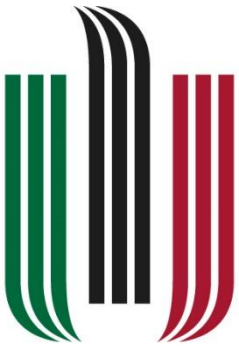


AKADEMIA GÓRNICZO-HUTNICZA IM. STANISŁAWA STASZICA W KRAKOWIE

WYDZIAŁ ZARZĄDZANIA

Analiza głównych składowych i skalowanie wielowymiarowe

Autor: *Izabela Guła*

Kierunek: *Informatyka i ekonometria (stacjonarne)*

Przedmiot: *Statystyczna analiza danych*

Kraków, 2024

Spis treści

Wstęp

Celem projektu jest przeprowadzenie analizy głównych składowych i skalowania wymiarowego dla danych z 2022 roku dotyczących powiatów Polski z wyłączeniem miast na prawach powiatów. Dane, które będą analizowane, pochodzą z Głównego Urzędu Statystycznego, udostępnione na oficjalnej stronie internetowej. Metody umożliwią uproszczenie analizy oraz lepsze zrozumienie kluczowych cech danych dotyczących powiatów, co pozwoli na skuteczniejszą wizualizację i dalszą interpretację wyników.

Analiza danych

Do przeprowadzenia badania wybrane zostały zmienne, które pozwalają na spełnienie założeń analizy głównych składowych oraz skalowania wymiarowego. Wybrane zmienne charakteryzują różnorodne aspekty społeczno-gospodarcze, takie jak warunki mieszkaniowe, aktywność gospodarcza, rynek pracy oraz demografia. Zbiór danych posiada 314 obserwacji o powiatach bez miast na prawach powiatu.

Zmienna	Opis zmiennej
X1	Mieszkania oddane do użytkowania, liczba podana na 1000 mieszkańców.
X2	Gęstość zaludnienia, czyli średnia ilość mieszkańców przypadającą na kilometr kwadratowy na danym terenie.
X3	Nowo zarejestrowane w rejestrze REGON podmioty gospodarki narodowej.
X4	Liczba osób poszkodowanych w wypadkach przy pracy.
X5	Stopa bezrobocia rejestrowanego, czyli stosunek liczby bezrobotnych zarejestrowanych do liczby cywilnej ludności aktywnej zawodowo.
X6	Ruch naturalny, czyli fakty urodzeń i zgonów a także zawierania związków małżeńskich oraz rozwodzenia się (i separacji).

Tabela 1. Zestawienie zmiennych

Przed rozpoczęciem analizy ograniczono wartości odstające w każdej zmiennej do zakresu górnego i dolnego wąsa na wykresach pudełkowych. Dzięki temu dane zostały oczyszczone z ekstremalnych wartości, które mogłyby wpływać na wyniki. Taki krok pomaga uniknąć zniekształceń w analizie i zapewnia, że wyniki lepiej odzwierciedlają rzeczywistość.

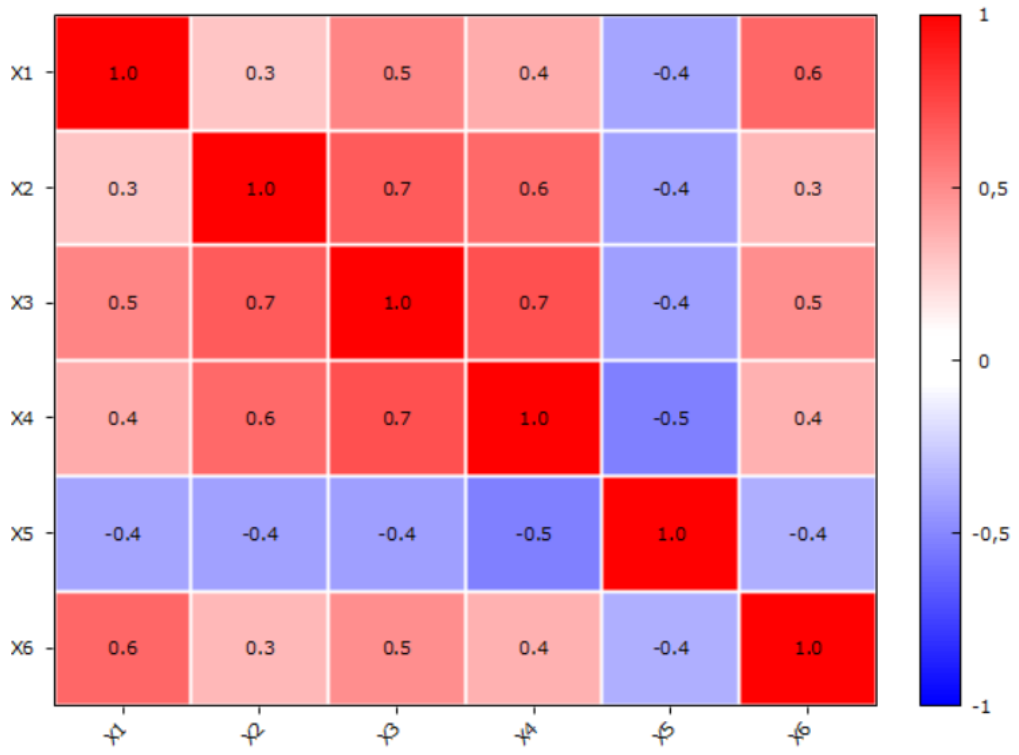
	X1	X2	X3	X4	X5	X6
Minimum	0,9	18,2	108	7	1	-11,7
Średnia	4,7	90,6	614,4	108,9	7,8	-4,4
Odchylenie standardowe	2,3	50,4	342	71,9	3,9	2,6
Mediana	4,4	75,5	530	91	7,1	-4,5
Współczynnik zmienności	0,5	0,6	0,6	0,7	0,5	0,6

Kurtoza	0,03	-0,04	0,02	0,22	-0,06	-0,09
Maksimum	10	205,6	1422	295	18,3	2,8

Tabela 2. Statystyki opisowe

Kolejny etap to obliczenie wybranych statystyk dla każdej zmiennej. Zmienna dotycząca gęstości zaludnienia cechuje się największym zróżnicowaniem, co wskazuje na znaczną różnicę między obszarami wiejskimi a bardziej zaludnionymi regionami, z minimum wynoszącym 18,2 i maksimum 205,6 mieszkańców na km². Ruch naturalny wykazuje średnią wartość ujemną, co sugeruje, że w większości powiatów liczba zgonów przewyższa liczbę urodzeń, choć w niektórych przypadkach saldo jest dodatnie (maksymalna wartość 2,8). Liczba osób poszkodowanych w wypadkach przy pracy wykazuje największe odchylenie standardowe, co może świadczyć o dużych różnicach w warunkach pracy w różnych częściach kraju. Dane nie posiadają żadnych brakujących wartości.

Korelacje między zmiennymi osiągają wartości świadczące o korelacji umiarkowanej bądź silnej. Jest to pożądana informacja, gdyż PCA ma sens, gdy dane są ze sobą dość istotnie skorelowane.



Rysunek 2. Macierz korelacji

Analiza głównych składowych

Analiza głównych składowych to technika statystyczna stosowana w celu redukcji wymiarowości danych, która przekształca zbiór oryginalnych zmiennych na nowy zestaw liniowo niezależnych zmiennych zwanych głównymi składowymi. Główne składowe są ułożone w kolejności malejącej

wariancji, dzięki czemu pierwszych kilka z nich zawiera większość informacji o zmienności danych, co pozwala na uproszczenie analizy, wizualizacji i przyspieszenie algorytmów uczących się, przy jednoczesnym ograniczeniu utraty informacji. PCA opiera się na obliczaniu wartości własnych i wektorów własnych macierzy korelacji lub kowariancji danych, co pozwala znaleźć kierunki maksymalnej wariancji w przestrzeni wielowymiarowej.

Analizę głównych składowych należy rozpocząć od sprawdzenia wszystkich założeń, aby móc stwierdzić, czy metoda może być użyta.

1) Rozkład wielowymiarowy normalny
   Po przeprowadzeniu testów normalności dla każdej ze zmiennej, wynika, że tylko jedna z nich posiada rozkład zbliżony do normalnego i jest to zmienna dotycząca ruchu naturalnego. Jednakże warunek ten nie jest konieczny, przy tak dużej liczbie obserwacji.
2) Brak danych oraz wartości odstające
   W wybranym zbiorze danych nie występują braki, ponieważ wszystkie wartości odstające zostały ograniczone.
3) Liczność próby i zmiennych
   Za rozsądną liczbę próby podaje się 100, w tym przypadku warunek ten jest spełniony, gdyż w wybranym zbiorze danych jest ponad 300 obserwacji. W przypadku liczby zmiennych pożądana liczba wynosi 5 i warunek ten jest również spełniony, ponieważ liczba zmiennych wynosi 6.

Pierwszym ważnym krokiem do przeprowadzenia analizy głównych składowych jest obliczenie macierzy kowariancji lub korelacji. W przypadku badanych danych musi zostać obliczona macierz korelacji, ponieważ zmienne mają różną skalę, przed tym należy zestandaryzować zmienne. Przed przystąpieniem do analizy należy pamiętać, że PCA ma wyłącznie sens, gdy zmienne są ze sobą dość istotnie skorelowane.

W celu potwierdzenia tej informacji przeprowadzony zostanie test Bartletta, oceniający istotność macierzy korelacji. Dane zostały zestandaryzowane, a następnie obliczona została macierz korelacji.

H_0: Brak istotnych korelacji między zmiennymi

H_1: Istnieją istotne korelacje między zmiennymi

P-value wyniosło 2.387917e-46 sugeruje, że jest praktycznie zerowe, co oznacza, że istnieją podstawy do odrzucenia hipotezy zerowej i stwierdzenia, że korelacje między zmiennymi są istotne.

Drugim testem, który potwierdzi istotność skorelowanych zmiennych, jest test Kaisera-Mayera-Olkina, w skrócie KMO. Miara KMO nazywana jest miarą adekwatności doboru zmiennych i przyjmuje

wartości z przedziału od 0 do 1. Im wyższa wartość miary KMO, tym w większym stopniu można uznać istnienie związków pomiędzy zmiennymi i tym lepiej dane nadają się do analizy czynnikowej.

```
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = macierz_korelacji)
Overall MSA =  0.79
MSA for each item =
  X1   X2   X3   X4   X5   X6
0.74 0.82 0.77 0.80 0.84 0.79
```

Rysunek 3. Wyniki KMO

Wszystkie zmienne wykazują wartości wynoszące ponad 0,74 i są do zadowalające wyniki, które sugerują, że w badanym zbiorze danych, nie ma przeciwwskazań do przeprowadzenia analizy głównych składowych.

Następnym etapem jest wyznaczenie wartości i wektorów własnych. W tym przypadku badanie zostanie przeprowadzone w języku R i wykorzystana zostanie funkcja *prcomp()* z biblioteki *stats*. Funkcja przeprowadzająca analizę składowych głównych wykorzystuje metodę wartości osobliwych do transformacji danych w nowy układ współrzędnych, maksymalizując zmienność wyjaśnianą przez kolejne składowe główne.

	PC1	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**
X1	-0,3723	0,5556	-0,5303	0,4361	-0,1337	0,2520
X2	-0,3923	-0,4723	0,4259	0,5801	-0,2222	0,2362
X3	-0,4361	-0,2023	-0,1537	0,0287	0,8547	-0,1179
X4	-0,4197	-0,3334	-0,2923	-0,6147	-0,2667	0,4222
X5	0,4542	-0,0001	0,0435	0,0772	0,3494	0,8147
X6	-0,3672	0,5623	0,6530	-0,2977	0,0949	0,1575

Tabela 3. Wektory własne

4,8062	0,8756	0,1531	0,1016	0,0633	0

Tabela 4. Wartości własne

Wartości własne w PCA odzwierciedlają ilość wariancji wyjaśnianej przez każdą składową główną i są uporządkowane malejąco. Oblicza się je jako wynik dekompozycji macierzy korelacji, gdzie każda wartość własna odpowiada określonej składowej głównej, pokazując, ile zmienności w danych można przypisać tej składowej.

Na tym etapie powinien nastąpić wybór liczby składowych. Istnieje wiele kryteriów, które pomagają wybrać odpowiednią liczbę, dlatego przeanalizowane zostanie kilka z nich w celu wyboru optymalnej liczby składowych.

1) Kryterium Kaisera (1960)

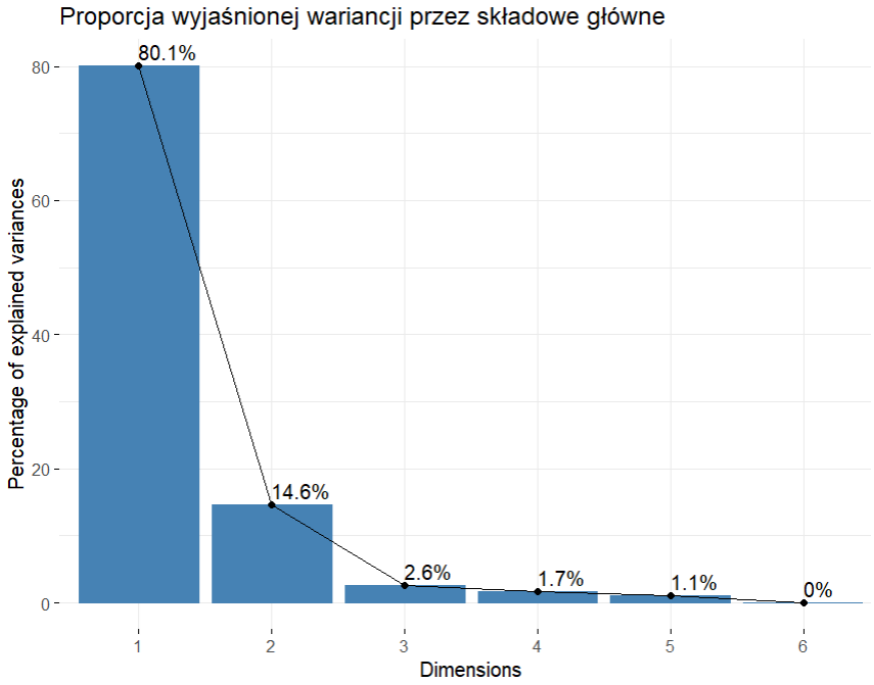
   Kryterium to mówi, że należy rozważyć tylko składowe, dla których wartości własne są większe niż 1. W tym przypadku zalecany byłby wybór tylko pierwszej składowej, gdyż jej wartość wynosi 4,8.

2) Kryterium Joliffe (1972)

   Jest to poprawione kryterium Kaisera mówiące, że należy rozważać składowe, których wartości własne są większe niż 0,7. To kryterium dawałoby podstawy do wyboru dwóch pierwszych składowych.

3) Wyjaśniana wariancja przez składowe

   Kryterium to mówi, że należy wybrać tyle składowych, aby wyjaśnić z góry ustaloną część zmienności. Najczęściej ustala się tę wartość na 80%, jednak może to być też wartość z przedziału 70% - 90%. Poniższy wykres przedstawia część wariancji wyjaśnianej przez poszczególne składowe. Można zauważyć ogromną przewagę pierwszej składowej, gdyż jej wariancja wyjaśnia ponad 80% danych. Wybór dwóch składowych pozwoliłby na zachowanie ponad 94% wariancji, co może być uznawane za bardzo dobrą reprezentację oryginalnych danych przy minimalnej utracie informacji.

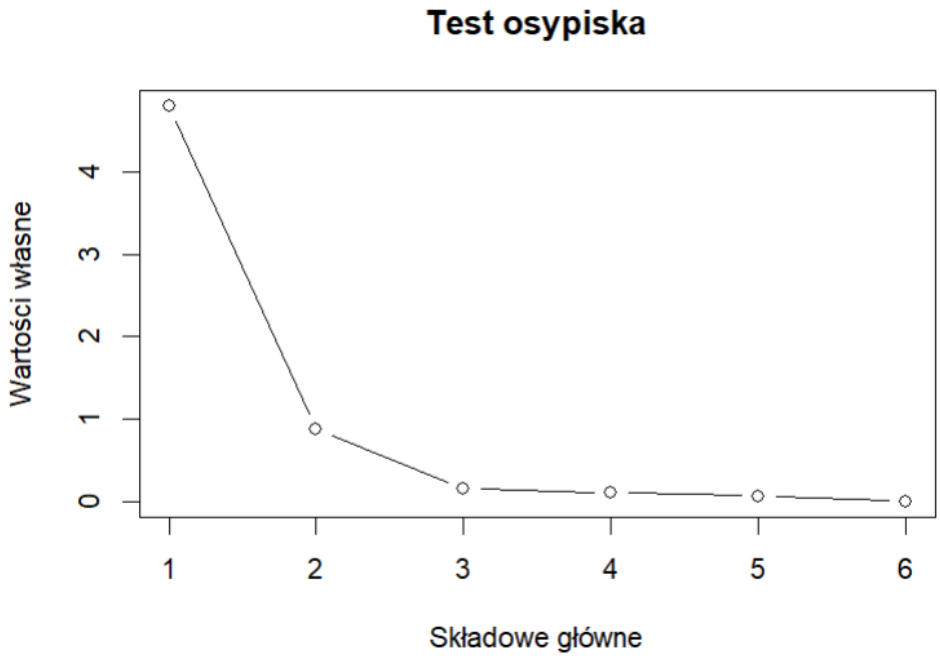


Rysunek 3. Wariancja wyjaśniana przez składowe

4) Test osypiska (1966)

   Test ten polega na stworzeniu wykresu przedstawiającego wartości własne i określeniu, w którym miejscu występuje łagodny spadek wartości własnych. Na poniższym wykresie można zaobserwować, że po drugiej składowej następuje znaczący spadek wartości własnych, co

wskazuje na brak istotnych informacji w kolejnych składowych. Wartości własne od trzeciej składowej stają się na tyle małe, że dalsze składowe nie wnoszą już istotnej zmienności.



Rysunek 4. Test osypiska

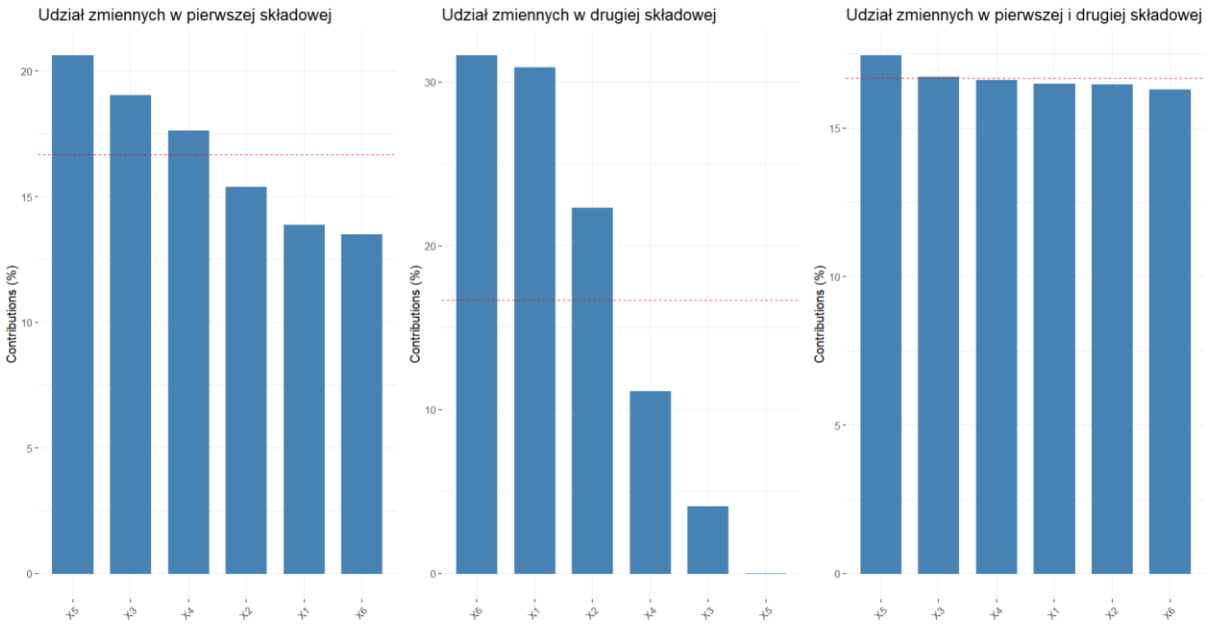
Po przeanalizowaniu każdego z powyższych kryteriów można stwierdzić, że optymalna liczba składowych będzie wynosić dwa. Dodatkowo taka liczba pozwoli na wizualizowanie wyników analizy głównych składowych w drugim wymiarze, co pozwoli przedstawić wyniki w łatwy i interesujący sposób.

	PC1	**PC2**
X1	-0,3723	0,5556
X2	-0,3923	-0,4723
X3	-0,4361	-0,2023
X4	-0,4197	-0,3334
X5	0,4542	-0,0001
X6	-0,3672	0,5623

Tabela 5. Wektory własne dla dwóch składowych

Składowe należy interpretować poprzez wartości wektora własnego. Większe wartości sugerują większy udział w wyjaśnianiu danej zmiennej, znaki mówią natomiast o odmiennych kierunkach, jednak w interpretacji nie mają one znaczenia. Pierwsza składowa można powiedzieć, że jest dość podobna. Nie biorąc pod uwagę znaków, można stwierdzić, że osiąga wartości z przedziału od 0,3672 dla zmiennej dotyczącej ruchu naturalnego do 0,4542 dla zmiennej dotyczącej stopy bezrobocia i oznacza to, że każda zmienna wyjaśniana jest w podobny sposób. Natomiast druga składowa wykazuje się już większe zróżnicowanie. Najmniejsza wartość bliska zeru występuje dla zmiennej dotyczącej

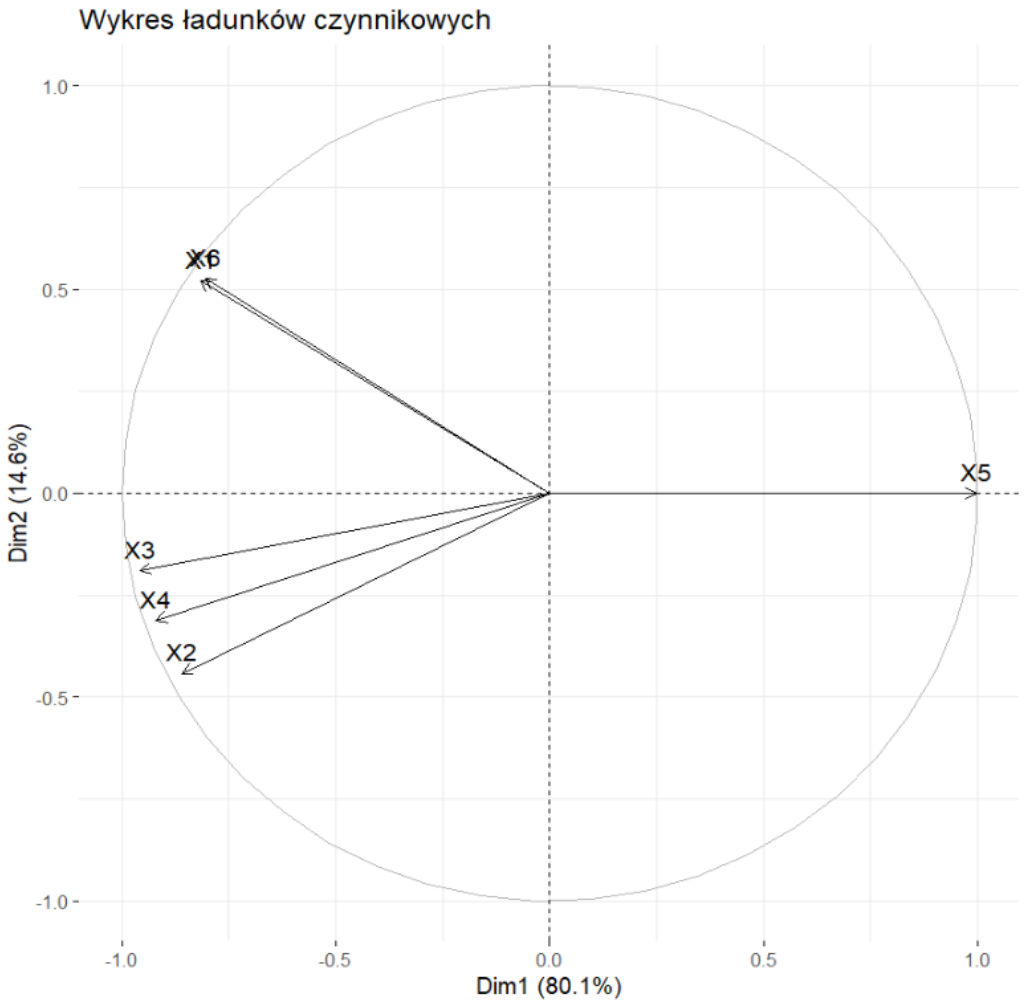
stopy bezrobocia i oznacza to, że bezrobocie rejestrowane ma minimalny wpływ na drugą składową. Największe wartości osiągają zmienne dotyczące liczby mieszkań oddanych do użytku i ruchu naturalnego i oznacza to, że te zmienne mają największy wpływ na składową.



Rysunek 5. Udział zmiennych

Powyższe wykresy przedstawiają udział zmiennych dla każdej ze składowych oraz łącznie dla obu. Na pierwszych dwóch wykresach można zauważyć, to co zostało już wcześniej wspomniane z wartości wektorów własnych. Trzeci wykres przedstawia, w jaki sposób zmienne mają wpływ na składowe po ich połączeniu. Na pierwszy rzut oka widać bardzo podobny udział każdej ze zmiennej, co oznacza, że składowe po połączeniu w pewien sposób dopełniają się. Taki efekt sugeruje, że obie składowe razem lepiej odzwierciedlają całość struktury danych, co może prowadzić do bardziej stabilnej i pełnej reprezentacji zmienności w danych.

Kolejnym etapem interpretacji jest analiza wykresu ładunków czynnikowych. Wykres ten przedstawia wektory o początku w środku układu współrzędnych, które reprezentują zmienne. Każdy z nich znajduje się na płaszczyźnie wyznaczonej przez pierwszą i drugą składową.



Rysunek 6. Wykres ładunków czynnikowych

Interpretacja geometryczna:

1) Współrzędne końca wektora

   Są to odpowiadające ładunki czynnikowe zmiennych, które opisują. jak zmienne są powiązane z poszczególnymi głównymi składowymi. Przykładowo zmienna mówiąca o stopie bezrobocia jest w bardzo dużym stopniu skorelowana z pierwszą składową (0,996), ale bardzo słabo skorelowana z drugą składową (-0,001). Wszystkie zmienne są mocno skorelowane z pierwszą składową i na średnim poziomie z drugą.

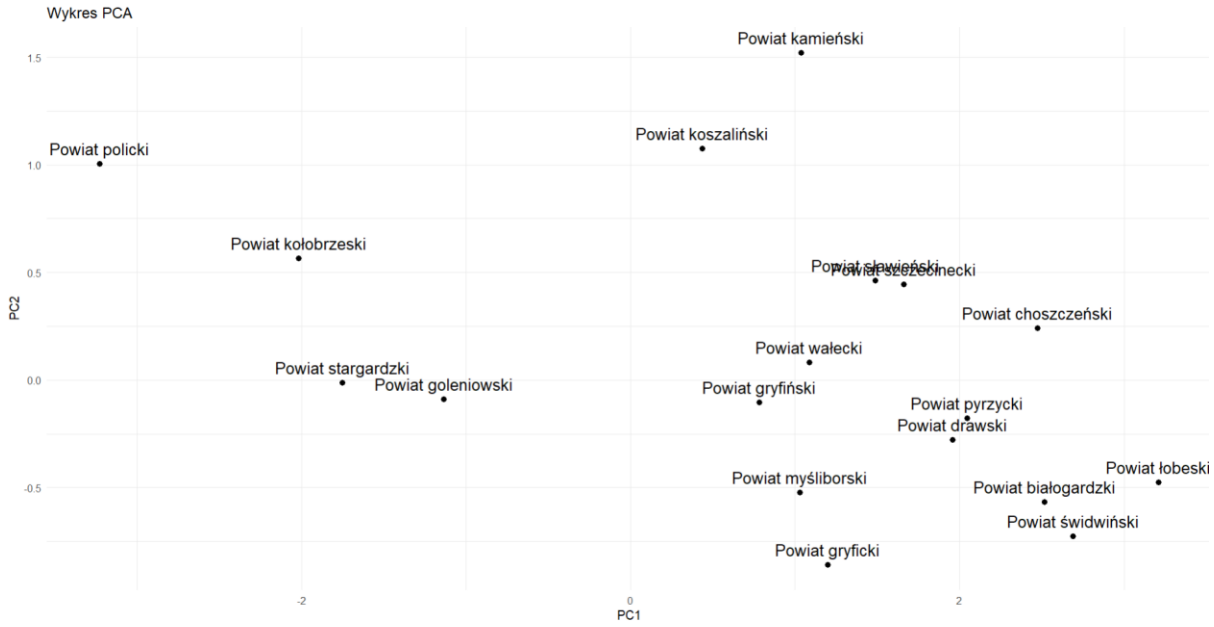
2) Długość wektora

   Informacja ta odnosi się do siły powiązania zmiennej z przestrzenią utworzoną przez główne składowe. Można ją zrozumieć w kontekście, jak mocno zmienna jest skorelowana z obiema głównymi składowymi. W tym przypadku każda ze zmiennych ma długość wektora wynoszącą praktycznie jeden. Oznacza to, że wektory wykazują silną zależność między zmiennymi a głównymi składowymi.

3) Kąt między wektorami

Wskazuje na skorelowanie zmiennych pierwotnych. Mniejszy kąt pomiędzy wektorami wskazuje na większą korelacje zmiennych, natomiast im większy kąt, tym mniejsza korelacja między zmiennymi. W przypadku poniższego wykresu można zauważyć bardzo silną korelację między zmienną dotyczącą liczby oddanych mieszkań do użytku, a ruchem naturalnym. Reszta zmiennych jest w pewnym stopniu skorelowana, a najmniejszą korelacją wykazują się zmienne o stopie bezrobocia i ruchu naturalnym.

W celu przedstawienia wyników analizy głównych składowych stworzony został wykres obrazujący rozkład powiatów w drugim wymiarze. Z uwagi na dużą liczbę obserwacji, wybrane zostały powiaty z województwa zachodniopomorskiego, w celu sprawdzenia w jakiś sposób analiza przyczyniła się do ich rozkładu i grupowania się na przestrzeni wszystkich badanych obiektów.



Rysunek 7. Wizualizacja analizy głównych składowych dla powiatów województwa zachodniopomorskiego

Na wykresie zostały wyraźnie wyodrębnione cztery grupy. Każda z nich charakteryzuje się różnymi cechami, które pozwolą na udowodnienie ich ugrupowania.

- Grupa 1

  Powiaty: policki.

  Powiat policki został zaliczony do osobnej grupy ze względu na jego odosobnione położenie na wykresie. Powiat ten charakteryzuje się największą gęstością zaludnienia spośród wszystkich badanych powiatów, co może być spowodowane położeniem powiatu przy powiecie szczecińskim i niemieckiej granicy, gdyż wiele osób może dojeżdżać stamtąd łatwo do pracy. Dodatkowo jest to powiat z najmniejszym ruchem naturalnym. Powiat policki jest ważnym ośrodkiem przemysłowym, gdzie znajdują się Zakłady Chemiczne „Police” S.A.,

będące jednym z największych producentów nawozów w Polsce. Jednocześnie region wyróżnia się bogactwem przyrody, m.in. Puszczą Wkrzańską i rezerwatem Świdwie, które są cenionymi obszarami ochrony środowiska i turystyki.

- Grupa 2

  Powiaty: goleniowski, kołobrzeski, stargardzki.

  Zaczynając od zbioru danych, powiaty te charakteryzują się najwyższymi wartościami dla liczby nowo zarejestrowanych przedsiębiorstw oraz liczby osób poszkodowanych przy wypadkach w pracy. Dodatkowo powiaty te charakteryzują się niską stopą bezrobocia rejestrowanego oraz niską wartością ruchu naturalnego. Powiaty goleniowski, kołobrzeski i stargardzki leżą w pobliżu Szczecina i Bałtyku, co sprzyja ich rozwojowi przemysłowemu, turystycznemu i transportowemu. Łączy je także rozbudowana infrastruktura komunikacyjna oraz znaczenie w regionalnej gospodarce – od przemysłu i logistyki w powiatach goleniowskim i stargardzkim po turystykę w kołobrzeskim.

- Grupa 3

  Powiaty: kamieński, koszaliński.

  W przypadku tej grupy oba powiaty osiągają podobne wartości w każdej z badanych zmiennych, co prawdopodobnie wyjaśnia ich wyodrębnienie spośród innych powiatów.

  Powiaty kamieński i koszaliński łączy bogactwo przyrodnicze – oba regiony posiadają liczne lasy, jeziora i tereny nadmorskie, co czyni je atrakcyjnymi dla turystyki przyrodniczej. Oba powiaty oferują także możliwości do uprawiania turystyki aktywnej jak piesze wędrówki, rowerowe wycieczki oraz sporty wodne, zwłaszcza w okolicach Mielna. Dodatkowo oba regiony są miejscem o dużym znaczeniu kulturowym i historycznym, z licznymi zabytkami, w tym pomnikami archeologicznymi, jak Kamienne Kręgi w Koszalinie.

- Grupa 4

  Powiaty: białogardzki, choszczeński, drawski, gryficki, gryfiński, myśliborski, pyrzycki, sławieński, szczeciński, wałecki, łobeski.

  Ostatnia grupa jest najliczniejsza, jednak mimo jej wielkości widoczne jest skumulowanie powiatów w jednej grupie. Co łączy te powiaty ? Wszystkie mają najmniejsze wartości dla liczby mieszkań oddanych do użytków na 1000 mieszkańców. Dodatkowo wskazują na niską gęstość zaludnienia, co może oznaczać, że są to powiaty wiejskie i stąd mała liczba mieszkań oddanych do użytku. Wszystkie z nich posiadają małą liczbę nowych przedsiębiorstw oraz najniższe wartości dla ruchu naturalnego. Większość z tych powiatów charakteryzuje się obecnością terenów rolniczych, a także bogactwem wodnym, takim jak jeziora i rzeki. Część z nich jest także związana z rozwojem przemysłu drzewnego, rolnictwa i turystyki, w tym turystyki wodnej.

Skalowanie wielowymiarowe

Skalowanie wielowymiarowe pozwala na wizualizacje obiektów n-wymiarowych w przestrzeni m-wymiarowej. Jest to technika redukcji wielowymiarowości, która polega na znalezieniu funkcji, która przekształca odległości rzeczywiste na skalowane przy najmniejszej utracie wiadomości. Metoda polega na zmianie liczbie wymiarów, aby w jak najmniejszym stopniu zniekształcić odległości i zminimalizować wartość popełnianego błędu. Założeniem do przeprowadzania skalowania wielowymiarowego jest wysoka korelacja pomiędzy zmiennymi, która została wykazana w analizie głównych składowych.

Pierwszym etapem badania jest zestandaryzowanie danych, a następie obliczenie odległości między obiektami, w tym przypadku zostanie użyta odległość Euklidesowa. Istnieje kilka rodzajów skalowania, w badaniu zostaną zastosowane dwie techniki skalowania metrycznego, ponieważ ta metoda pozwala na badanie zmiennych, które są mierzone na skali przedziałowej lub ilorazowej.

Klasyczne skalowanie wielowymiarowe

Klasyczne skalowanie wielowymiarowe polega na odwzorowaniu odległości między obiektami z macierzy odległości w przestrzeni o mniejszej liczbie wymiarów, tak aby zachować, jak największą zgodność między oryginalnymi i odwzorowanymi relacjami. Do jego interpretacji zostanie użyty współczynnik STRESS, który mierzy stopień dopasowania odwzorowanych odległości do oryginalnych odległości w macierzy. Najlepsza wartość współczynnika STRESS powinna być bliska zeru i oznacza idealna dopasowanie. Skalowanie wymiarowe rozpoczyna się od stworzenia macierzy odległości. W celu przedstawienia wyników prezentowane będą tylko pięć powiatów z województwa zachodniopomorskiego, gdyż całość jest zdecydowanie za duża, a najważniejsze jest przedstawienie, jak odległości będą się zmieniać przy następujących zmianach.

	P. białogardzki	**P. choszczeński**	**P. drawski**	**P. goleniowski**	**P. gryficki**
P. białogardzki	0	1,045	1,393	4,663	2,587
P. choszczeński	1,045	0	1,331	4,562	2,575
P. drawski	1,393	1,331	0	3,629	1,454
P. goleniowski	4,663	4,562	3,629	0	3,009
P. gryficki	2,587	2,575	1,454	3,009	0

Tabela 6. Początkowa macierz odległości

W tym etapie nastąpi obliczenie macierzy odległości dla pierwszego wymiaru. W tym celu wykorzystana zostanie funkcja *cmdscale()* z pakietu z *stats*, która przekształca macierz odległości w zestaw współrzędnych w przestrzeni o zredukowanej wymiarowości, zachowując jak najlepiej oryginalne odległości między punktami dla wybranego wymiaru.

	P. białogardzki	P. choszczeński	P. drawski	P. goleniowski	P. gryficki
P. białogardzki	0	0,011	0,503	3,752	1,245
P. choszczeński	0,011	0	0,491	3,741	1,234
P. drawski	0,503	0,491	0	3,249	0,743
P. goleniowski	3,752	3,741	3,249	0	2,607
P. gryficki	1,245	1,234	0,743	2,607	0

Tabela 7. Macierz odległości dla pierwszego wymiaru

Różnice w macierzy odległości dla pierwszego wymiaru a pierwotnej macierzy odległości są spore, ponieważ widać zmniejszenie każdej z odległości. Wartość współczynnika STRESS wyniosła 0,39, co oznacza bardzo słabe dopasowanie, dlatego nastąpi zwiększenie wymiaru do drugiego.

	P. białogardzki	P. choszczeński	P. drawski	P. goleniowski	P. gryficki
P. białogardzki	0	0,860	0,556	3,769	1,318
P. choszczeński	0,860	0	0,792	3,774	1,786
P. drawski	0,556	0,792	0	3,252	1,000
P. goleniowski	3,769	3,774	3,252	0	2,629
P. gryficki	1,318	1,786	1,000	2,629	0

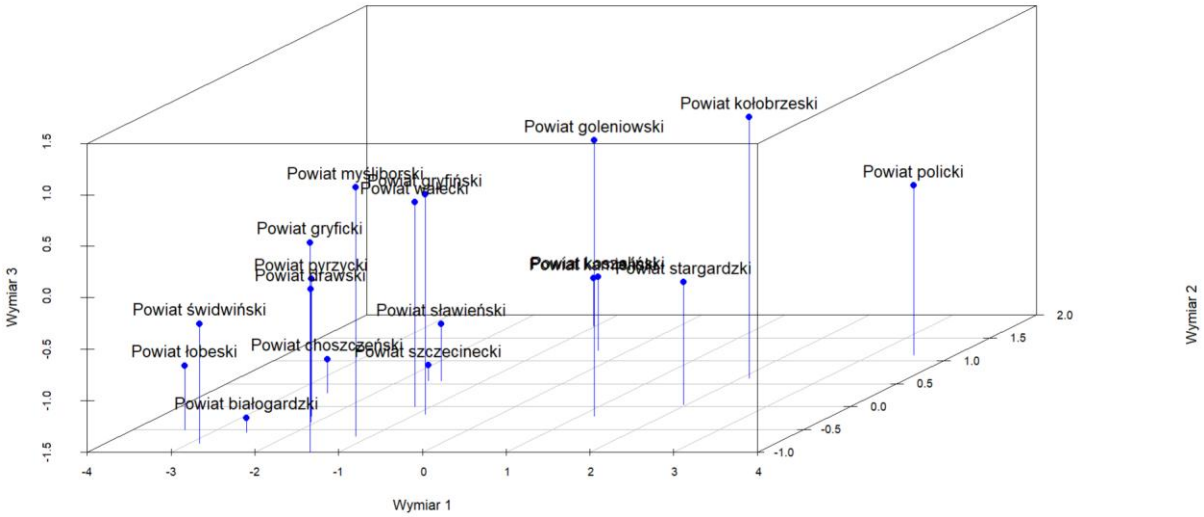
Tabela 8. Macierz odległości dla drugiego wymiaru

W przypadku drugiego wymiaru również widać zmianę wartości są mniejsze niż w początkowej macierzy odległości, lecz większe niż w tej z pierwszego wymiaru. Wartość współczynnika STRESS spadła, ponieważ wynosi 0,23, jednak wartość ta wciąż nie jest zadowalająca. W celu znalezienia lepszego dopasowania, w kolejnym etapie nastąpi przekształcenie do trzeciego wymiaru.

	P. białogardzki	P. choszczeński	P. drawski	P. goleniowski	P. gryficki
P. białogardzki	0	0,881	1,268	4,542	2,307
P. choszczeński	0,881	0	1,239	4,444	2,470
P. drawski	1,268	1,239	0	3,538	1,253
P. goleniowski	4,542	4,444	3,538	0	2,706
P. gryficki	2,307	2,470	1,253	2,706	0

Tabela 9. Macierz odległości dla trzeciego wymiaru

Trzeci wymiar przynosi podobne wnioski i w tym przypadku również nastąpił spadek współczynnika STRESS, dla trzeciego wymiaru wyniósł on 0,14 i oznacza słabe dopasowanie. Wyniki te nie są zadowalające prawdopodobnie przez dużą liczbę badanych obiektów. Miara dopasowania byłaby też lepsza, przy większym wymiarze, jednak wymiar trzecim jest najbardziej rozsądnym przez to, że istnieje możliwość zwizualizowania go.



Rysunek 8. Wizualizacja klasycznego skalowania wielowymiarowego dla powiatów województwa zachodniopomorskiego

Klasyczne skalowanie wielowymiarowe pozwoliło zwizualizować położenie powiatów województwa zachodniopomorskiego w trzech wymiarach. W takim wymiarze ciężej wyodrębnić grupy jednak na wykresie widoczne są pewne podobieństwa, które pozwalają na klasyfikacje powiatów.

- Grupa 1

  Powiaty: policki

  Podobnie jak przy analizie głównych składowych, powiat policki został wyodrębniony jako osobna grupa. Jest to spowodowane wyraźnym oddaleniem na wykresie i trudem przypisania go do którejś innej grupy. Jego charakterystyki zostały opisane już wcześniej, dlatego punkt ten zostanie pominięty.

- Grupa 2

  Powiaty: goleniowski, kołobrzeski

  Powiaty te znajdują się w centralnej części wykresu. Charakteryzują się podobną liczbą nowo zarejestrowanych przedsiębiorstw oraz liczbą osób poszkodowanych w wypadkach przy pracy. Poza tym oba z nich znajdują się przy większych aglomeracjach miejskich, jednak gęstość zaludnienia w powiecie kołobrzeskim jest dwa razy większa. Powiat goleniowski i kołobrzeski łączy bliskość Morza Bałtyckiego oraz rozwinięta infrastruktura turystyczna, która przyciąga odwiedzających z Polski i zagranicy. Oba powiaty charakteryzują się także znaczącym potencjałem gospodarczym, w tym przemysłem i usługami związanymi z transportem i portami (lotniczym w Goleniowie i morskim w Kołobrzegu).

- Grupa 3

  Powiaty: kamieński, koszaliński, stargardzki.

  Wszystkie trzy obiekty mają stosunkowo podobne wartości dla mieszkań oddanych do użytku. Jednak w przypadku pozostałych zmiennych, ciężko wysnuć jakieś podobieństwa. Powiaty

kamieński, koszaliński i stargardzki łączy rolniczo-turystyczny charakter oraz bliskość atrakcyjnych przyrodniczo terenów, takich jak Wybrzeże Bałtyckie i Pojezierze Pomorskie. Wspólną cechą jest również rozwój lokalnych inicjatyw gospodarczych opartych na rolnictwie, przetwórstwie oraz turystyce wiejskiej i rekreacyjnej.

- Grupa 4

  Powiaty: białogardzki, choszczeński, drawski, gryficki, gryfiński, myśliborski, pyrzycki, sławieński, szczeciński, wałecki, łobeski.

  Czwarta grupa jest największą, ponieważ zalicza się do niej aż jedenaście powiatów. Jego charakterystyki zostały opisane już wcześniej, dlatego punkt ten zostanie pominięty.

Skalowanie Sammona

Metoda skalowania Sammona to technika redukcji wymiarowości, która minimalizuje różnice między rzeczywistymi a odwzorowanymi odległościami w przestrzeni o niższej liczbie wymiarów. Optymalizuje układ punktów tak, aby odległości krótkie były lepiej zachowane niż długie, co pozwala lepiej uchwycić lokalną strukturę danych.

	P. białogardzki	**P. choszczeński**	**P. drawski**	**P. goleniowski**	**P. gryficki**
P. białogardzki	0	0,212	1,079	5,403	1,941
P. choszczeński	0,212	0	0,868	5,191	1,729
P. drawski	1,079	0,868	0	4,323	0,861
P. goleniowski	5,403	5,191	4,323	0	3,462
P. gryficki	1,941	1,729	0,861	3,462	0

Tabela 10. Macierz odległości Sammona w pierwszym wymiarze

Wyniki w skalowaniu Sammona reprezentują współrzędne obiektów w zredukowanej przestrzeni jedno wymiarowej. Podobnie jak w klasycznym skalowaniu wielowymiarowym, dla każdego obiektu została obliczona macierz odległości. W przypadku metody Sammona jedne odległości w pierwszym wymiarze są mniejsze a inne większe niż z początkowej macierzy odległości. Współczynnik dopasowania STRESS wynosi 0,13, co oznacza słabe dopasowanie, jednak jest lepiej niż w przypadku klasycznego skalowania w pierwszym wymiarze.

	P. białogardzki	**P. choszczeński**	**P. drawski**	**P. goleniowski**	**P. gryficki**
P. białogardzki	0	0,705	1,174	5,253	2,543
P. choszczeński	0,705	0	1,404	5,135	2,776
P. drawski	1,174	1,404	0	4,142	1,394
P. goleniowski	5,253	5,135	4,142	0	3,149
P. gryficki	2,543	2,776	1,394	3,149	0

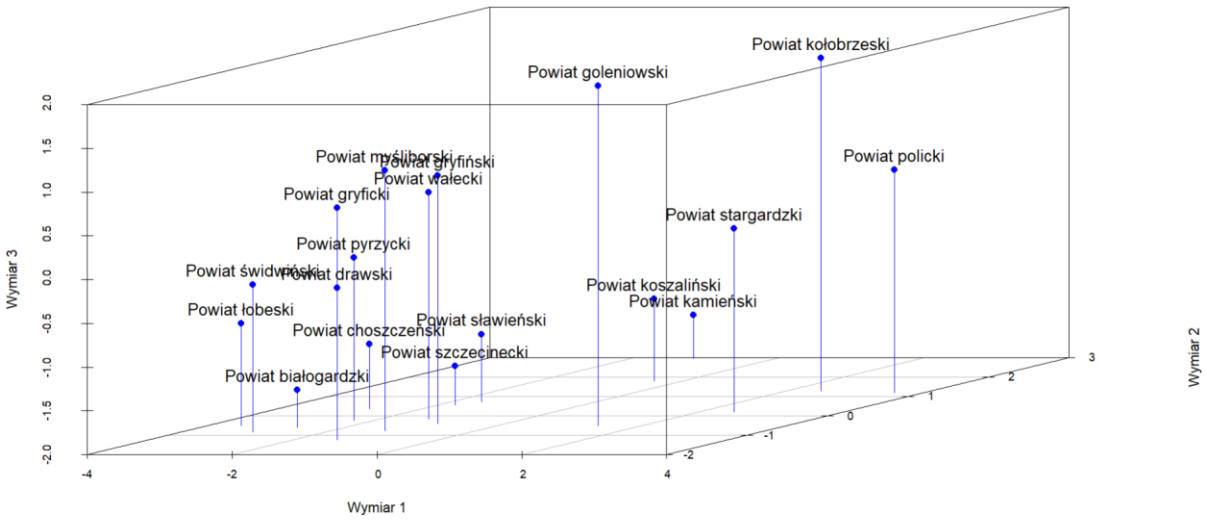
Tabela 11. Macierz odległości Sammona w drugim wymiarze

Przy wzroście wymiaru jedne wartości rosną, a inne maleją. Najbardziej interesujący jest współczynnik STRESS, który wynosi 0,04 i wskazuje na dobre dopasowanie. W celu przeprowadzenia podobnej analizy do klasycznego skalowania a przedstawione zostaną jeszcze wyniki z trzeciego wymiaru.

	P. białogardzki	**P. choszczeński**	**P. drawski**	**P. goleniowski**	**P. gryficki**
P. białogardzki	0	0,988	1,259	5,330	2,638
P. choszczeński	0,988	0	1,337	5,233	2,803
P. drawski	1,259	1,337	0	4,235	1,487
P. goleniowski	5,330	5,233	4,235	0	3,136
P. gryficki	2,638	2,803	1,487	3,136	0

Tabela 12. Macierz odległości Sammona w trzecim wymiarze

Po podniesieniu wymiaru do trzeciego współczynnik STRESS wynosi 0,01, co wskazuje na bardzo dobre dopasowanie, ponieważ zwiększenie wymiaru pozwoliło na lepsze odwzorowanie relacji między punktami, co w przypadku skalowania Sammona dało lepsze rezultaty niż metoda klasycznego skalowania, która mogła mieć trudności z uchwyceniem tych zależności.



Rysunek 9. Wizualizacja skalowania Sammona dla powiatów województwa zachodniopomorskiego

Ponieważ trzeci wymiar dla skalowania Sammona wykazał najlepsze dopasowanie, został on zwizualizowany w taki sam sposób, jak w metodzie klasycznego skalowania. Skalowanie Sammona, podobnie jak metoda klasyczna, pozwala na zobrazowanie relacji między powiatami.

- Grupa 1

  Powiaty: policki
- Grupa 2

  Powiaty: goleniowski, kołobrzeski

- Grupa 3
  Powiaty: kamieński, koszaliński, stargardzki.
- Grupa 4
  Powiaty: białogardzki, choszczeński, drawski, gryficki, gryfiński, myśliborski, pyrzycki, sławieński, szczeciński, wałecki, łobeski.

Ponieważ wszystkie powiaty zostały pogrupowane tak samo, jak w metodzie klasycznego skalowania wielowymiarowego, pominięte zostało ponowne opisywanie podobieństw między nimi. W przypadku tej metody można zauważyć drobne różnice w ułożeniu obiektów prawdopodobnie jest to spowodowane zauważeniem przez metodę pewnych różnic i podobieństw w zmiennych, które nie zostały wykryte wcześniej. Ta wizualizacje można uznać na lepszą i dokładniejszą sugerując się lepszą wartością współczynnika dopasowania.

Podsumowanie

Analiza PCA spełniła wymagane założenia, a wybrane dwie główne składowe wyjaśniły ponad 94% zmienności, z pierwszą obejmującą wszystkie zmienne równomiernie, a drugą wyodrębniającą zmienne związane z mieszkaniami i ruchem naturalnym. Wizualizacja przedstawiła rozmieszczenie powiatów przy informacjach dostarczonych przez pierwsze dwie składowe.

Skalowanie wielowymiarowe odwzorowało relacje między powiatami w zredukowanej przestrzeni, przy czym metoda Sammona osiągnęła lepsze dopasowanie (STRESS 0,01) niż klasyczne (STRESS 0,14). Wyniki podkreśliły różnice między powiatami rolniczymi a zurbanizowanymi, z trzecim wymiarem jako najlepszym do interpretacji i wskazania różnic demograficznych i gospodarczych.

Analiza głównych składowych i skalowanie wymiarowe wykazują się pewnymi różnicami w wynikach, jednak nie różnią się one drastycznie. Na pewno różnice wynikały z różnych wymiarów w wizualizacjach, jednak można uznać, że wyniki grupowały badane powiaty w bardzo podobny sposób.

Wizualizacje przeprowadzono dla wybranych powiatów województwa zachodniopomorskiego, co ułatwiło analizę zależności i wyciąganie wniosków, choć przy mniejszej próbie należy uwzględnić pewne rozbieżności wynikające z obliczeń.

Zalety analizy głównych składowych:

- Większa ilość kryteriów znajdujących optymalną liczbę głównych składowych, co pozwala łatwiejsze i dokładniejsze ustalenie liczby głównych składowych.
- Wyjaśnianie zmienności za pomocą wariancji.
- Łatwość obliczeń , aby poprawnie obliczyć PCA (osobista preferencja).

Wady analizy głównych składowych:

- Duża liczba założeń, która w jakimś stopniu może zniekształcać wyniki lub blokować wykorzystywanie metody.
- Wrażliwość na różne jednostki, w przypadku różnych jednostek PCA wymaga standaryzacji.
- Interpretacja głównie przy pomocy wizualizacji.
- Problemy z wizualizacją przy większych zbiorach danych.

Zalety skalowania wielowymiarowego:

- Możliwość interpretacji za pomocą wizualizacji, ale również macierzy odległości.
- Mniejsza liczba założeń, która pozwala na analizę różnorodnych zbiorów danych.
- Istnienie współczynników determinacji, które pozwalają w bardzo łatwy sposób ocenić dopasowanie powstałego skalowania.

Wady skalowania wielowymiarowego:

- Obliczeniowo bardziej wymagające od PCA.
- Problemy z wizualizacją przy większych zbiorach danych.
- Duże różnice w miarach dopasowania przy różnych metodach.

Podsumowując, wyniki analizy głównych składowych i skalowania wielowymiarowego wykazały pewne różnice, jednak można zauważyć także istotne podobieństwa. Obie metody dostarczyły zbliżonych wniosków dotyczących struktury danych, choć każda z nich uwypukliła inne aspekty relacji między zmiennymi. Dzięki temu analiza była bardziej wszechstronna i pozwoliła na pełniejsze zrozumienie zależności w danych.

Spis Tabel

Spis rysunków