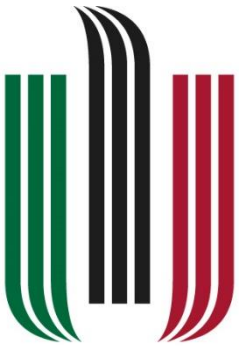


AKADEMIA GÓRNICZO-HUTNICZA IM. STANISŁAWA STASZICA W KRAKOWIE

WYDZIAŁ ZARZĄDZANIA

Zastosowanie metod wielowymiarowej analizy porównawczej do oceny stanu środowiska w województwach Polski

Autor: *Izabela Guła*

Kierunek: *Informatyka i ekonometria (stacjonarne)*

Przedmiot: *Statystyczna analiza danych*

Kraków, 2024

Wstęp

Ocena stanu środowiska naturalnego w Polsce jest istotnym elementem polityki zrównoważonego rozwoju, mającym na celu monitorowanie i poprawę jakości życia mieszkańców. Różnice w poziomie zanieczyszczeń, gospodarce odpadami czy emisjach szkodliwych substancji sprawiają, że konieczna jest kompleksowa analiza danych środowiskowych na poziomie regionalnym. Jednym z narzędzi, które umożliwia takie badania, jest wielowymiarowa analiza porównawcza. Dzięki zastosowaniu metod takich jak analiza skupień oraz porządkowanie liniowe, możliwe jest nie tylko porównanie poszczególnych województw, ale także grupowanie ich ze względu na podobieństwo w zakresie wybranych zmiennych diagnostycznych. Celem niniejszej pracy jest wykorzystanie metod wielowymiarowej analizy porównawczej do oceny stanu środowiska we wszystkich województwach Polski. Badanie to bazuje na publikacji naukowej „Zastosowanie metod wielowymiarowej analizy porównawczej do oceny stanu środowiska w województwie dolnośląskim", a także na najnowszych dostępnych danych statystycznych. Analiza porządkowania liniowego zostanie przeprowadzona z użyciem dwóch podejść: wzorcowego i bez wzorcowego, co pozwoli na pełniejsze zrozumienie zróżnicowania regionalnego w Polsce pod kątem wyzwań związanych ze środowiskiem naturalnym. W ramach projektu zostaną zastosowane różnorodne metody analizy skupień (grupowanie podziałowe i hierarchiczne), aby pogrupować województwa na podstawie ich cech środowiskowych. Dodatkowo, techniki porządkowania liniowego pozwolą na stworzenie rankingów województw, co umożliwi zidentyfikowanie regionów o najlepszej i najgorszej kondycji środowiskowej. W ostateczności, praca dąży do wyciągnięcia wniosków dotyczących sytuacji środowiskowej oraz polityki środowiskowej na poziomie regionalnym.

Analiza danych

Zestawienie zmiennych obrazujących stan i ochronę środowiska w województwach Polski w 2022 roku

Zmienne	Opis zmiennej	Charakter zmiennej
X1	Powierzchnia lasów gminnych na 100 km^2 [ha]	Stymulanta
X2	Emisja zanieczyszczeń gazowych [t/r]	Destymulanta
X3	Zużycie wody na potrzeby przemysłu [dam^3]	Destymulanta

X4	Nakłady na środki trwałe służące gospodarce wodnej [tys. zł]	Stymulanta
X5	Nakłady na środki trwałe służące ochronie środowiska [tys. zł]	Stymulanta
X6	Powierzchnia obszarów prawnie chronionych [ha]	Stymulanta
X7	Masa utworzonych odpadów komunalnych na 1 mieszkańca [kg]	Destymulanta
X8	Odpady wytworzone i dotychczas składowane poza komunalnymi [tys. ton]	Destymulanta
X9	Ścieki komunalne oczyszczane na 100 km^2	Destymulanta
X10	Liczba pomników przyrody na 100 km2	Stymulanta

Zestawienie zmiennych obrazujących stan i ochronę środowiska w województwach Polski w 2022 roku obejmuje zarówno stymulanty, jak i destymulanty, co pozwala na kompleksową ocenę sytuacji ekologicznej. Powierzchnia lasów gminnych X1 oraz powierzchnia obszarów prawnie chronionych X6 działają jako stymulanty, pozytywnie wpływając na środowisko. Emisja zanieczyszczeń gazowych X2 i masa odpadów komunalnych na mieszkańca X7 to destymulanty, które wskazują na problemy środowiskowe, obciążające ekosystem. Nakłady na ochronę środowiska X5 i gospodarkę wodną X4 są z kolei kluczowymi inwestycjami stymulującymi poprawę warunków ekologicznych w regionach. W późniejszej części badania destymulanty zostaną zamienione na stymulanty.

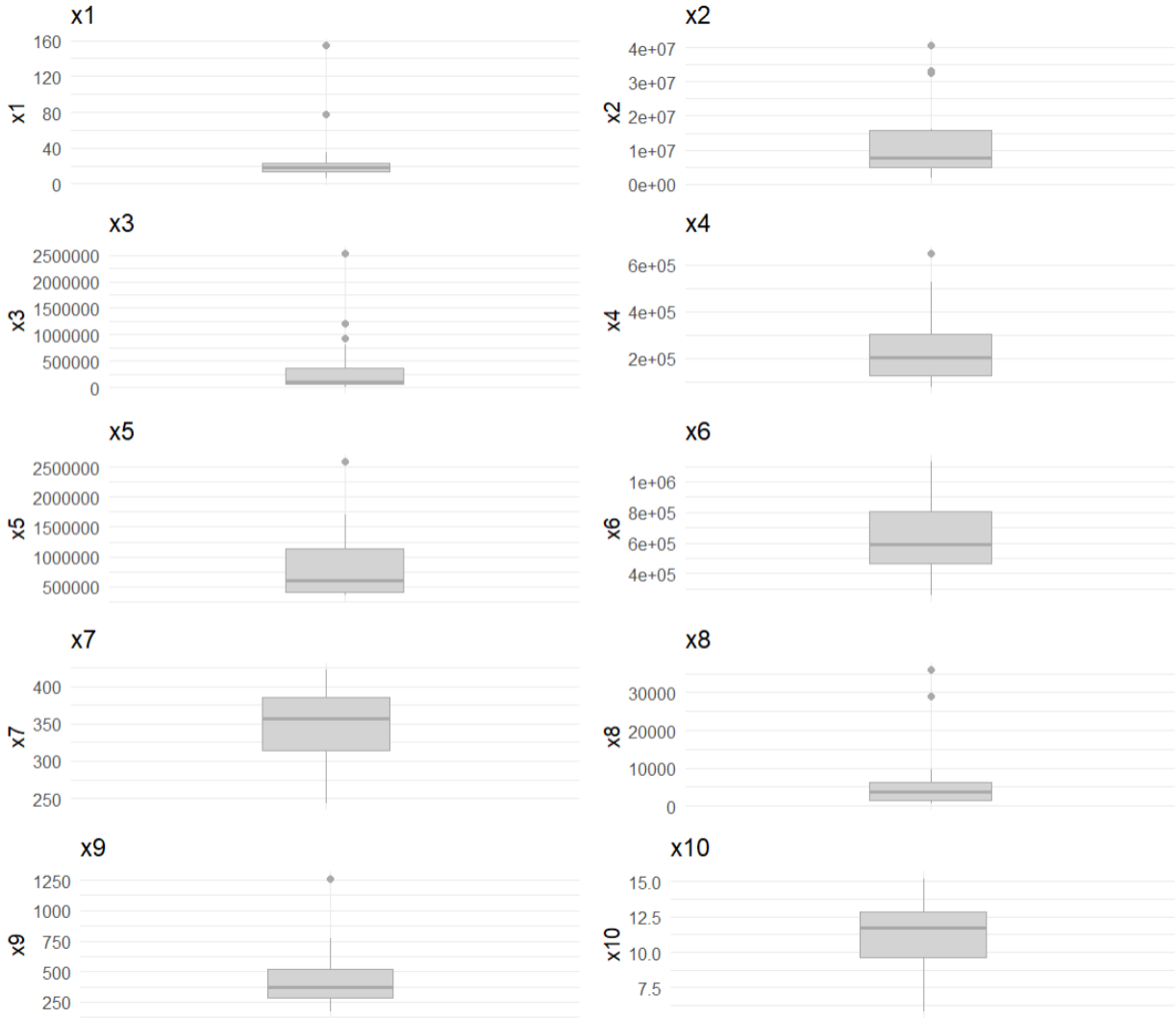
Statystyki opisowe

Statistic	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10
Median	18.5	7560000.0	101000.0	205000.0	597000.0	592000.0	357.5	3770.0	375.0	12.0
Mean	29.438	12700000.0	405000.0	249000.0	870000.0	632000.0	348.063	7190.0	448.063	11.125
Std. Deviation	37.587	12400000.0	678000.0	166000.0	628000.0	260000.0	55.521	10300.0	276.149	2.63
Variance	1410.0	153000000000000.0	460000000000.0	27600000000.0	394000000000.0	67700000000.0	3080.0	106000000.0	76300.0	6.92
Kurtosis	9.031	0.616	6.255	0.832	2.526	-0.522	-0.657	4.463	4.359	-0.4
Std. Error of Kurtosis	1.091	1.091	1.091	1.091	1.091	1.091	1.091	1.091	1.091	1.091
Minimum	6.0	1610000.0	10800.0	79100.0	358000.0	260000.0	243.0	655.0	170.0	-4.0
Maximum	155.0	40800000.0	2530000.0	647000.0	2590000.0	1130000.0	422.0	35900.0	1260.0	15.0
25th percentile	13.5	4850000.0	72000.0	126000.0	414000.0	469000.0	314.0	1610.0	284.75	9.75
50th percentile	18.5	7560000.0	101000.0	205000.0	597000.0	592000.0	357.5	3770.0	375.0	12.0
75th percentile	22.75	15700000.0	370000.0	305000.0	1140000.0	803000.0	385.5	6150.0	518.75	13.0
Coeff. of Variation	1.2768	0.97286	1.6755	0.66842	0.72115	0.41166	0.15997	1.4293	0.61727	0.24528

Powyższa tabela przedstawia wybrane statystyki opisowe dotyczące wybranych zmiennych. Dla każdej zmiennej obliczona została: średnia, mediana, odchylenie standardowe, wariancja, kurtoza, błąd standardowy kurtozy, wartość minimalna, wartość maksymalna, percentyle oraz współczynnik zmienności. Zmienne nie mają brakujących wartości. Wszystkie zmienne poza X6, X7 i X10 posiadają rozkład platykurtyczny, który sugeruje spłaszczenie rozkładu. Reszta zmiennych charakteryzuje się rozkładem leptokurtycznym, który oznacza skupienie wartości wokół średniej. Wszystkie zmienne poza X6, X7 i X10 mają wysoką wartość współczynnika zmienności, co może sugerować, że zmienne

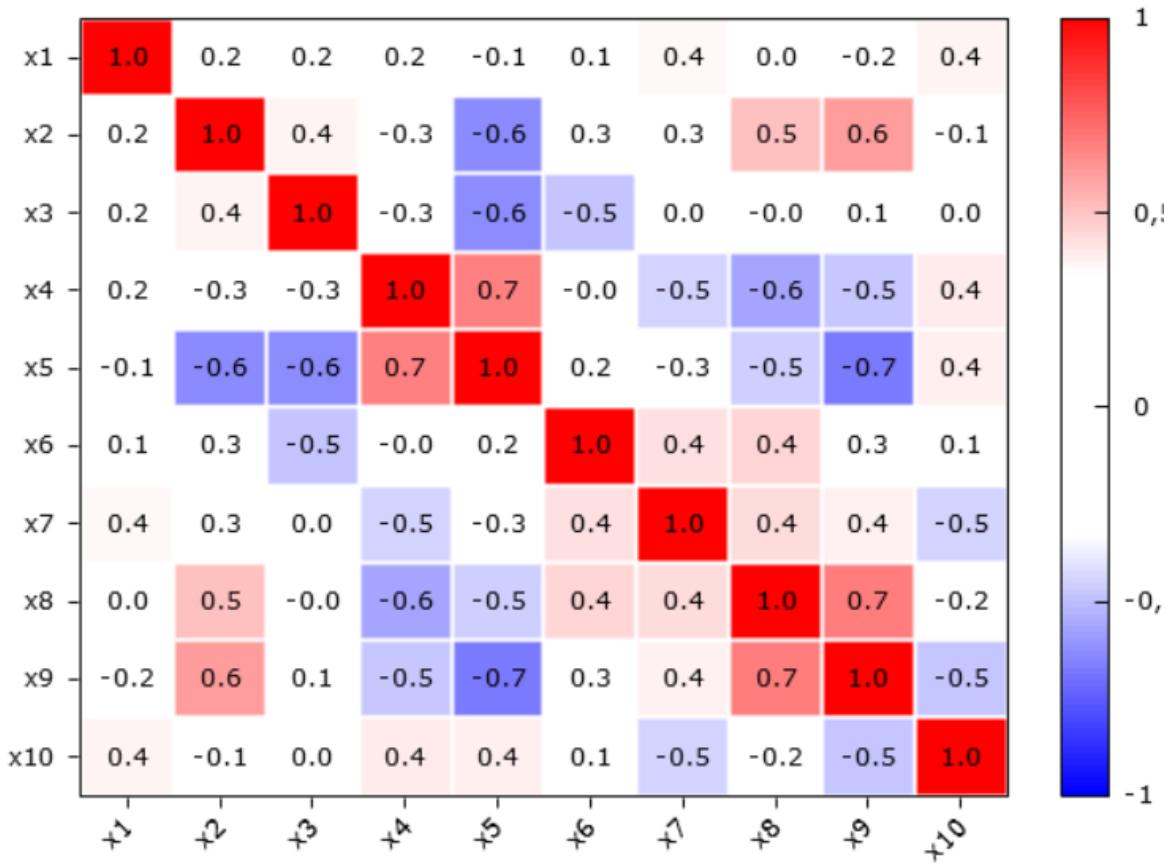
wykazują dużą rozbieżność, co oznacza, że wartości są bardziej rozproszone wokół średniej. Wysoka zmienność może świadczyć o niestabilności lub większym ryzyku w analizowanych danych.

Wykresy pudełkowe



Powyższe wykresy przedstawiają rozkład statystyczny każdej ze zmiennych. Niektóre ze zmiennych mają wartości odstające, w postaci jednego lub dwóch województw. Po sprawdzeniu okazuje się, że w przypadku każdej zmiennej wartości odstające są dla innych województw. Wszystkie zmienne mogą okazać się bardzo istotne w badaniu, dlatego wszystkie zmienne i obserwacje zostają zachowane, mając nadzieje, że nie zaburzy to późniejszego badania.

Korelacja między zmiennymi



Macierz korelacji przedstawia, że korelacje między wszystkimi zmiennymi są niskie. Jest to pożądana informacja, gdyż oznacza to, że nie ma problemu współliniowości, co mogłoby prowadzić do zniekształceń wyników w modelach regresyjnych. Niska korelacja sugeruje, że zmienne nie są silnie powiązane ze sobą, co oznacza, że każda zmienna wnosi unikalną informację do modelu. Dodatkowo brak współliniowości zwiększa stabilność oszacowań parametrów oraz poprawia predykcyjną moc modelu.

Porządkowanie liniowe

Porządkowanie liniowe to metoda, która pozwoli na uszeregowanie województw według wartości cechy lub wskaźnika, umożliwiając ich porównanie pod względem stanu środowiska. Dzięki tej technice można zidentyfikować regiony o najkorzystniejszych i najmniej korzystnych warunkach środowiskowych, co ułatwia analizę przestrzenną i podejmowanie decyzji dotyczących polityki ekologicznej. W tym badaniu wykorzystane zostaną dwie metody: metoda Hellwiga oraz metoda sumy rang. Metoda Hellwiga polega na ocenie obiektów poprzez porównanie ich do tzw. wzorca idealnego, czyli teoretycznego obiektu o najlepszych możliwych wartościach cech. Metoda sumy rang polega na przypisaniu rang każdej zmiennej w zależności od wartości, a następnie obliczeniu średniej z tych rang

dla każdego obiektu. W przypadku każdej z nich przedstawione zostaną wyniki porządkowania liniowego, ranking oraz klasyfikacja województw w formie graficznej.

Wynik porządkowania liniowego województw polski

Województwo	Metoda porządkowania liniowego	
	Hellwiga	Sumy rang
dolnośląskie	0,19	7,7
kujawsko – pomorskie	0,24	9,1
lubelskie	0,17	7,8
lubuskie	0,21	9,4
łódzkie	0,14	6,8
małopolskie	0,47	10,3
mazowieckie	0,23	7,5
opolskie	0,13	7,4
podkarpackie	0,49	12,1
podlaskie	0,21	9,3
pomorskie	0,25	9
śląskie	0,06	6,7
świętokrzyskie	0,16	6,6
warmińsko - mazurskie	0,23	9,5
wielkopolskie	0,37	11
zachodniopomorskie	0,17	6,8

W przypadku obu metod należy zwrócić szczególną uwagę na skalę, w jakiej są podawane wyniki, gdyż ma ona kluczowe znaczenie dla interpretacji otrzymanych wartości i porównywalności między analizowanymi obiektami. W metodzie sumy rang wyniki opierają się na przypisaniu odpowiednich rang, które nie są ograniczone do konkretnego przedziału, podczas gdy w metodzie Hellwiga skala wyników również nie jest jednoznacznie zdefiniowana. W obu przypadkach wyższe wartości oznaczają województwa wyróżniające się najlepszym stanem ochrony środowiska.

Ranking województw na podstawie metod porządkowania liniowego

Pozycja w rankingu	Metoda Hellwiga	Metoda sumy rang
1	podkarpackie	podkarpackie

2	małopolskie	wielkopolskie
3	wielkopolskie	małopolskie
4	pomorskie	warmińsko - mazurskie
5	kujawsko - pomorskie	lubuskie
6	warmińsko - mazurskie	podlaskie
7	mazowieckie	kujawsko - pomorskie
8	lubuskie	pomorskie
9	podlaskie	lubelskie
10	dolnośląskie	dolnośląskie
11	lubelskie	mazowieckie
12	zachodniopomorskie	opolskie
13	świętokrzyskie	łódzkie
14	łódzkie	zachodniopomorskie
15	opolskie	śląskie
16	śląskie	świętokrzyskie

Wyniki wykazują niewielkie zróżnicowanie, jednak w obu rankingach te same województwa zajmują czołowe pozycje. Pozostałe województwa pojawiają się w podobnych miejscach, choć z drobnymi różnicami. Można stwierdzić, że obie metody skutecznie stworzyły wiarygodny ranking porządkowania liniowego, co potwierdza również wysoki współczynnik korelacji Tau Kendalla, wynoszący 0,69. Kolejnym krokiem jest pogrupowanie województw w odpowiednie grupy. Klasyfikację województw do odpowiednich klas przeprowadzono na podstawie analizy kwartylowej, co pozwoliło na wyodrębnienie czterech kategorii. W każdej klasie znajdują się wyniki porządkowania liniowego odpowiedniej metody. W pierwszej od zera do Q1, w drugiej grupie od Q1 do Q2, w trzeciej od Q2 do Q3, a w czwartej pozostałe wartości. Każda z tych kategorii odzwierciedla różne poziomy jakości ochrony środowiska, przy czym województwa przypisane do pierwszych klas odznaczają się lepszym ogólnym stanem ochrony środowiska. Taki sposób klasyfikacji umożliwia łatwiejsze zrozumienie zróżnicowania w stanie ochrony środowiska na obszarze całego kraju oraz wskazanie regionów, które wymagają szczególnej uwagi w działaniach proekologicznych.

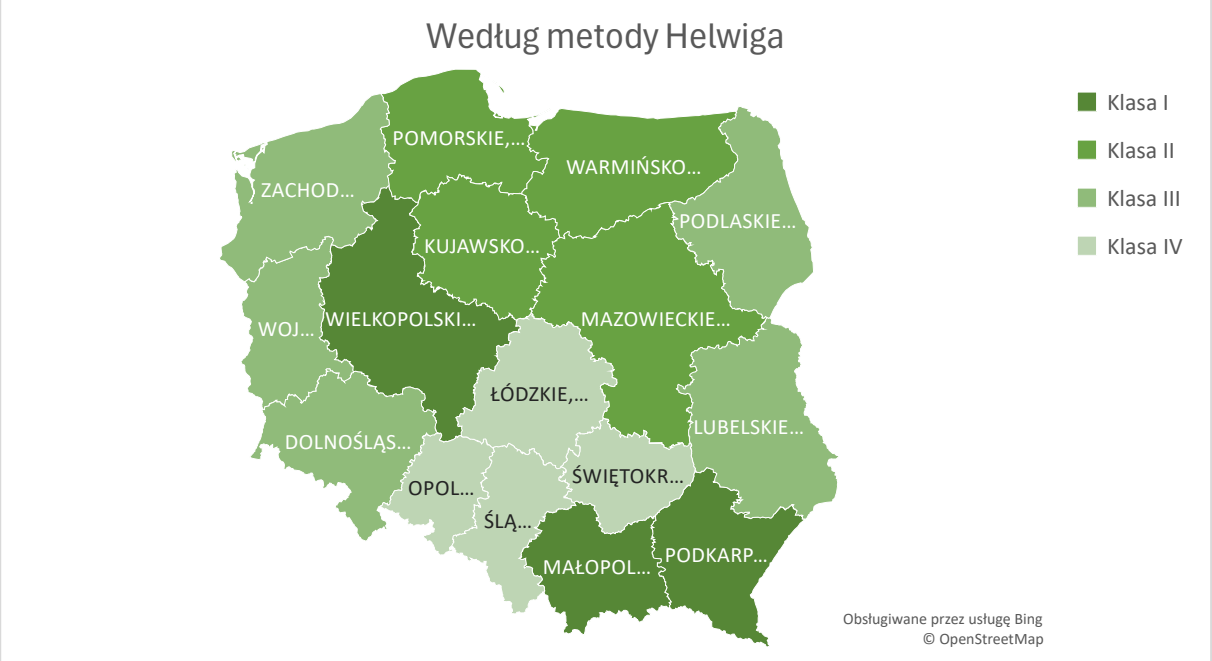
Zestawienie klasyfikacji województw

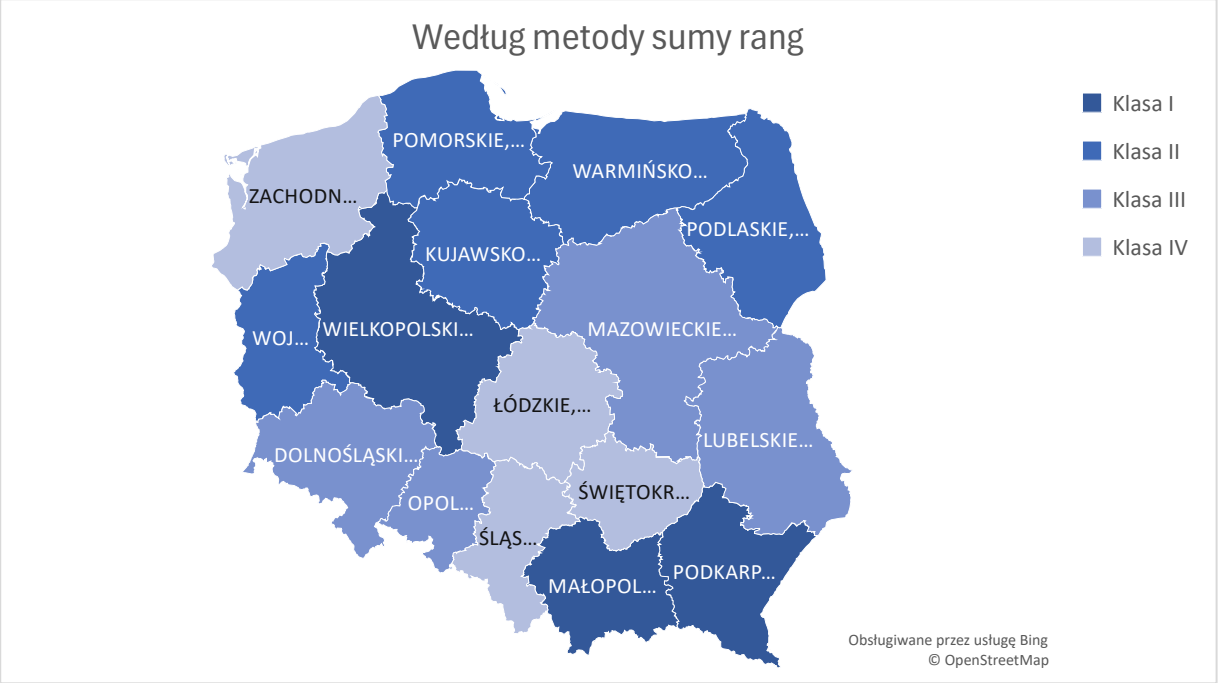
Klasa	Hellwig	Suma rang
I	podkarpackie, małopolskie, wielkopolskie	podkarpackie, wielkopolskie, małopolskie

II	pomorskie, kujawsko – pomorskie, warmińsko – mazurskie, mazowieckie	warmińsko – mazurskie, lubuskie, podlaskie, kujawsko – pomorskie, pomorskie
III	lubuskie, podlaskie, dolnośląskie, lubelskie, zachodniopomorskie	lubelskie, dolnośląskie, mazowieckie, opolskie
IV	świętokrzyskie, łódzkie, opolskie, śląskie	łódzkie, zachodniopomorskie, śląskie, świętokrzyskie

Na poniższych mapach przedstawiono klasyfikację województw w Polsce w kontekście stanu ochrony środowiska. Intensywniejsze kolory na mapie wskazują na klasy, które charakteryzują się lepszymi warunkami ochrony środowiska, mniej intensywne kolory to województwa posiadające gorsze warunki ochrony środowiska. Takie graficzne przedstawienie wyników pozwala lepiej zobrazować wyniki stanu ochrony środowiska we wszystkich województwach.

Stan ochrony środowiska w województwach polski w 2022 roku według wybranych metod porządkowania liniowego



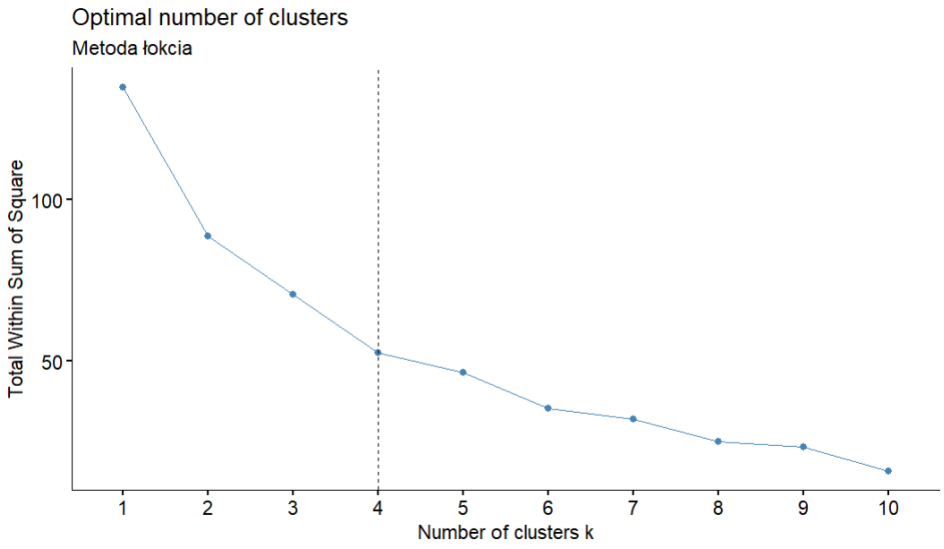


Analiza skupień

Analiza skupień jest to proces przypisywania obiektów do grup na podstawie analizy struktury danych. Założeniem jest, że obiekty w skupieniach wykazują tendencję do wzajemnego podobieństwa, a obiekty w różnych skupieniach wykazują tendencję do niepodobieństwa. Istnieją dwie metody klasyfikowania, pierwszą z nich jest grupowanie podziałowe. Ten rodzaj grupowania polega na podziale zbioru obiektów na określoną przed początkiem badania liczbę K rozłącznych skupień. W tym przypadku zostanie wykorzystana metoda k – średnich, jest to popularna technika, która jest używana do grupowania danych w klastry. Działa to tak, że najpierw musi nastąpić wybór liczby klastrów, a następnie algorytm losowo wybiera punkty, które będą reprezentować centra tych klastrów. Następnie przypisuje każdy punkt danych do najbliższego centrum, a potem aktualizuje położenie centrów na podstawie średnich wartości punktów w każdym klastrze. Proces powtarza się, aż centra przestaną się zmieniać, co oznacza, że grupowanie osiągnęło stabilność. Drugą metodą analizy skupień, jaka zostanie wykorzystana, jest grupowanie hierarchiczne. Grupowanie hierarchiczne to technika, która tworzy drzewo klastrów, łącząc dane w sposób hierarchiczny, gdzie każdy punkt początkowo stanowi oddzielny klaster, a następnie grupy są łączone na podstawie ich podobieństwa, aż do uzyskania jednego, zbiorczego klastra lub określonej liczby klastrów. Wykorzystana zostanie jedna z metod grupowania hierarchicznego, metoda Warda dla jednej z wybranych funkcji odległości, która łączy klastry w taki sposób, aby minimalizować wariancję pomiędzy punktami w klastrach a ich centroidami, co prowadzi do bardziej zwartych i jednorodnych grup danych.

Grupowanie podziałowe

Mając na uwadze wrażliwość grupowania podziałowego na wartości odstające, do obu metod grupowania wykorzystana została metoda ograniczenia wartości zmiennych skrajnych do wartości górnego lub dolnego wąsa. Celem tego jest uzyskanie jak najbardziej podobnych i porównywalnych wyników analizy skupień. Pierwszym ważnym krokiem jest wybór odpowiedniej liczby klastrów.



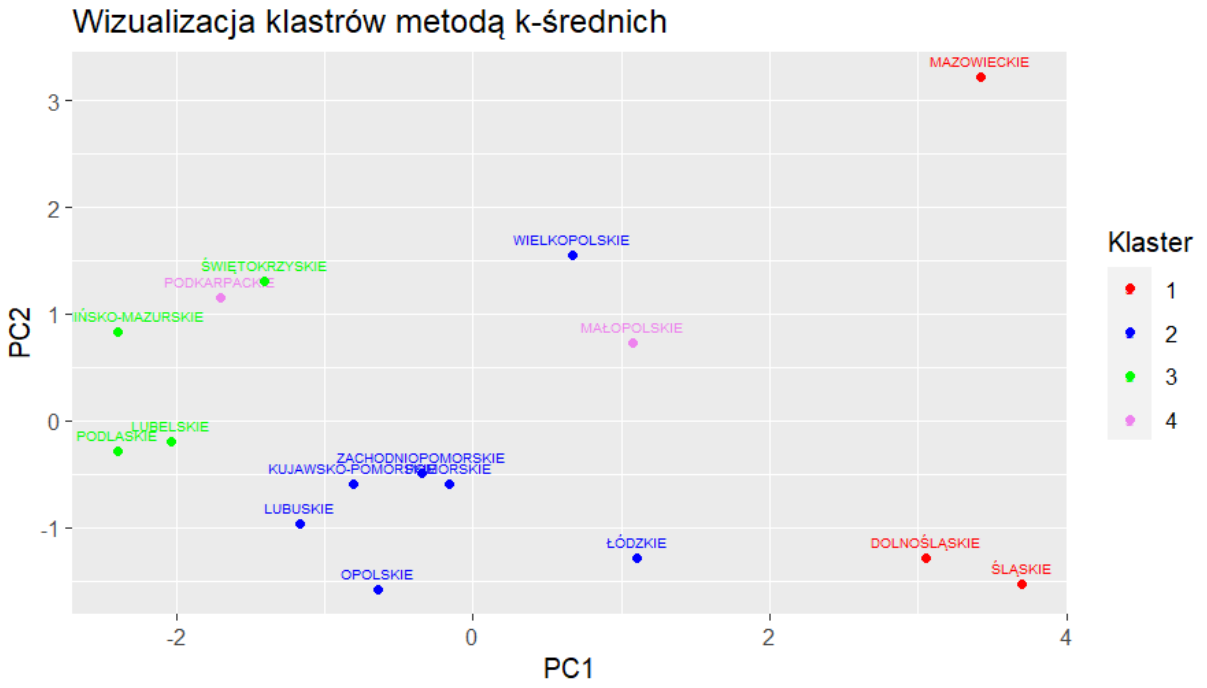
Istnieje wiele metod, które pozwalają wybrać najefektywniejszą liczbę, W tej analizie wykorzystana została metoda łokcia. Metoda wskazuje najlepszą liczbę klastrów w miejscu, w którym osiąga pewne załamanie, będące łokciem. W tym wypadku pierwsze załamanie występuje dla dwóch klastrów, jednak jest to za mała liczba, dlatego zostaje wybrana liczba cztery.

Wyniki metody k – średnich

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
Klaster 1	-0,14	-1,30	-0,51	1,42	1,55	-0,25	-0,77	-1,62	-1,40	0,43
Klaster 2	-0,30	0,06	0,09	-0,22	-0,25	0,39	-0,59	0,33	0,25	0,22
Klaster 3	-0,53	0,54	0,02	-0,92	-0,68	0,55	1,10	0,39	0,90	-1,17
Klaster 4	2,32	0,68	0,39	0,48	-0,09	0,66	1,05	0,46	-0,56	0,92

Wyniki metody k-średnich pokazują, że klaster pierwszy charakteryzuje się znaczącymi wartościami dodatnimi dla zmiennych X4 i X5, co może sugerować, że obiekty w tym klastrze mają wyraźnie wyższe wyniki w tych kategoriach w porównaniu do pozostałych klastrów. Klaster 2, zbliżony do średnich wartości w większości zmiennych, może wskazywać na grupę obiektów o umiarkowanych cechach, co może być korzystne w kontekście analiz, gdzie potrzebne są zrównoważone profile. Z kolei

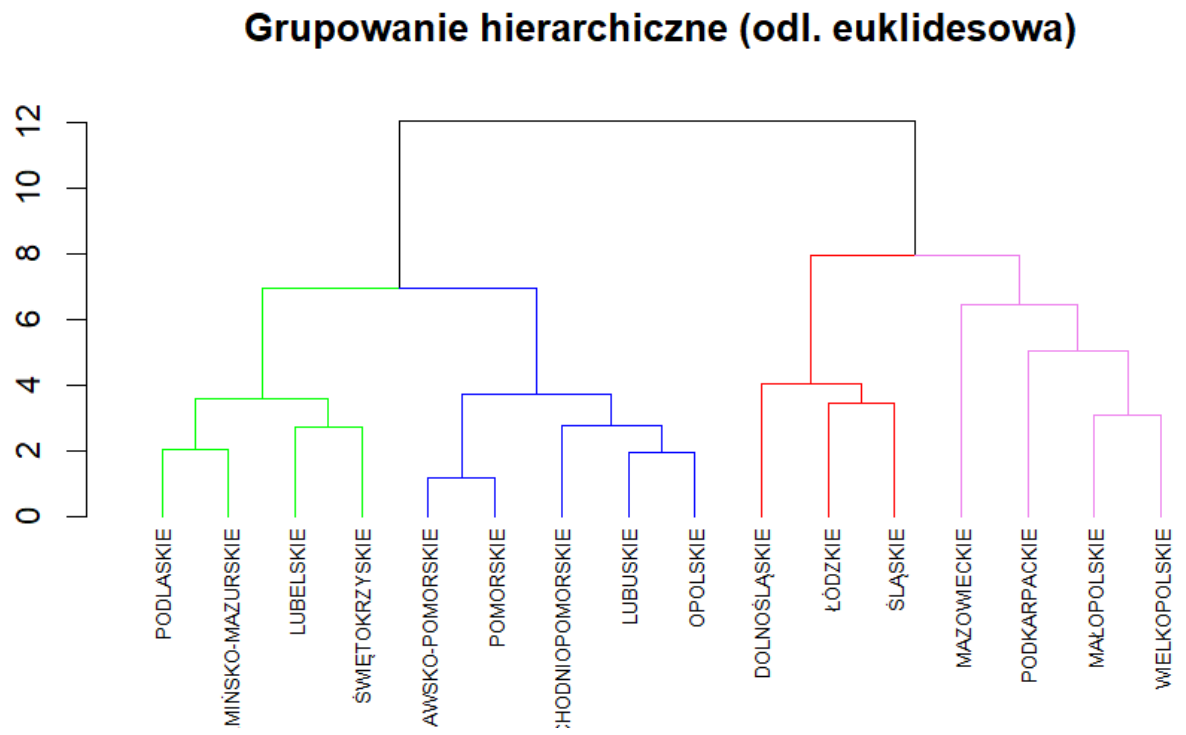
klaster 4 wyróżnia się zdecydowanie wysokimi wartościami w zmiennej X1, co może sugerować, że obiekty w tym klastrze są istotnie różne od pozostałych, a ich unikalne cechy mogą mieć znaczący wpływ na dalsze analizy i decyzje strategiczne.



Kolejny krok to przejście do stworzenia wizualizacji stworzonych klastrów. W celu wizualizacji wykorzystana została metoda analizy głównych składowych, która redukuje wymiarowość danych, przekształcając oryginalne zmienne w nowe, które zachowują najwięcej informacji o wariancji. Oś X reprezentuje największą wariancję, a oś Y – drugą co do wielkości wariancję, umożliwiając wizualizację rozmieszczenia obiektów i identyfikację klastrów w przestrzeni tych składowych. Wykres przedstawia rozmieszczenie wszystkich czterech klastrów. Można zauważyć, że nie są one równowartościowe, jednak jest to najlepsza możliwa opcja. Klaster pierwszy zawiera trzy obiekty i są rozmieszczone po prawej części wykresu. Klaster drugi posiada najwięcej obiektów, jest ich aż siedem. Położone są w centralnej części wykresu, rozciągając się od środka do dołu. Klaster trzeci posiada cztery obiekty i położone są w lewej części wykresu. Podobne położenie jest w klastrze czwartym, gdzie są dwa obiekty. Ciekawą rzeczą jest to, że obiekty klastra trzeciego i czwartego nachodzą się, co oznacza, że mogą być do siebie podobne pod względem cech, co może sugerować, że te klastry są bliskie sobie lub że istnieją wspólne atrybuty, które mogą prowadzić do ich częściowego nałożenia. Klastry w tym wypadku są ważne, ponieważ pozwalają na grupowanie obiektów o podobnych cechach. Dzięki temu łatwiej zrozumieć i analizować dane, zwłaszcza w przypadku dużych zbiorów.

Grupowanie hierarchiczne

W celu stworzenia grupowania hierarchicznego należy wybrać metodę obliczania oraz odległość, z jakiej będzie się korzystać. W tym projekcie obliczane są odległości euklidesowe między obserwacjami w zestandaryzowanych danych, a następnie stosowana jest metoda Warda, jest uznawana za skuteczną technikę grupowania. W rezultacie powstaje dendrogram ilustrujący struktury klastrowe, a obserwacje są przypisywane do czterech grup. Do wyboru odpowiedniej liczby skupień, wykorzystana została naturalna metoda dedukcji sprawdzająca wybór podziału, dla którego między poziomem k, a k+1, jest dostatecznie duża różnica na dendrogramie. Dla liczby skupień równej pięć powstawała nowa grupa zawierająca dwa województwa, ale pozostawiała jedna grupa z jednym województwem. W przypadku liczby trzy, jedna grupa wychodziła z o wiele większą liczba województw niż pozostałe. W przypadku zwiększenia liczby grup, nie były to dostatecznie duże różnice na dendrogramie, dlatego zachowana została liczba skupień wynosząca cztery.



Powyższy wykres ilustruje wyniki grupowania hierarchicznego opartego na odległości euklidesowej, zastosowanego metodą Warda. Każda z powstałych grup została oznaczona innym kolorem, co ułatwia wizualne rozróżnienie klastrów i interpretację wyników. Klaster pierwszy został oznaczony kolorem czerwonym, drugi niebieskim, trzeci zielonym, a czwarty fioletowym. Etykiety obiektów zostały odpowiednio przypisane do liści dendrogramu, co pozwala na dokładną identyfikację elementów w poszczególnych grupach.

Interpretacja

Porządkowanie liniowe

Analizując wyniki porządkowania liniowego oraz ranking województw opracowany za pomocą metody Hellwiga i metody sumy rang, zauważamy wysokie podobieństwo obu klasyfikacji. W obu przypadkach te same województwa zajmują czołowe miejsca, z wyjątkiem zamiany miejsc drugiego i trzeciego. W dalszej części rankingu występują pewne podobieństwa, jednak końcowe pozycje różnią się nieznacznie. Takie zbieżności można wyjaśnić silną dodatnią korelacją, której wartość współczynnika Tau Kendalla wynosi 0,69. To wskazuje na wyraźną tendencję, że wzrost jednej zmiennej idzie w parze ze wzrostem drugiej. Mimo że korelacja nie jest idealna, sugeruje ona silny związek, który wskazuje na podobne trendy między analizowanymi zmiennymi.

W celu głębszej analizy warto przyjrzeć się województwom zajmującym zarówno wysokie, jak i niskie pozycje w rankingu, aby zrozumieć, co wpływa na ich wyniki. Województwo podkarpackie zajmuje najwyższą pozycję w obu rankingach. Znane z malowniczych Bieszczad, województwo to charakteryzuje się największą powierzchnią lasów gminnych na 100 km² spośród wszystkich województw. Lasy te przyczyniają się do usuwania dwutlenku węgla z atmosfery, co wpływa na poprawę jakości powietrza. Podkarpacie nie musi również obawiać się zanieczyszczeń powietrza, gdyż posiada jedną z najniższych emisji gazów cieplarnianych w Polsce. Dodatkowo region ten cechuje się wysoką powierzchnią obszarów prawnie chronionych, w tym Bieszczadzkim Parkiem Narodowym, trzecim co do wielkości parkiem narodowym w kraju. Warto również zauważyć, że województwo to ma najniższy wskaźnik urbanizacji (41,09%), co może prowadzić do mniejszej emisji zanieczyszczeń w porównaniu do bardziej zurbanizowanych regionów.

Kolejne w rankingu jest województwo małopolskie. Jego wysoka pozycja może wydawać się zaskakująca, biorąc pod uwagę problemy z jakością powietrza, jednak w 2022 roku emisja zanieczyszczeń gazowych nie była w nim najwyższa. Co więcej, województwo to plasuje się w czołówce pod względem inwestycji w ochronę środowiska i gospodarkę wodną, a także posiada największą liczbę parków narodowych w Polsce. Współczynnik urbanizacji wynosił tu 47,83%, co wskazuje na umiarkowany poziom zurbanizowania, a udział przemysłu w PKB (22,9%) również był jednym z niższych. Większa urbanizacja i przemysł mogą prowadzić do zwiększonej emisji zanieczyszczeń, co wpływa negatywnie na stan środowiska.

Zaskakująco wysoko w rankingu znalazło się również województwo wielkopolskie. Jego pozycja może być wynikiem wysokich nakładów na gospodarkę wodną oraz ochronę środowiska, oraz dużą powierzchnią obszarów prawnie chronionych, mimo że województwo to posiada tylko jeden park narodowy. Wielkopolska jest umiarkowanie zurbanizowanym regionem (53,3% w 2022 roku), a udział

przemysłu w PKB wyniósł 27,8%. Zajmowane wysokie miejsce w rankingu potwierdza, że region ten skutecznie inwestuje w ochronę środowiska.

Z kolei wśród województw zajmujących niższe pozycje, powtarzają się województwa śląskie, świętokrzyskie i łódzkie. Województwo świętokrzyskie w 2022 roku charakteryzowało się wysoką emisją zanieczyszczeń gazowych oraz dużym zużyciem wody przez przemysł. Mimo to udział przemysłu w PKB (27,5%) był niższy niż w innych regionach, co może sugerować, że przemysł w tym województwie jest bardziej zasobochłonny. Województwo świętokrzyskie jest także jednym z najmniej zurbanizowanych w Polsce (44,83%) i ma najniższe nakłady na ochronę środowiska.

Województwo łódzkie, bardziej zurbanizowane (61,75%), miało najwyższą emisję zanieczyszczeń gazowych oraz stosunkowo wysoki udział przemysłu w PKB (29,2%). Brak parków narodowych oraz niska powierzchnia lasów i obszarów chronionych również wpływają na jego niską pozycję w rankingu.

Najgorzej wypadło województwo śląskie, które ma jedną z najwyższych emisji zanieczyszczeń, wysoką produkcję odpadów przemysłowych oraz największy udział przemysłu w PKB (35,4%). Jest to także najbardziej zurbanizowany region w Polsce (75,89%), co wiąże się z dużym obciążeniem środowiskowym. Niemniej jednak, województwo to przeznacza wysokie nakłady na ochronę środowiska, co może wskazywać na próby minimalizowania negatywnych skutków działalności przemysłowej.

Dodatkowo wykresy przedstawiające podział województw według stanu ochrony środowiska wyraźnie ukazują przewagę województw z południowej Polski, które radzą sobie wyjątkowo dobrze. Warto jednak wspomnieć, że województwa takie jak łódzkie, świętokrzyskie, śląskie i zachodniopomorskie, położone odpowiednio w centralnej, południowo-wschodniej, południowo-zachodniej i północno-zachodniej części kraju, borykają się z poważnymi wyzwaniami w zakresie ochrony środowiska. Te regiony wymagają szczególnej uwagi i dodatkowych działań, aby poprawić stan środowiska i zbliżyć się do wyników osiąganych przez liderów rankingu.

Podsumowując, analiza wyników porządkowania liniowego potwierdza, że istnieje wiele czynników wpływających na pozycję województw w rankingu. Warto mieć na uwadze, że istnieją inne czynniki, które również mogą wpływać na stan ochrony środowiska w Polsce, jednak przedawniona zmienne w naprawdę dobry sposób opisują przedstawiony problem i pokrywają się z rzeczywistością.

Analiza skupień

Interpretacja analizy skupień rozpocznie się od grupowania podziałowego. Jak już wcześniej zostało wspomniane, do tej analizy została wykorzystana metoda k – średnich, a działanie metody zostało

opisane w badaniu. W obu metodach grupowania stworzone zostały cztery klastry, których zawartości przedstawia poniższa tabela.

Wyniki podziału analizy skupień

	Grupowanie podziałowe	Grupowanie hierarchiczne
Klaster 1	mazowieckie, dolnośląskie, śląskie	dolnośląskie, śląskie, łódzkie
Klaster 2	zachodniopomorskie, łódzkie, opolskie, lubuskie, kujawsko – pomorskie, pomorskie, wielkopolskie	Zachodniopomorskie, opolskie, lubuskie, kujawsko – pomorskie, pomorskie
Klaster 3	świętokrzyskie, warmińsko – mazurskie, lubelskie, podlaskie	warmińsko – mazurskie, podlaskie, lubelskie, świętokrzyskie
Klaster 4	małopolskie, podkarpackie	mazowieckie, wielkopolskie, podkarpackie, małopolskie

W przypadku każdego klastra zostały zaznaczone najwyższe i najniższe wartości średniej i odchylenia standardowego. Kolor czerwony oznacza najniższą średnią, zielony najwyższa. Kolor żółty oznacza najniższe odchylenie, czyli sytuacje, gdy wartości są skupione wokół średniej, niebieski oznacza natomiast największe rozproszenie danych.

Średnia i odchylenie standardowe zmiennych dla klastrów z grupowania podziałowego

	Średnia				Odchylenie standardowe			
	Klastry							
	1	2	3	4	1	2	3	4
X1	24,3	18	9,8	51,9	15,1	1,8	3,9	0,0
X2	27960728	11840825	6098402	4474800	10412221	11137455	6277321	3683281
X3	598746	297434	336483	151919	852519	397065	580948	99926
X4	479950	211771	96202	328217	179981	90606	17149	107095
X5	1774610	715427	444968	808263	584217	407134	143578	421966
X6	567012	529390	774778	804031	427078	198888	249311	3468
X7	391,0	381,3	286,8	290,0	32,0	20,2	27,5	66,5
X8	12936,8	3458,8	3147,3	2831,1	4552,6	3122,1	2339,0	2844,5

X9	775,1	376,8	219,2	572,0	260,6	90,5	57,7	285,3
X10	12,3	11,7	7,9	13,6	0,8	2,4	1,9	1,8

Analiza poszczególnych klastrów w metodzie grupowania podziałowego

Klaster 1

Województwa dolnośląskie i śląskie wykazują duże podobieństwo, co można zauważyć na wykresie wizualizacji klastrów metodą k - średnich, na którym znajdują się blisko siebie. Natomiast województwo mazowieckie, mimo przynależności do tego samego klastra, znajduje się znacznie wyżej, co sugeruje obecność wartości odstających, z którymi analiza skupień nie poradziła sobie. Mimo to warto poszukać wspólnych cech tych regionów. Charakterystyczną cechą tego klastra jest największa średnia dla wielu zmiennych spośród wszystkich klastrów. Klaster pierwszy posiada najwyższą średnią dla zmiennych: X2, X3, X4, X5, X7, X8 i X9. Oznacza to, że zgrupował województwa osiągające najwyższe wartości w wielu zmiennych. Klaster ten zgromadził także najbardziej podobne wartości dotyczące liczby pomników przyrody, gdyż odchylnie standardowe jest najniższe dla tej zmiennej. Dodatkowo posiada kilka zmiennych, które są najbardziej rozproszone od średniej X1, X3, X4, X5, X6 i X8. Warto wspomnieć także, że wszystkie województwa w tym klastrze w 2022 roku znalazły się w czołówce najbardziej zurbanizowanych obszarów w Polsce, a ich łączny udział w produkcie krajowym brutto wyniósł aż 43,8%. Czynniki te, związane z intensywną urbanizacją, mogą być przyczyną przynależności tych województw do jednego klastra.

Klaster 2

Ten klaster obejmuje obiekty skoncentrowane w jednym obszarze, które posiada najniższą średnią dla zmiennej mówiącej o powierzchni obszarów chronionych. Dodatkowo najbardziej rozproszone wartości wokół średniej posiada dla emisji zanieczyszczeń gazowych i liczbie pomników przyrody. Obszar ten natomiast posiada najmniejsze odchylenie standardowe dla masy wytworzonych odpadów komunalnych. Jest to klaster posiadający największą liczbę województw, dlatego rozpoznanie wspólnych cech mogło być trudne i może to być wytłumaczenie, dlaczego nie posiada on nic więcej wspólnego.

Klaster 3

Na podstawie wykresu można stwierdzić, że jest to najbardziej „zgrupowany” klaster, w którym obiekty znajdują się w bliskich odległościach. Posiada on najniższe średnie dla zmiennych X1, X4, X5, X7, X9, X10. Większość z tych zmiennych to stymulanty, dla których oczekiwane są jak najwyższe wartości (poza X7, X9), dlatego jest to dość pesymistyczna informacja, gdyż prawdopodobnie w tych

województwach stan ochrony środowiska jest słaby. Warto również zaznaczyć wysokie bezrobocie, oraz że te województwa w 2022 roku miały najniższy produkt krajowy brutto na mieszkańca.

Klaster 4

Klaster czwarty posiada tylko dwa województwa i jest dość rozproszony, jednak należy mieć na uwadze, że skoro powstał, to musi posiadać jakieś ciekawe wspólne cechy. Największe średnie występują dla zmiennych X1, X6 i X10 i jest to pozytywna informacja, gdyż każda z tych zmiennych może pozytywnie wpływać na stan ochrony środowiska w Polsce. Negatywną informacją są najniższe średnie dla X2, X3 i X8, każda z tych zmiennych jest destymulantą, ponieważ ich wysoka wartość wpływa negatywne na środowisko naturalne. Warto też wspomnieć o najmniejszym odchyleniu standardowym we wcześniej wymienionych zmiennych, oznacza to, że dane wartości są jak najbardziej podobne do średniej.

Średnia i odchylenie standardowe zmiennych dla klastrów z grupowania hierarchicznego

	Średnia				Odchylenie standardowe			
	1	2	3	4	1	2	3	4
X1	28,3	17	9,8	32,6	8,3	1,9	3,9	22,7
X2	27960728	6929066	6098402	13078555	10412221	5089905	6277321	15074501
X3	104775	175916	336483	705784	3927	314020	580948	694346
X4	349872	217627	96202	379562	254065	84205	17149	116509
X5	1290700	491952	444968	1379891	389311	148879	143578	789147
X6	333442	498363	774778	887130	53004	123395	249311	119232
X7	390,3	388,3	286,8	329,7	33,1	18,9	27,5	60,6
X8	1370,5	2088,7	3147,3	4048,1	3455,1	1827,5	2339,0	2925,6
X9	718,2	330,9	219,2	550,1	317,8	97,1	57,7	189,3
X10	12,3	11,2	7,9	12,9	0,8	2,6	1,9	1,4

Analiza poszczególnych klastrów w metodzie grupowania hierarchicznego

Klaster 1

Jak już wcześniej wspomniano, województwa śląskie i dolnośląskie są do siebie zbliżone pod wieloma względami. Charakteryzują się wysokim stopniem urbanizacji oraz istotnym udziałem w PKB,

zwłaszcza w przemyśle, województwo łódzkie też można zaliczyć tej kategorii. Klaster ten posiada najwyższe średnie dla zmiennej X2, X7 i X9, gdzie każda z tych zmiennych ma negatywny wpływ na stan środowiska naturalnego. Kolejną negatywną informacją są niskie średnie dla zmiennych informujących o powierzchni obszarów chronionych. Co ciekawe wspominane zostało, że są to województwa posiadające wysoki udział przemysłu w PKB, jednak klaster ten posiada najniższe średnie zużycie wody w przemyśle oraz odpadów składowanych poza odpadami komunalnymi. Ta grupa województw posiada trzy zmienne (X3, X6 i X10), które posiadają największe skupienie wokół średniej.

Klaster 2

Klaster drugi posiada najwięcej województw w grupowaniu hierarchicznym. Interesujące jest to, że nie posiada on żadnej najwyższej i najniższej średniej dla jakiejkolwiek zmiennej. Jednak dla zmiennej X1, X2, X7 i X8 posiada najmniejsze odchylenie standardowe, które może świadczyć o największym skupieniu danych wokół średniej i o niejakiej stabilności tego klastra. Co ciekawe jest to bardzo podobny klaster do tego z grupowania podziałowego.

Klaster 3

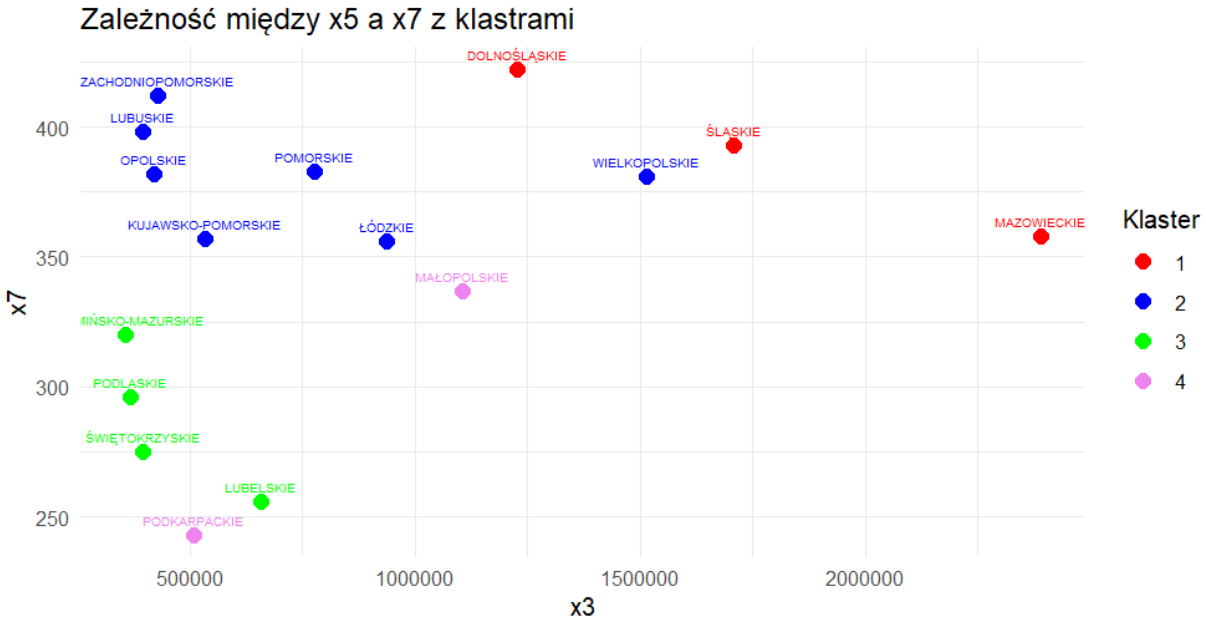
Klaster trzeci posiada najniższe średnie aż dla kilku zmiennych i są nimi zmienne X1, X2, X4, X5, X7, X9 i X10. Oznacza to, że gromadzi on naprawie wiele cennych informacji. Część zmiennych to stymulanty, oznaczając, że ich niska wartość nie sprzyja ochronie środowiska np. powierzchnia lasów, nakłady na gospodarkę wodną, nakłady na środowisko, czy liczba pomników przyrody. Reszta zmiennych to destymulanty i ich wysoka średnia nie jest pozytywną informacją. Jednak tak wiele informacji w tym klastrze świadczy, że powstał, ponieważ gromadzi województwa posiadające dużo wspólnych cech.

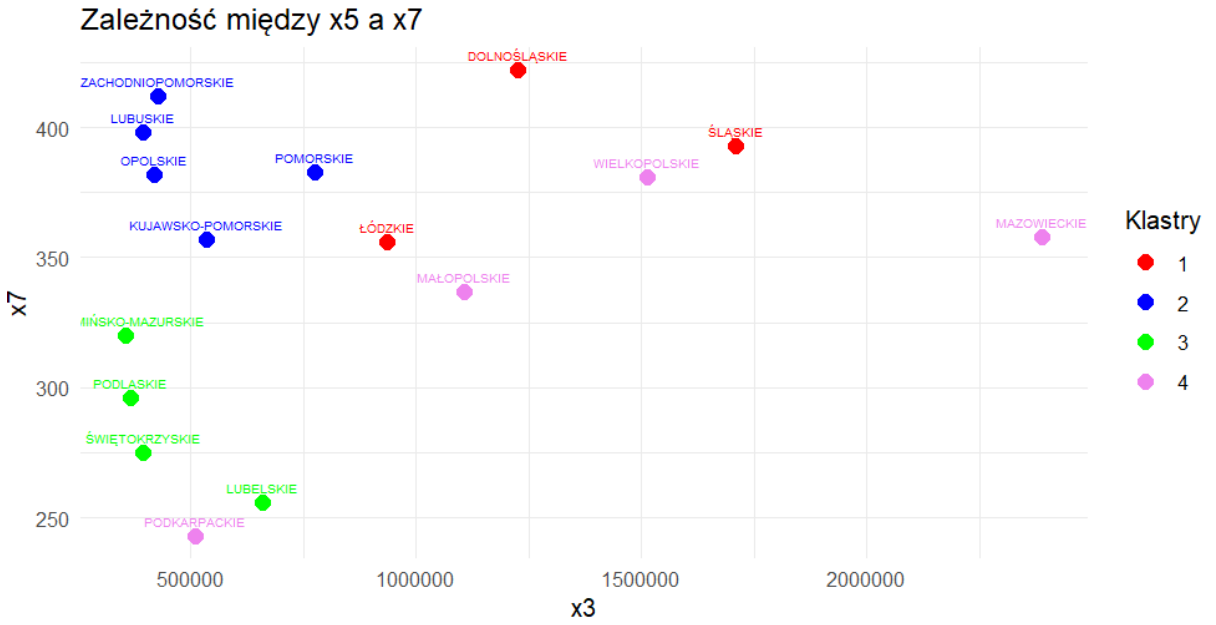
Klaster 4

Ta grupa posiada najwyższe średnie dla zmiennych: X1, X3, X4, X5, X6, X8, X10. Podobnie jak w przypadku poprzedniego klastra są tu stymulanty, jak i destymulanty, co oznacza, że zmienne te przynoszą złe oraz dobre rzeczy dla środowiska naturalnego. Do stymulant w tym klastrze należą zmienne: X1, X4, X5, X6 i X10, reszta to stymulanty. Jednak zostały wyodrębnione, ponieważ zostały zauważone w nich właśnie te cechy. Grupa ta jest zróżnicowana, również, ponieważ obejmuje zarówno bardzo rozwinięte, mocno zurbanizowane województwa, jak i województwo podkarpackie, które jest najsłabiej zurbanizowanym województwem.

Ostatnim etapem analizy jest stworzenie wykresów ilustrujących zależności między wybranymi zmiennymi i obserwacja, jak grupują się województwa w przypadku dwóch różnych metod grupowania. Wykres przedstawia zależność między zużyciem wody na potrzeby przemysłowe a masą utworzonych

odpadów komunalnych na jednego mieszkańca. Dodatkowo zastosujemy kolorowanie województw przypisanych do grup utworzonych w wyniku grupowania hierarchicznego, co pozwoli na lepszą wizualizację wyników. Pierwszy wykres zawiera klastry z grupowania podziałowego, a drugi z grupowania hierarchicznego.





Na podstawie wyników klasteryzacji, rozmieszczenie województw różni się w zależności od zastosowanej metody. Na pierwszy rzut oka oba grupowania tworzą bardzo podobne wykresy. Grupowanie podziałowe tworzy bardziej wyraźne i oddzielne grupy, podczas gdy grupowanie

hierarchiczne jest trochę bardziej rozproszone przez obecność klastra pierwszego, który jest rozproszony przez większość wykresu. Jednak w większości widoczne jest bardzo duże podobieństwo.

Porównanie wyników grupowania podziałowego i hierarchicznego ujawnia różnice w przyporządkowaniu województw do klastrów, wynikające z odmiennych założeń obu metod. Grupowanie hierarchiczne tworzy bardziej równoliczne grupy, co ułatwia identyfikację podobieństw między województwami w zakresie ochrony środowiska i struktury gospodarczej. Dzięki pomiarowi odległości między obiektami metoda ta skutecznie identyfikuje regiony o zbliżonych cechach. Metoda grupowania podziałowego natomiast może prowadzić do bardziej zróżnicowanych klastrów, w których województwa różnią się istotnie pod względem kluczowych wskaźników, co umożliwia lepsze zrozumienie lokalnych uwarunkowań i potrzeb rozwojowych poszczególnych regionów. Wybranie, która metoda analizy skupień poradziła sobie lepiej, jest w tym przypadku bardzo trudne. Finalnie oba grupowania stworzyły bardzo podobne klastry, różniące się często jednym województwem w każdym z nich, a każdy z nich wnosił jakieś cenne informacje do grupowania. Ostatecznie, aby uzyskać pełniejszy obraz i lepiej zrozumieć różnice oraz podobieństwa między województwami, warto rozważyć szerszą analizę wyników obu metod, co może prowadzić do jeszcze bardziej precyzyjnych i użytecznych wniosków dla dalszych analiz i działań strategicznych.

Podsumowanie

Podsumowując, analiza metod Hellwiga i sumy rang wykazuje wysoką zgodność, szczególnie w przypadku czołowych województw. Podkarpackie zajmuje najwyższą pozycję dzięki m.in. dużym obszarom chronionym i niskiej emisji zanieczyszczeń. Małopolskie i Wielkopolskie również plasują się wysoko, głównie dzięki inwestycjom w ochronę środowiska, mimo problemów z jakością powietrza.

W przeciwieństwie do nich, województwa śląskie, świętokrzyskie i łódzkie zajmują niskie pozycje z powodu dużej emisji zanieczyszczeń i wysokiej industrializacji. Wyniki te pokazują silną zależność między rozwojem przemysłu a stanem środowiska, podkreślając potrzebę działań w regionach o słabszych wynikach.

Porównanie metod grupowania podziałowego i hierarchicznego dla wszystkich klastrów ujawnia zarówno podobieństwa, jak i różnice w rozkładzie województw i ich cechach. W przypadku klastra 1, w obu metodach, województwa dolnośląskie i śląskie są zbliżone. Jednak w grupowaniu hierarchicznym dodatkowo włączono województwo łódzkie, a klaster ten posiada wyższe wartości dla zmiennych mających negatywny wpływ na środowisko. Klaster 2 w obu metodach obejmuje najwięcej województw, ale w metodzie hierarchicznej nie posiada wyraźnie najwyższych ani najniższych średnich dla zmiennych, co wskazuje na bardziej zrównoważony profil, natomiast w podziałowym trudniej było zidentyfikować wspólne cechy dla wszystkich województw. Klaster 3 w obu metodach gromadzi

województwa o najniższych średnich dla wielu zmiennych, głównie stymulantów ochrony środowiska, co sugeruje, że te regiony mają problemy z ochroną środowiska i niższe wskaźniki rozwoju. Klaster 4 natomiast w metodzie podziałowej obejmuje dwa województwa o wysokich średnich dla pozytywnych stymulantów środowiskowych, podczas gdy w grupowaniu hierarchicznym również grupuje województwa z wysokimi wynikami, ale z większym zróżnicowaniem, w tym województwa zarówno zurbanizowane, jak i mniej rozwinięte. Widoczne podobieństwa i niewielkie różnice pokazują, że ciężko jest zdecydować, które grupowanie poradziło sobie lepiej, ponieważ każde z nich ma swoje lepsze i gorsze strony.

Bibliografia

1. https://www.datanovia.com/en/lessons/determining-the-optimal-number-of-clusters-3-must-know-methods/
2. http://eregion.wzp.pl/wskaznik/wskaznik-urbanizacji
3. https://stat.gov.pl/obszary-tematyczne/rachunki-narodowe/rachunki-regionalne/produkt-krajowy-brutto-i-wartosc-dodana-brutto-w-przekroju-regionow-w-2022-r-,7,7.html
4. https://psz.praca.gov.pl/documents/10828/20671751/2022%20Bezrobocie%20rejestrowane.pdf/30c897d2-2e67-4a82-92f9-2c6ced357874?t=1681202335897
5. James.G, Witten.D, Hastie.T, Tibshirani. R, *An Introduction to Statistical Learning with Applications in R,* 2023