Analiza jakości życia w województwach Polski

oraz porównanie jej wyników w latach 2008 i 2022

Agnieszka Olechnowicz (211023), Patrycja Pobuta (211028),

Patrycja Król (230077), Radosław Polak (229500)

Spis treści

1. Streszczenie

Praca poświęcona jest analizie jakości życia w województwach Polski w latach 2008 oraz 2022. Badanie opiera się na definicji jakości życia przyjętej przez Światową Organizację Zdrowia (WHO), która uwzględnia zarówno subiektywne oceny jednostek, jak i obiektywne wskaźniki gospodarcze, społeczne i środowiskowe. Analiza koncentruje się na mierzalnych aspektach jakości życia, takich jak poziom zatrudnienia, dostęp do opieki zdrowotnej, bezpieczeństwo, jakość środowiska, infrastruktura transportowa, warunki mieszkaniowe oraz inwestycje publiczne.

Wykorzystano metody analizy skupień, w tym metodę Warda i k-średnich, aby pogrupować województwa na podstawie podobieństw i różnic w poziomie życia mieszkańców. Wyniki badania pozwoliły na identyfikację regionów o wysokiej oraz niskiej jakości życia oraz wskazanie kluczowych zmian zachodzących między rokiem 2008 a 2022. Otrzymane rezultaty mogą stanowić podstawę do planowania polityki regionalnej oraz działań na rzecz poprawy warunków życia w Polsce.

2. Słowa kluczowe

Jakość życia, Analiza regionalna, Województwa Polski, Porównanie międzyroczne, Metoda Warda, Metoda k-średnich, Poziom życia, Różnice regionalne, Lata 2008 i 2022, Analiza skupień, Polityka regionalna.

3. Wprowadzenie

Tematem badania jest analiza jakości życia w województwach Polski w latach 2008 oraz 2022. Według definicji WHO, jakość życia (QOL, ang. quality of life) to “subiektywna ocena przez jednostkę jej sytuacji życiowej w kontekście do kultury i systemu wartości, w których oraz związana z jej celami, oczekiwaniami, standardami oraz obawami. Jest to szerokie pojęcie, na które w złożony sposób wpływa zdrowie fizyczne człowieka, jego stan psychiczny, osobiste wierzenia, relacje społeczne oraz charakterystyczne cechy ich środowiska.” Na poziom jakości życia mają więc wpływ zarówno kwestie czysto osobiste (np. więzi społeczne, osobowość czy odczuwane emocje) jak i czynniki bardziej obiektywne, które są łatwiejsze do zmierzenia (np. dochody czy warunki mieszkaniowe). Zagadnienie to jest kluczowe w ocenie, który region Polski najbardziej sprzyja mieszkańcom, co jest szczególnie istotne dla osób młodych w kontekście planowania przyszłości. Dodatkowo identyfikacja obszarów o niższym poziome jakości. z życia może być sygnałem do podjęcia działań mających na celu poprawę warunków życia i dobrostanu mieszkańców tych regionów.

W badaniu ograniczono się do oceny zjawiska w kontekście ogólnych czynników, które można w łatwy sposób zmierzyć, a pominięto aspekty całkowicie indywidualne. Jako poziom jakości życia rozumiemy więc analizę czynników z różnych pól gospodarki tj. rynek pracy i dochody, opieka zdrowotna, bezpieczeństwo, łączność, środowisko naturalne, edukacja, handel i inflacja, czynniki społeczne, mieszkalnictwo oraz inwestycje publiczne.

4. Cel i zakres badania

Badanie ma na celu przeanalizowanie jakości życia ludności Polski w różnych województwach przy wykorzystaniu takich czynników jak zatrudnienie, warunki społeczne, środowiskowe, transportowe, finansowe oraz kwestie bezpieczeństwa panujące w danym województwie w latach 2008 i 2022 oraz ich klasteryzację ze względu na te cechy. Stworzenie skupisk pozwoli zauważyć, które jednostki są na zbliżonym poziomie badanego zjawiska, a które znacznie się od siebie różnią.

Zakres badania obejmuje analizę skupień województw Polski na podstawie danych z lat 2008 i 2022 pochodzących z GUS dotyczących różnych czynników, które składają się na poziom jakości życia. Wybrane metody zastosowano dla każdego roku oddzielnie, oceniono jakość uzyskanych skupień, a także dokonano porównania wyników otrzymanych na przestrzeni lat.

5. Przegląd literatury

Analiza jakości życia jest często poruszanym tematem w badaniach, biorąc pod uwagę nie tylko różne obszary geograficzne, ale także różnorodne czynniki na nią wpływające. W kontekście tej pracy szczególną uwagę zwrócono na próby badania problemu i jego rezultaty na poziomie polskich województw.

Nowak (2017) do oceny jakości życia wykorzystała 21 zmiennych z różnych dziedzin życia, takich jak warunki materialne, aktywność ekonomiczna, zdrowie, edukacja oraz jakość środowiska zamieszkania, gdzie dane dotyczyły 2010 oraz 2015 roku. Dla każdego z województw obliczyła syntetyczną miarę rozwoju Hellwiga i na jej podstawie w dokonała ich podziału na cztery grupy. W obu latach w zbiorze najlepszych pod względem poziomy jakości życia województw znalazły się śląskie, mazowieckie oraz dolnośląskie, natomiast do najsłabszych należały podkarpackie oraz warmińsko-mazurskie. Dodatkowo przeprowadziła także analizę skupień metodą Warda, która znacznie wyodrębniła w obu latach województwo mazowieckie (ze względu na najwyższe wartości cech), a w 2010 także podkarpackie (tym razem z powodu niskich wartości na tle kraju).

Majecka i Nowak (2019) do oceny jakości życia również wykorzystały metodę Hellwiga, tym razem dla 2016 roku. Na podstawie 22 zmiennych z różnych obszarów tematycznych dla 2016 roku obliczyły miernik syntetyczny, a następnie korzystając z jej średniej arytmetycznej i odchylenia standardowego wyznaczyły cztery grupy województw o zbliżonym poziomie jakości życia. W grupie I znalazły się obiekty o najwyższym poziomie (woj. mazowieckie, lubuskie i wielkopolskie), a grupie IV – najniższym (łódzkie i lubelskie).

6. Zmienne wybrane do analizy

Do badania jakości życia wybrane zostały zmienne, które w znaczny sposób wpływają na poziom tego zjawiska. Pierwotnie pod uwagę wziętych zostało 14 zmiennych, dla których przeprowadzono analizę zmienności oraz korelacji, na podstawie której odrzucono część z nich. Ostatecznie wybranych zostało 9 cech, które przedstawiono poniżej (na czerwono zostały zaznaczony destymulanty, a na zielono stymulanty):

X1 – Stopa bezrobocia rejestrowanego [%]
X2 – Przeciętny miesięczny dochód rozporządzalny na 1 osobę [zł]
X3 – Przestępstwa stwierdzone przez Policję ogółem na 1000 mieszkańców [l. przestępstw]
X4 – Linie kolejowe ogółem na 100 km2 [km]
X5 – Udział powierzchni terenów zieleni w powierzchni ogółem [%]
X6 – Emisja zanieczyszczeń powietrza z zakładów szczególnie uciążliwych [t/r]
X7 – Absolwenci na 10 tys. ludności [liczba osób]
X8 – Rozwody na 10 tys. ludności [liczba osób]
X9 – Wypadki drogowe na 100 tys. ludności [liczba wypadków]

Poniżej przedstawiono szczegółowy opis każdej zmiennej oraz uzasadnienie ich klasyfikacji.

1. Destymulanty
Destymulanty to zmienne, których wzrost wpływa negatywnie na jakość życia, co objawia się w formie problemów społecznych, ekonomicznych lub środowiskowych.

- **X1 – Stopa bezrobocia rejestrowanego [%]**
  - **Opis:** Stopa bezrobocia rejestrowanego wskazuje odsetek osób bez pracy, które zarejestrowały się jako bezrobotne.
  - **Uzasadnienie:** Wysoka stopa bezrobocia świadczy o trudnej sytuacji na rynku pracy, co prowadzi do ograniczenia dochodów i zwiększenia ubóstwa w danym regionie. Bezrobocie negatywnie wpływa na poczucie stabilności życiowej i dobrobyt mieszkańców.
- **X3 – Przestępstwa stwierdzone przez Policję ogółem na 1000 mieszkańców [liczba przestępstw]**
  - **Opis:** Liczba przestępstw zarejestrowanych przez policję w przeliczeniu na 1000 mieszkańców.
  - **Uzasadnienie:** Wysoka przestępczość wpływa na poczucie bezpieczeństwa mieszkańców. W regionach z dużą liczbą przestępstw, mieszkańcy mogą doświadczać większego stresu i obaw o własne bezpieczeństwo, co obniża ogólną jakość życia.
- **X6 – Emisja zanieczyszczeń powietrza z zakładów szczególnie uciążliwych [t/r]**
  - **Opis:** Wskaźnik ten mierzy ilość zanieczyszczeń emitowanych przez zakłady przemysłowe w danym regionie.
  - **Uzasadnienie:** Wysoka emisja zanieczyszczeń powietrza ma negatywny wpływ na zdrowie mieszkańców oraz na jakość środowiska. Problemy zdrowotne związane z zanieczyszczeniem powietrza mogą prowadzić do obniżenia komfortu życia oraz wzrostu wydatków na opiekę zdrowotną.

- **X8 – Rozwody na 10 tys. ludności [liczba osób]**
    - **Opis:** Wskaźnik liczby rozwodów w przeliczeniu na 10 tys. mieszkańców.
    - **Uzasadnienie:** Wysoka liczba rozwodów może świadczyć o problemach w relacjach międzyludzkich, co wpływa na dobrostan psychiczny mieszkańców. Problemy rodzinne mogą prowadzić do stresu, a także obniżenia jakości życia.
- **X9 – Wypadki drogowe na 100 tys. ludności [liczba wypadków]**
    - **Opis:** Liczba wypadków drogowych w przeliczeniu na 100 tys. mieszkańców.
    - **Uzasadnienie:** Wysoka liczba wypadków drogowych wskazuje na niebezpieczne warunki na drogach, co wpływa na poczucie bezpieczeństwa mieszkańców oraz ich komfort życia. Częste wypadki mogą skutkować stratami w ludziach oraz zwiększonymi wydatkami na opiekę zdrowotną i rehabilitację.

2. Stymulanty

Stymulanty to zmienne, których wyższe wartości przyczyniają się do poprawy jakości życia, wskazując na pozytywne aspekty życia społecznego, gospodarczego lub środowiskowego.

- **X2 – Przeciętny miesięczny dochód rozporządzalny na 1 osobę [zł]**
    - **Opis:** Wartość przeciętnego miesięcznego dochodu, jakim dysponuje jedna osoba w danym regionie.
    - **Uzasadnienie:** Wyższy dochód rozporządzalny oznacza lepsze warunki materialne mieszkańców, co sprzyja zaspokajaniu ich potrzeb i zwiększa możliwości inwestycyjne, zarówno w sferze osobistej, jak i zawodowej.
- **X4 – Linie kolejowe ogółem na 100 km² [km]**
    - **Opis:** Wskaźnik ilości linii kolejowych w przeliczeniu na 100 km² powierzchni.
    - **Uzasadnienie:** Rozwinięta sieć kolejowa zwiększa dostępność transportową, co wpływa na mobilność mieszkańców. Umożliwia łatwiejszy dostęp do pracy, edukacji oraz innych usług, co pozytywnie wpływa na jakość życia.
- **X5 – Udział powierzchni terenów zieleni w powierzchni ogółem [%]**
    - **Opis:** Procentowy udział terenów zielonych w całkowitej powierzchni danego regionu.
    - **Uzasadnienie:** Większy udział terenów zielonych przyczynia się do poprawy jakości środowiska oraz zdrowia mieszkańców. Tereny zielone sprzyjają rekreacji, poprawiają jakość powietrza i estetykę przestrzeni miejskiej.
- **X7 – Absolwenci na 10 tys. ludności [liczba osób]**
    - **Opis:** Liczba absolwentów na 10 tys. mieszkańców danego regionu.
    - **Uzasadnienie:** Wyższy wskaźnik absolwentów sugeruje lepszy dostęp do edukacji, co wpływa na rozwój społeczny i ekonomiczny regionu. Wysoki poziom wykształcenia przekłada się na lepsze możliwości zatrudnienia i wzrost innowacyjności.

7. Wstępna analiza danych

5.1. Statystyki opisowe

W pierwszej kolejności zostały obliczone statystyki opisowe wszystkich zmiennych dla obu lat. Poniższe tabele przedstawiają, jak kształtowały się ich wartości średnie, odchylenie standardowe, mediana, wartości minimalne i maksymalne, zakres, skośność, kurtoza, błąd standardowy oraz 1 i 3 kwartyl.

	mean	sd	median	min	max	range	skew	kurtosis	se	Q0.25	Q0.75
X1	10.56	2.98	9.90	6.40	16.80	10.40	0.40	-0.56	0.74	8.18	13.07
X2	1014.58	123.44	1015.93	791.27	1336.46	545.19	0.76	2.36	30.86	946.32	1064.16
X3	28.31	5.26	27.78	18.14	37.59	19.45	0.06	-0.12	1.32	25.99	31.87
X4	6.95	3.16	6.40	3.80	17.40	13.60	2.62	8.46	0.79	5.22	7.23
X5	0.58	0.44	0.43	0.18	1.75	1.57	1.75	2.40	0.11	0.34	0.54
X6	13519938.50	12478701.05	10563720.50	1381026.00	42672053.00	41291027.00	1.26	0.88	3119675.26	4537089.50	17285131.25
X7	104.56	16.89	106.00	72.00	138.00	66.00	-0.24	0.40	4.22	96.00	114.50
X8	16.79	4.26	17.10	8.90	22.40	13.50	-0.53	-0.59	1.07	14.88	19.92
X9	124.88	27.60	119.65	89.80	187.10	97.30	0.63	-0.13	6.90	103.50	144.10

Tabela 1. Statystyki opisowe dla 2008 roku

	mean	sd	median	min	max	range	skew	kurtosis	se	Q0.25	Q0.75
X1	5.90	1.87	5.70	2.90	8.80	5.90	0.13	-1.33	0.47	4.40	7.42
X2	2185.71	211.39	2188.58	1829.05	2685.93	856.88	0.49	0.80	52.85	2060.67	2305.17
X3	21.65	5.19	20.91	12.93	35.08	22.15	1.03	2.20	1.30	19.26	23.26
X4	6.62	2.65	6.25	3.80	15.10	11.30	2.35	7.07	0.66	5.05	6.80
X5	0.63	0.46	0.47	0.21	1.85	1.64	1.71	2.28	0.12	0.37	0.61
X6	12726459.62	12381079.73	7561180.00	1610597.00	40817473.00	39206876.00	1.30	0.62	3095269.93	4854056.50	15667014.75
X7	69.06	22.61	65.00	29.00	108.00	79.00	0.19	-0.68	5.65	56.50	85.50
X8	15.59	2.36	15.75	11.60	20.20	8.60	0.00	-0.52	0.59	13.97	17.30
X9	54.79	14.90	53.05	29.10	92.60	63.50	0.75	1.74	3.73	46.10	64.67

Tabela 2. Statystyki opisowe dla 2022 roku

Powyższe wyniki zinterpretowane zostały dla każdej ze zmiennych, porównując przy tym ich poziomy na przestrzeni lat.

X1 – Stopa bezrobocia rejestrowanego

Średnio w 2008 roku stopa bezrobocia wynosiła 10,56%, a w 2022 roku spadła do 5,9%. Obserwujemy zatem znaczny spadek bezrobocia, co świadczy o dużej poprawie rynku pracy. Podobnie zmienia się odchylenie standardowe wynoszące w 2008 roku 2,98, a w 2022 – 1,87, więc sytuacja zaczęła się wyrównać we wszystkich województwach. Można zauważyć, że najwyższy poziom bezrobocia w 2008 wynosił 16,8% - w warmińsko-mazurskim, a w 2022, 8,8% w podkarpackim. W 2022 roku najniższa wartość wynosiła 2,9% co jest zdecydowanie niższą wartością od minimalnej z 2008 roku, która wynosiła 6,4%, co wskazuje na zdecydowaną poprawę na tle ostatnich lat. Mediana stopy bezrobocia wyniosła kolejno 9,9% oraz 5,7%, coraz bardziej zbliżając się do wartości średnich w tych latach. Dodatnie skośności wskazują, że rozkład prawdopodobieństwa cechy charakteryzuje się asymetrią prawostronną, jednak ich poziom jest na tyle niski, szczególnie w roku 2022 (0,13), że nie jest ona znaczna. Kurtoza natomiast w obu przypadkach jest ujemna, co oznacza, że wartości są bardziej rozproszone wokół średniej niż w rozkładzie normalnym.

X2 – Przeciętny miesięczny dochód rozporządzalny na 1 osobę

Średnio w 2008 roku przeciętny miesięczny dochód rozporządzalny na 1 osobę wynosił 1014,58 zł, natomiast w 2022 roku wzrósł do 2185,71 zł. Możemy zauważyć wyraźny trend wzrostowy dochodu w badanym okresie, co może świadczyć o ogólnej poprawie sytuacji ekonomicznej. Również odchylenie standardowe wzrosło wraz z dochodem, co może sugerować coraz większe zróżnicowanie dochodów w populacji. Najniższy przeciętny dochód w 2008 roku wyniósł 791,27 zł, natomiast najwyższy w 2022 roku – 2685,93 zł. Możemy zauważyć, że różnica między najniższym a najwyższym dochodem znacznie się zwiększyła w ciągu lat, a zestawiając to z odchyleniem, może wskazywać to na rosnącą dysproporcję dochodową w społeczeństwie. Dodatnie skośności wskazują, że rozkład cechy charakteryzuje się asymetrią prawostronną, natomiast dodatnia kurtoza, że jest on leptokurtyczny, a wartości są bardziej skupione wokół średniej niż w rozkładzie normalnym.

X3 – Przestępstwa stwierdzone przez Policję ogółem na 1000 mieszkańców

W 2008 roku średnia liczba przestępstw stwierdzonych przez Policję na 1000 mieszkańców wynosiła 28,31. Po latach zaobserwowano spadek tej liczby do 21,65 w 2022 roku. Odchylenie standardowe w badanych okresach utrzymywało się na podobnym poziomie. Zatem średnia liczba przestępstw przez 15 lat spadła natomiast zróżnicowanie nie zmieniło się w sposób znaczący. Najwyższa odnotowana liczba przestępstw w 2008 roku należała do woj. lubuskiego, natomiast w 2022 roku już do województwa śląskiego. Najniższą natomiast liczbą przestępstw w obu latach charakteryzuje się woj. podkarpackie. W roku 2008 rozkład prawdopodobieństwa cechy był praktycznie równy rozkładowi symetrycznemu, natomiast w 2022 zaobserwowano silniejszą asymetrię prawostronną. Kurtoza dla pierwszego roku miała poziom ujemny (-0,12), natomiast dla drugiego znacznie wyższy dodatni (2,20).

X4 – Linie kolejowe ogółem na 100 km2

Średnia długość linii kolejowych na 100 km2 w 2008 roku wynosiła 6,95 km, w 2022 spadła do 6,62 km. Oznacza to, że liczba linii kolejowych spadła, niektóre linie zostały usunięte, inne wybudowane, co można wnioskować przez spadek odchylenia standardowego na przestrzeni tych lat. Zróżnicowanie i marginalizacja między województwami zmniejszyła się, jednak nadal podlaskie pozostaje regionem z najmniejszym zagęszczeniem linii kolejowych, a śląskie z największym. Skośność w obu przypadkach na poziomie ok. 2,5, a kurtoza silnie dodatnia (8,46 i 7,07).

X5 – Udział powierzchni terenów zieleni w powierzchni ogółem

Udział terenów zielonych, generalnie utrzymuję się na podobnym poziomie. Średnia w latach 2008 i 2022 wynosiła kolejno 0,58% oraz 0,63%. Odchylenie standardowe również praktycznie się nie zmienił. Zatem procent terenów zielonych jest utrzymany w miarę na równym poziomie, a zróżnicowanie nie zmienia się znacząco. Największą wartość udziału powierzchni terenów zielonych utrzymuje woj. podkarpackie, natomiast najniższą woj. Podlaskie. Skośność w obu przypadkach na poziomie ok. 1,7, tak samo jak kurtoza ok. 2,30.

X6 – Emisja zanieczyszczeń powietrza z zakładów szczególnie uciążliwych

W 2008 roku średnia emisja zanieczyszczeń powietrza z zakładów szczególnie uciążliwych wynosiła 13519938,5 t/r. W kolejnych latach obserwowano tendencję spadkową osiągając 12726459,6 t/r w 2022 roku. Odchylenie standardowe również spadło. Można wnioskować, że spadek emisji na przestrzeni lat jest zwiększeniem uwagi społeczeństwa na ekologię oraz wprowadzania przez państwo i Unię Europejską ustaw na ten temat. Województwo śląskie w 2008 roku było liderem w tym zakresie, natomiast w 2022 to łódzkie było największym emitentem zanieczyszczeń. Wiąże się to z faktu, że są to duże obszary przemysłowe – pierwsze z nich od razu przywodzi na myśl kopalnie węgla i huty jako źródło tej emisji, a także rozwój przemysłu ciężkiego, transport oraz największa gęstość zaludnienia, drugie również charakteryzuje się dużym udziałem przemysłu włókienniczego i odzieżowego czy chemicznego, ale także również emisją z transportu i gospodarstw domowych. Najczystszym powietrzem charakteryzuje się woj. warmińsko-mazurskie. Skośność dla tej cechy wyniosła w obu latach ok. 1,30, a kurtoza ok. 0,7.

X7 – Absolwenci na 10 tys. ludności

W 2008 roku średnia liczba absolwentów wynosiła 104,56 na 10 tys. mieszkańców, natomiast w 2022 zanotowano spadek do poziomu 69,06. Odchylenie zwiększyło się z 16,89 do 22,61. Widzimy zatem znaczne obniżenie ilości absolwentów na przestrzeni lat, a także duże zróżnicowanie między województwami. W 2008 najwięcej absolwentów było w mazowieckim, a w 2022 w małopolskim. Natomiast najniższa ilość absolwentów w 2008 należała do opolskiego, a w 2022 do lubuskiego. Spadek może wynikać z podejścia społeczeństwa do ukończenia uczelni wyższych, ale również wybuchu pandemii w roku 2020, który pośrednio został objęty okresem badania. Różnice występują także w skośności i kurtozie, które w obu latach przyjęły przeciwne znaki. Mimo że wartości są bliskie zeru to w 2008 zaobserwowano lekką asymetrię lewostronną, a w 2022 prawostronną. W przypadku kurtozy znaki były odwrotne – dla 2008 roku uzyskano wartość dodatnią, a dla 2022 ujemną, jednak ponownie nie był to wysoki poziom (+/- 0,5).

X8 – Rozwody na 10 tys. ludności

Średnia liczba rozwodów w 2008 roku wynosiła 16,79 na 10 tys. rozwodów, a w 2022 roku 15,59. Odchylenie standardowe stopniowo spadało od poziomu 4,26 w 2008 do 2,36 w 2022 roku. Podobnie z najwyższymi i najniższymi wartościami, które spadały z roku na rok. Najwięcej rozwodów miało miejsce w województwie zachodniopomorskim, a najmniej w 2008 roku w lubelskim, a następnie w podkarpackim. Powody takiego spadku mogą mieć różne przyczyny, jedną z głównych może być ogólny spadek małżeństw, co logicznie prowadzi do mniejszej ilości rozwodów, ale również mogą to być czynniki takie jak większy dostęp do psychologów i terapii dla małżeństw. Kurtoza w obu latach była na poziomie ok. -0,5, natomiast skośność w 2008 roku wyniosła -0,53, a w 2022 – 0,00, czyli miała rozkład symetryczny.

X9 – Wypadki drogowe na 100 tys. ludności

Średnia liczba wypadków drogowych znacząco spadła od 2008 roku z poziomu 124,88 do 54,79 w 2022 roku. Odchylenie standardowe również spadło w podobny sposób od wartości 27,6 do 14,9. Najwięcej wypadków odnotowano w województwie łódzkim, a najmniej najpierw w lubuskim, następnie kujawsko-pomorskim i w 2022 roku w podlaskim. Spadek tej liczby może być rezultatem różnych działań na rzecz poprawy bezpieczeństwa drogowego, takich jak poprawa infrastruktury drogowej, wprowadzenie środków ograniczających prędkość, czy też edukacja kierowców. Największa liczba wypadków w łódzkim, może być spowodowana dużą urbanizacją i siecią dróg takich jak rozpościerająca się autostrada A2. Skośność rozkładu prawdopodobieństwa cechy lekko prawostronna, natomiast kurtoza w pierwszym roku nieznacznie ujemna, a w drugim znacznie wyższa i dodatnia (-0,13 i 1,74).

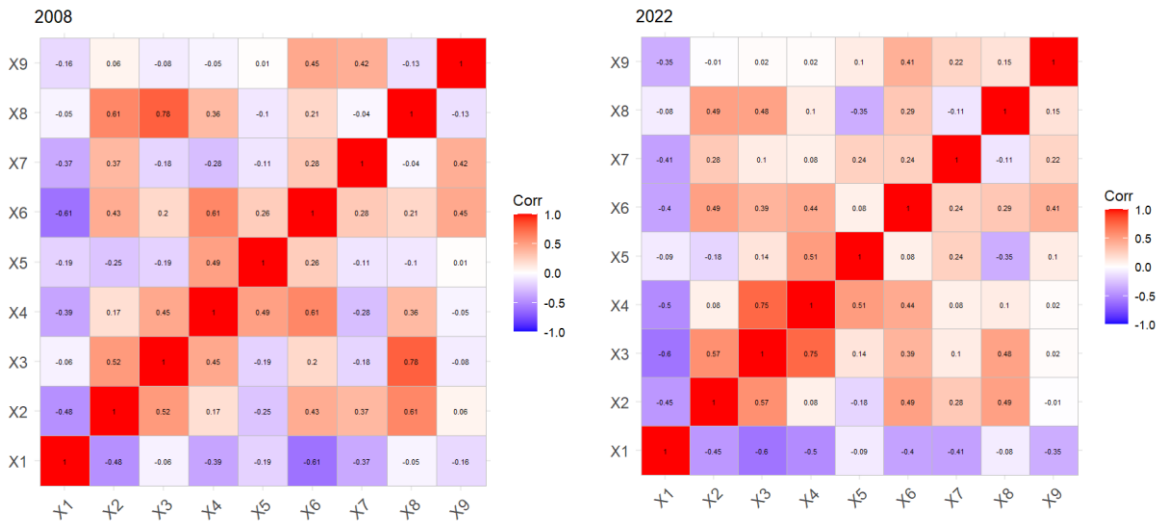
Obliczone wcześniej współczynniki zmienności dla ostatecznego zestawu zmiennych prezentują się następująco:

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]
wsp_zm08	28.16760	12.166304	18.58689	45.44273	76.22136	92.29850	16.15463	25.40134	22.10047
wsp_zm22	31.63035	9.671537	23.98699	40.03908	73.33880	97.28613	32.73154	15.16009	27.19508

Tabela 3. Wartości współczynnika zmienności dla lat 2008 i 2022

Jak można zauważyć dla prawie wszystkich zmiennych mają one wartości powyżej 10%. Jedynym przypadkiem, dla którego wartość ta jest niższa jest przeciętny miesięczny dochód rozporządzalny na 1 osobę w roku 2022, jednak ze względu na zbliżoną wartość do progu odrzucenia oraz istotne znaczenie w kontekście badanego zjawiska nie została ona usunięta.

Wyniki również przeprowadzonego wcześniej badania korelacji między zmiennymi przedstawione zostały na poniższych wykresach:

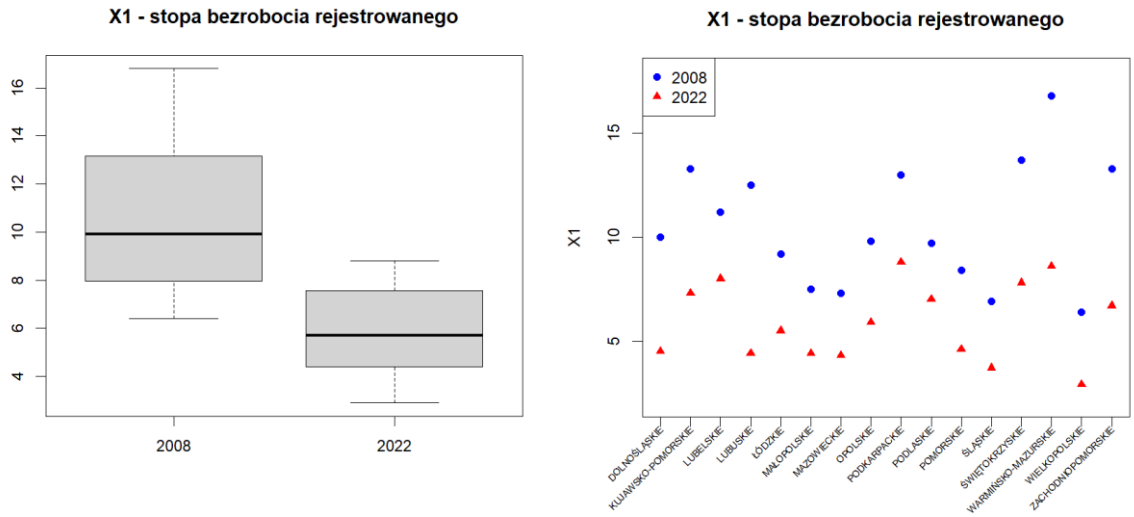


Wykres 1. Macierz korelacji między zmiennymi dla roku 2008 oraz 2022

Wśród ostatecznie dobranych zmiennych nie znajdują się więc takie, które byłyby silnie ze sobą skorelowane.

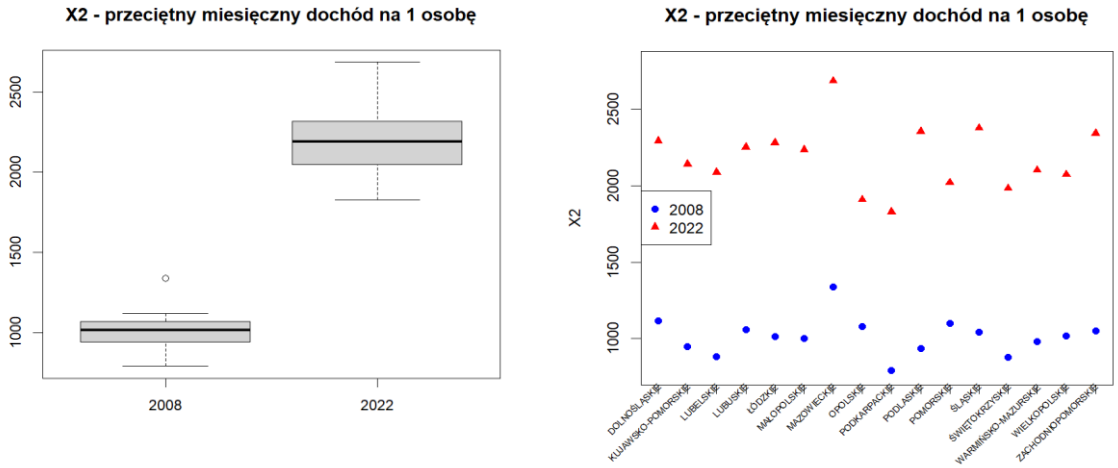
5.2. Wizualizacja danych

Każda ze zmiennych została przedstawiona na wykresie pudełkowym oraz zwykłym punktowym w zestawieniu obu badanych lat. Pozwoliło to wyraźnie ukazać jak ich wartości zmieniły się na przestrzeni lat. Boxploty umożliwiły także łatwą identyfikację obserwacji odstających. Poniżej zaprezentowane zostały wyniki tej wizualizacji.



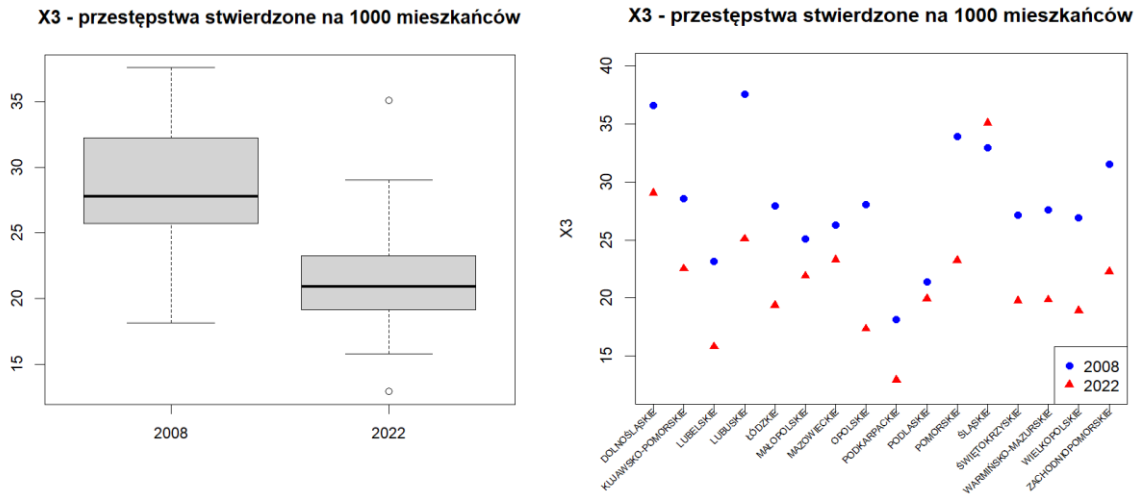
Wykres 2. Wykres pudełkowy i punktowy dla zmiennej X1 w latach 2008 i 2022

Widać wyraźny spadek stopy bezrobocia, jednak kształt wykresu punktowego w obu przypadkach jest zbliżony, co wskazuje, że różnice między województwami pozostały w większości zachowane.



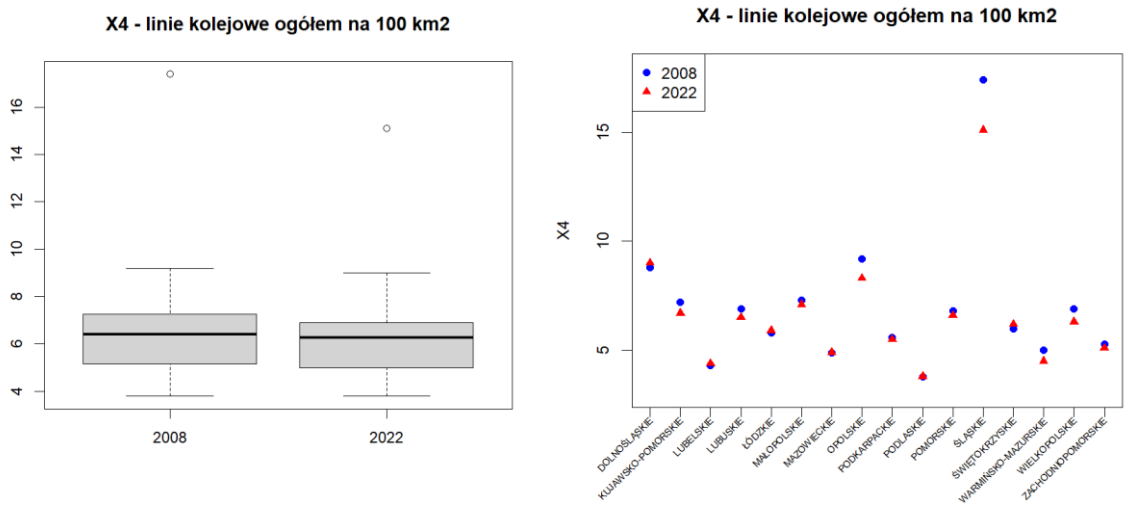
Wykres 3. Wykres pudełkowy i punktowy dla zmiennej X2 w latach 2008 i 2022

Już na tym etapie możemy zauważyć, że dla dochodu w roku 2008 występuje obserwacja odstająca o wyższej wartości od pozostałych. Na drugim wykresie można odczytać, że jest to województwo mazowieckie, co jest zgodne z intuicją związaną ze stolicą. Ponownie sytuacja na przestrzeni lat znacznie się polepszyła, a zależności pozostały bardzo podobne.



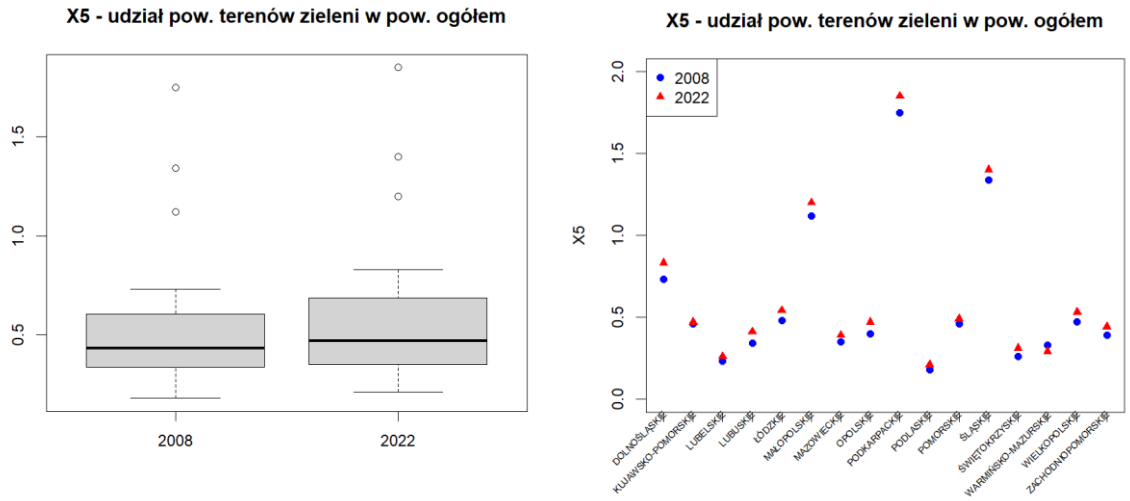
Wykres 4. Wykres pudełkowy i punktowy dla zmiennej X3 w latach 2008 i 2022

Jest zauważalne, że dla 2022 roku wystąpiła wartość odstająca znacznie wyższa od pozostałych wartości. Odpowiada ona województwu śląskiemu. Inną wartością odstającą również dla roku 2022 jest liczba przestępstw dla województwa podkarpackiego, która jest najmniejsza spośród wszystkich województw. Poza przypadkiem województwa śląskiego widać, że liczba przestępstw w poszczególnych województwach jest mniejsza dla roku 2022 w porównaniu z rokiem 2008. Zatem pod tym względem sytuacja na ogół także uległa znacznej poprawie wraz z upływem lat.



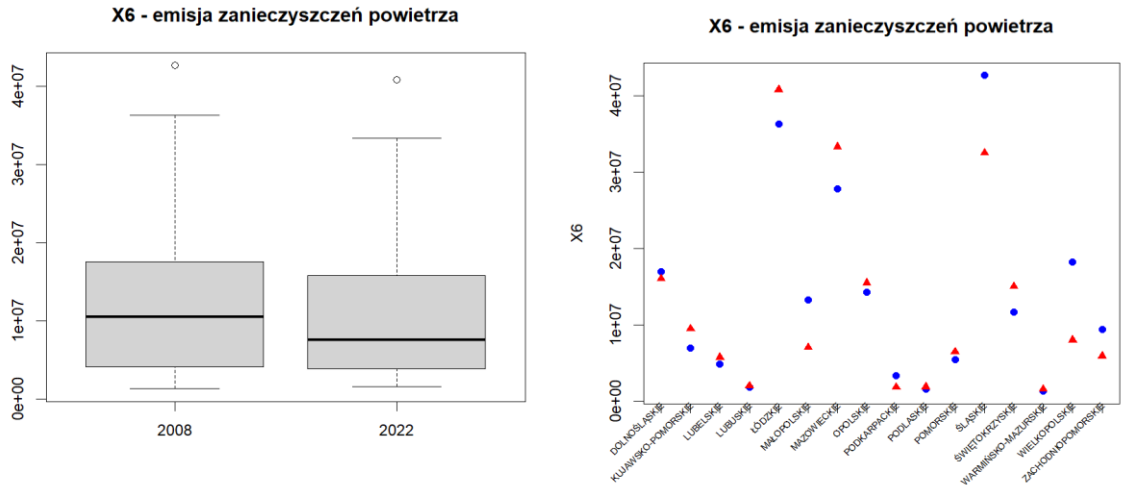
Wykres 5. Wykres pudełkowy i punktowy dla zmiennej X4 w latach 2008 i 2022

Na pierwszy rzut oka można zaobserwować, że długość linii kolejowych na 100 km^2 dla większości województw jest na dość zbliżonym poziomie w obu analizowanych latach. Z kolei dla województwa śląskiego zarówno dla roku 2008 jak i 2022 występują wartości odstające znacznie wyższe od pozostałych. Ponadto dla tej obserwacji wartość dla roku 2022 jest zauważalnie mniejsza od wartości dla 2008. Biorąc pod uwagę powyższe obserwacje można stwierdzić, że sytuacja na przestrzeni lat nie uległa znaczącym zmianom, poza jednym z województw.



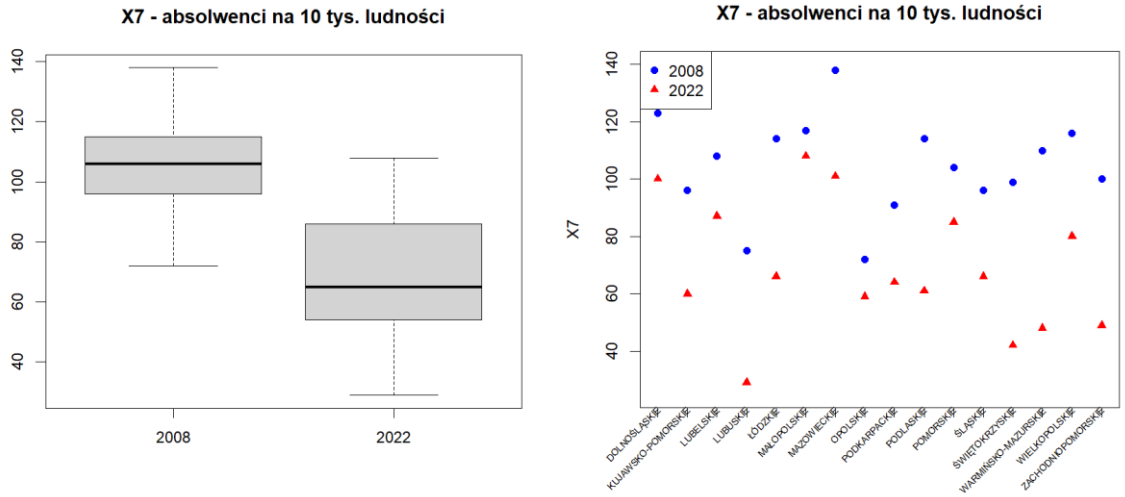
Wykres 6. Wykres pudełkowy i punktowy dla zmiennej X5 w latach 2008 i 2022

W tym przypadku występują po trzy wartości odstające znacznie większe od pozostałych dla obu lat i dla tych samych województw (małopolskie, podkarpackie, śląskie). Udział powierzchni terenów zielonych w powierzchni ogółem jest dla większości województw (w tym dla województw odstających) nieco wyższy dla roku 2022, więc można powiedzieć, że ogólna sytuacja uległa poprawie.



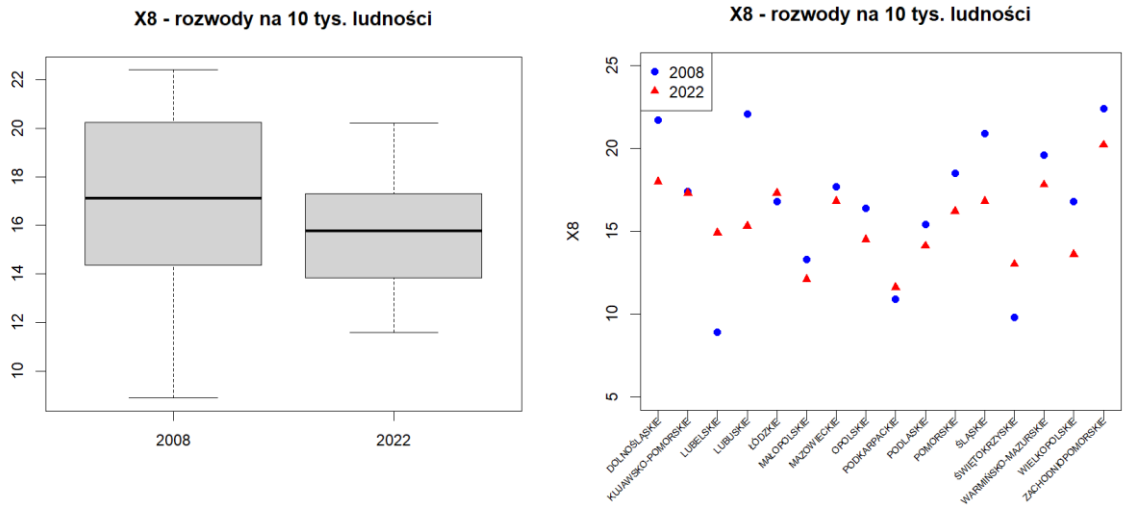
Wykres 7. Wykres pudełkowy i punktowy dla zmiennej X6 w latach 2008 i 2022

Na podstawie powyższych wizualizacji widać, że zależność między emisją zanieczyszczeń powietrza nie ma jasnej uogólnionej tendencji. Występują województwa, gdzie emisja zanieczyszczeń zmalała (np. małopolskie, śląskie, wielkopolskie) oraz takie, gdzie ta emisja wzrosła wraz z upływem lat (np. łódzkie, mazowieckie, świętokrzyskie). Są także przypadki województw, gdzie emisja zanieczyszczeń na przestrzeni lat nie uległa znacznej zmianie (np. lubuskie, podlaskie, warmińsko-mazurskie). Wartości odstające odnoszą się do województwa śląskiego w roku 2008 i do województwa łódzkiego w 2022.



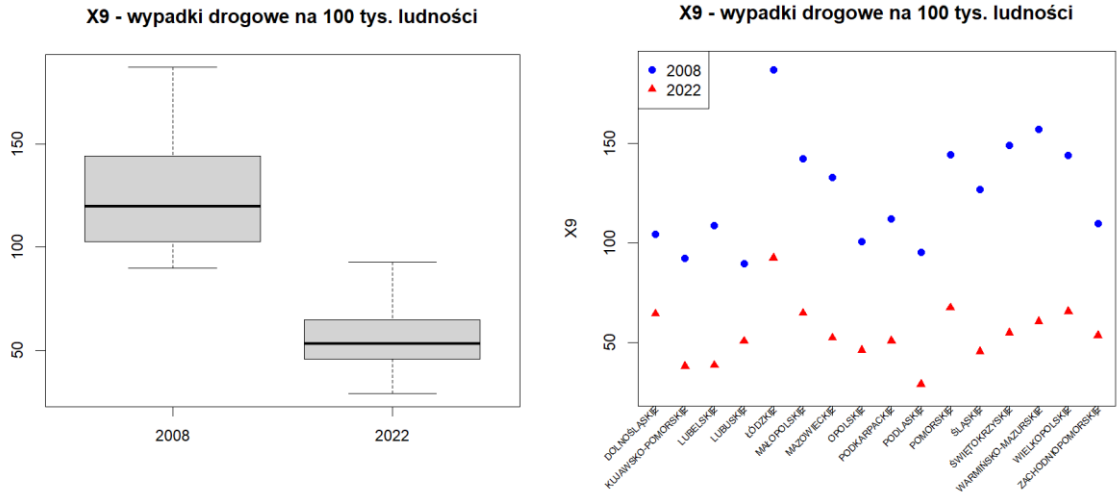
Wykres 8. Wykres pudełkowy i punktowy dla zmiennej X7 w latach 2008 i 2022

Dla wszystkich województw bez wyjątku obserwuje się znaczny spadek liczby absolwentów na 10 tys. ludności, czyli sytuacja w tej kwestii mocno się pogorszyła na przestrzeni lat.



Wykres 9. Wykres pudełkowy i punktowy dla zmiennej X8 w latach 2008 i 2022

Na podstawie obu wykresów odnośnie liczby rozwodów widać, że sytuacja w 2022 roku uległa poprawie oraz pewnej stabilizacji, co widać po zauważalnie mniejszym rozrzucie wartości. Na ogół w większości województw liczba ta zmalała względem roku 2008, w pojedynczych przypadkach wzrosła a w innych pozostała na zbliżonym poziomie.



Wykres 10. Wykres pudełkowy i punktowy dla zmiennej X9 w latach 2008 i 2022

We wszystkich województwach sytuacja dotycząca wypadków drogowych uległa znacznej poprawie wraz z upływem lat. Dodatkowo także w tym przypadku sytuacja ta bardziej się ustabilizowała co charakteryzuje o wiele mniejsze rozproszenie wartości z roku 2022 względem wartości z 2008.

5.3. Braki danych

Wybrane do badania dane pochodzące z Banku Danych Lokalnych GUS były kompletne i nie występowały w nich żadne braki.

5.4. Obserwacje odstające

Obserwacje odstające zostały zidentyfikowane na poziomie wizualizacji danych poprzez zastosowanie wykresów pudełkowych. Jeszcze raz przedstawione zostały w tabeli:

Zmienna	2008	2022
X2	mazowieckie (1336,46)	
X3		podkarpackie (12,93), śląskie (35,08)
X4	śląskie (17,4)	śląskie (15,1)
X5	małopolskie (1,12), podkarpackie (1,75), śląskie (1,34)	małopolskie (1,2), podkarpackie (1,85), śląskie (1,4)
X6	śląskie (42672053)	łódzkie (40817473)

Tabela 4. Zestawienie obserwacji odstających dla 2008 i 2022 roku

Ze względu na charakter danych żadna z obserwacji nie została jednak usunięta, ponieważ są one związane z cechami charakterystycznymi danego regionu. Mimo że dla dochodów (X2) mazowieckie zostało sklasyfikowane jako obserwacja odstająca tylko dla roku 2008, to również w roku 2022 wartość ta była znacznie wyższa od pozostałych. Duży wpływ ma na to obecność stolicy w tym województwie. Podobnie w przypadku przestępczości (X3) – wg GUS województwo podkarpackie charakteryzuje się najniższą przestępczością od lat. Długość linii kolejowych (X4) znacznie dominuje w województwie śląskim, co jest spowodowane głównie przemysłowym charakterem obszaru oraz gęstym zaludnieniem, a także znaczeniem historycznym. Od razu można zauważyć, że także dla procentu terenów zieleni (X5)

obserwacje nietypowe pozostają te same, co wskazuje, że wynika to z pewnych cech terenów. Niezwykle intuicyjna jest również obecność województw śląskiego i łódzkiego jako odstających emitentów zanieczyszczeń powietrza (X6) w kraju, co związane jest z szeroko rozwiniętym przemysłem na tych obszarach. Dlatego też, pomimo wartości znacznie różniących się na tle kraju, żadna z obserwacji nie została usunięta, gdyż dostarczają one ważnych informacji na temat konkretnych województw.

8. Opis metod wykorzystanych w pracy

W badaniu wykorzystano metody analizy danych wielowymiarowych polegające na klasteryzacji, które zostały opisane poniżej.

8.1. Metoda Warda

Jak podaje Basiura (2013) metoda Warda to hierarchiczna metoda aglomeracyjna klasyfikacji obiektów, w której kryterium wyboru pary zbiorów łączonych w danym kroku jest wartością optymalną pewnej funkcji celu, gdzie najbardziej popularną jest suma kwadratów odchyleń poszczególnych elementów skupienia od środka ciężkości tego skupienia. Sama metoda zaproponowana została przez Warda (1963), natomiast za jej rekurencyjną implementację odpowiadają Lance i Williams (1967). Algorytm ten wykonywany jest następująco:

1. Wyznaczenie macierzy odległości D pomiędzy wszystkimi obiektami przy założeniu, że każdy obiekt stanowi osobną grupę.
2. Znalezienie minimalnego elementu macierzy d_{pq}, który wskaże obiekty o najmniejszej odległości, czyli najbardziej do siebie podobne. Kolejno następuje połączenie tych obiektów w nową grupę $A_r = A_p \cup A_q$, gdzie A_p oznacza obiekt pierwszy, a A_q drugi.
3. Obliczenie odległości pozostałych obiektów od nowopowstałej grupy według wzoru, wpisując je w kolejnej iteracji do wiersza i kolumny obiektu pierwszego, a usuwając z macierzy D wiersz i kolumnę obiektu drugiego. Wzór do obliczenia odległości dla metody Warda wygląda następująco:

$$d_{ir} = \frac{N_i + N_p}{N_i + N_r} d_{ip} + \frac{N_i + N_q}{N_i + N_r} d_{iq} - \frac{N_i}{N_i + N_r} d_{pq}$$

gdzie $d_{ir}, d_{iq}, d_{iq}\ i\ d_{pq}$ to odległość między poszczególnymi grupami, a $N_r, N_i, N_p i\ N_q$ liczebności grup.

4. Należy powtarzać kroki 2-3 aż wszystkie obiekty znajdą się w jednej klasie.

8.2. Metoda k-średnich

Jak podają Sobolewski i Sokołowski A. (2017) metoda k-średnich jest szeroko stosowana w analizach danych z różnych dziedzin. Polega na znalezieniu optymalnego podziału zbioru obiektów na k podzbiorów, przy czym kryterium jakości podziału jest maksymalizacja sumy wariancji międzygrupowej zmiennych, co jest równoważne minimalizacji wariancji wewnątrzgrupowej. W każdej iteracji dochodzi do stopniowej poprawy wstępnego podziału poprzez przenoszenie obiektów między grupami. W metodzie k-średnich odległości

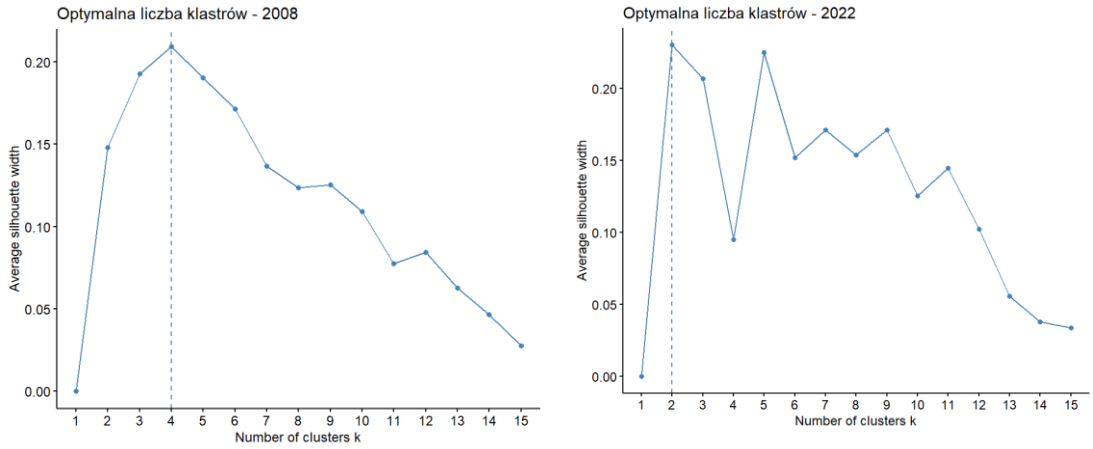
pomiędzy obiektami określa się za pomocą odległości euklidesowej lub jej kwadratu. Procedura metody przedstawia się następująco:

1. Podzielenie obiektów na k grup. Mogą być on dobrane losowo, kierując się opiniami ekspertów lub wyznaczone na podstawie rankingów stworzonych dla tych danych. Zadaje się również z góry kryterium stopu, czyli górny kres iteracji.
2. Wyznaczenie środka ciężkości każdej grupy w przestrzeni zmiennych (np. średnia).
3. Obliczenie odległości każdego obiektu od środka każdej grupy oraz przypisanie go do grupy, której środek leży najbliżej.
4. Należy powtarzać kroki 2-3 do momentu, aż w kolejnej iteracji nie nastąpi żadne przemieszczenie się obiektów między grupami lub gdy osiągnięte zostanie zadane kryterium stopu.

9. Wyniki przeprowadzonych badań

9.1. Wybór liczby klastrów

Istnieje wiele metod optymalnego wyboru liczby klastrów, a uzyskane nimi wyniki często nie są jednoznaczne. W celu dokonania wyboru przedstawiono wartości wskaźnika Silhouette dla grupowania obiektów od 2 do 15 klastrów na wykresie, gdzie jego najwyższa wartość jest sugerowaną liczbą skupień. Wyniki działania zamieszczono poniżej.

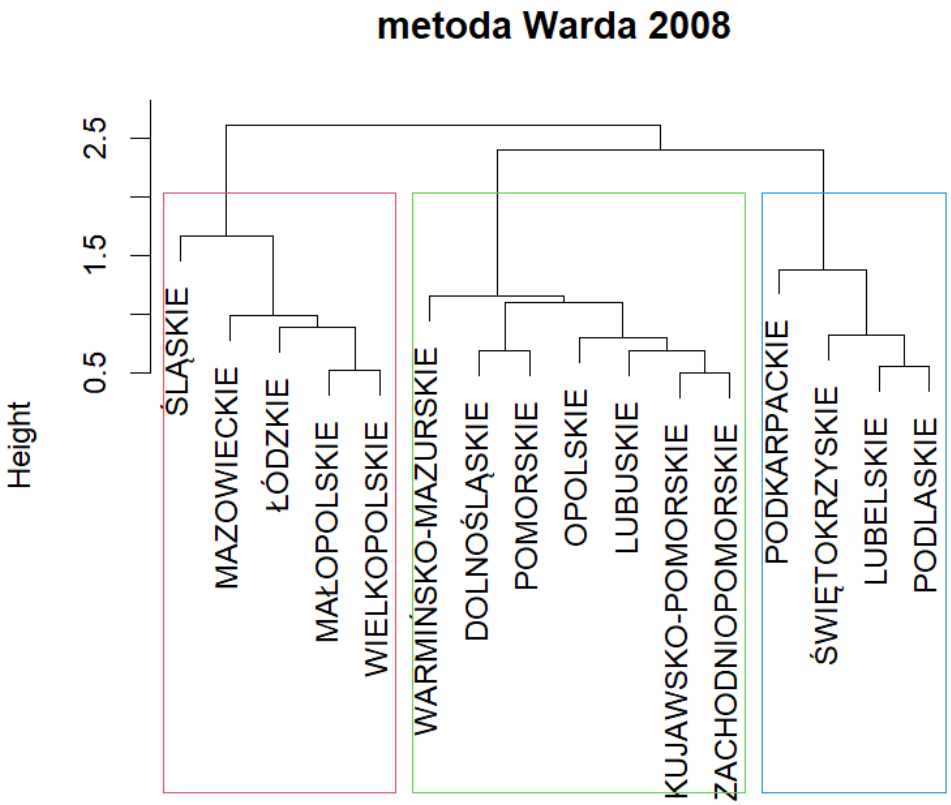


Wykres 11. Wartości wsk. Silhouette dla kolejnych liczb klastrów dla roku 2008 i 2022

Niestety wynik nie jest jednoznaczny, a wskaźniki podpowiadają różne liczby dla każdego roku – 4 klastry dla 2008 oraz 2 dla 2022. Z tego powodu obliczono także wartość heurystyki $k \approx \sqrt{m/2}$, gdzie m to liczba obiektów poddawanych grupowaniu. Otrzymany wynik działania to $k \approx 2{,}828427$, co w przybliżeniu wynosi 3. Jest to jednocześnie wartość znajdująca się dokładnie pomiędzy sugerowanymi przez poprzednią metodę, a na wykresie można zauważyć, że wartości wskaźnika Silhouette dla tej liczby klastrów jest nadal stosunkowo wysoka (kolejno druga i trzecia w kolejności dla danych lat). Dodatkowo podział województw na trzy grupy dostarczy więcej informacji o strukturze kraju, bez obniżania jakości wyników (co mogłoby się zdarzyć dla roku 2022, w przypadku wybrania 4 skupień). Dlatego też ostatecznie wybrano 3 klastry do dalszej analizy.

9.2. Metoda Warda

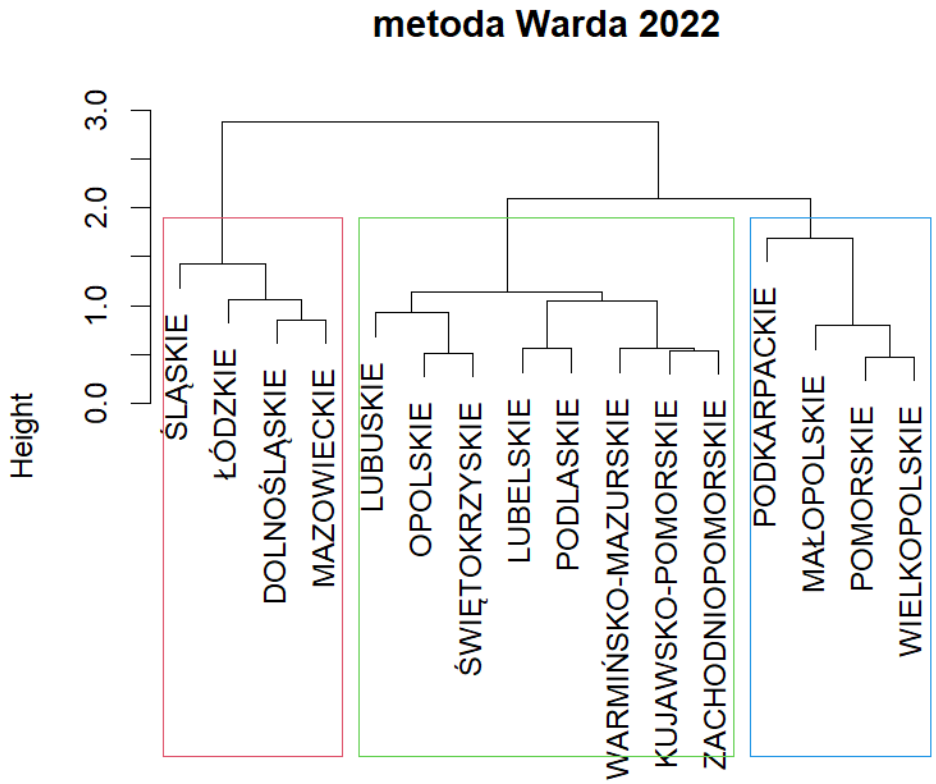
Za pomocą metody Warda podzielono obiekty na trzy klastry o zbliżonym poziomie badanych cech. Uzyskany podział województw przedstawiono na poniższych dendrogramach.



Wykres 12. Podział województw na grupy metodą Warda dla roku 2008

Metoda Warda charakteryzuje się zbliżoną liczebnością klastrów, dlatego też uzyskane grupy liczą kolejno 5, 7 oraz 4 województwa. W pierwszym klastrze znalazły się województwa śląskie, mazowieckie, łódzkie, małopolskie i wielkopolskie, czyli region centralny i południowy oraz województwo wielkopolskie. Są to obszary charakteryzujące się niską stopą bezrobocia, dużą ilością absolwentów oraz średnimi dochodami (poza mazowieckim, dla których zarobki są najwyższe w kraju), ale również wysokim poziomem emitowanych zanieczyszczeń (szczególnie województwa śląskie, łódzkie, co już wcześniej zostało zidentyfikowane ze względy na przemysłowy charakter obszarów, ale także mazowieckie) i wyższą od średniej kraju liczbę wypadków. Najliczniejszy klaster stanowią województwa warmińsko-mazurskie, dolnośląskie, pomorskie, opolskie, lubuskie, kujawsko-pomorskie oraz zachodniopomorskie, czyli północ oraz zachód Polski. Dla większości badanych zmiennych przyjmują średnie wartości na tle kraju, jednak można również zauważyć, że poza województwem kujawsko-pomorskim oraz warmińsko-mazurskim, pozostałe obiekty tej grupy mają najwyższe dochody zaraz po województwie mazowieckim oraz niską liczbę rozwodów, ale jednocześnie są to obszary o wysokiej przestępczości. Ostatnia grupa składa

się z województwa podkarpackiego, świętokrzyskiego, lubelskiego i podlaskiego, czyli regionu wschodniego. Odrębność tych obszarów powodują najniższe dochody, niski udział terenów zieleni w powierzchni całkowitej (poza województwem podkarpackim, które ma najwyższy procent w Polsce), ale także niski poziom przestępczości oraz rozwodów. Ciekawą cechą uzyskanego podziału jest fakt, że wszystkie grupy stanowią sąsiadujące ze sobą województwa i widać wyraźny podział kraju na trzy części.



Wykres 13. Podział województw na grupy metodą Warda dla roku 2022

Dla 2022 roku można zaobserwować zmiany struktury klastrów uzyskanych metodą Warda – tym razem pierwsza grupa liczy 4 obiekty, natomiast druga 8. Aż 9 województw pozostało w tych samych grupach, jednak w każdym klastrze zaszły zmiany. W grupie pierwszej nadal pozostają województwa śląskie, mazowieckie oraz łódzkie, ale zamiast małopolskiego i wielkopolskiego znalazło się tam tym razem dolnośląskie. Cała grupa wykazuje lepsze dochody niż w roku 2008, ale uzyskuje niższe, choć nadal powyżej średniej, wartości stopy bezrobocia. Kolejną grupę tworzą województwa lubuskie, opolskie, świętokrzyskie, lubelskie, podlaskie, warmińsko-mazurskie, kujawsko-pomorskie oraz zachodniopomorskie, przy czym ponad połowa z nich znalazła się w tej grupie również w roku poprzednim. Ze względu na dużą liczebność grupy ciężko jednoznacznie określić poziomy cech, ale większość w nich charakteryzuje się niską ilością wypadków oraz emisją zanieczyszczeń i wysokim poziomem stopy bezrobocia, ale również najniższymi liczbami absolwentów czy udziałem terenów zieleni. Ostatnią grupę stanowią natomiast województwa małopolskie, wielkopolskie,

pomorskie oraz podkarpackie. Podobnie jak ostatnio charakteryzują je dość niskie dochody (chociaż na przestrzeni kraju lepsze niż poprzednio) oraz niska liczba rozwodów, ale tym razem także wysoka liczba wypadków. Skupienia te nie tworzą już też trzech zamkniętych grup, jeśli spojrzy się na problem pod kątem geograficznym, ale są bardziej "rozrzucone" po mapie.

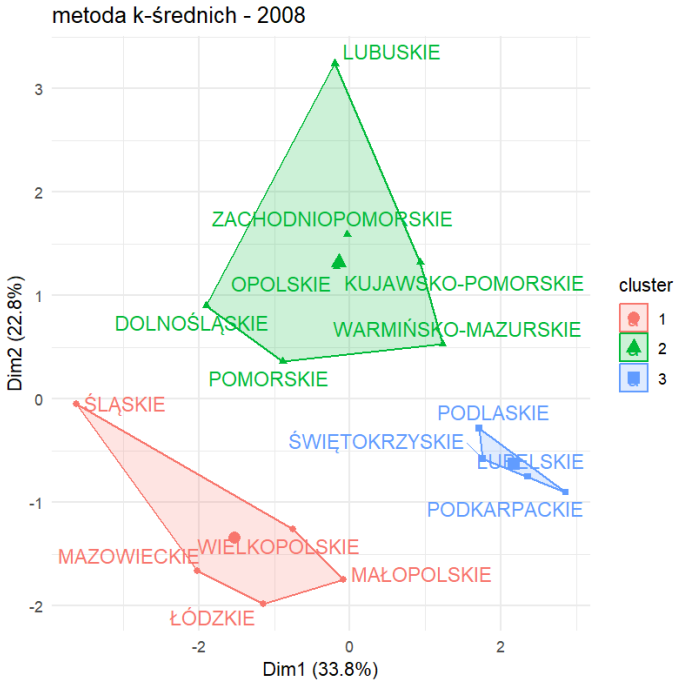
województwo	grupa 2008	grupa 2022
ŚLĄSKIE	1	1
MAZOWIECKIE	1	1
ŁÓDZKIE	1	1
MAŁOPOLSKIE	1	3
WIELKOPOLSKIE	1	3
DOLNOŚLĄSKIE	2	1
WARMIŃSKO-MAZURSKIE	2	2
ZACHODNIOPOMORSKIE	2	2
OPOLSKIE	2	2
LUBUSKIE	2	2
KUJAWSKO-POMORSKIE	2	2
POMORSKIE	2	3
ŚWIĘTOKRZYSKIE	3	2
LUBELSKIE	3	2
PODLASKIE	3	2
PODKARPACKIE	3	3

Tabela 5. Zestawienie grup według metody Warda w latach 2008 i 2022

Zestawiają wyniki klasteryzacji dla obu lat łatwo zauważyć, że ponad połowa województw pozostała w przypisanych im grupach, a jeśli zachodziły zmiany, to najczęściej o jedną grupę. Jedynym przypadkiem, kiedy doszło do przemieszczenia o dwie grupy są województwa małopolskie i wielkopolskie, u których na przestrzeni lat można zaobserwować duży spadek emisji zanieczyszczeń, co mogło spowodować wykluczenie z grupy pierwszej. Biorąc pod uwagę długi okres czasu między badanymi latami oraz świadomość jak bardzo zmieniło się przez ten czas, podział ten można uznać za całkiem stały w czasie.

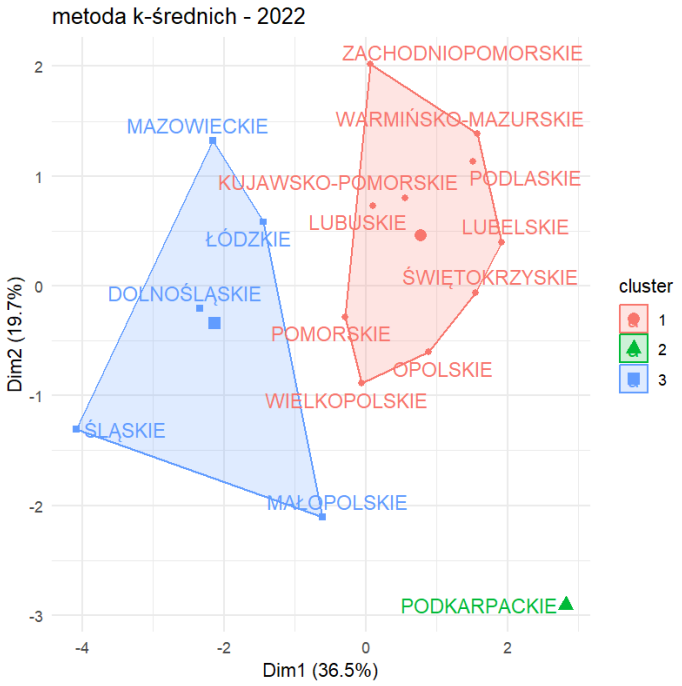
9.3. Metoda k-średnich

Dla takiej samej liczby klastrów przeprowadzono także grupowanie metodą k-średnich.



Wykres 14. Podział województw na grupy metodą k-średnich w 2008 roku

W 2008 roku analiza k-średnich podzieliła województwa na 3 klastry. Pierwszy klaster obejmuje województwa śląskie, mazowieckie, łódzkie, małopolskie oraz wielkopolskie. Są to regiony charakteryzujące się stosunkowo niską stopą bezrobocia, wysoką liczbą absolwentów oraz wysokimi dochodami, przy czym mazowieckie wyróżnia się najwyższymi dochodami w kraju. Obszary te są również obarczone wysoką emisją zanieczyszczeń, co jest zgodne z ich bardziej uprzemysłowionym charakterem. Drugi klaster składa się z województw: lubuskiego, zachodniopomorskiego, opolskiego, kujawsko-pomorskiego, warmińsko-mazurskiego, dolnośląskiego oraz pomorskiego. Te regiony przyjmują średnie wartości dla większości analizowanych zmiennych. Wyróżniają się dużym udziałem terenów zielonych i stosunkowo niską liczbą rozwodów, ale także większą przestępczością. Trzeci klaster tworzą województwa lubelskie, podlaskie, podkarpackie i świętokrzyskie. Są to obszary o najniższych dochodach i niskim poziomie rozwodów, ale cechują się także niskim poziomem przestępczości. Województwo podkarpackie cechuje się najwyższym udziałem terenów zieleni w kraju. Cały klaster charakteryzuję się niższą liczbą absolwentów oraz wyższą stopą bezrobocia.



Wykres 15. Podział województw na grupy metodą k-średnich w 2022 roku

W 2022 roku struktura klastrów uległa zmianie. W pierwszym klastrze liczba województw nie uległa zmianie natomiast zmieniła się jej skład, ponieważ dolnośląskie zostało dołączone do tej grupy w zamian za województwo wielkopolskie, które przeniosło się do drugiego klastra. Drugi klaster powiększył się znacznie w stosunku do roku 2008, bo aż o 3 województwa: lubelskie, podlaskie i świętokrzyskie co może sugerować poprawę jakości życia w tych regionach. Ostatni klaster zmienił się równie znacząco, ponieważ obejmuje tylko województwo podkarpackie.

województwo	grupa 2008	grupa 2022
ŚLĄSKIE	1	1
MAZOWIECKIE	1	1
ŁÓDZKIE	1	1
MAŁOPOLSKIE	1	1
WIELKOPOLSKIE	1	2
DOLNOŚLĄSKIE	2	1
WARMIŃSKO-MAZURSKIE	2	2
ZACHODNIOPOMORSKIE	2	2
OPOLSKIE	2	2
LUBUSKIE	2	2
KUJAWSKO-POMORSKIE	2	2
POMORSKIE	2	2
ŚWIĘTOKRZYSKIE	3	2
LUBELSKIE	3	2
PODLASKIE	3	2
PODKARPACKIE	3	3

Na przestrzeni lat 2008-2022 analiza metodą k-średnich wykazała zmiany w strukturze klastrów, co odzwierciedla zmieniającą się sytuację społeczno-ekonomiczną w polskich województwach. W 2008 roku klaster pierwszy obejmował regiony centralne i południowe – były to obszary o stosunków niskiej stopie bezrobocia, wysokiej liczbie absolwentów i średnich dochodach, ale jednocześnie charakteryzowały się wysokim poziomem emisji zanieczyszczeń i większą liczbą wypadków. Drugi klaster skupiał województwa głównie północne i zachodnie, które przyjmowały średnie wartości analizowanych zmiennych, z wyższym udziałem terenów zielonych i stosunkowo wyższą przestępczością. Trzeci klaster obejmował wschodnie regiony, gdzie obserwowano najniższe dochody, niską przestępczość i niski poziom rozwodów, ale także wyższą stopę bezrobocia.

W 2022 roku zmiany w strukturze klastrów były widoczne – województwo podkarpackie stało się odrębnym klastrem ze względu na swoje specyficzne cechy, takie jak wysoki udział terenów zielonych i niski poziom przestępczości. Inne województwa przesunęły się między klastrami, co wskazuje na zmiany w ich sytuacji społeczno-ekonomicznej, np. wzrost dochodów czy spadek stopy bezrobocia.

W obu metodach Warda i k-średnich obserwujemy podział Polski na grupy województw o wspólnych cechach, jednak metoda Warda daje bardziej spójne geograficznie klastry, co widać zwłaszcza w roku 2022, gdzie k-średnich pokazuje bardziej rozproszoną strukturę, przykładem jest odrębność podkarpackiego w tej metodzie. Zatem metoda Warda pozwala na uzyskanie bardziej zrównoważonych podziałów z uwzględnieniem wariancji wewnątrzgrupowej, co prowadzi do bardziej intuicyjnych wyników. Natomiast metoda k-średnich lepiej wychwytuje regionalne odrębności i zmiany w rozkładzie cech, ale jej wyniki mogą być bardziej zróżnicowane pod względem liczebności klastrów i mogą wskazywać na większe zróżnicowanie między grupami.

9.4. Ocena jakości skupień

Współczynnik Silhouette'a (sylwetki) jest miarą podobieństwa obiektu do klastra, do którego został przyporządkowany w porównaniu do innych klastrów. Osiąga wartości z zakresu (-1, 1), gdzie wysoka wartość wskazuje, że obiekt jest dobrze dopasowany do swojego klastra i słabo dopasowany do sąsiednich klastrów. Innymi słowy im wartość tego współczynnika jest bliższa 1 to znaczy, że dany obiekt został jednoznacznie przyporządkowany do jednego z klastrów i tym samym znajduje się daleko od innych sąsiednich klastrów. Wartość bliższa 0 odpowiada sytuacji, gdzie obiekt znajduje się na granicy decyzyjnej między dwoma klastrami co utrudnia jednoznaczne przyporządkowanie takiego obiektu do grupy, bo odległości do sąsiednich klastrów są zbliżone. Z kolei wartości ujemne świadczą o potencjalnie nieprawidłowym przyporządkowaniu obiektu do danego klastra (Belyadi, 2021).

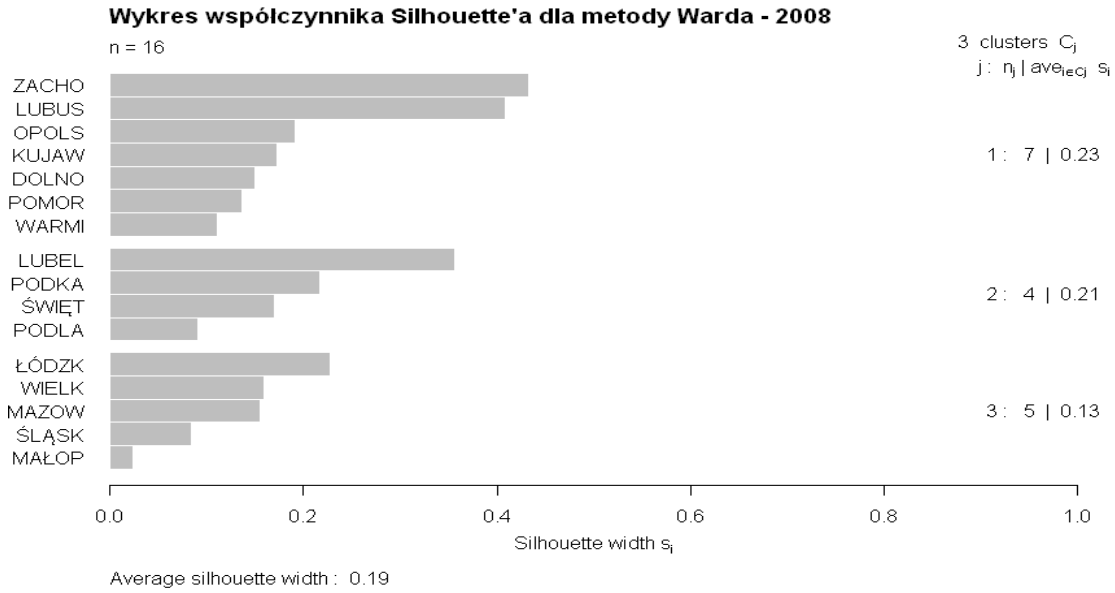
Aby obliczyć współczynnik Silhouette'a należy wyznaczyć dwa inne wskaźniki, wskaźnik spójności i separacji. Pierwszy z nich określa średnią odległość danego obiektu (próbki) od innych obiektów w tym samym klastrze (im mniejsza wartość tym lepiej). Drugi natomiast określa średnią odległość danego obiektu (próbki) od obiektów w obrębie sąsiedniego

klastra (im wyższa wartość tym lepiej) (Belyadi, 2021). Zależność do obliczenia współczynnika Silhouette'a:

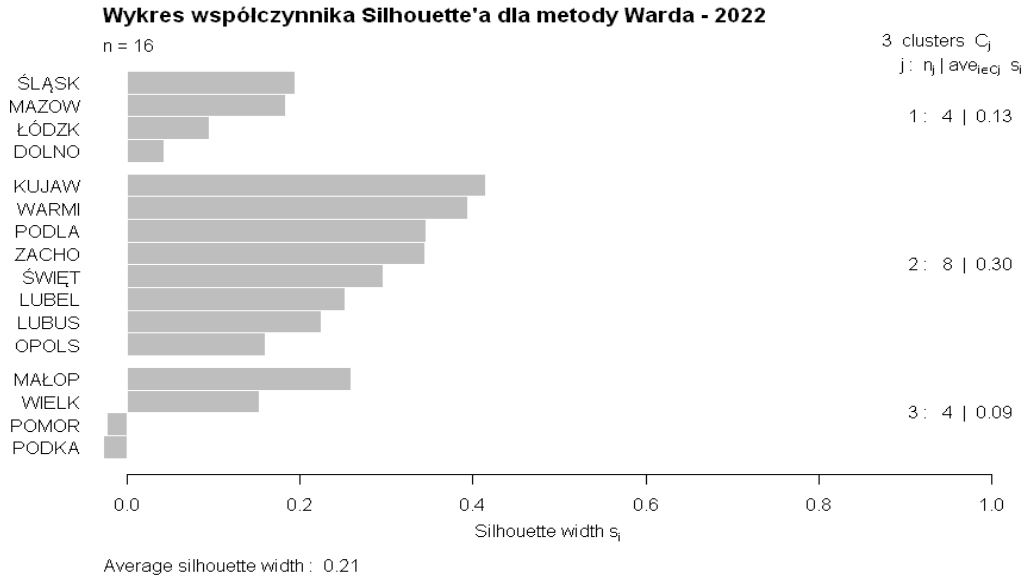
$$S = \frac{b - a}{\max(a, b)}$$

gdzie: a - wskaźnik spójności, b - wskaźnik separacji

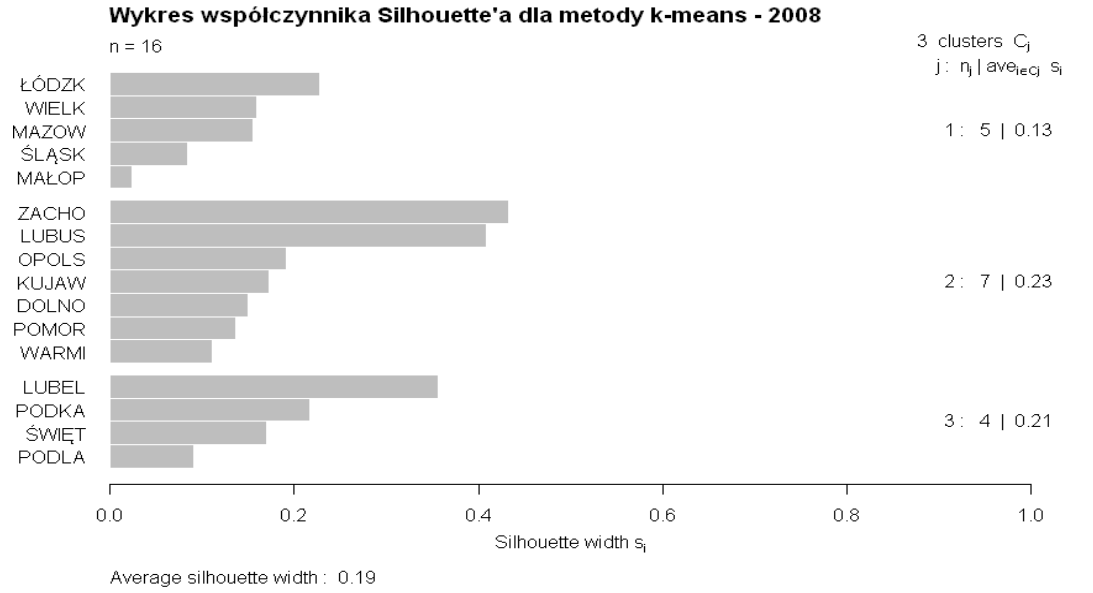
Poniżej przedstawiono wykresy wartości współczynników sylwetki (Silhouette'a) dla badanych lat i metod Warda oraz k-średnich dla poszczególnych województw. Dla każdej z grup zostały obliczone średnie wartości tego współczynnika. Dodatkowo została również obliczona wartość średnia współczynnika na podstawie wszystkich obserwacji w danym roku dla danej metody.



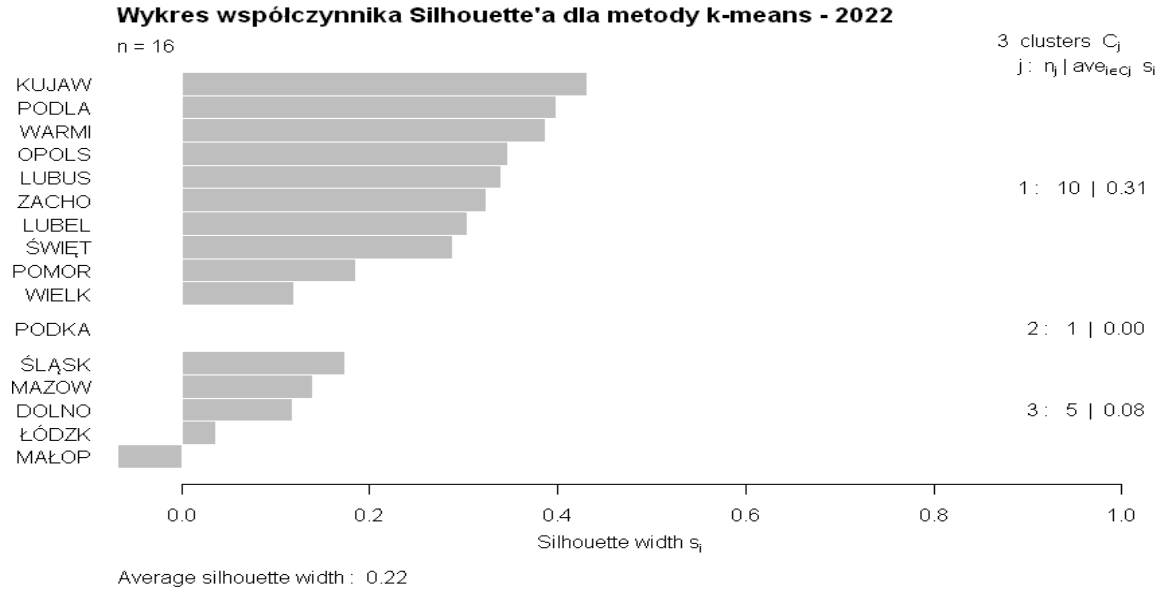
Wykres 16. Wartości współczynnika sylwetki (Silhouette'a) dla poszczególnych województw dla metody Warda w 2008 roku



Wykres 17. Wartości współczynnika sylwetki (Silhouette'a) dla poszczególnych województw dla metody Warda w 2022 roku



Wykres 18. Wartości współczynnika sylwetki (Silhouette'a) dla poszczególnych województw dla metody k-średnich w 2008 roku



Wykres 19. Wartości współczynnika sylwetki (Silhouette'a) dla poszczególnych województw dla metody k-średnich w 2022 roku

Uśrednione według wszystkich województw wartości współczynnika sylwetki dla badanych lat i wykorzystywanych metod są na zbliżonym poziomie, dla powyższych wykresów są to odpowiednio wartości 0,19; 0,21; 0,19; 0,22. Dla metody Warda przy roku 2008 (Wykres 16) wartość współczynnika dla województwa małopolskiego jest bardzo bliska zeru co oznacza, że próbka ta znajduje się niedaleko granicy między dwoma sąsiednimi klastrami. Podobna sytuacja ma miejsce przy metodzie k-średnich dla roku 2008 (Wykres 18) również dla tego województwa. Z kolei dla metody k-średnich w roku 2022 (Wykres 19) taki przypadek występuje dla województwa łódzkiego, natomiast dla województwa małopolskiego występuje wartość ujemna współczynnika co sygnalizuje o potencjalnie nieprawidłowym

przyporządkowaniu tej próbki. Ponadto w tym przypadku dla województwa podkarpackiego wystąpiła wartość zerowa współczynnika, która wynika z powstania jednoelementowego klastra co tak naprawdę informuje o niejednoznaczności tej próbki względem pozostałych. Dla metody Warda w roku 2022 (Wykres 17) wartości ujemne współczynnika uzyskano dla województw pomorskiego i podkarpackiego. Na ogół wartości średnie współczynnika sylwetki w obrębie konkretnych klastrów niezależnie od roku i metody nie przekraczają wartości 0,3 (maksymalna występująca średnia wartość to 0,31). Niekiedy są one także o wiele bliższe zeru (np. 0,09; 0,13). Takie wyniki świadczą raczej o dość słabej klasteryzacji, ponieważ uzyskiwane wartości są znacznie bardziej bliższe 0 niż 1. Wynikać może to m.in. z nieregularnego rozrzutu próbek, co utrudnia utworzenie klastrów o dużym zagęszczeniu, czyli o jednoznacznie przyporządkowanych do nich elementach, co jest pożądaną cechą przy klastrowaniu danych (wtedy współczynniki Silhouette'a osiągają duże wartości bliższe 1).

10. Podsumowanie i wnioski

W pracy przeanalizowana została sytuacja województw pod względem jakości życia dwoma metodami analizy skupień, metodą Warda oraz k-średnich, a na ich podstawie wyznaczono trzy grupy obiektów. Głównym celem badania było wskazanie regionów, które są do siebie najbardziej podobne, a które znacznie różnią się od pozostałych, biorąc pod uwagę czynniki wpływające na poziom jakości życia. Analizując wyniki uzyskane dla obu metod, można zauważyć, że są one do siebie bardzo zbliżone, szczególnie dla roku 2008, gdzie powstały takie same skupiska. W tym przypadku można zaobserwować także widoczny podział geograficzny – pierwszy klaster stanowi środek i południe kraju (śląskie, mazowieckie, łódzkie, małopolskie i wielkopolskie), drugi północ oraz zachód (dolnośląskie, warmińsko-mazurskie, zachodniopomorskie, opolskie, lubuskie, kujawsko-pomorskie i pomorskie), a trzeci wschód i południe (świętokrzyskie, lubelskie, podlaskie i podkarpackie). Każdy z nich zawiera charakterystyczne cechy, które posiadają znajdujące się w nich województwa, wskazując na podobieństwo obiektów wewnątrz tych samych grup, np. niska stopa bezrobocia i wysoka emisja zanieczyszczeń dla grupy pierwszej czy niskie dochody i niski poziom przestępczości dla grupy trzeciej. Wyniki klasteryzacji dla roku 2022 przy pomocy obu metod natomiast nieznacznie się od siebie różnią. Szczególnie rzuca się w oczy jednoelementowa grupa w metodzie k-średnich, którą stanowi województwo podkarpackie, ze względu na swoje specyficzne cechy. Pokrywa się to z przytoczonym artykułem Nowak (2017). Pomimo dużych zmian wartości cech na przestrzeni badanych lat, struktury klastrów nie zmieniły się drastycznie i również były do siebie zbliżone. W każdym podziale województwa śląskie, mazowieckie i łódzkie znajdowały się w obrębie tego samego skupienia, tak samo jak warmińsko-mazurskie, zachodniopomorskie, opolskie, lubuskie i kujawsko-pomorskie. Ze względu na obecność zarówno pozytywnych jak i negatywnych cech każdej grupy, nie da się jednoznacznie ocenić, która z nich jest "lepsza", a która "gorsza", jednak zastosowanie metod klasteryzacji obiektów może dostarczyć wielu informacji na temat danego regionu oraz ułatwić ich subiektywną ocenę osobom, które posiadają konkretną hierarchię wartości. Uzyskany podział okazał się jednak dość słabej jakości przy badaniu współczynnikiem Silhouette, którego wartości były bliższe 0 niż 1.

11. Bibliografia

Basiura, B. (2013). Metoda Warda w zastosowaniu klasyfikacji województw Polski z różnymi miarami odległości. *Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu, Research Papers of Wrocław University of Economics.* nr 279, s. 209-216

Belyadi H., Haghighat A. (2021), *Machine Learning Guide for Oil and Gas Using Python*, p.139

Lance G., Williams W.T. (1967) A general theory of classificatory storing strategies i hierarchical systems. *Computer Journal*, nr 9

Majecka, A., & Nowak, P. (2019). Uwarunkowania jakości życia w polskich województwach. *Nierówności Społeczne a Wzrost Gospodarczy*, 3(59), 149–161.

Nowak, P. (2017). Zróżnicowania regionalne jakości życia w Polsce. *Institute of Economic Research Working Papers*, No. 85/2017

Sobolewski, M. & Sokołowski A. (2017). Grupowanie metodą k-średnich z warunkiem spójności. *Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu, Research Papers of Wrocław University of Economics* nr 468

Ward J.H. (1963) Hierarchical grouping to optimize an objective function. J*ournal of the American Statistical Association.* no. 58, pp. 236-244.

World Health Organization. (1997). „WHOQOL, Measuring Quality of Life" *Division of Mental Health and Prevention of Substance Abuse*

Bank Danych Lokalnych GUS; https://bdl.stat.gov.pl/bdl/dane/podgrup/temat