zawodów. Zawężona grupa wyników ułatwia przeprowadzenie tejże analizy, umożliwia obserwację trendów obecnych na rynku pracy oraz pozwala przewidzieć jego przyszłą strukturę.

Rezultaty badań sytuacji absolwentów stanowią również podstawę do tworzenia rankingów uczelni wyższych. Jednym z zestawień jest „Ranking Perspektywy", który co roku opracowuje taką listę w oparciu o typ uczelni oraz konkretną grupę kryteriów. Celem rankingu jest wskazanie szkół wyższych o najwyższej jakości kształcenia, pokazanie aktualnego stanu polskiego szkolnictwa wyższego a także wskazanie kierunku, w jakim powinny rozwijać się polskie ośrodki akademickie. Badanie to opiera się na analizie określonych kryteriów oraz ich wagi, co pozwala na otrzymanie następujących wskaźników (Perspektywy, 2019):

- liczbę publikacji naukowych w stosunku do ogólnej liczby nauczycieli akademickich,
- liczbę cytowań w stosunku do liczby publikacji, z wykluczeniem autocytowań,

- liczbę cytowań otrzymanych przez publikację do średniej liczby cytowań otrzymanych przez podobne publikacje naukowe,
- ekonomicznych losów absolwentów, czyli pomiaru zatrudnienia absolwentów oraz wysokości ich zarobków,
- jakości przyjętych na studia, czyli pomiaru wyników egzaminów maturalnych osób, które podjęły studia na I roku studiów stacjonarnych.

Pomimo, iż ostatnie dwa wskaźniki wykorzystywane są w tworzeniu rankingu dopiero od kilku lat, stanowią one jego ważny element, ponieważ bazują na danych zewnętrznych, a nie tych pozyskanych bezpośrednio z uczelni wyższych. Potwierdza to sens prowadzenia regularnych badań ekonomicznych losów absolwentów oraz monitorowania zmian w obszarze polskiego kształcenia wyższego.

1.4 Podsumowanie

W ostatnich latach obserwuje się znaczny wzrost zainteresowania monitorowaniem ekonomicznych losów absolwentów, co przyczynia się do rozwoju badania a także pozwala na uzyskanie bardziej miarodajnych wyników o wyższej jakości. Regularna analiza otrzymanych raportów pozwala uczelniom wprowadzać zmiany zarówno w programie nauczania, jak i samej metodzie pozyskiwania danych w przyszłych badaniach. W następnym rozdziale omówione zostanie badanie, w którym przedstawiono jego metodykę oraz wykorzystywane wskaźniki, które posłużą do dalszych analiz.

Rozdział 2

Ogólnopolski System Monitorowania Ekonomicznych Losów Absolwentów - Jakub Skotarek

2.1 Szczegółowy opis metodyki badania

System monitoringu Ekonomicznych Losów Absolwentów ELA stworzony przez polskich badaczy jest jednym z najnowocześniejszych systemów generujących automatyczne raporty-przewodniki dla każdego kierunku studiów wszystkich uczelni w kraju, który w ciągu ostatnich lat poddał analizie sytuację ekonomiczną ponad 1,1 mln absolwentów (MNiSW, 2019a). Projekt ELA został stworzony przez Państwowy Instytut Badawczy na zlecenie Ministerstwa Nauki i Szkolnictwa Wyższego wykorzystując przy tym unikalnego oprogramowania oraz metodologii stworzonej przez naukowców z Uniwersytetu Warszawskiego (MNiSW, 2019a).

Raporty generowane przez system ELA opierają się na danych administracyjnych pochodzących ze źródeł Zakładu Ubezpieczeń Społecznych oraz systemu POL-on, które ze względu na państwowy charakter pozwalają na (MNiSW, 2019b):

- przeprowadzenie analizy porównawczej sytuacji ekonomicznej absolwentów w różnych okresach,
- wyeliminowanie błędów wynikających z indywidualnej odpowiedzi respondentów poprzez bazowanie na danych statystycznych,
- interpretację otrzymanych wyników dzięki precyzyjnie skonstruowanym wskaźnikom,

- brak konieczności przeprowadzania badań bezpośrednich, tj. ankiet czy kwestionariuszy,
- objęcie badaniem wyselekcjonowanej grupy absolwentów ze względu na wybraną metodę badawczą,
- analizę dynamiki zmian zachodzących na rynku pracy.

Mimo korzyści, jakie niesie ze sobą korzystanie z powyższych źródeł danych administracyjnych, istnieją pewne ograniczenia, do których należy zaliczyć (MNiSW, 2019b):

- administracyjny charakter danych, nieuwzględniający specyfiki badań, w których są one wykorzystywane,
- ograniczenie analizy do posiadanych danych, nie uwzględniających opinii absolwentów czy pracodawców,
- nierejestrowanie umowy o dzieło oraz umowy zlecenia w danych zbieranych przez ZUS, a także umów zawieranych za granicą oraz przez osoby prowadzące własną działalność gospodarczą.
- brak ocen, a także innych szczegółowych informacji na temat absolwenta w danych dostarczanych przez system POL-on.

Informacje analizowane przez system ELA pozbawione są danych osobowych, przez co prywatność badanych osób pozostaje nienaruszona. Wyniki zebrane dla zbiorowości poniżej 10 osób nie są uwzględniane w publikowanych raportach, co uniemożliwia ich powiązanie z osobami objętymi badaniem. Rezultaty analiz prezentowane są w formie raportów, które dzielą się na następujące kategorie tematyczne (MNiSW, 2019b):

- poszukiwanie pracy i bezrobocie,
- wynagrodzenia,
- praca, a dalsze studia,
- geograficzne zróżnicowanie losów absolwentów.

Raporty publikowane są ponadto na trzech poziomach analizy dla (MNiSW, 2019b):

- programu studiów,

- szkoły wyższej,
- całego kraju.

Pierwszy z poziomów analizy, a mianowicie program studiów stanowi podstawę badania, ponieważ pozwala na zrealizowanie jego celu, jakim jest dostarczenie informacji na temat danego kierunku studiów oraz oczekiwanej sytuacji ekonomicznej po jego ukończeniu. Dodatkowo zawężony obszar analizy pozwala na wyodrębnienie grup absolwentów odmiennych kierunków studiów, co niemożliwe jest w przypadku dalszych poziomów analizy, czyli danej uczelni czy całego kraju.

Każdy z raportów publikowanych w danej kategorii tematycznej dzieli się ze względu na rodzaj studiów (MNiSW, 2019b):

- studia pierwszego stopnia,
- studia drugiego stopnia,
- studia jednolite magisterskie.

Pierwszy z raportów, a mianowicie poszukiwanie pracy i bezrobocie skupia się na średnim czasie od ukończenia studiów do podjęcia pracy przez badanych absolwentów, a także analizie stabilności zatrudnienia. W raporcie uwzględniony jest współczynnik bezrobocia wyliczany dla badanego okresu i dotyczy on wszystkich absolwentów, którzy w tym czasie co najmniej raz nie znajdowali się w stanie aktywnego zatrudnienia. Sprawozdanie to zawiera również informacje na temat ryzyka bezrobocia, doświadczenia zawodowego absolwentów w oparciu o rodzaj umowy, a także średnią liczbę pracodawców.

Drugi z publikowanych raportów dotyczy wynagrodzeń uzyskiwanych przez absolwentów w pierwszych latach po ukończeniu studiów. Sprawozdanie uwzględnia wszystkie formy zatrudnienia z wyłączeniem samozatrudnienia, ponieważ informacje te nie znajdują się w bazach danych ZUS. Średnie wynagrodzenia obliczane są dla całego badanego okresu, a dodatkowo uzupełnione są o informację na temat średnich wynagrodzeń w Polsce. Ponieważ powyższe informacje mają dość ogólny charakter, raport poszerzony jest o szczegółowe informacje, w których otrzymane wyniki podawane są dla pięciu grup w zależności od kwoty otrzymywanego wynagrodzenia.

Kolejne sprawozdanie dotyczy pracy oraz dalszych studiów i informuje jaka liczba absolwentów kontynuowała naukę na innym kierunku studiów po uzyskaniu dyplomu, a także średni

okres w jakim studiowali. Raport ten pokazuje również jaka część absolwentów studiów I stopnia zdecydowała się kontynuować naukę oraz ilu z nich ukończyło kolejne studia.

Ostatni z raportów dotyczy geograficznego zróżnicowania losów absolwentów i dostarcza informacji na temat ich aktualnego miejsca zamieszkania oraz jego odległości od ukończonej uczelni. Dane te prezentowane są w zależności od wielkości miejscowości zamieszkania i uwzględniają wybrane cechy społeczno-demograficzne.

2.2 Opis wskaźników wykorzystywanych w raporcie ELA

W raportach generowanych przez system ELA wykorzystuje się wiele wskaźników dla absolwentów zatrudnionych na umowę o pracę oraz w służbach mundurowych, które umożliwiają zbadanie (MNiSW, 2019b):

- liczby i procentu absolwentów występujących w rejestrach ZUS,
- procentu absolwentów kontynuujących naukę po uzyskaniu dyplomu, którzy:
  - studiowali na innych programach studiów,
  - podjęli studia II stopnia,
  - podjęli i ukończyli studia II stopnia,
  - ukończyli kolejne studia,
- czasu poszukiwania pracy, w tym:
  - czasu poszukiwania pracy na umowę o pracę,
- liczby absolwentów zatrudnionych w służbach mundurowych:
  - procent osób zatrudnionych w służbach mundurowych,
  - czas poszukiwania pracy w służbach mundurowych,
- poziomu bezrobocia:
  - procent absolwentów, którzy mieli doświadczenie bycia bezrobotnym,
  - ryzyko bezrobocia,
  - względny wskaźnik bezrobocia,

- doświadczenia zawodowego dla absolwentów, którzy mieli doświadczenie:
  - w jakiejkolwiek pracy,
  - w pracy na umowę o pracę,
  - samozatrudnienia,
- procentu przepracowanych miesięcy:
  - w jakiejkolwiek formie,
  - na umowę o pracę,
  - w ramach samozatrudnienia,
- średniej miesięcznej liczby równoczesnych pracodawców,
- liczby przypadków zakończenia pracy etatowej,
- wynagrodzenia:
  - średnie miesięczne wynagrodzenie,
  - średnie miesięczne wynagrodzenie z tytułu umowy o pracę,
  - względny wskaźnik zarobków.

2.2.1 Zatrudnienie wśród absolwentów

Pierwszy wskaźnik wykorzystuje dane ZUS dotyczące ogólnego zatrudnienia absolwentów. Należy przy tym pamiętać, iż nie uwzględniają one absolwentów, którzy nie podjęli jeszcze żadnej aktywności na rynku pracy, pracowali w formach, które nie są rejestrowane w ZUS, przebywają za granicą lub są zarejestrowani w KRUS.

2.2.2 Kontynuacja studiów przez absolwentów

Kolejna grupa wskaźników informuje o tym, jaki procent absolwentów zdecydował się na kontynuowanie nauki na innych kierunkach lub na studiach II stopnia i uwzględniają one również studia rozpoczęte przed uzyskaniem dyplomu przez absolwenta. Wskaźniki te mają kluczowe znaczenie w badaniu losów ekonomicznych absolwentów, ponieważ dostarczają one informacji na temat ich pozycji na rynku pracy, a także ich aktywności gospodarczej.

Czas poszukiwania pracy, w tym na umowę o pracę informuję o średnim okresie, w jakim absolwenci podjęli pierwszą pracę od momentu uzyskania dyplomu z wykluczeniem osób, które nie podjęły jakiejkolwiek pracy. Wskaźnik ten analizowany dla poszczególnych kierunków studiów, może wskazać które z nich cieszą się największą popularnością i pozwalają na najszybsze znalezienie pracy.

2.2.3 Absolwenci w służbach mundurowych

Osobną grupą wskaźników wykorzystywanych w raportach ELA są indeksy dotyczące absolwentów zatrudnionych w służbach mundurowych. Dostarczają one informacji na temat procentu absolwentów, którzy zostali zatrudnieni w tym obszarze, a także jaki był średni czas poszukiwania przez nich pracy od momentu uzyskania dyplomu.

2.2.4 Wskaźnik bezrobocia wśród absolwentów

Kolejną, ważną grupą wskaźników są te dotyczące poziomu bezrobocia. Informują one ilu absolwentów było zarejestrowanych jako bezrobotnych w czasie trwania badania, nawet jeśli aktywność zawodowa była zawieszona w krótkim okresie. Wyróżnia się tutaj wskaźnik ryzyka bezrobocia, który obrazuje prawdopodobieństwo bezrobocia dla określonej grupy studentów. Wyliczany jest on na podstawie średniego okresu bezrobocia dla tej grupy i po ilu miesiącach zawieszonej aktywności zawodowej podjęli oni kolejną pracę. W tej grupie wykorzystywany jest również względny wskaźnik bezrobocia, który bada indywidualne prawdopodobieństwo wystąpienia bezrobocia w oparciu o dane średniej stopy bezrobocia w analizowanych powiatach.

2.2.5 Doświadczenie zawodowe absolwentów

Doświadczenie zawodowe dla absolwentów badane jest poprzez poziom składek wpływających do ZUS z tytułu jakiekolwiek zatrudnienia, umowy o pracę, a także z tytułu samozatrudnienia. Wskaźniki te informują o typie zatrudnienia, a uzupełnione o kolejną grupę wskaźników dotyczących liczby przepracowanych miesięcy dostarczają szczegółowych informacji na temat stabilności podjętego zatrudnienia i długotrwałości pracy absolwentów.

W raportach uwzględniany jest również wskaźnik średniej miesięcznej liczby równoczesnych pracodawców, który uwzględnia zapłatę za wykonaną pracę przy jednoczesnym zatrudnieniu na umowę o pracę. Analiza wskaźnika wymaga indywidualnej interpretacji, ponieważ może świad-

czyć zarówno o korzystnej jak i niekorzystnej sytuacji na rynku pracy. Podjęcie dodatkowego zatrudnienia może być związane z trudną sytuacją ekonomiczną absolwenta, ale może również być jego dobrowolnym wyborem związanym z chęcią kumulacji zarobków, czy też odpowiedzią na duże zapotrzebowanie rynku na pracowników z danymi umiejętnościami.

Wskaźnik liczby przypadków zakończenia pracy etatowej informuje o rocznej średniej, kiedy absolwent zdecydował się odejść z pracy. Rezygnacja z zatrudnienia może mieć różny charakter i wynikać zarówno z decyzji pracodawcy jak i absolwenta, który może zdecydować się na dobrowolną zmianę pracodawcy lub też typu czy obszaru zatrudnienia.

2.2.6 Miesięczne wynagrodzenie absolwentów

Ostatnia grupa wskaźników dotyczy wynagrodzenia i pozwala na analizę średnich miesięcznych zarobków, z uwzględnieniem tych uzyskanych z tytułu umowy o pracę. Wskaźniki te wyliczane są na podstawie łącznych zarobków, a także liczby przepracowanych miesięcy. Dane te uzupełnione są dodatkowo o względny wskaźnik wynagrodzenia, który pozwala na indywidualną interpretację wyników z wyszczególnieniem poszczególnych powiatów.

2.3 Eksploracyjna analiza danych

Eksploracyjna analiza danych (EDA – exploratory data analysis) stanowi początek każdego badania. Jej celem jest poznanie i zrozumienie analizowanych procesów oraz generowanych przez nią danych. (Morzy, 2013)

Do przeprowadzenia analiz posłużono się raportem badania ELA, przedstawiającym ekonomiczne losy absolwentów, w ramach ukończonej uczelni w 2016 roku. Okresem badawczym był rok po uzyskaniu dyplomu. Selekcja danych odbyła się dla absolwentów studiów licencjackich i inżynierskich I stopnia. Dodatkowo przedmiotem prowadzonych analiz były uczelnie, które w swojej ofercie edukacyjnej posiadają profil ekonomiczny. Dzięki temu ograniczeniu wykluczone zostały Uniwersytety Medyczne, Akademie Sztuk Pięknych, Akademie Wychowania Fizycznego oraz grupa Wyższych Szkół i innych uczelni, na których nie są prowadzone kierunki kształcenia ekonomicznego. Oprócz powyższego narzucony został również warunek liczby absolwentów na kierunkach ekonomicznych, który powinien być większy niż 200. Ograniczenie to zostało nałożone ze względu na stosunkowy mały procent studentów będących na profilu ekonomicznym w stosunku do całkowitej ich liczby na uczelniach. Dzięki temu możliwe było wyróż-

nienie szkół wyższych o charakterze ekonomicznym, zachowując przy tym omawiany wcześniej średni stosunek na poziomie ponad 20%. Na podstawie powyższych ograniczeń wyselekcjonowane zostało 118 rekordów.

Do analizy głównych składowych oraz analizy skupień wybrano 6 ze 183 dostępnych wskaźników badania ELA.

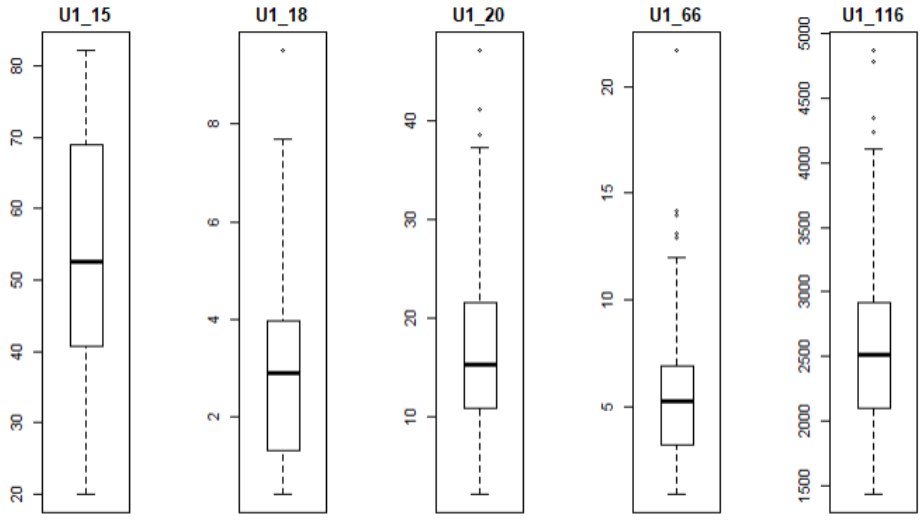
Tabela 2.1. Wskaźniki wyselekcjonowane do dalszych analiz

Nazwa zmiennej	Etykieta
U1_2	Nazwa szkoły wyższej
U1_15	Procent absolwentów, którzy po uzyskaniu dyplomu podjęli studia II stopnia
U1_18	Średni czas (w miesiącach) od uzyskania dyplomu do podjęcia pierwszej pracy po uzyskaniu dyplomu
U1_20	Procent absolwentów, którzy mieli doświadczenie bycia bezrobotnym po uzyskaniu dyplomu
U1_66	Procent absolwentów, którzy mieli doświadczenie samozatrudnienia po uzyskaniu dyplomu
U1_116	Średnie miesięczne wynagrodzenie ze wszystkich źródeł po uzyskaniu dyplomu

Źródło: Opracowanie własne na podstawie danych https://ela.nauka.gov.pl/.

Pierwszym krokiem eksploracyjnej analizy danych jest sprawdzenie czy w zbiorze znajdują się braki danych dla każdego ze wskaźników, jednak procedura analizy braków danych udowodniła, że nasz zbiór jest kompletny, dzięki czemu nie ma podstaw do wypełniania brakujących wartości. Oprócz kontroli braków danych w zbiorze warto zwrócić również uwagę na obserwacje odstające, które mogły zostać omyłkowo zawyżone lub zaniżone przy wprowadzaniu danych. W tym celu wykorzystane zostały wykresy pudełkowe, które przedstawiają wartości minimalne oraz maksymalne dla każdego wskaźnika, kwantyle wraz z medianą oraz obserwacje odstające.

Według załączonego wykresu 2.1, zmienna U1_15 tj. podjęcie studiów drugiego stopnia, nie posiada obserwacji odstających, natomiast podjęcie pierwszej pracy (U1_18), doświadczenie bycia bezrobotnym (U1_20), doświadczenie samozatrudnienia(U1_66) oraz średnie miesięczne wynagrodzenie(U1_116) takie obserwacje zawierają. Są to jednak wartości, które w rzeczywistości mają prawo zaistnieć i nie był to błąd wprowadzania danych. Ponadto ze względu na charakterystyki uczelni, które reprezentują powyższe obserwacje, stwierdzone zostało iż ich występowanie jest całkowicie uzasadnione.



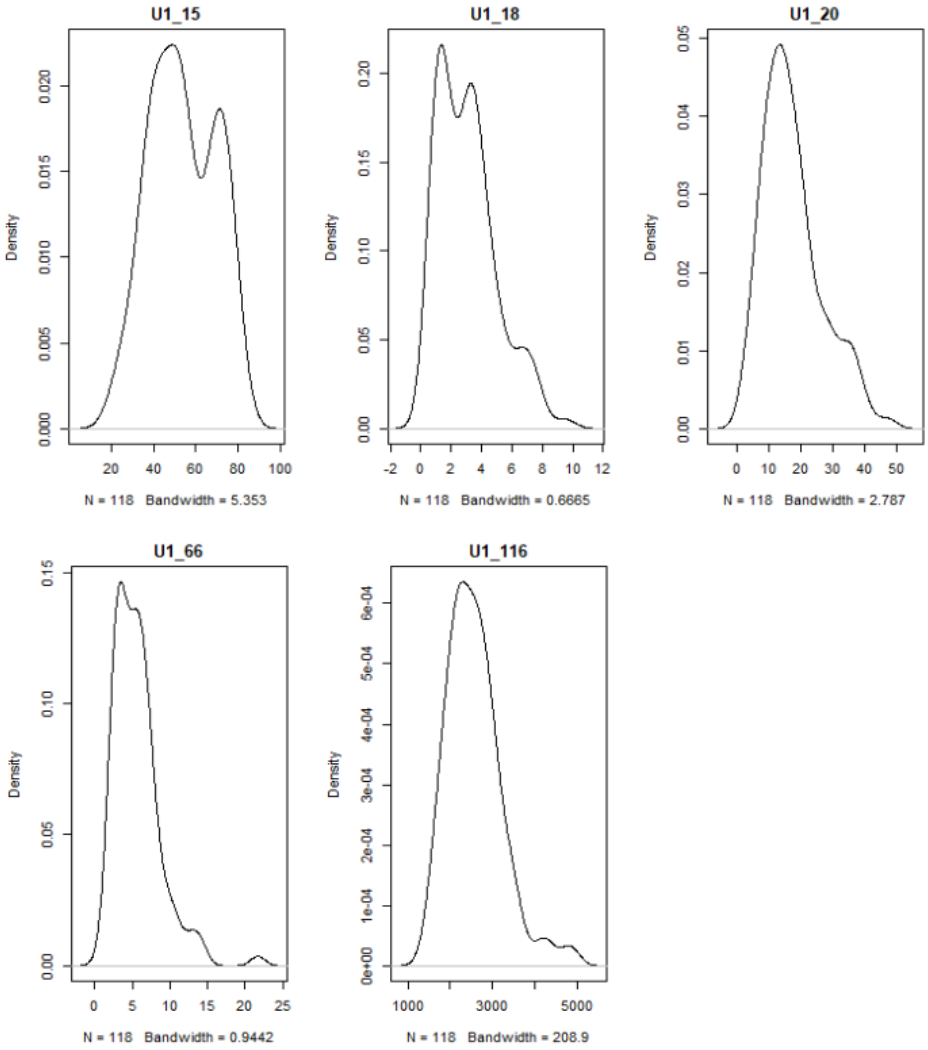
Rysunek 2.1. Wykres obserwacji odstających w zbiorze

Źródło: Opracowanie własne.

Kolejnym etapem jest sprawdzenie jakim rozkładem prawdopodobieństwa charakteryzują się poszczególne wskaźniki. W tym celu policzone zostały parametry opisowe dla każdego z nich oraz stworzone zostały wykresy skośności-kurtozy omawianych wskaźników (Delignette-Muller & Dutang, 2015).

W celu stwierdzenia jakim rozkładem cechują się wskaźniki na wykresie 2.2 posłużono się kryterium informacyjnym Akaikego, który służy do porównania dopasowania danych empirycznych do teoretycznych między różnymi rozkładami prawdopodobieństw. Wobec tego wskaźnikami, które mają rozkłady zbliżone do rozkładów normalnych są U1_15 - podjęcie studiów drugiego stopnia oraz U1_116 - średnie miesięczne wynagrodzenie. Pozostałe charakteryzują się rozkładem Weibulla. Dzięki temu jesteśmy wstanie dowiedzieć się jakie jest prawdopodobieństwo wystąpienia konkretnej wartości w zbiorze. By dokładniej zrozumieć dane, warto zwrócić uwagę na występujące w zbiorze zależności, dlatego istotnym jest by przeanalizować współczynnik korelacji między zmiennymi.

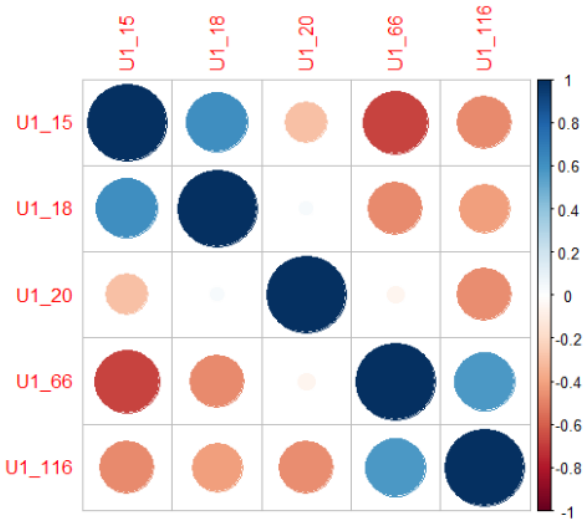
Wykres 2.3 ilustruje współczynnik korelacji Pearsona między poszczególnymi wskaźnikami. Niebieski kolor oznacza dodatnią zależność liniową między dwoma zmiennymi, natomiast czerwony ujemną. W związku z tym, dodatnią umiarkowaną zależnością charakteryzują się zmienne U1_15, czyli podjęcie studiów drugiego stopnia i średni czas do podjęcia pierwszej pracy (U1_18). Oznacza to, że wraz ze wzrostem procentu absolwentów, którzy po uzyskaniu dy-



Rysunek 2.2. Wykresy rozkładów empirycznych wskaźników

Źródło: Opracowanie własne.

plomu podjęli studia drugiego stopnia umiarkowanie rośnie wartość średniego czasu od uzyskania dyplomu do podjęcia pierwszej pracy. Drugą parą zmiennych, które są umiarkowanie, dodatnio skorelowane są U1_66 - doświadczenie samozatrudnienia oraz średnie miesięczne wynagrodzenie (U1_116). W tym przypadku, wraz ze wzrostem procentu absolwentów, którzy mieli doświadczenie samozatrudnienia po uzyskaniu dyplomu jednocześnie rośnie umiarkowanie średnie miesięczne wynagrodzenie ze wszystkich źródeł po uzyskaniu dyplomu. Z kolei między procentem absolwentów, którzy kontynuowali naukę na drugim stopniu (U1_15) a procentem absolwentów, którzy mieli doświadczenie samozatrudnienia (U1_66) istnieje silna ujemna zależność liniowa. Według tej korelacji, im więcej studentów decydowało się na edukację na drugim stopniu tym mniejsza z nich część decydowała się na samozatrudnienie. Oprócz powyż-



Rysunek 2.3. Wykres korelacji między wskaźnikami

Źródło: Opracowanie własne.

szego, występują również inne ujemnie skorelowanie zmienne takie jak:

- podjęcie studiów drugiego stopnia i średnie miesięczne wynagrodzenie (U1_15 - U1_116)
- podjęcie studiów drugiego stopnia i doświadczenie bycia bezrobotnym (U1_15 - U1_20)
- średni czas do podjęcia pierwszej pracy i doświadczenie samozatrudnienia (U1_18 - U1_66)
- średni czas do podjęcia pierwszej pracy i średnie miesięczne wynagrodzenie (U1_18 - U1_116)
- doświadczenie bycia bezrobotnym i średnie miesięczne wynagrodzenie (U1_20 - U1_116)

2.4 Podsumowanie

System monitoringu Ekonomicznych Losów Absolwentów, jest nieocenionym źródłem danych, który umożliwia przeprowadzenie analizy sytuacji ekonomicznej absolwentów. Dzięki raportom opartym na danych administracyjnych pochodzących z Zakładu Ubezpieczeń Społecznych oraz systemu POL-on, dane które są w nich zawarte są wiarygodne oraz przedstawiają realną sytuację absolwentów szkół wyższych. Umożliwił również określenie przedmiotu badań oraz selekcję wskaźników, które są istotne z punktu widzenia dalszej analizy.

Stworzony zbiór danych zawiera uczelnie o charakterze ekonomicznym, co oznacza, że znajdują się w nim tylko i wyłącznie szkoły wyższe, które w swojej ofercie edukacyjnej posiadają profil ekonomiczny. Ich dobór polegał na ograniczeniu liczby absolwentów na kierunkach ekonomicznych dla konkretnych uczelni do dwustu. Przełożyło się to na średni 20% stosunek ich liczby do liczby wszystkich absolwentów danej uczelni, co świadczy o jej ekonomicznym charakterze. Z kolei wskaźniki zostały wyselekcjonowane w taki sposób, by opisywały najistotniejsze, ekonomiczne aspekty losów absolwentów oraz by interpretacja wyników analiz była intuicyjna i zrozumiała. Przeprowadzona eksploracyjna analiza danych udowodniła również zależność między wybranymi wskaźnikami, w tym dodatnią korelację niektórych zmiennych oraz silnie ujemną korelację między procentem absolwentów, którzy kontynuowali naukę na drugim stopniu, a procentem absolwentów, którzy mieli doświadczenie samozatrudnienia. Oprócz powyższego, zbiór nie posiada żadnych braków danych oraz nie występują nieuzasadnione obserwacje odstające, które mogłyby negatywnie wpłynąć na wiarygodność analiz.

Tak przygotowany zbiór danych posłuży nam do przeprowadzenia analizy głównych składowych, który na celu ma zredukowanie liczby wymiarów, zachowując przy tym jak największą część informacji zbioru podstawowego. Istota tej metody oraz cały proces został omówiony w rozdziale trzecim.

Rozdział 3

Analiza głównych składowych w badaniu losów absolwentów - Krzysztof Sukiennicki

3.1 Teoretyczne podstawy analizy głównych składowych

Bardzo często dane statystyczne, które analizujemy mają charakter wielowymiarowy. Oznacza to, że są one opisywane dużą, często silnie skorelowaną liczbą zmiennych. Interpretacja staje się wówczas bardziej skomplikowana, a ze względu na złożoność zagadnienia nie można ograniczyć się do jednej lub dwóch cech. Nawet jeśli obserwacja może być prowadzona tylko dla jednej z nich to przeważnie dalsza analiza prowadzi do sytuacji wielowymiarowej. Aby ją uprościć i jednocześnie sprawić, że będzie ona zrozumiała, stosuje się statystyczne metody wielowymiarowej analizy (Balicki, 2013). Jedną z najpopularniejszych jest analiza głównych składowych, która analizuje strukturę zależności zmiennych oraz sprowadza dużą liczbę badanych cech do znacznie mniejszych, niezależnych tzw. składowych głównych.

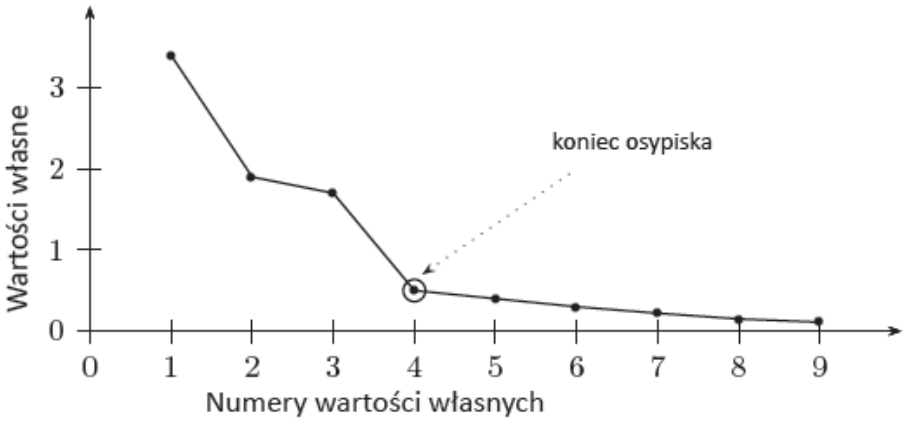
Przyjmuje się, że początki teoretycznych podstaw analizy głównych składowych pochodzą od K. Pearsona (1901). Uważał on, że w wielowymiarowej elipsoidzie w p-wymiarowej przestrzeni pomiarów, najlepszymi pomiarami są te, które korespondują z jej pionowymi osiami. Jednak dalszy rozwój tej metody zawdzięczamy H. Hotellingowi (1933), który wykorzystywał ją w swoich pracach dotyczących analizy testów osiągnięć szkolnych (Czopek, 2013). Jego technika wyznaczania głównych składowych jest wykorzystywana aż do dzisiaj.

Głównym powodem więc, dla którego wykorzystywana jest ta metoda jest redukcja wymiarowości złożonego zjawiska. Polega ona na utworzeniu całkowicie nowego zbioru zmiennych z istniejących już zmiennych wejściowych, który w mniejszym lub większym stopniu mógłby za-

stąpić zbiór pierwotny. Po transformacji otrzymano taką samą liczbę głównych składowych co liczba zmiennych w zbiorze początkowym, jednak ostatecznie wybrano kilka z nich, które są najbardziej istotne pod względem dalszej analizy. Pożądana jest jak najmniejsza liczba głównych składowych, opisująca złożoną strukturę zależności, przy zachowaniu jak najwyższej zmienności danych (Czopek, 2013).

Istnieje kilka kryteriów, którymi należy się sugerować przy ich doborze, m.in (Czopek, 2013):

- procent wyjaśnionej wariancji – liczbę składowych głównych określa na podstawie zawartej w nich wariancji zmiennych pierwotnych. Przyjmuje się, że ich wariancja powinna odzwierciedlać w ponad 80 procentach zmienne pierwotne,
- kryterium Kaisera – pod uwagę powinny być brane tylko takie składowe główne, których wartość własna jest równa lub bliska dowolnej, wystandaryzowanej zmiennej pierwotnej,
- wykres osypiska – liczba składowych głównych przedstawia wykres tempa spadku procentu wyjaśnionej wariancji.



Rysunek 3.1. Wykres osypiska

Źródło: *https://pqstat.pl/?mod_f=test_pca*

Punkt zaznaczony na wykresie 3.1 jest końcem osypiska, co oznacza, że proces stabilizuje się i linia malejąca przechodzi w poziomą. Na tej podstawie możemy stwierdzić ile głównych składowych powinniśmy uwzględnić w naszej analizie.

Jeżeli badane zmienne oznaczymy jako $X_j(j = 1, ..., p)$, gdzie X jest macierzą danych, dla której zakładamy, że wszystkie wiersze są liniowo niezależne, a p oznacza liczbę zmiennych to

nowe, nieobserwowane zmienne Y_l, które są liniowymi kombinacjami zmiennych wejściowych, możemy zapisać w postaci układu równań (Balicki, 2013):

$$\begin{aligned} Y_1 &= w_{11}X_1 + w_{21}X_2 + ... + w_{p1}X_p \\ Y_2 &= w_{12}X_1 + w_{22}X_2 + ... + w_{p2}X_p \\ &... \\ Y_m &= w_{1m}X_1 + w_{2m}X_2 + ... + w_{pm}X_p \end{aligned} \tag{3.1}$$

Zapis układu w postaci uogólnionej:

$$Y_l = w_{1l}X_1 + w_{2l}X_2 + ... + w_{pl}X_p = \sum_{j=1}^{p} w_{jl}X_j \quad \text{dla } l = 1,...,m = p. \tag{3.2}$$

Zapis w formie macierzowej:

$$y = W^{'}x. \tag{3.3}$$

Gdzie W jest (p×m) macierzą współczynników:

$$w = (w_{jl}) = \begin{bmatrix} w_{11} & w_{12} & ... & w_{1m} \\ w_{21} & w_{22} & ... & w_{2m} \\ ... & ... & ... & ... \\ w_{p1} & w_{p2} & ... & w_{pm} \end{bmatrix}. \tag{3.4}$$

Natomiast $\mathbf{x}$ jest kolumnowym wektorem badanych zmiennych, a $\mathbf{y}$ kolumnowym wektorem zmiennych transformowanych.

Utworzone zmienne Y_l w układzue równań 3.1 nazywamy głównymi składowymi zmiennych X_j, natomiast współczynniki w_{jl} w macierzy są to tak zwane ładunki składowe. Zazwyczaj liczba głównych składowych jest równa liczbie zmiennych wyjściowych, jednak zdarza się w drodze wyjątku, że ich liczba jest mniejsza z powodu redukcji wymiarów problemu. Ogólnie rzecz biorąc zmienne te stanowią mieszaninę cech obserwowanych, a każda z nich opisuje pewną część informacji wnoszonej przez cechy wyjściowe. Ze względu na ich charakter, mówi się, że są to cechy syntetyczne, ponieważ zastępują one dużą liczbę zmiennych, często powiązanych ze

sobą, w mniejszą liczbę zorganizowanych zmiennych w formie wskaźników (w_{jl}). Dodatkowo składowe Y_l są ortogonalne, co oznacza, że są nieskorelowane między sobą, unormowane, czyli suma kwadratów ładunków składowych danej kombinacji jest równa jeden oraz suma ich wariancji jest taka sama jak ogólna wariancja zmiennych X_j (Balicki, 2013).

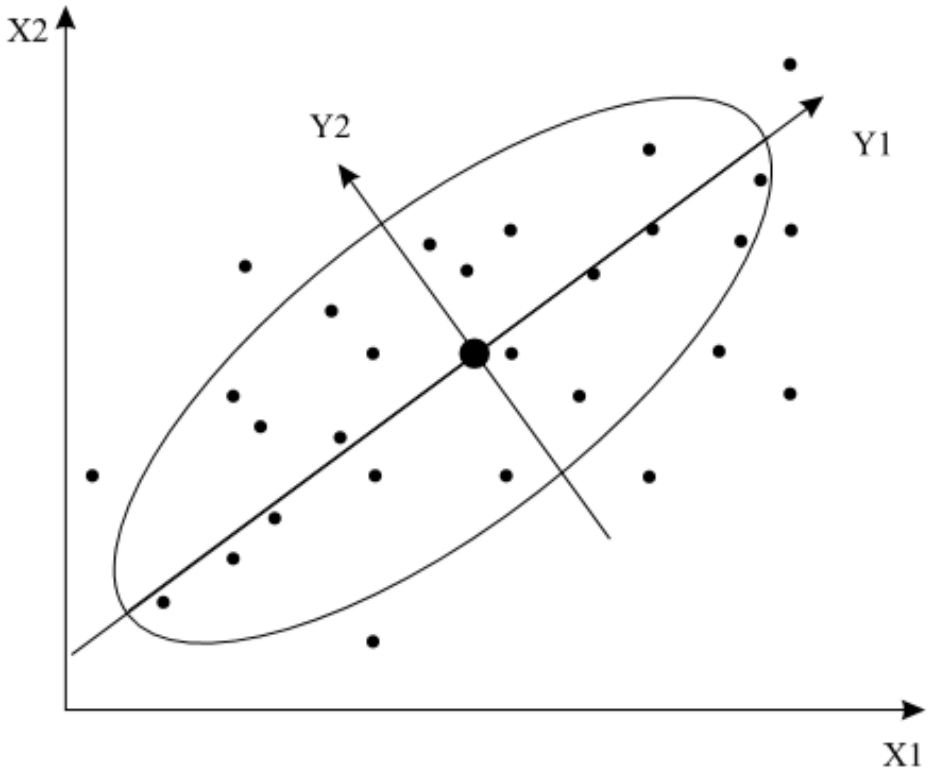
Punktem wyjściowym jest macierz kowariancji lub macierz korelacji, jednak w określonych przypadkach wynikiem może być też macierz odległości. Zawarte w nich informacje są istotą wyznaczania składowych głównych. Wartości takiej macierzy, chcemy przedstawić w postaci jak najmniejszej liczby składowych głównych, która wyjaśniałaby jak największą część zmienności. Oznacza to, że jeżeli mamy całkowitą zmienność danego wektora zmiennych i każda składowa główna wyjaśnia jakąś jego część to naszym celem jest doprowadzenia do takiego podziału, że jedna lub parę pierwszych składowych wyjaśniałoby maksymalną część całkowitej zmienności. Jeżeli znajdziemy liczbę głównych składowych, która wyjaśnia prawie całą zmienność to możemy zaniechać dalszego ich wyznaczania, ponieważ będę one mało istotne z punktu widzenia zmienności. Dlatego też składowe porządkujemy według ich udziału w wyjaśnianiu zmienności (Balicki, 2013).

W ujęciu geometrycznym, interpretacja głównych składowych polega na opisaniu rozproszenia układu n punktów w p-wymiarowej przestrzeni cech. Możemy tego dokonać wprowadzając nowe układy liniowych, ortogonalnych współrzędnych, względem których wariancje danych punktów są uporządkowane malejąco. Transformacja polega więc na rzutowaniu obserwacji na utworzone osie główne elipsoidy. W przypadku dwumiarowym każdy punkt w elipsoidalnym zbiorze ma dwie składowe X_1 i X_2. Jednak zwracając uwagę na kierunek osi wielkiej elipsoidy możemy zaobserwować kierunek, w którym dane są rozproszone. Kierunek ten jest pierwszą główną składową Y_1. Natomiast druga główna składowa jest ortogonalna do pierwszej, co oznacza, że biegnie pod kątem 90 stopni względem niej i jej zróżnicowanie jest mniejsze (Balicki, 2013).

Ogólnie rzecz biorąc transformowanie polega na przesunięciu środka układu do punktu średnich badanych zmiennych oraz rotacji wokół tego środa, dzięki czemu otrzymujemy współrzędne Y_1 oraz Y_2 głównych składowych.

Oprócz tworzenia zmiennych składowych w celu redukcji wymiarowości, wyróżnia się inne zastosowania analizy głównych składowych, m.in. (Balicki, 2013):

- redukcja danych poprzez usunięcie części zmiennych pierwotnych, które nie są istotne, ponieważ nie wnoszą dodatkowej informacji. Dzięki temu ograniczamy duży zbiór danych



Rysunek 3.2. Graficzna ilustracja wyodrębniania głównych składowych

Źródło: A. Balicki, Statystyczna analiza wielowymiarowa i jej zastosowania społeczno-ekonomiczne, s. 77, rys. 2.1

i w dalszej analizie wykorzystujemy mniejszą ich liczbę. Od podstawowego celu różni się tym, że nie tworzymy tutaj nowych zmiennych lecz usuwamy zmienne z wejściowego zbioru,

- badanie korelacji między zmiennymi wyspecyfikowanego zbioru;
- badanie grupowania się i porządkowanie jednostek w przestrzeni p-wymiarowej;
- badanie złożonych zjawisk i przyporządkowanie odpowiednich wag dla każdej zmiennej,
- ortogonalizacja oszacowań regresyjnych, czyli przekształcenie układu liniowo niezależnych wektorów oszacowań w układ wektorów ortogonalnych.

3.2 Implementacja metod analizy głównych składowych w języku R

Pakiet R jest bardzo dobrym narzędziem do szeroko pojętych analiz statystycznych, ze względu na ogromną listę funkcji umożliwiającą ich prowadzenie. Dodatkowo możliwość wizualizacji danych oraz prezentacji wynik w postaci różnych modyfikowanych wykresów, grafów i ogólnie pojętych rysunków skłoniła nas do skorzystania z tego właśnie pakietu. W niniejszym podrozdziale zostaną przedstawione implementacje najpopularniejszych metod analizy głównych składowych oraz na podstawie porównania, wybierzemy jedną z najefektywniejszych. Posłuży nam do tego zbiór danych opisany w podrozdziale 2.3. Metody, które zostaną przedstawione różnią się sposobem wyznaczania składowych głównych. Jako pierwsze opisane zostaną funkcje prcomp oraz princomp z pakietu podstawowego stats.

W przypadku funkcji prcomp() nowe zmienne wyznaczane są z użyciem dekompozycji na wartości osobliwe SVD(ang. Singular Value Decomposition) (Biecek & Trajkowski, 2011), co oznacza, że funkcja ta dokonuje rozkładu macierzy na iloczyn trzech specyficznym macierzy. Jest ona szczególnie zalecana ze względu na dużą dokładność numeryczną. Implementacja metody prcomp() w pakiecie R wygląda następująco:

```
dataset_num_pca <- prcomp(dataset_num, center = TRUE, scale. = TRUE)
```

Program 3.1. Implementacja metody prcomp()

Pierwszym parametrem funkcji `prcomp( )` jest zbiór danych, który zawiera tylko wartości ciągłe. Ograniczenie to wynika z faktu, że gdy mamy do czynienia ze zmiennymi dyskretnymi o więcej niż dwóch kategoriach, zakodowanymi przy pomocy zestawu zero-jedynkowych zmiennych wskaźnikowych, stosowanie analizy głównych składowych nie jest poprawnym podejściem (Górniak, 1998). Drugi parametr `center = TRUE` określa, czy zmienne powinny być przesunięte i wyśrodkowane na wartość zero. Polega ona na odjęciu od każdego elementu kolumny odpowiedniej wartości średniej. Powoduje to usunięcie stałych elementów, które nic nie wnoszą do wiedzy o zróżnicowaniu danych. Trzeci parametr `scale. = TRUE` jest bardzo istotny w ujęciu analizy głównych składowych. Skalowanie jest zwykle zalecane, aby uniknąć zdominowania wyników przez zmienne o dużych wartościach w stosunku do pozostałych. (Gramacki & Gramacki, 2009)

Funkcja prcomp() umożliwia również dostęp do poszczególnych obiektów:

Tabela 3.1. Komponenty metody prcomp(stats)

$rotation	Wektor odchyleń standardowych dla obserwacji. Kolejne zmienne odpowiadają odchyleniom standardowym liczonym dla kolejnych składowych głównych. Macierz obrotu przekształcająca oryginalny układ współrzędnych w nowy układ współrzędnych.
$center	Wektor wartości wykorzystanych przy centrowaniu obserwacji.
$scale	Wektor wartości wykorzystanych przy skalowaniu obserwacji.
$x	Macierz współrzędnych kolejnych obserwacji w nowym układzie współrzędnych, macierz ta ma identyczne wymiary co oryginalny zbiór zmiennych

Źródło: Opracowanie własne na podstawie
https://www.rdocumentation.org/packages/stats/versions/3.6.0/topics/prcomp

Drugą metodą z pakietu stats jest funkcja `princomp`. W teorii właściwości wyznaczonych składowych głównych nie będą się różniły od składowych wyznaczonych metodą `prcomp`, jednak zdarza się, że w określonych sytuacjach wyniki mogą się różnić. Zasadniczą różnicą tych funkcji jest sposób ich wyznaczania. W tym przypadku składowe główne wyznaczane są poprzez wektory własne macierzy kowariancji pomiędzy zmiennymi, a nie z użyciem dekompozycji na wartości osobliwe SVD jak w przypadku funkcji `prcomp`. Implementacja metody `princomp` w pakiecie R wygląda następująco:

```
dataset_num_copy.pca <- princomp(dataset_num_copy, cor = TRUE)
```

Program 3.2. Implementacja metody princomp()

Tak jak w przypadku funkcji `prcomp()` pierwszym argumentem funkcji `princomp()` jest zbiór danych numerycznych. Drugim argumentem jest wartość logiczna wskazująca, że do obliczenia wartości głównych składowych ma zostać wykorzystana macierz korelacji. Możemy posłużyć się tutaj macierzą korelacji lub kowariancji. W tej metodzie centrowanie jest wykonywane automatycznie. W przypadku użycia macierzy korelacji jest w pewnym sensie tożsame z wykorzystaniem skalowania danych.

Kolejną metodą jest metoda `PCA()` z pakietu FactoMineR, który został stworzony w celu wielowymiarowej analizy danych eksploracyjnych. Zaletą tej metody jest fakt, że nie musimy standaryzować danych przed analizą, ponieważ metoda `PCA()` robi to automatycznie. Implementacja metody `PCA()` w pakiecie R wygląda następująco:

```
res.pca <- PCA(dataset_num, scale.unit = TRUE, ncp = 5, graph = TRUE)
```

Program 3.3. Implementacja metody PCA()

Tak jak we wcześniejszych metodach pierwszym argumentem funkcji `PCA()` jest zbiór da-

nych numerycznych. Kolejny, `scale.unit` skaluje wariancję danych. Argument `ncp` określa liczbę wymiarów, które mają być zachowane w końcowym wyniku. Mamy również możliwość wyświetlenia grafu za pomocą parametru `graph`. Wynikiem funkcji `PCA()` jest lista, zawierająca m.in. takie komponenty jak:

Tabela 3.2. Komponenty metody PCA(FactoMineR)

Nazwa	Opis
$eig	wartości własne
$var	wyniki dla zmiennych
varcoord	współrzędne zmiennych
varcor	korelacje zmiennych - wymiary
varcos2	cos2 dla zmiennych
varcontrib	składowe zmiennych
$ind	wyniki dla cech
indcoord	współrzędne dla cech
indcos2	cos2 dla cech
indcontrib	składowe cech

Źródło: Opracowanie własne na podstawie *http://factominer.free.fr/index.html*

Kolejna analiza metody głównych składowych może być wdrożona z pomocą pakietu pcurve. Za twórcę uważa się Chrisa Walsha, który przeniósł dany pakiet z biblioteki S-Plus Glenna De'Atha. Implementacja opiera się na pojedynczym lub dominującym gradiencie. Wykorzystywana jest przeważnie dla danych ze świata ekologii. Poniżej implementacja danej metody w środowisku RStudio.

```
dataset_num.pca <- pca(dataset_num, cent = FALSE, scle = TRUE)
```

Program 3.4. Implementacja metody pca() pakietu pcurve

Argumenty:

- mat - macierz numeryczna
- cent - wartość logiczna odnosząca się do centralnego argumentu w skali
- scle - wartość logiczna odnosząca się do argumentu skala w celu skalowania

Wynikiem funkcji `pca` pakietu pcurve jest lista, zawierająca następujące komponenty:

Tabela 3.3. Komponenty metody pca(pcurve)

$psc	macierz ładunków głównych składowych
$d	macierz zawierająca wartość pojedynczą (wartość własną) każdego komponentu głównego na jego przekątnej
$v	macierz wektorów własnych

Źródło: Opracowanie własne na podstawie
https://www.rdocumentation.org/packages/mixOmics/versions/6.3.2/topics/pca

3.3 Wyniki analizy głównych składowych

Do przeprowadzania analizy głównych składowych wykorzystana została metoda `prcomp()` pakietu `stats` ze względu na charakter danych, które zawierają tylko wartości numeryczne z wyjątkiem kolumny z nazwami szkół wyższych. W celu lepszej prezentacji wyników wykorzystany został pakiet `factoextra`, dzięki któremu w łatwy i efektywny sposób można wydobyć i zwizualizować wyniki analizy danych wielowymiarowych. Przed przystąpieniem do analizy głównych składowych opuszczona została wspomniana wcześniej kolumna z nazwami szkół wyższych z podstawowego zbioru i stworzony został nowy podzbiór z poszczególnymi wskaźnikami. W tym celu wykorzystana została metoda `select_if()` z pakietu `dplyr`, by wyodrębnić tylko zmienne numeryczne, co jest jednoznaczne z pominięciem jedynej występującej w zbiorze zmiennej znakowej.

Tak przygotowany zbiór, można wykorzystać do przeprowadzenia analizy głównych składowych.

```
dataset_num.pca <- prcomp(dataset_num, center = TRUE, scale. = TRUE)
```

Program 3.5. Implementacja metody prcomp()

Zmienna `dataset_num.pca` zawiera wyniki funkcji `prcomp()` wywołanej na przygotowanym wcześniej zbiorze. Argumenty logiczne `center` oraz `scale.` zostały ustalone na wartość TRUE, co oznacza, że dane w zbiorze zostały wyśrodkowane oraz wyskalowane, by uniknąć zdominowania wyników przez zmienne o dużych wartościach. Za pomocą funkcji `summary()` można uzyskać wyniki dotyczące m.in. procentu wyjaśnionej wariancji przez każdą składową:

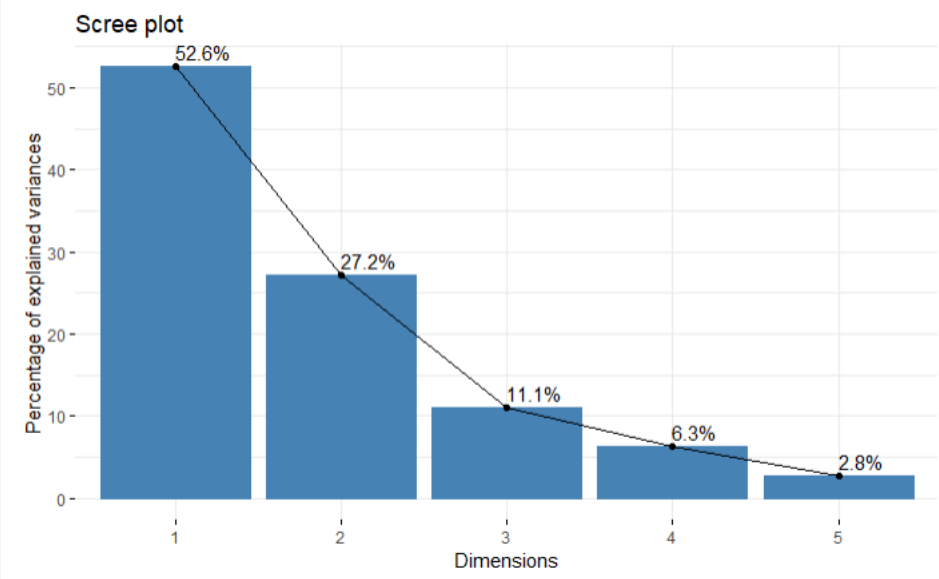
Funkcja zwraca pięć głównych składowych, co jest równe liczbie zmiennych w zbiorze danych. Pierwsza główna składowa wyjaśnia w przybliżeniu 53% wariancji, natomiast druga 27%. Funkcja `summary()` zwraca również wartość odchylenia standardowego każdej składowej głównej oraz skumulowaną wariancję.

Tabela 3.4. Procent wyjaśnionej wariancji przez główne składowe

	PC1	PC2	PC3	PC4	PC5
Standard deviation	1.6213	1.1658	0.7457	0.5623	0.3742
Proportion of Variance	0.5257	0.2718	0.1112	0.0632	0.0280
Cumulative Proportion	0.5257	0.7976	0.9088	0.9720	1.0000

Źródło: Opracowanie własne.

W celu zwizualizowania wyników możemy skorzystać z funkcji `fviz_eig()` z pakietu `factoextra`:



Rysunek 3.3. Procent wyjaśnionej wariancji przez główne składowe

Źródło: Opracowanie własne.

Wykres 3.3 przedstawia wartości własne wszystkich pięciu składowych głównych. Są one uporządkowane w kolejności malejącej. Dzięki zobrazowaniu wyników łatwo można dostrzec proporcje wyjaśnionej wariancji przez każdy wymiar oraz stwierdzić jakąś liczbę głównych składowych powinno się przyjąć do dalszych analiz.

Aby uzyskać szczegółowe wyniki analizy PCA dla zmiennych, koniecznie jest skorzystanie z funkcji `get_pca_var()`:

```
var.pca <- get_pca_var(dataset_num.pca)
```

Program 3.6. Implementacja metody get_pca_var()

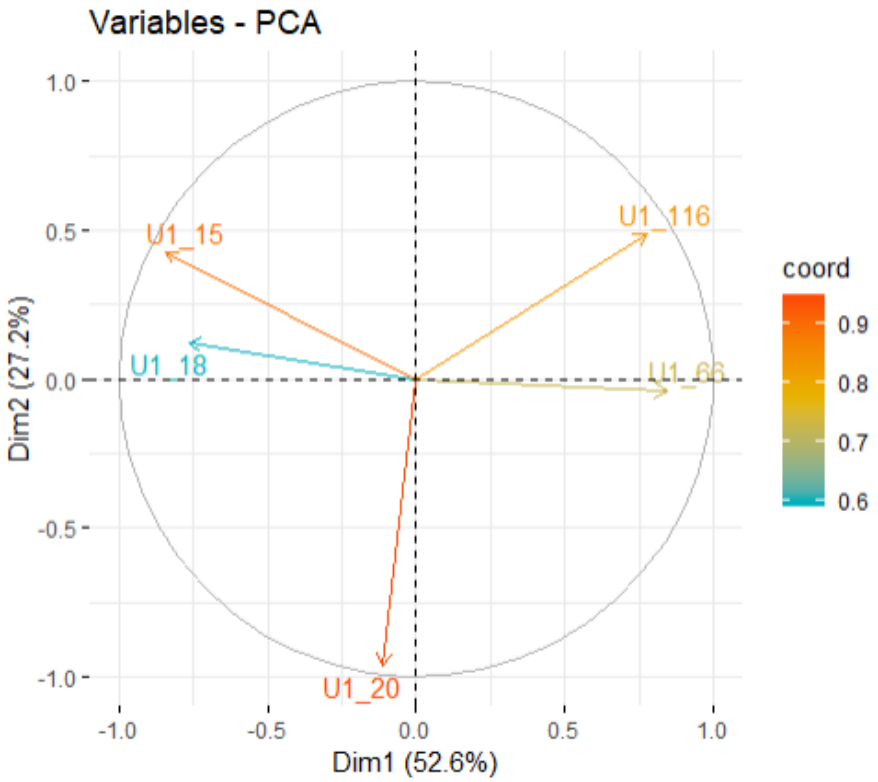
Teraz za pomocą komponentu `$coord` obiektu `var.pca` mamy możliwość m.in. wyświetlenia korelacji między zmiennymi a głównymi składowymi:

Tabela 3.5. Współczynniki korelacji między zmiennymi a głównymi składowymi

	Dim.1	Dim.2
U1_15	-0.84	0.42
U1_18	-0.76	0.12
U1_20	-0.11	-0.96
U1_66	0.85	-0.04
U1_116	0.78	0.49

Źródło: Opracowanie własne.

Wyświetlone zostały współczynniki korelacji pierwszych dwóch składowych głównych, ponieważ posłużą nam one do opisywania relacji między wymiarami. W celu ich wizualizacji możemy stworzyć koło korelacji za pomocą funkcji `fviz_pca_var()` pakietu `factoextra`:



Rysunek 3.4. Koło korelacji

Źródło: Opracowanie własne.

Przedstawione koło 3.4 opisuje korelację między zmiennymi a dwoma pierwszymi głównymi składowymi. Moduł wektora wskazuje w jakim stopniu dana zmienna przyczyniła się do wyjaśnienia wariancji w danych składowych. Zmienne znajdujące się z dala od środka są dobrze reprezentowane na mapie czynnikowej. Zmienne pozytywnie skorelowane między sobą

są grupowane razem, natomiast skorelowane negatywnie znajdują się po przeciwnych stronach ćwiartek względem pozytywnie skorelowanych. Zastosowanie koloru jest opcjonalne i opisuje te same właściwości co moduły wektorów. Według przedstawionego wykresu zmienne, które są skorelowane z wymiarem pierwszym mają moduł wektora skierowany w poziomie. Wskaźnikami, które są silnie dodatnie skorelowane z tym wymiarem są U1_66, tj. procent absolwentów, którzy mieli doświadczenie samozatrudnienia po uzyskaniu dyplomu oraz U1_116, czyli średnie miesięczne wynagrodzenie ze wszystkich źródeł po uzyskaniu dyplomu. Występują również zmienne silnie ujemnie skorelowane takie jak procent absolwentów, którzy po uzyskaniu dyplomu podjęli studia drugiego stopnia (U1_15) i średni czas (w miesiącach) od uzyskania dyplomu do podjęcia pierwszej pracy po uzyskaniu dyplomu (U1_18). Z kolei doświadczenie bycia bezrobotnym (U1_20), co widać na załączonym kole korelacji, jest słabo ujemnie skorelowana. W przypadku wymiaru drugiego, zmienne, które są z nim skorelowane mają moduły wektorów skierowane w pionie. Widać wyraźnie, że wskaźnikiem, który jest ujemnie bardzo silnie skorelowany jest U1_20, tj. procent absolwentów, którzy mieli doświadczenie bycia bezrobotnym po uzyskaniu dyplomu. Jest zwrot znajduje się prawie na brzegu koła, co oznacza, że jego wartość jest bliska -1. Pozostałe wskaźniki są umiarkowanie skorelowane z wymiarem drugim.

Aby dokładnie zrozumieć relację między zmiennymi a głównymi składowymi należy przeanalizować procent udziału poszczególnych zmiennych w wyjaśnianiu wariancji głównych składowych. Dostęp do tych informacji dostarcza nam komponent `$contrib`: Udział zmiennych

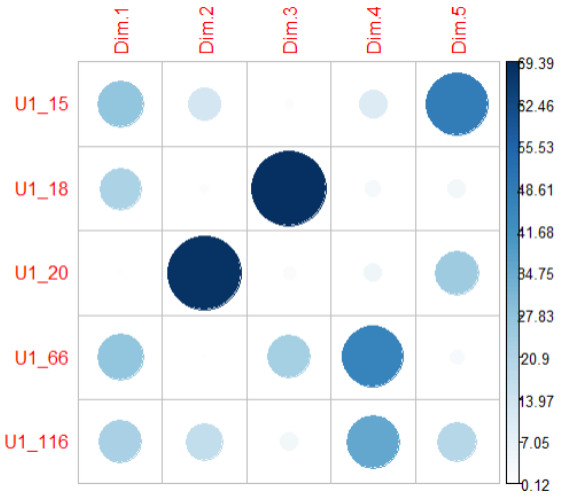
Tabela 3.6. Tabela udziału zmiennych w objaśnianiu głównych składowych

	Dim.1	Dim.2
U1_15	27.15	13.27
U1_18	22.20	1.10
U1_20	0.50	68.18
U1_66	27.30	0.12
U1_116	22.85	17.33

Źródło: Opracowanie własne.

w objaśnianiu zmienności w danej składowej głównej jest wyrażany w procentach. Funkcja `corrplot()` z pakietu o takiej samie nazwie umożliwia zilustrowanie omawianej relacji:

Suma udziału zmiennych na wykresie 3.5 w objaśnianiu zmienności danej głównej składowej jest równa jeden. Każdy punkt pokazuje jaki procent stanowi zmienna w wyjaśnianiu wa-



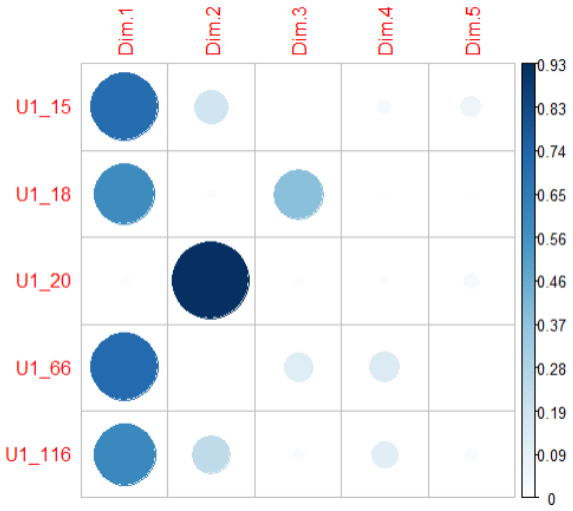
Rysunek 3.5. Wykres udziału zmiennych w objaśnianiu zmienności głównych składowych

Źródło: Opracowanie własne.

riancji danego wymiaru. Te, które są najbardziej skorelowane z pierwszą z nich są najistotniejsze w wyjaśnianiu zmienności w zbiorze danych. Zmienne, które nie są skorelowane z żadną składową główną lub są skorelowane z ostatnimi wymiarami, są zmiennymi o niskim wkładzie i mogą zostać usunięte.

Kolejną interesującą miarą jest jakość reprezentacji zmiennych na mapie czynnikowej, co jest oznaczane jako `cos2`. Jest to współczynnik korelacji między zmienną a główną składową podniesiony do potęgi. Za pomocą funkcji `corrplot()` możemy również zwizualizować wartość `cos2` zmiennych dla wszystkich głównych składowych:

Suma wyjaśnionej wariancji na wykresie 3.6 dla danej zmiennej `cos2` we wszystkich głównych składowych jest równa jeden. Wysoka wartość `cos2` wskazuje na dobrą reprezentację zmiennej w danym wymiarze i w związku z tym zostaje umieszczona blisko obwodu okręgu korelacji. W przypadku, gdy suma `cos2` dla dwóch pierwszych wymiarów jest równa jeden, wówczas zmienna znajduję się na samej krawędzi koła. Z kolei niski `cos2` wskazuje, że nie jest ona idealnie reprezentowana przez główne składowe oraz że znajduje się blisko środka okręgu. Jak widać zmienna, która nie została wyjaśniona przez pierwszą główną składową, czyli procent absolwentów, którzy mieli doświadczenie bycia bezrobotnym po uzyskaniu dyplomu(U1_20), została wyjaśnione przez drugą. Podobna sytuacja wystąpiła w przypadku zmiennej odpowiadającej za średni czas od uzyskania dyplomu do podjęcia pierwszej (U1_18), której duża część została wyjaśniona dopiero w trzeciej składowej głównej.



Rysunek 3.6. Wykres jakości zmiennych

Źródło: Opracowanie własne.

3.4 Podsumowanie

Analiza głównych składowych umożliwiła zredukowanie wielowymiarowości danych. Implementacja tej analizy ograniczyła liczbę zmiennych, generując dwie główne składowe, które w 80% opisują rozproszenie danych w zbiorze podstawowym. Przy wyborze liczby wymiarów w celach dalszej analizy grupowania kierowaliśmy się skumulowanym procentem wyjaśnionej wariancji analizowanych zmiennych oraz kryterium Cattella. Nie uwzględniliśmy kryterium Kaisera ze względu na stosunkowo niewystarczającą liczbę zmiennych, ponieważ przy rozpatrywanej przez nas ilości wskaźników (tj. 5) istniała szansa wyodrębnienia zbyt małej liczby czynników (Czopek, 2013).

Na podstawie kryterium Cattella, osypiskiem czynnikowym, czyli elementem po którym występuje łagodny spadek wartości własnych, będą wartości o numerach od 3 do 5. Z tego względu czynniki, które znajdują się po lewej stronie wykresu tj. numer 1 oraz 2, są naszymi głównymi składowymi. Następnym kryterium doboru liczby czynników jest skumulowany procent wariancji. Zakładając, że powinien być on w granicach 80% na podstawie książki „Statystyczna analiza wielowymiarowa i jej zastosowania społeczno-ekonomiczne" A. Balicki, wybieramy dwie składowe główne, które w 79,76% wyjaśniają zbiór pierwotny.

Pierwsza z nich opisuje w przybliżeniu 53% zmienne pierwotne. Zawiera ona zmienne skorelowane dodatnio takie jak: procent absolwentów, którzy mieli doświadczenie samozatrudnienia po uzyskaniu dyplomu (U1_66), średnie miesięczne wynagrodzenie ze wszystkich źródeł

po uzyskaniu dyplomu (U1_116). Występują również zmienne skorelowane ujemnie: procent absolwentów, którzy po uzyskaniu dyplomu podjęli studia II stopnia (U1_15), średni czas (w miesiącach) od uzyskania dyplomu do podjęcia pierwszej pracy po uzyskaniu dyplomu (U1_18), procent absolwentów, którzy mieli doświadczenie bycia bezrobotnym po uzyskaniu dyplomu (U1_20). Korelacja dodatnie oraz ujemne są ze sobą powiązane, co oznacza, że wraz ze wzrostem tych pierwszych maleją drugie.

Druga składowa uzupełnia w 27% wariancję całego zbioru podstawowego. Wyjaśnia przede wszystkim procent absolwentów, którzy mieli doświadczenie bycia bezrobotnym po uzyskaniu dyplomu (U1_20). Zmienna ta została opisana w wymiarze pierwszym zaledwie w 5%, jednak w wymiarze drugim jest reprezentowana już w 93%, stanowiąc zarazem 68% jej wyjaśnionej wariancji.

Przeprowadzona analiza umożliwia nam wizualizację obserwacji ze zbioru danych w dwuwymiarowej przestrzeni, której odpowiadają dwie pierwsze główne składowe. Jest to istotne z punktu widzenia analizy skupień, która na celu ma identyfikacje uczelni o podobnych charakterystykach i zestawieniu ich w ramach pojedynczych grup. Metodyka analizy skupień oraz wykorzystanie w niej wyników analizy głównych składowych została opisana w rozdziale czwartym.

Rozdział 4

Analiza skupień w badaniu losów absolwentów - Jakub Skotarek

4.1 Teoretyczne podstawy analizy skupień

4.1.1 Idea analizy skupień

Grupowanie, szeregowanie oraz klasyfikowanie obiektów było znane od wieków. W przeszłości metody te stosowano, aby móc odróżnić pewne specyficzne cechy danych zjawisk od innych. Zastosowanie ich w praktyce pozwoliło usprawnić oraz ujednolicić wiele różnych obszarów naukowych m.in. poprzez rozwój języka oraz gramatyki, stworzenie tablicy Mendelejewa w dziedzinie chemii oraz zbadanie historycznej zmienności danej jednostki w oparciu o brak lub jej dominującą cechę w badaniach antropologicznych.

W ogólnym znaczeniu klasyfikacja jest efektywną metodą, która pozwala pozyskać informacje z dużych zbiorów danych. Stosując pewne wzorce podobieństw, a także zwracając uwagę na zróżnicowanie badanych obiektów oraz wykorzystując w tym celu etykiety ich klas, można uzyskać finalny opis danych Powszechnie stosowanym terminem opisującym to zjawisko jest grupowanie, jednak w ostatnich latach na znaczeniu coraz bardziej zyskuje termin analiza skupień. Łączy on zarówno techniki eksploracyjne oraz te o podstawach probabilistycznych. Mówi się, że początek ich rozwoju to rok 1973, kiedy to ukazały się prace R.F. Linga (Balicki, 2013).

Analiza skupień (ang. *cluster analysis*) jest działem wielowymiarowej analizy statystycznej, która posiada metody służące do wyodrębnienia jednorodnych podzbiorów badanych obiektów. Jest to badanie, które służy szukaniu oraz znajdywaniu wyżej wymienionych grup w nie-

jednorodnym zbiorze obiektów. Wyselekcjonowane obiekty tworzą grupę ze względu na unikalne cechy względem elementów z odrębnej grupy, inaczej mówiąc im obiekty są bardziej do siebie podobne tym istnieje większe prawdopodobieństwo, że będą tworzyć grupę. Wyszukiwanie skupień obiektów jest oparte na zmiennych, które charakteryzują analizowane obiekty, dlatego bardzo ważna rolę w tym procesie będzie ich dobór. Metoda jest także bardzo wrażliwa na przypadki odstające oraz takie zmienne, które słabo rozróżniają badany obiekty.

4.1.2 Miary niepodobieństwa

Odległość i niepodobieństwa obserwacji to kluczowy element we wszystkich technikach związanych z analizą skupień. Miary będą obliczone ze względu na wartości cech badanych jednostek. Wyniki będą zdeterminowane od tego jak będzie zdefiniowana odległość między obserwacją a klasą oraz jak będzie ona mierzona.

W analizie skupień są brane pod uwagę 2 pomiary. Stopień podobieństwa *(ang. similarity)* oraz niepodobieństwa *(ang. dissimiliarity)* elementów w wielomiarowej przestrzeni cech. Do ich oceny stopnia podobieństwa lub zróżnicowania obiektów wielocechowych są stosowane miary m.in. miary zróżnicowania lub odległości oraz miary podobieństwa.

Miary te w literaturze występują pod różnymi symbolami i nazwami, które pochodzą często od nazwisk pomysłodawców lub badaczy, którzy daną miarę stosowali w oparciu o swoje badania. Poniżej zostaną wymienione miary odległości, które są oparte na ogólnej metryce potęgowej.

Pierwszą z nich jest metryka miejska (ang. *city block*), która jest definiowana przez wykładnik m = 1

$$d_{rs}^{(m)} = \left[\sum_{j=1}^{p} \mid x_{rj} - x_{sj} \mid^{\frac{1}{m}} \right] \quad r,s = 1,...,n; r \neq s, \tag{4.1}$$

gdzie m jest dowolną liczbą naturalną (zwaną stałą Minkowskiego), specyfikującą wykładnik potęgi, której wartości zmieniają wagi dużych i małych różnic, natomiast symbol rs będzie reprezentował w tym przypadku skupienia.

Metryka ta była uważana przez Anderberga (Balicki, 2013) jako miara najbardziej naturalna. Powinna być używana w oparciu o nieskorelowane zmienne, co w praktyce może okazać się ograniczeniem, które nie będzie szczególnie przestrzegane.

Drugą z nich jest metryka euklidesowa (ang. euclidian metric), która jest definiowana przez

wykładnik m = 2

$$d_{rs}^{(2)} = \sqrt{\sum_{j=1}^{p} \left(x_{rj} - x_{sj}\right)^2} = \left[\left(x_r - x_s\right)' \left(x_r - x_s\right)\right]^{0,5} \tag{4.2}$$

Jest ona znana jako odległość dwóch punktów w przestrzeni p-wymiarowej. Jej ograniczeniem również jest nieskorelowanie ze sobą zmiennych. Aby wyróżnić małe różnice między obiektami stosowany jest kwadrat tej metryki.

$$\left(d_{rs}^{(2)}\right)^2 = \sum_{j=1}^{p} \left(x_{rj} - x_{sj}\right)^2 \tag{4.3}$$

Wykorzystanie jej będzie istotne przy grupowaniu hierarchicznym, w którym wykorzystywane są metody: centroidalna, Warda, średniego zróżnicowania oraz sumy kwadratów.

Korzystanie z wyżej wymienionych metryk może odbyć się tylko wtedy, kiedy różne zmienne są ze sobą porównywalne. Dlatego pierwszym krokiem powinno być standaryzowanie zmiennych.

Kiedy zostanie wybrana odpowiednia metryka, pozwoli to na utworzenie kwadratowej macierzy odległości.

$$D = (d_{rs}) = \begin{bmatrix} 0 & d_{12} & \ldots & d_{1m} \\ d_{21} & 0 & \ldots & d_{2m} \\ \ldots & \ldots & \ldots & \ldots \\ d_{n1} & w_{n2} & \ldots & 0 \end{bmatrix} \tag{4.4}$$

Macierz ta jest symetryczna (d_{rs}=d_{sr}) oraz na głównej przekątnej ma zera.

4.1.3 Metody hierarchiczne i optymalizacyjne

Metody klasyfikuje się na:

- Metody hierarchiczne, do których należą:
  - procedury aglomeracyjne
  - procedury podziału
- Metody niehierarchiczne, do których należą:
  - metoda Hartigana

- metoda kul
- metoda kostek
- metoda podziału przestrzennego
- metoda taksonomii stochastycznej
- metoda k-średnich

Metody hierarchiczne są najczęściej wykorzystywaną metodą wielowymiarowego grupowania obiektów. Proces grupowania hierarchicznego przebiegający od pojedynczych obiektów do ostatecznej grupy skupiającej je wszystkie nazywamy aglomeracyjnym (Balicki, 2013). Główną cechą tego grupowania jest traktowanie poszczególnych elementów jako odrębnych grup, a następnie ich sekwencyjne klasyfikowanie na podstawie ich różnic czy podobieństw. Grupowanie rozpoczyna się od połączenia obiektów wykazujących względem siebie największe podobieństwo, dołączając w kolejnych etapach kolejne elementy. Każdy kolejny krok wymaga od grupującego pomiarów między nowopowstałymi skupieniami, a tymi już powstałymi, tworząc w ten sposób macierze odległości. Są one kluczowe na kolejnym etapie procesu, w którym przeliczane są powstałe odległości.

Wielokrotne powtarzanie powyższego procesu pozwala na uzyskanie ostatecznej grupy obiektów oraz wykresu hierarchicznego uporządkowania, nazywanym drzewkiem połączeń lub dendrogramem (Balicki, 2013).

Jedną z metod grupowania aglomeracyjnego jest metoda najbliższego sąsiada. Polega ona na przeliczeniu odległości między obiektami jednego skupienia, a obiektami innego skupienia według kryterium najmniejszej odległości (Balicki, 2013).

$$d_{ip} = min(d_{ir}, d_{is})(i = 1, 2, ..., n; i \neq r; i \neq s). \tag{4.5}$$

Pozwala ona na wyznaczenie odległości między nowym, a starym zgrupowaniem wykorzystując do tego jego najbliższe obiekty. Ponieważ w początkowym etapie każdy obiekt traktowany jest jako odrębna grupa, łączy się elementy sobie najbliższe. Odległość między obiektami nazywana jest najniższym progiem prawdopodobieństwa i pozwala połączyć jednostki, które są względem siebie najbardziej podobne. W kolejnych etapach określane są odległości między pierwszym skupieniem, a pozostałymi, jeszcze niezgrupowanymi obiektami, a następnie zgrupowanie tych będących najbardziej do siebie zbliżonymi. Wyliczenie miary takiej odległości

może być wykonywane na podstawie poniższego algorytmu:

$$d_{ip} = \frac{1}{2}d_{ir} + \frac{1}{2}d_{is} - \frac{1}{2}\left|d_{ir} - d_{is}\right|. \tag{4.6}$$

Czynność ta powtarzana jest wielokrotnie aż do momentu uzyskania finalnego grupowania.

Kolejną z metod wykorzystywanych podczas grupowania hierarchicznego jest metoda najdalszego sąsiada. Jest ona przeciwieństwem metody opisanej powyżej i polega na łączeniu ze sobą najbardziej oddalonych obiektów. W powyższej metodzie elementy macierzy odległości są sukcesywnie przekształcane zgodnie z kryterium:

$$d_{ip} = max(d_{ir}, d_{is}), \tag{4.7}$$

natomiast algorytm przeliczania odległości jest w postaci:

$$d_{ip} = \frac{1}{2}d_{ir} + \frac{1}{2}d_{is} + \frac{1}{2}\left|d_{ir} - d_{is}\right|. \tag{4.8}$$

W kolejnych etapach powstają skupiska, między którymi odległości te są coraz mniejsze. Pomijając odmienny sposób przeliczania odległości, każdy kolejny etap jest taki sam, co stanowi podstawę grupowania aglomeracyjnego. W przeciwieństwie do poprzedniej metody, metoda najdalszego sąsiada skupia się na tworzeniu grup spójnych wewnętrznie pozwalając na uzyskanie odmiennych rezultatów.

Jedną z niehierarchicznych metod grupowania obiektów jest grupowanie podziałowe, które polega na podziale zbioru na poszczególne grupy. Podział na określoną liczbę grup nazywany jest K-podziałem, a liczba grup określana jest przez badacza (Balicki, 2013). Istnieje wiele możliwych podziałów, a ich celem jest znalezienie najbardziej podobnych obiektów w obrębie tej samej klasy względem wybranego kryterium grupowania. Wzór rekurencyjny na niebanalną liczbę podziałów został opracowany przez Fortier'a i Solomon'a $N(n, K)$

$$N(n,K) = \frac{1}{K!}\left[K^n - \sum_{g=1}^{K-1}\frac{K!}{(K-1)!}N(n,g)\right], \tag{4.9}$$

gdzie $N(n, g)$ jest liczbą podziałów n obiektów na g grup $(g=1, \ldots, K-1)$ Celem grupowania jest znalezienie optymalnego podziału badanych obiektów poprzez wstępny podział obiektów na określoną liczbę klas, a następnie ulepszyć go poprzez zastosowanie odpowiednich przekształceń.

Ze względu na możliwie dużą liczbę potencjalnych podziałów pozwalają na wykorzystanie wielu sposobów podziału. Kryterium jego adekwatności stanowi ważny element grupowania i możliwe jest dzięki wykorzystaniu miar heterogeniczności oraz izolacji. Miary te oparte są na własnościach mierzalnych poszczególnych cech obiektów lub też na odległościach pomiędzy nimi. Miary heterogeniczności nazywane są również wewnątrzgrupową sumą kwadratów i pozwalają zbadać odchylenia wartości danych obiektów od średniej wartości badanego skupienia. Miary izolacji polegają natomiast na sumowaniu wartości poszczególnych grup obiektów, a następnie wyznaczeniu optymalnego podziału. W przeciwieństwie do miar heterogeniczności opierają się ona na badaniu odległości pomiędzy poszczególnymi obiektami badanej grupy, a obiektami spoza niej. Każde z powyższych ugrupowań może stanowić podstawę do analizy skupień, będących użytecznym narzędziem analizy wybranych danych. Analiza skupień pozwala wykryć jednorodne grupy obiektów w celu poznawczym lub praktycznym. Pozwala również na wyodrębnienie naturalnej struktury grupowej dla badanych obiektów. Dodatkowo analiza skupień umożliwia zbadanie oraz ocenę wymiarowości badanego zjawiska poprzez ustalenie zmiennych najlepiej odzwierciedlających własności badanych obiektów. Analiza skupień pozwala zbadać prawidłowości wewnątrz badanych grup, zredukować duże zbiory obiektów skupiając się na średnich wartościach poszczególnych podzbiorów, a także umożliwić ich podział w celu przeprowadzenia dalszych, wielowymiarowych analiz.

4.1.4 Miary oceny jakości analizy skupień

Przy uwzględnianiu jakości przeprowadzonego grupowania stosuje się wiele różnorodnych funkcji oceny. Najpopularniejszą oraz najczęściej wykorzystywaną w praktyce jest funkcja o nazwie: suma błędów kwadratowych *(ang. square error criterion)*:

$$E(C) = \sum_{k=1}^{K} \sum_{p \in C_k} d^2(p, m_k), \tag{4.10}$$

gdzie m_k jest środkiem ciężkości skupienia Ck, a d(. , .) jest funkcją odległości. Wyniki grupowania mogą przybierać różne wartości oraz zmieniać się, gdzie jest to uwarunkowane wyborem definicji odległości punktów.

4.2 Implementacja metod analizy skupień w R

W środowisku R istnieje wiele możliwości implementacji algorytmów analizy skupień. W niniejszym rozdziale zostaną przedstawione jedne z najczęściej wykorzystywanych z grup niehierarchicznych oraz hierarchicznych. Do metod niehierarchicznych zaliczamy funkcję *kmeans*, która odpowiada za algorytm k-średnich oraz funkcję *pam* (ang.Partitioning Around Medoid) z pakietu pakietu *cluster* obsługująca algorytm k-medoidów. Z kolei grupę metod hierarchicznych reprezentuje funkcja *hclust*, czyli hierarchiczny algorytm grupowania (Nowak-Brzezińska, 2012) oraz funkcja *diana* (DIvisive ANAlysis Clustering) z pakietu *cluster*.

Funkcja *kmeans* jest często wykorzystywana, ze względu na małą złożoność oraz względnie nieskomplikowaną implementację. Polega ona na wstępnym podzieleniu zbioru na *k* skupień, następnie dla każdego z nich liczony jest tzw. centroid oraz przypisywany jest każdy element ze zbioru do najbliższej mu grupy. Każda obserwacja jest przypisana do danego skupiska w taki sposób, że suma kwadratów odległości obserwacji do przypisanych im skupień jest jak najmniejsza. Implementacja tej metody w pakiecie R wygląda następująco:

```
k3 <- kmeans(x = comp, centers = 3, nstart = 6)
```

Program 4.1. Implementacja metody kmeans()

Pierwszym argumentem funkcji *kmeans* jest macierz danych numerycznych - *comp*, ponieważ funkcja ta wymaga, aby wszystkie zmienne były zmiennymi ilościowymi. Jest to macierz, która podlega grupowaniu. W tym przypadku są to czynniki wyodrębnione za pomocą analizy głównych składowych. Wartość *centers* określa liczbę skupień, które chcemy uzyskać lub może być to również liczba początkowych centrów skupień. Ostatni atrybut *nstart* określa liczbę losowych zbiorów danych branych pod uwagę w grupowaniu, jeśli w parametrze *centers* podano liczbę grup. Do przeprowadzenia grupowania funkcja *kmeans* domyślnie korzysta z algorytmu Hartigana i Wonga (1979), jednak za pomocą opcjonalnego atrybutu *method* możemy posłużyć się również algorytmem MacQueena, Lloyda lub Forgy'ego (R Core Team, 2018). Zaletą tej funkcji jest fakt, że nie zależy ona w dużym stopniu od wartości skrajnych oraz łączy skupienia o stosunkowo niskich wariancjach.

Punktem wyjściowym powyższej metody jest lista zwracająca następujące komponenty:

Drugą metodą z grupy algorytmów niehierarchicznych jest funkcja *pam* obsługująca algorytm k-medoidów. W odróżnieniu od algorytmu k-średnich, funkcja *pam* w kolejnych iteracjach wyznacza nowe grupy na podstawie obiektów, które aktualnie należą do rozpatrywanego

Tabela 4.1. Komponenty metody kmeans

Nazwa	Opis
$cluster	wektor liczb całkowitych (od 1:k) wskazujący skupienie, do którego przypisany jest każdy punkt
$centers	macierz zawierająca środki skupień
$withinss	wektor sumy kwadratów wewnątrz skupienia
$tot.withinss	łączna suma kwadratów wnętrz skupienia
$size	liczba punktów w każdym skupieniu

Źródło: Opracowanie własne na podstawie
https://www.rdocumentation.org/packages/stats/versions/3.6.0/topics/kmeans

zbioru (Nowak-Brzezińska, 2012).

Jej celem jest znalezienie k reprezentatywnych skupień, które minimalizują sumę zróżnicowania obserwacji do ich najbliższego skupienia. Jej implementacja wygląda następująco:

```
k3 <- pam(x = comp, k = 3, diss = FALSE, metric = "euclidean")
```

Program 4.2. Implementacja metody pam()

Tak jak w przypadku funkcji *kmeans*, pierwszym argumentem jest macierz danych numerycznych. Za pomocą logicznego atrybutu *diss*, możemy ustalić czy do obliczeń posłuży macierz numeryczna czy macierz podobieństw. Parametr *k* określa liczbę skupień, które mają zostać wyodrębnione. Ostatni z nich – *metric*, mówi o tym, która metryka powinna zostać użyta: euklidesowa lub manhattan. Oprócz wyżej wymienionych atrybutów, funkcja *pam* umożliwia również standaryzację zmiennych za pomocą logicznego atrybutu *stand*. Dotyczy on tylko i wyłącznie macierzy o danych numerycznych. W przypadku macierzy podobieństw atrybut ten jest ignorowany (Maechler, Rousseeuw, Struyf, Hubert & Hornik, 2019). Funkcja *pam* zwraca m.in. takie komponenty jak:

Tabela 4.2. Komponenty metody pam

Nazwa	Opis
$medoids	obiekt reprezentujący dane skupienie
$id.med	wektor liczb zawierający liczbę obserwacji medoidów
$clustering	wektor zawierający numer skupienia dla każdego obiektu
$objective	wartość funkcji celu po pierwszym i drugim etapie algorytmu *pam*

Źródło: Opracowanie własne na podstawie
https://www.rdocumentation.org/packages/cluster/versions/2.0.9/topics/pam.object

Kolejną metodą jest funkcja *hclust* z pakietu podstawowego *stats*, którą zaliczamy do metod

hierarchicznych. Stosuje ona podejście aglomeracyjne, co oznacza, że grupowanie rozpoczyna się od każdej obserwacji w ramach pojedynczych skupień, a następnie sukcesywnie łączy je, aż do momentu spełnienia kryterium zatrzymania (Pathak, 2018). Odległości między skupieniami są rozumiane jako maksymalna odległość między poszczególnymi składnikami w danych grupach. Na każdym etapie procesu, skupienia które znajdują się najbliżej są łączone w jeden. Implementacja tej metody w pakiecie R wygląda następująco:

```
cluster <- hclust(dist(comp), method = "complete")
```

Program 4.3. Implementacja metody hclust()

Metoda *hclust* wymaga, aby dane, które podlegają grupowania były w formie macierzy odległości. Z tego względu użyta dostała funkcja *dist*, która oblicza dystans pomiędzy wierszami w danej macierzy. Domyślnie używana ona odległości euklidesowej, jednak za pomocą atrybutu *method* istnieje możliwość zmiany metryki odległości. Parametr *method* funkcji *hclust* odpowiada za specyfikację metody łączenia skupień (Kodali, 2017). W tym przypadku użyta została metoda *complete*, co oznacza, że łączone są sąsiedztwa, które są najbardziej oddalone od siebie. Rezultatem tej metody jest wektor prezentujący przydział poszczególnych obserwacji zbioru wejściowego do utworzonej liczby grup, natomiast wyniki są prezentowane w postaci dendrogramu.

Aby wykonać grupowanie hierarchiczne strategią z góry do dołu, można wykorzystać funkcję *diana* z pakietu *cluster*. Zaliczana jest ona do grupy metod podziału, w których grupowanie zaczyna się od korzenia, a wszystkie obiekty znajdują się w jednym skupisku. Cały proces tej funkcji jest więc odwrotnością funkcji *hclust*. Grupowanie kończy się, gdy wszystkie obiekty znajdą się w swoim pojedynczym skupieniu (Statistical Tools For High-Throughput Data Analysis, 2019). Implementacja metody *diana* w pakiecie R:

```
res.diana <- diana(x = comp, metric = "euclidean")
```

Program 4.4. Implementacja metody diana()

Pierwszym argumentem jest macierz danych lub macierz podobieństw w zależności w jaki sposób określimy parametr *diss*. W tym przypadku użyta została macierz danych, która zawiera tylko wartości numeryczne. Tak jak w przypadku funkcji niehierarchicznej – *pam*, możemy określić metrykę, jaką użyjemy do obliczanie różnic między obserwacjami: Euklidesowa lub Manhattan. Obiektem klasy *diana* jest lista z następującymi komponentami:

Tabela 4.3. Komponenty metody diana

Nazwa	Opis
$order	wektor określający permutację oryginalnych obserwacji
$order.lab	wektor zawierający etykiety obserwacji
$heigh	wektor o średnicy skupisk przed ich rozszczepieniem
$dc	współczynnik podziału, mierzący strukturę skupień zbioru danych *pam*
$merge	opisuje podział na etapie n-i grupowania
$data	macierz zawierająca oryginalne lub znormalizowane pomiary

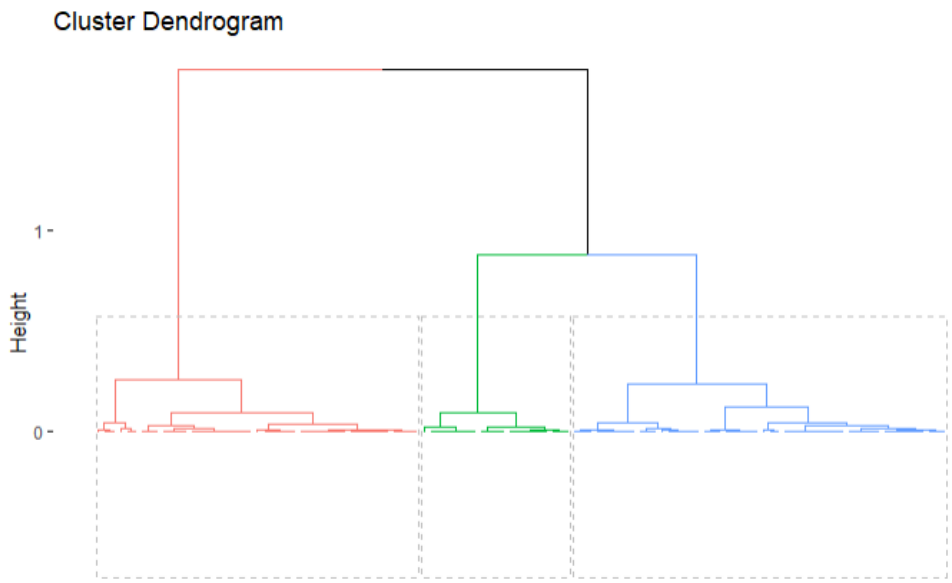
Źródło: Opracowanie własne na podstawie
https://www.rdocumentation.org/packages/cluster/versions/2.0.7-1/topics/diana

4.3 Wyniki analizy skupień

Do przeprowadzenia analizy skupień posłużyliśmy się grupowaniem hierarchicznym wyników analizy głównych składowych oraz metodą k-średnich. Jest to podejście zwane *HCPC* (Hierarchical Clustering on Principal Components), które pozwala nam łączyć trzy standardowe metody używane przy analizie danych wielowymiarowych, takie jak: analiza głównych składowych, grupowanie hierarchiczne oraz grupowanie niehierarchiczne, szczególnie k-średnich („HCPC - Hierarchical Clustering on Principal Components: Essentials", 2017). Jako zbiór danych określiliśmy zatem macierz współrzędnych pierwszych dwóch głównych składowych, zachowując przy tym w przybliżeniu 80% informacji ze zbioru podstawowego. W celu grupowania zbioru skorzystaliśmy z metody *kmeans* z pakietu podstawowego. Istotnym dla tej metody jest, aby przed dokonaniem grupowania zdefiniować liczbę skupień, które chcemy otrzymać. W tym celu wykonaliśmy grupowanie hierarchiczne za pomocą funkcji *HCPC* z pakietu *FactoMineR* oraz użyliśmy funkcji *NbClust* z pakietu o tej samej nazwie, która wykorzystuje ponad dwadzieścia różnych wskaźników do obliczenia optymalnej liczby skupień. Po wywołaniu funkcji *HCPC* na zbiorze danych zawierającym nowe współrzędne punktów po analizie głównych składowych, możemy wywołać funkcję *fviz_dend* z pakiet *factoextra*, która w estetyczny sposób wyświetli nam dendrogram oraz pokaże optymalną liczbę skupień dla naszego zbioru.

W rezultacie otrzymujemy trzy skupienia, które zostały wyliczone na podstawie aglomeracyjnej metody grupowania. Wykorzystana została również odległość euklidesowa oraz metoda Warda w funkcji *HCPC*.

Następną metodą jest kompleksowa funkcja *NbClust* z pakietu o tej samej nazwie. Jej zaletą jest przeprowadzenie testów na bardzo dużej liczbie wskaźników, które sugerują jaką liczbę



Rysunek 4.1. Dendrogram skupień

Źródło: Opracowanie własne.

skupień powinno się założyć w celach dalszych analiz. W wyniku zwracana jest liczba głosów za poszczególną liczbą skupień oddanych przez konkretne wskaźniki. Przedział poszukiwanych skupień został ograniczony w funkcji od dwóch do sześciu. Poniższy wydruk prezentuje wyniki omawianej metody:

```
NbClust(comp, distance = "euclidean", method = "kmeans", min.nc = 2, max.nc
    = 6)

    ******************************************************************
* Among all indices:
* 5 proposed 2 as the best number of clusters
* 13 proposed 3 as the best number of clusters
* 1 proposed 4 as the best number of clusters
* 2 proposed 5 as the best number of clusters
* 2 proposed 6 as the best number of clusters

                   ***** Conclusion *****

* According to the majority rule, the best number of clusters is  3


*******************************************************************
```

Program 4.5. Implementacja metody NbClust

Również w tym przypadku, według reguły większości, najkorzystniejszą liczbą skupień jest liczba trzy. Wśród wszystkich wskaźników, aż trzynaście sugerowało liczbę w naszym wyniku, natomiast zaledwie pięć wskazywało na liczbę o jeden mniejszą.

Następnym krokiem jest przeprowadzenie analizy skupień z wykorzystaniem algorytmu k-

średnich. W związku z powyższymi wynikami jako liczbę skupień podana została liczba trzy, natomiast zbiorem danych są współrzędne punktów dwóch pierwszych składowych głównych. Dodatkowo cały proces powtórzony został dwadzieścia pięć razy, by uzyskać jak najdokładniejszy wynik. Do stworzenia wykresu wykorzystana została funkcja *fviz_cluster* [factoextra]. Jako argumenty przyjmuje ona wyniki algorytmu k-średnich oraz oryginalne dane z podstawowego zbioru danych. Po wywołaniu funkcji wyświetlony zostaje wykres, w którym punkty reprezentują poszczególne szkoły wyższe. Ich rozproszenie opisane jest w wymiarach, które odpowiadają wynikom analizy głównych składowych.

Rysunek 4.2. Wykres analizy skupień

Źródło: Opracowanie własne.

Każda uczelnia w danej grupie na wykresie 4.2 ma przypisany numer identyfikacyjny, który odpowiada konkretnej uczelni ze zbioru podstawowego. Poniżej zamieszczone zostały tabele, które opisują powyższe przyporządkowanie numerów.

W celu zbadania charakterystyk wyodrębnionych skupień zbadany został rodzaj uczelni wyższej w każdej z trzech grup. Podziału dokonano na Uniwersytety, Politechniki, Wyższe Szkoły oraz Inne Szkoły Wyższe takie jak Akademie, Szkoły Główne i Uczelnie.

W skupieniu pierwszym większość stanowią Wyższe Szkoły. Warto zaznaczyć, że aż osiemnaście z dwudziestu dwóch uczelni w tej grupie są Państwowymi Wyższymi Szkołami, w tym szesnaście stanowią Państwowe Wyższe Szkoły Zawodowe. Oprócz wymienionych znajdują się

Numer <int>	Nazwa <chr>
16	Politechnika Rzeszowska im. Ignacego Łukasiewicza
36	Politechnika Świętokrzyska
49	Uniwersytet Technologiczno-Humanistyczny im. Kazimierza Pułaskiego w Radomiu
52	Politechnika Koszalińska
56	Państwowa Wyższa Szkoła Techniczno-Ekonomiczna im. ks. Bronisława Markiewicza w Jarosławiu
58	Państwowa Wyższa Szkoła Zawodowa w Nowym Sączu
59	Państwowa Wyższa Szkoła Zawodowa w Tarnowie
63	Państwowa Wyższa Szkoła Zawodowa im. Witelona w Legnicy
64	Państwowa Wyższa Szkoła Zawodowa im. Stanisława Pigonia w Krośnie
66	Państwowa Wyższa Szkoła Zawodowa w Płocku
68	Wyższa Szkoła Ekonomii, Prawa i Nauk Medycznych im. prof. Edwarda Lipińskiego w Kielcach
71	Państwowa Wyższa Szkoła Zawodowa w Koninie
75	Państwowa Wyższa Szkoła Zawodowa im. Stanisława Staszica w Pile
76	Państwowa Wyższa Szkoła Zawodowa w Nysie
78	Państwowa Wyższa Szkoła Zawodowa im. prof. E. Szczepanika w Suwałkach
79	Państwowa Wyższa Szkoła Informatyki i Przedsiębiorczości w Łomży
80	Państwowa Wyższa Szkoła Zawodowa w Ciechanowie
87	Państwowa Wyższa Szkoła Zawodowa w Elblągu
91	Wyższa Szkoła Handlowa w Radomiu
99	Państwowa Wyższa Szkoła Zawodowa w Skierniewicach
106	Państwowa Wyższa Szkoła Zawodowa im. Jana Amosa Komeńskiego w Lesznie
115	Państwowa Wyższa Szkoła Zawodowa we Włocławku

Rysunek 4.3. Uczelnie i szkoły wyższe wchodzące w skład pierwszego skupienia

Źródło: Opracowanie własne.

Numer <int>	Nazwa <chr>
1	Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie
2	Uniwersytet Warszawski
3	Politechnika Wrocławska
4	Politechnika Śląska
5	Uniwersytet im. Adama Mickiewicza w Poznaniu
6	Uniwersytet Łódzki
7	Politechnika Warszawska
8	Uniwersytet Warmińsko-Mazurski w Olsztynie
9	Uniwersytet Jagielloński w Krakowie
10	Uniwersytet Gdański
11	Politechnika Gdańska
12	Politechnika Poznańska
13	Uniwersytet Marii Curie-Skłodowskiej w Lublinie
14	Uniwersytet Mikołaja Kopernika w Toruniu
15	Uniwersytet Ekonomiczny w Krakowie
17	Szkoła Główna Gospodarstwa Wiejskiego w Warszawie
18	Uniwersytet Wrocławski
19	Uniwersytet Rzeszowski
20	Uniwersytet Śląski w Katowicach
21	Politechnika Łódzka
23	Uniwersytet Rolniczy im. Hugona Kołłątaja w Krakowie
25	Uniwersytet Jana Kochanowskiego w Kielcach
27	Uniwersytet Szczeciński
28	Uniwersytet w Białymstoku

Numer <int>	Nazwa <chr>
29	Uniwersytet Zielonogórski
30	Politechnika Lubelska
31	Uniwersytet Ekonomiczny w Katowicach
32	Uniwersytet Przyrodniczy w Poznaniu
33	Uniwersytet Ekonomiczny we Wrocławiu
34	Politechnika Częstochowska
35	Uniwersytet Kardynała Stefana Wyszyńskiego w Warszawie
37	Politechnika Białostocka
38	Zachodniopomorski Uniwersytet Technologiczny w Szczecinie
39	Uniwersytet Przyrodniczy w Lublinie
40	Uniwersytet Ekonomiczny w Poznaniu
41	Uniwersytet Przyrodniczy we Wrocławiu
42	Szkoła Główna Handlowa w Warszawie
43	Uniwersytet Opolski
44	Uniwersytet Technologiczno-Przyrodniczy im. Jana i Jędrzeja Śniadeckich w Bydgoszczy
45	Politechnika Opolska
46	Katolicki Uniwersytet Lubelski Jana Pawła II w Lublinie
47	Uniwersytet Przyrodniczo-Humanistyczny w Siedlcach
50	Wojskowa Akademia Techniczna im. Jarosława Dąbrowskiego
61	Akademia im. Jana Długosza w Częstochowie
65	Państwowa Szkoła Wyższa im. Papieża Jana Pawła II w Białej Podlaskiej
70	Akademia Wychowania Fizycznego im. Jerzego Kukuczki w Katowicach
103	Akademia im. Jakuba z Paradyża
104	Akademia Wojsk Lądowych imienia generała Tadeusza Kościuszki

Rysunek 4.4. Uczelnie i szkoły wyższe wchodzące w skład drugiego skupienia

Źródło: Opracowanie własne.

również trzy politechniki oraz jeden uniwersytet.

W skupieniu drugim 64,6% ze wszystkich znajdujących się uczelni stanowią Uniwersytety. Jednocześnie jest ona jedyną grupą, w której występują one tak licznie. Analogiczna sytuacja ma miejsce w przypadku Politechnik, które w grupie drugiej pojawiają się aż dziesięć razy, co daje wynik 20,8%. Wyższe Szkoły stanowią 4,2% całego zbioru, natomiast Inne Wyższe Szkoły, w tym przypadku Akademie, 12,5%.

W skupieniu trzecim wyróżniamy tylko dwie kategorie szkół. Pierwszą z nich są Wyższe

Numer <int>	Nazwa <chr>	Numer <int>	Nazwa <chr>
22	Wyższa Szkoła Bankowa we Wrocławiu	89	Szkoła Wyższa im. Pawła Włodkowica w Płocku
24	Społeczna Akademia Nauk z siedzibą w Łodzi	90	Gdańska Szkoła Wyższa z siedzibą w Gdańsku
26	Wyższa Szkoła Bankowa w Poznaniu	92	Kujawska Szkoła Wyższa we Włocławku
48	Wyższa Szkoła Bankowa w Toruniu	93	Wyższa Szkoła Finansów i Zarządzania w Warszawie
51	Krakowska Akademia im. Andrzeja Frycza Modrzewskiego w Krakowie	94	Wszechnica Polska Szkoła Wyższa w Warszawie
53	Wyższa Szkoła Informatyki i Zarządzania z siedzibą w Rzeszowie	95	Kujawsko-Pomorska Szkoła Wyższa w Bydgoszczy
54	Akademia Techniczno-Humanistyczna w Bielsku-Białej	96	Wyższa Szkoła Zarządzania w Gdańsku
55	Wyższa Szkoła Bankowa z siedzibą w Gdańsku	97	Staropolska Szkoła Wyższa w Kielcach
57	Wyższa Szkoła Biznesu w Dąbrowie Górniczej	98	Wyższa Szkoła Handlowa im. Bolesława Markowskiego w Kielcach w likwidacji
60	Wyższa Szkoła Ekonomii i Innowacji w Lublinie	100	Wyższa Szkoła Administracji i Biznesu im. Eugeniusza Kwiatkowskiego w Gdyni
62	Wyższa Szkoła Bezpieczeństwa w Poznaniu	101	Szczecińska Szkoła Wyższa Collegium Balticum w Szczecinie
67	Uczelnia Techniczno-Handlowa im. Heleny Chodkowskiej	102	Wyższa Szkoła Zarządzania Edukacja we Wrocławiu
69	Uczelnia Warszawska im. Marii Skłodowskiej-Curie w Warszawie	105	Uczelnia Łazarskiego w Warszawie
72	Akademia Morska w Gdyni	107	Uczelnia Jana Wyżykowskiego
73	Akademia Leona Koźmińskiego w Warszawie	108	Wyższa Szkoła Humanitas w Sosnowcu
74	Wyższa Szkoła Zarządzania i Administracji w Opolu	109	Wyższa Szkoła Gospodarki Euroregionalnej im. Alcide De Gasperi w Józefowie
77	Wyższa Szkoła Gospodarki w Bydgoszczy	110	Gdańska Wyższa Szkoła Humanistyczna
81	Górnośląska Wyższa Szkoła Handlowa im. Wojciecha Korfantego w Katowicach	111	Wyższa Szkoła Przedsiębiorczości i Administracji w Lublinie
82	Wyższa Szkoła Zarządzania i Bankowości w Krakowie	112	Wyższa Szkoła Finansów i Zarządzania w Białymstoku
83	Polsko-Japońska Akademia Technik Komputerowych	113	Collegium Mazovia Innowacyjna Szkoła Wyższa w Siedlcach
84	Wyższa Szkoła Logistyki w Poznaniu	114	Wyższa Szkoła Biznesu - National Louis University z siedzibą w Nowym Sączu
85	Wyższa Szkoła Zarządzania Ochroną Pracy w Katowicach	116	Wielkopolska Wyższa Szkoła Społeczno-Ekonomiczna w Środzie Wielkopolskiej
86	Akademia Finansów i Biznesu Vistula	117	Europejska Uczelnia Informatyczno-Ekonomiczna w Warszawie
88	Wyższa Szkoła Ekologii i Zarządzania w Warszawie	118	Wyższa Szkoła Turystyki i Hotelarstwa w Gdańsku

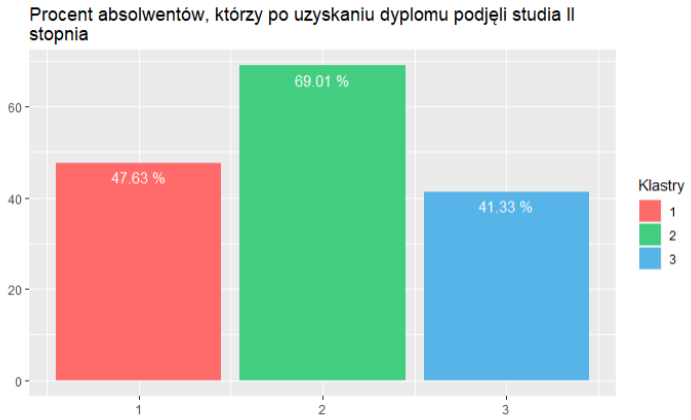
Rysunek 4.5. Uczelnie i szkoły wyższe wchodzące w skład trzeciego skupienia

Źródło: Opracowanie własne.

Szkoły, które stanowią 75% całego zbioru oraz sklasyfikowane jako Inne Wyższe Szkoły – Akademie oraz Uczelnie, które pojawiały się łącznie dwanaście razy.

Oprócz opisania występujących kategorii szkół wyższych w grupach, obliczona została również średnia wartość dla każdego wskaźnika w poszczególnych skupienia. Ze względu na rozproszenie obserwacji względem wybranych wskaźników, ich wyniki znacznie różnią się w obrębie powstałych skupisk.

Poniżej zwizualizowane zostały omawiane wyniki w postaci wykresów słupkowych.

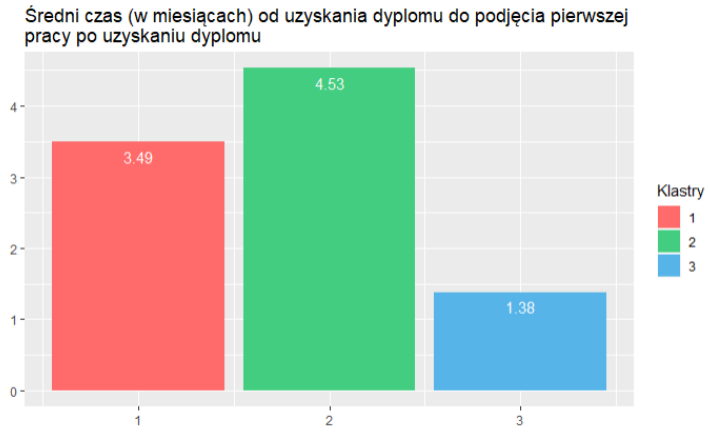


Rysunek 4.6. Wykres średniej wartości wskaźnika U1_15 w skupieniach

Źródło: Opracowanie własne.

Na przykładzie wykresu 4.6 zmiennej U1_15, tj. procent absolwentów, którzy po uzyskaniu dyplomu podjęli studia drugiego stopnia, znacząco wyróżnia się skupienie drugie, osiągając

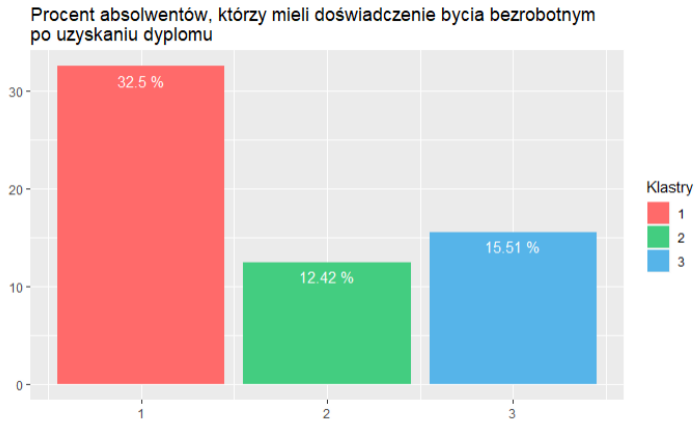
wartość 69,01%. Znacznie mniej skłonni do kontynuowania nauki na studiach drugiego stopnia byli absolwenci w grupie pierwszej – 47,63% oraz trzeciej – 41,33%.



Rysunek 4.7. Wykres średniej wartości wskaźnika U1_18 w skupieniach

Źródło: Opracowanie własne.

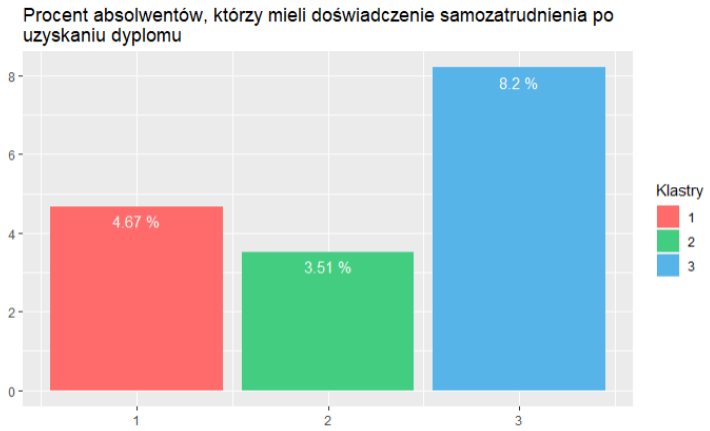
Drugim wskaźnikiem w naszym zbiorze danych był średni czas (w miesiącach) od uzyskania dyplomu do podjęcia pierwszej pracy po uzyskaniu dyplomu (U1_18). W tym przypadku średnio najdłuższym czasem oczekiwania charakteryzowało się skupienie drugie – 4.53 miesiąca. Kolejno mniejszą wartością odznaczała się grupa pierwsza i oscyluje ona w granicach 3.49 miesiąca. Znacząco wyraźny spadek obserwujemy w skupieniu trzecim, który jest ponad trzykrotnie mniejszy od skupienia drugiego i wynosi zaledwie 1.38 miesiąca.



Rysunek 4.8. Wykres średniej wartości wskaźnika U1_20 w skupieniach

Źródło: Opracowanie własne.

Na wykresie 4.8 przedstawiony został wskaźnik U1_20, tj. procent absolwentów, którzy mieli doświadczenie bycia bezrobotnym po uzyskaniu dyplomu. Analizując powyższy wykres, zdecydowanie można zaobserwować wysoką wartość tego wskaźnika w skupieniu pierwszym, który wynosi 32,5%. W pozostałych skupieniach wartość ta jest znacznie niższa i oscyluje w granicach 15,51% w przypadku skupienia trzeciego i 12,42% w przypadku drugiego.

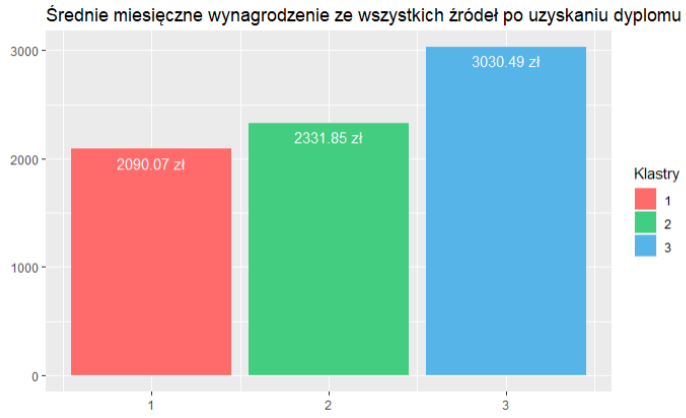


Rysunek 4.9. Wykres średniej wartości wskaźnika U1_66 w skupieniach

Źródło: Opracowanie własne.

Kolejnym wykresem jest procent absolwentów, którzy mieli doświadczenie samozatrudnienia po uzyskaniu dyplomu (U1_66). Skutkiem tego, iż obserwacje w wkupieniu trzecim są na wykresie skupień najbardziej rozproszone w kierunku omawianej zmiennej (w prawo), jest dominująca wartość tego czynnika w grupie w stosunku do pozostałych. Osiągnęła ona wartość 8,2%, natomiast w grupie pierwszej zaledwie 4.67%. Z kolei w skupieniu drugim tylko 3.51% absolwentów miało doświadczenia samozatrudnienia po uzyskaniu dyplomu.

Ostatnim przedstawionym wskaźnikiem jest średnie miesięczne wynagrodzenie ze wszystkich źródeł po uzyskaniu dyplomu. Najwyższą średnią wartość na przedstawionym wykresie osiągnęło skupienie trzecie – 3030,49 zł. Kolejno drugą, średnią najwyższą, skupienie drugie ze średnim miesięcznym wynagrodzeniem o wartości 2331,85zł, natomiast najniższym, bo 2090,07 zł średnim miesięcznym wynagrodzeniem cechuje się skupienie pierwsze.



Rysunek 4.10. Wykres średniej wartości wskaźnika U1_116 w skupieniach

Źródło: Opracowanie własne.

4.4 Wnioski

Zastosowanie analizy skupień z wykorzystaniem analizy głównych składowych pozwoliło na wyodrębnienie trzech charakterystycznych grup. Wyselekcjonowane zbiory opisują właściwości dla konkretnych rodzajów szkół wyższych, ponieważ zgodnie z wynikami analizy skupień, każde skupienie zawiera w znacznej większości uczelnie jednego typu. W związku z tym można stwierdzić, że Państwowe Wyższe Szkoły, Uniwersytety czy Szkoły Wyższe różnią się między sobą nie tylko nazewnictwem, ale również ze względu na analizowane w pracy wskaźniki, które w ramach wyodrębnionych skupisk wykazywały znaczące różnice.

W pierwszym skupieniu znajdują się dwadzieścia dwie uczelnie. Jest to grupa, w której występują w znacznej większości Państwowe Wyższe Szkoły Zawodowe. Kierunek rozproszenia obserwacji należących do tego skupienia na wykresie skupień wskazuje, że są to uczelnie, po których absolwenci najczęściej rejestrowali się jako bezrobotni w Urzędach Pracy po uzyskaniu dyplomu. Średni procent dla tej zmiennej wynosi aż 32,5%. Uczelnią, która charakteryzuje się najwyższym średnim procentem tej wartości jest Państwowa Wyższa Szkoła Zawodowa we Włocławku z wynikiem 47,2%. Jednak mimo iż grupa pierwsza cechowała się najwyższym średnim procentem absolwentów, którzy mieli doświadczenie bycia bezrobotnym to nie osiągnęła średniego najdłuższego czasu od uzyskania dyplomu do podjęcia pracy przez absolwentów. W tym przypadku średni czas wyniósł 3.49 miesiąca, co stanowi drugą pozycję wśród wszystkich skupień. Może mieć to związek z częstotliwością rejestracji absolwentów w Urzędzie Pracy, który przyczynił się do szybszego podjęcia pracy przez zarejestrowanych. Z kolei średnio 4,67%

absolwentów w ramach omawianej grupy miało doświadczenie samozatrudnienia po uzyskaniu dyplomu oraz spośród wszystkich wyodrębnionych skupień charakteryzowała się najniższą średnią miesięcznych wynagrodzeń ze wszystkich źródeł po uzyskaniu dyplomu, która wyniosła 2090,07 zł. Kolejną charakterystyką danego skupienia jest procent absolwentów, którzy po uzyskaniu dyplomu podjęli studia drugiego stopnia. Według przeprowadzonej analizy, blisko co drugi absolwent nie kontynuował edukacji na drugim stopniu nauczania. W związku z powyższym z ekonomicznego punktu widzenia, jest to grupa szkół wyższych, która w świetle przedstawionych statystyk wypada najmniej korzystnie.

Skupienie drugie składa się w ponad 85% z Uniwersytetów oraz Politechnik i zawiera czterdzieści osiem obserwacji. Cechą charakterystyczną dla tej grupy jest wysoki średni procent absolwentów, którzy po uzyskaniu dyplomu podjęli studia drugiego stopnia. Wyniósł on w przybliżeniu 69%, przez co możemy stwierdzić, że absolwenci Uniwersytetów oraz Politechnik przejawiają większe skłonności do kontynuowania nauki na drugim stopniu w przeciwieństwie do absolwentów Państwowych Wyższych Szkół, Akademii czy Wyższych Szkół. Osoby, które ukończyły szkoły w danej grupie były również najmniej skorzy do samozatrudnienia i zaledwie 3,51% z nich odważyło się na tę formę pracy. Dodatkowo w grupie tej występuje najniższy średni procent absolwentów, którzy mieli doświadczenie bycia bezrobotnym po uzyskaniu dyplomu i wynosi on 12,42%. Spośród wszystkich uczelni w zbiorze danych najmniejszą wartością mogła pochwalić się Szkoła Główna Handlowa w Warszawie z wynikiem 2,1%. Nie przełożyło się to jednak na średni wynik czasu od uzyskania dyplomu do podjęcia pierwszej pracy, ponieważ w ramach omawianej grupy wyniósł on najwięcej spośród wszystkich, czyli 4.53 miesiąca. Z kolei średnie miesięczne wynagrodzenie ze wszystkich źródeł po uzyskaniu dyplomu wynosi 2331,85 zł, które jest większe o 241,78 zł w porównaniu z grupą pierwszą. Zaistniała sytuacja może sugerować, że absolwenci Uniwersytetów i Politechnik stawiają pracodawcom większe wymagania dotyczące wynagrodzenia, w związku z czym okres poszukiwania pierwszej pracy wydłuża się średnio o ponad miesiąc w porównaniu z absolwentami Państwowych Wyższych Szkół Zawodowych.

Liczebność szkół w skupieniu trzecim jest równoważna skupieniu drugiemu, tj. czterdzieści osiem obserwacji. W grupie tej również przeważa określony rodzaj szkół wyższych i są nią Szkoły Wyższe, które stanowią 75% całego zbioru. Kolejne 25% są Innymi Wyższymi Szkołami do których zaklasyfikowaliśmy Akademie oraz Uczelnie. Analizując statystyki wyników analizy skupień można stwierdzić, że grupa ta została wyodrębniona ze względu na procent absolwentów,

którzy mieli doświadczenie samozatrudnienia po uzyskaniu dyplomu, który wyniósł aż 8,2%. Najwyższe średnie zarobki spośród wszystkich grup, które wyniosły 3030,49 zł oraz najniższy średni czas od uzyskania dyplomu do podjęcia pierwszej pracy, czyli 1,38 miesiąca, może świadczyć o tym, że absolwenci Szkół Wyższych studiów stopnia pierwszego są poszukiwanymi osobami przed pracodawców na rynku pracy. Z drugiej strony średnie wysokie zarobki mogą być powiązane z wysokim procentem absolwentów, którzy mieli doświadczenie samozatrudnienia. Uczelnie, które miały na bardzo wysokim poziomie ten drugi wskaźnik charakteryzowały się wynagrodzeniami absolwentów znacznie odbiegającymi od średniej. Przykładem jest Polsko-Japońska Akademia Technik Komputerowych, która procent samozatrudnienia przez absolwentów miała na poziomie 21,7% i średnimi wynagrodzeniami absolwentów w wysokości 4240,83 zł. Kolejnym przykładem jest Europejska Uczelnia Informatyczno-Ekonomiczna w Warszawie ze wskaźnikiem samozatrudnienia 12% i wynagrodzeniami 4787,79 zł. Statystycznym dowodem opisanej sytuacji jest umiarkowana zależność korelacji Pearsona na poziomie 0,57 między tymi dwoma zmiennymi. Na podstawie powyższych wyników można stwierdzić, że absolwenci Szkół Wyższych mają możliwość najszybszego podjęcia pierwszej pracy po uzyskaniu dyplomu, mogą pochwalić się również najwyższymi średnimi zarobkami oraz najczęściej decydują się na pracę w formie samozatrudnienia w porównaniu z innymi szkołami wyższymi.

Podsumowanie (Jakub Skotarek, Krzysztof Sukiennicki)

Monitorowanie ekonomicznych losów absolwentów jest bardzo ważnym procesem, pozwalającym na stałe poprawianie jakości kształcenia i lepsze dopasowanie studentów do sytuacji panującej na rynku pracy. Dzięki regularnie prowadzonym badaniom i coraz większej bazie wyników możliwa jest również analiza trendów nie tylko dla pojedynczych jednostek edukacyjnych, ale również dla całych zbiorowości.

Wykorzystanie w pracy analizy głównych składowych oraz analizy skupień pozwoliło zrealizować cel, którym była segmentacja uczelni wyższych. Na podstawie badania wyodrębniono trzy skupienia, w większości zawierających uczelnie jednego typu. Dzięki przeprowadzonemu badaniu obliczono poziom bezrobocia dla każdej z grup, a także okres pomiędzy ukończeniem studiów przez absolwentów, a rozpoczęciem przez nich pracy. Dla poszczególnych skupień przedstawiono również liczbę absolwentów kontynuujących naukę, a także średnie zarobki dla absolwentów badanych zbiorowości.

W pierwszym skupieniu znajdują się dwadzieścia dwie uczelnie. Jest to grupa, w której występują w znacznej większości Państwowe Wyższe Szkoły Zawodowe. Kierunek rozproszenia obserwacji należących do tego skupienia na wykresie skupień wskazuje, że są to uczelnie, po których absolwenci najczęściej rejestrowali się jako bezrobotni w Urzędach Pracy po uzyskaniu dyplomu. Jednak mimo iż grupa pierwsza cechowała się najwyższym średnim procentem absolwentów, którzy mieli doświadczenie bycia bezrobotnym to nie osiągnęła średniego najdłuższego czasu od uzyskania dyplomu do podjęcia pracy przez absolwentów. W tym przypadku średni czas wyniósł 3.49 miesiąca, co stanowi drugą pozycję wśród wszystkich skupień. W związku z powyższym z ekonomicznego punktu widzenia, jest to grupa szkół wyższych, która w świetle przedstawionych statystyk wypada najmniej korzystnie.

Skupienie drugie składa się w ponad 85% z Uniwersytetów oraz Politechnik i zawiera czter-

dzieści osiem obserwacji. W tym skupieniu znajduje się również Uniwersytet Ekonomicznzy w Poznaniu. Cechą charakterystyczną dla tej grupy jest wysoki średni procent absolwentów, którzy po uzyskaniu dyplomu podjęli studia drugiego stopnia. Z kolei średnie miesięczne wynagrodzenie ze wszystkich źródeł po uzyskaniu dyplomu wynosi 2331,85 zł, które jest większe o 241,78 zł w porównaniu z grupą pierwszą. Zaistniała sytuacja może sugerować, że absolwenci Uniwersytetów i Politechnik stawiają pracodawcom większe wymagania dotyczące wynagrodzenia, w związku z czym okres poszukiwania pierwszej pracy wydłuża się średnio o ponad miesiąc w porównaniu z absolwentami Państwowych Wyższych Szkół Zawodowych.

Liczebność szkół w skupieniu trzecim jest równoważna skupieniu drugiemu, tj. czterdzieści osiem obserwacji. W grupie tej również przeważa określony rodzaj szkół wyższych i są nią Szkoły Wyższe, które stanowią 75% całego zbioru. Kolejne 25% są Innymi Wyższymi Szkołami do których zaklasyfikowaliśmy Akademie oraz Uczelnie. Analizując statystyki wyników analizy skupień można stwierdzić, że grupa ta została wyodrębniona ze względu na procent absolwentów, którzy mieli doświadczenie samozatrudnienia po uzyskaniu dyplomu, który wyniósł aż 8,2%. Najwyższe średnie zarobki spośród wszystkich grup, które wyniosły 3030,49 zł. Na podstawie powyższych wyników można stwierdzić, że absolwenci Szkół Wyższych mają możliwość najszybszego podjęcia pierwszej pracy po uzyskaniu dyplomu, mogą pochwalić się również najwyższymi średnimi zarobkami oraz najczęściej decydują się na pracę w formie samozatrudnienia w porównaniu z innymi szkołami wyższymi.

Źródło danych stanowiły zbiory badania ELA, które zostały zaimplementowane z wykorzystaniem środowiska pakietów statystycznych w języku R. Wykorzystanie różnych metod badania pozwoliło na uzyskanie wartościowych i miarodajnych wyników, dzięki którym możliwa była właściwa kwalifikacja badanych ośrodków naukowych, a także wyciągnięcie z analizy prawidłowych wniosków.

Uzyskane interpretacja na podstawie opracowanych wyników może być cenną wskazówką przy wyborze rodzaju oraz profilu uczelni.

Bibliografia

Ławrynowicz, M. & Michoń, P. (2011). Badanie losów absolwentów. Metodologia i Metodyka. Poznań.

Balicki, A. (2013). *Statystyczna analiza wielowymiarowa i jej zastosowania społeczno-ekonomiczne*. Gdańsk: Wydawnictwo Uniwersytetu Gdańskiego.

Biecek, P. & Trajkowski, K. (2011). Na przełaj przez Data Mining.

Bożykowski, M., Dwórznik, M., Giermanowska, E., Izdebski, A., Jasiński, M., Konieczna-Sałamatin, J., … Zając, T. (2014). *Monitorowanie losów absolwentów uczelni wyższych z wykorzystaniem danych administracyjnych zakładu ubezpieczeń społecznych*. Instytut Badań Edukacyjnych. Warszawa.

Czopek, A. (2013). Analiza porównawcza efektywności metod redukcji zmiennych - analiza składowych głównych i analiza czynnikowa. *Studia Ekonomiczne / Uniwersytet Ekonomiczny w Katowicach*.

Delignette-Muller, M. L. & Dutang, C. (2015). fitdistrplus: An R Package for Fitting Distributions. *Journal of Statistical Software*, *64*(4), 1–34.

Edusfera. (2019). Jak powstają rankingi uczelni. Dostęp z http://www.edusfera.pl/artykuly,46.php

Górniak, J. (1998). Analiza czynnikowa i analiza głównych składowych. Instytut Socjologii UJ.

Gaebel, M., Hauschildt, K., Muehleck, K. & Smidt, H. (2012). *Tracking Learners' and Graduates' Progression Paths TRACKIT*. Bruksela.

Gramacki, J. & Gramacki, A. (2009). Redukcja wymiarowości oraz wizualizacja danych wielowymiarowych z wykorzystaniem projektu R. Uniwersytet Zielonogórski.

HCPC - Hierarchical Clustering on Principal Components: Essentials. (2017).

Higher Education Statistics Agency. (2019). Badania losow absolwentow UK.

INCHER-Kassel. (N.d.). Research Area Students and Graduates. Dostęp z https://www.uni-kassel.de/einrichtungen/en/incher/research/research-area-students-and-graduates.html

Jasiński, M., Bożykowski, M., Zając, T., Chłoń-Domińczak, A. & Żołtak, M. (2017). Who gets a job after graduation? Factors affecting the early career employment chances of higher education graduates in Poland.

Kodali, T. (2017). Hierarchical Clustering in R.

Macioł, S., Miniewicz, B. & Moskalewicz-Ziółkowska, E. (2013). Monitorowanie losów absolwentów SGH. Warszawa.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M. & Hornik, K. (2019). *cluster: Cluster Analysis Basics and Extensions*. R package version 2.0.9 — For new features, see the 'Changelog' file (in the package source).

MNiSW. (2019a). Ekonomiczne Losy Absolwentów. Dostęp z https://www.gov.pl/web/nauka/ekonomiczne-losy-absolwentow

MNiSW. (2019b). Materiały informacyjne na temat zasobów udostępnianych w ramach trzeciej edycji ogólnopolskiego systemu monitorowania Ekonomicznych Losów Absolwentów szkół wyższych(ELA). Dostęp z https://ela.nauka.gov.pl/includes/pdf/opis_raportow.pdf

Morzy, T. (2013). Eksploracja danych. Warszawa: Wydawnictwo Naukowe PWN.

Nowak-Brzezińska, A. (2012). Analiza skupień.

Pałys, K., Pałys, M., Prokopowicz, M. & Sykuła, A. (2013). Absolwent Uniwersytetu Ekonomicznego we Wrocławiu – metodyka badań losów zawodowych absolwentów. Wrocław.

Pathak, M. (2018). Hierarchical Clustering in R.

Perspektywy. (2019). Klucz do lepszej przyszłości. Dostęp z http://www.perspektywy.pl/RSW2018/o-rankingu

Prawo o szkolnictwie wyższym: Dz. U. 2005 Nr 164 poz. 1365. (2005). Prawo o szkolnictwie wyższym: Dz. U. 2005 Nr 164 poz. 1365.

R Core Team. (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.

Statistical Tools For High-Throughput Data Analysis. (2019). Hierarchical Clustering Essentials - Unsupervised Machine Learning. Dostęp z http://www.sthda.com/english/wiki/print.php?id=237#agnes-and-diana-functions

Uniwersytet Śląski w Katowicach. (N.d.). Badanie Losów zawodowych absolwentów Uniwersytetu Śląskiego.

Uniwersytet Śląski w Katowicach. (2019). Badanie losów zawodowych absolwentów UŚ. Dostęp z https://student.us.edu.pl/badanie-losow-zawodowych-absolwentow-us

Uniwersytet Ekonomiczny w Poznaniu. (2019). Badanie losów absolwentów. Zespół ds. Badania Losów Absolwentów.

Spis tabel

Spis rysunków

Spis Programów