AKADEMIA GÓRNICZNO-HUTNICZA
im. Stanisława Staszica w Krakowie

WYDZIAŁ ZARZĄDZANIA

SRUDIA STACJONARNE

KIERUNEK: Informatyka i Ekonometria

PRACA DYPLOMOWA

(licencjacka)

Klaudia Krysińska

Klasyfikacja województw ze względu na przyczyny zgonów znanych jako „garbage codes”

Classification of provinces in terms of the causes of deaths known as “garbage codes”

Promotor: dr Beata Basiura

„Zatwierdzam do rejestracji i dopuszczam do obrony”

……..

Data i podpis promotora

Kraków, 2020

„Oświadczam, świadomy(-a) odpowiedzialności karnej za poświadczenie nieprawdy, że niniejszą pracę dyplomową wykonałem(-am) osobiście i samodzielnie i że nie korzystałem(-am) ze źródeł innych niż wymienione w pracy.”

Spis treści

Wstęp

Główny Urząd Statystyczny bada strukturę demograficzną Polski. Wiąże się z tym prowadzenie statystyk urodzeń i zgonów. Jej skuteczność oraz jasność zapisu zależy w dużej mierze od personelu medycznego w kraju, gdyż na podstawie kart zgonu, gdzie lekarz ma obowiązek opisać na skutek czego zmarł pacjent, lekarze kodujący nadają odpowiedni Kod ICD-10.[1] Zdarza się , że występują przypadki, gdzie opis medyczny nie przedstawia precyzyjnie przyczyny zgonu. Są to informacje niepełne i niepozwalające przypisać określonego przypadku do konkretnego kodu określającego co przyczyniło się do śmierci osoby. Światowa organizacja Zdrowia (WHO) wprowadziła kody, które zostają przyporządkowane do niejasnych przyczyn zgonów. Są to przyczyny znane jako „garbage codes", czyli tzw. „kody śmieciowe". Nie wnoszą one żadnej istotnej informacji do statystyk oraz nie można na ich podstawie decydować o kierunku rozwoju państwa w dziedzinie medycyny. WHO wydało dwie listy kodów: podstawową i rozszerzoną.[2] Według zaleceń „kodów śmieciowych" nie powinno się stosować. Świadczą one wręcz o zaniedbaniu lub braku kompetencji pracowników medycznych w kraju. W 2017 roku w Polsce aż 28.1% przypadków zgonów zostało określonych za pomocą „kodów śmieciowych". Przez brak dokładności opisu przyczyn umieralności, a zatem wysoki odsetek „garbage codes", Polska nie jest uwzględniana w statystykach międzynarodowych przez WHO.[3] Główny Urząd Statystyczny prowadzi ewidencję przyczyn zgonów opisywanych przez „kody śmieciowe" z podziałem na województwa. Dostępne są listy podstawowe i rozszerzone. Najczęstszymi kryteriami klasyfikacja jednostek administracyjnych w Polsce są m. in.: poziom życia, rozwój społeczno-gospodarczy, struktura gospodarki. W pracy zdecydowano się na podział województw ze względu na przyczyny zgonów znanych jako „garbage codes". Stanowią one bowiem znaczą część przyczyn śmierci w Polsce. Pozwalają określić między innymi poziom świadczeń medycznych w danym województwie, a także kompetencje i rzetelność lekarzy. Nie w każdym województwie opieka medyczna jest na takim samym poziomie,

[1] *Umieralność w Polsce. Najczęstsze przyczyny zgonów*, https://www.politykazdrowotna.com/48211, umieralnosc-w-polsce-najczestsze-przyczyny-zgonow, 12.08.2020

[2] WASILEWSKI, W.; GAŁĘCKI, J.;STAWIŃSKA-WITOSZYŃSKA, B., *Poradnik szkoleniowy dla lekarzy orzekających o przyczynach zgonów i wystawiających kartę zgonów.* Warszawa, 2019 (aktualizacja). [15-16]

[3] WASILEWSKI, W.; GAŁĘCKI, J.;STAWIŃSKA-WITOSZYŃSKA, B., *Poradnik szkoleniowy dla lekarzy orzekających o przyczynach zgonów i wystawiających kartę zgonów.* Warszawa, 2019 (aktualizacja). [6]

co może wpływać na rzetelność opisów przyczyn zgonów a tym samym na liczbę „kodów śmieciowych". Istotą badania było wyodrębnienie grup województw, które świadczą o wysokim poziomie opieki medycznej i takich dla których poziom ten i rzetelność pracy lekarzy są stosunkowo niższe. W przeprowadzonych badaniach uwzględniono dziesięć wybranych losowo „kodów śmieciowych".

Celem pracy jest klasyfikacja województw ze względu na przyczyny zgonów znanych jako „garbage codes" przy użyciu wybranych trzech różnych metod grupowania: metody k-średnich, metody Ward'a oraz klasyfikacji spektralnej. Uzyskane wyniki pozwolą na porównanie klastrów powstałych w skutek użycia metod „tradycyjnych" z klastrami powstałymi po zastosowaniu klasyfikacji spektralnej. Jedną z przyczyn takich zgonów są utonięcia, zanurzenia o nieokreślonym zamiarze. Można przypuszczać, iż w województwie pomorskim i zachodnio pomorskim, które są położone nad morzem, powinien występować większy odsetek osób zmarłych z tej przyczyny w porównaniu z innymi województwami. W Polsce jedną z najczęstszych przyczyn zgonów są choroby związane z układem krążenia i pracą serca. Takie przyczyny zgonów są wprost wpisywane do karty jako „zatrzymanie krążenia", „ustanie oddychania" „niewydolność serca", a także „ustanie krążenia i oddychania" i są bezużyteczne dla statystyk WHO. Niestety to właśnie jest przyczyną wyższych wartości w statystykach. Jak to wygląda w poszczególnych województwach zostało przedstawione w trzecim empirycznym rozdziale niniejszej pracy.

W pierwszym rozdziale pracy została wyjaśniona ogólna idea analizy skupień. Przedstawiono jej cel oraz zastosowania. Wskazano także podział metod wykorzystywanych w procesie klasyfikacji, a następnie zwięźle je objaśniono. Rozdział drugi zawiera dokładny opis zastosowanych metod w badaniu: metody k-średnich, metody Ward'a oraz klasyfikacji spektralnej. Przedstawiono m.in. ich cechy charakterystyczne oraz algorytmy działania. Na początku trzeciego rozdziału zostały zaprezentowane dane wykorzystane do badania. W dalszej jego części przedstawione zostały kolejne etapy klasyfikacji za pomocą metod opisanych w rozdziale drugim oraz interpretacja otrzymanych klastrów. Na końcu umieszczono podsumowanie, zawierające ogólne uwagi na temat klas oraz wyników otrzymanych z przeprowadzonych w pracy badań.

1. Podstawowe metody grupowania obiektów

Wraz z intensywnym wzrostem ilości informacji i rozwojem różnorodnych systemów przetwarzania danych metody grupowania danych stanowią punkt wyjścia do różnorodnych analiz i strukturyzacji zebranych danych.

Celem analizy skupień (ang. *clustering*) jest wyznaczenie podziału obiektów (obserwacji) na podzbiory (z reguły nie krzyżowane) nazywane klastrami, tak aby każdy klaster składał się z podobnych obiektów, a obiekty z różnych klastrów znacznie się różniły od siebie.

Podział obserwacji na grupy podobnych do siebie obiektów może stać się podstawą do wyciągania wniosków i zestawiania wyników. Może też stanowić wstęp do dalszego przetwarzania danych i podejmowania decyzji. Jest to jedna z metod analizy - „divide et impera" (tj. strategię „dziel i zwyciężaj").

1.1. Wstęp do analizy skupień

Analiza skupień należy do rodzaju klasyfikacji nazywanej klasyfikacją bez nadzoru lub bez nauczyciela (ang. *unsupervised learning*). Skład oraz liczba klastrów nie jest z góry znana. Są one ustalane w trakcie przeprowadzenia badania. Podział odbywa się na podstawie danych dotyczących grupowanych obiektów, zatem grupy zależą tylko od nich. Poszczególne klastry powinny składać się z podobnych do siebie elementów i jednocześnie znacznie różnych od wchodzących w skład innych skupisk.[4]

Można wyróżnić trzy warunki, które powinny zostać spełnione, aby analiza skupień była lepiej zdefiniowana. Dotyczą one zbioru A, który jest poddany klasyfikacji na grupy P_j, gdzie j=1,2,…,u. Zakłada się , że zbiór A składa się z elementów A_i, gdzie i=1,2,…,n. Podział P_i powinien spełniać następujące warunki:

- zupełność (suma grup P_j tworzy zbiór A) $\bigcup_{k=1}^{u} P_k = A$;
- rozłączność (części wspólne dla grup P_j są zbiorami pustymi) jeśli $(k, k' = 1, \dots, u;\ k \neq k')$ to $P_k \cap P_{k'} = \emptyset$;
- niepustość (nie ma takiej grupy P_j, która byłaby pusta) $P_k \neq \emptyset$, gdzie $k = 1, \dots, u$.

[4] GNANADESIKAN, R., *Discriminant analysis and clustering*. National Academies, 1988.

Przy spełnieniu powyższych warunków, powstałe grupy są bardziej precyzyjnie wydzielone, a otrzymane wyniki dogodniejsze do interpretacji.[5]

Celem analizy skupień jest podział badanych obiektów na grupy. Klasyfikacja odbywa się na podstawie zmiennych (danych). Każdy klaster powinien składać się z obiektów posiadających podobne cechy. Odrębne grupy natomiast muszą być jak najbardziej różne od siebie.[6] Innymi słowy, głównym celem jest znalezienie takiego podziału zbioru obiektów, gdzie podobieństwa wewnątrzgrupowych obiektów są jak największe, a międzygrupowe najmniejsze. W przypadku gdy podobieństwem jest szeroko rozumiana odległość pomiędzy obiektami odległości obiektów z jednej grupy powinny być jak najmniejsze, natomiast odległości pomiędzy obiektami z różnych grup jak największe (w szczególności większe od wszystkich odległości wewnętrznych).[7] Wybór odpowiedniej ilości grup oraz ich odpowiednia identyfikacja nie zawsze jest łatwa do uzyskania. Grupowanie bez nadzoru wydobywa wiedzę z danych i wiele zależy od danych a także od wyboru metody pomiaru podobieństwa, metody podziału i metody ustalenia liczby klastrów. Wiele zależy od indywidualnej interpretacji i decyzji osoby analizującej dane.[8]

Analiza skupień jest wykorzystywana w wielu dziedzinach nauki. Przykładem może być podział klientów banku. Pod uwagę mogą być brane cechy mające znaczenie z jego perspektywy, np. dochody, rodzaj umowy o pracę. Jest to zjawisko typu społeczno-ekonomicznego.[9] Kolejnymi dziedzinami zastosowań są m. in. medycyna - klasyfikacja pacjentów ze względu na ich objawy, branża biznesowa – klasyfikacja firm ze względu na ich rodzaj i formę działalności, media społecznościowe – klasyfikacja opinii itp.[10]

Metody klasyfikacji można podzielić na metody tradycyjne i metody „nowej generacji". Do tej drugiej zalicza się między innymi klasyfikacje spektralną (ang. *spectral*

[5] WALESIAK, M.; GATNAR, E. (ed.)., *Metody statystycznej analizy wielowymiarowej w badaniach marketingowych*. Wydaw. Akademii Ekonomicznej im. Oskara Langego, 2004. [317]
[6] WALESIAK, M.; GATNAR, E. (ed.)., *Metody statystycznej analizy wielowymiarowej w badaniach marketingowych*. Wydaw. Akademii Ekonomicznej im. Oskara Langego, 2004. [318]
[7] KORZENIEWSKI, J., *Metody selekcji zmiennych w analizie skupień. Nowe Procedury*. Wydawnictwo Uniwersytetu Łódzkiego, 2012. [15]
[8] GACZEK, W. M.; HEŁPA, M.; KASPRZYK, A., Niehierarchiczna analiza skupień—nowa metoda klasyfikacji zjawisk społeczno-gospodarczych. 1980.
[9] KORZENIEWSKI, J., *Metody selekcji zmiennych w analizie skupień. Nowe Procedury*. Wydawnictwo Uniwersytetu Łódzkiego, 2012. [5]
[10] GNANADESIKAN, R., *Discriminant analysis and clustering*. National Academies, 1988. [29-30]

clustering), która została ona opisana w rozdziale 2.[11] W pracy przyjęto następujący podział metod tradycyjnych:[12]

- metody hierarchiczne, w których wyróżnia się dwa rodzaje:
  - aglomeracyjna (ang. *agglomerative*),
  - deglomeracyjna (ang. *divisive*);
- metody optymalizacyjne, w których można wyróżnić metody metodę k-średnich, k-medoids.
- inne metody na przykład metody obszarowe oraz gęstościowe.

1.2. Grupowanie z wykorzystaniem metod optymalizacyjnych

Metody optymalizacyjne w analizie skupień mają na celu podział obiektów w ten sposób, aby zmienność w klasach była jak najmniejsza (jej optymalizacja), natomiast zmienność między grupami była jak największa. Wszystkie elementy muszą zostać poddane klasyfikacji. Do metod optymalizacyjno-iteracyjnych zalicza się m.in.: metodę k-średnich, metodę k-centroid.[13]

Metody optymalizujące w przeciwieństwie do hierarchicznych muszą mieć z góry podaną liczbę skupień. Z założenia nie mogą powstać klastry puste, zatem ich ilość musi być mniejsza od liczby badanych elementów. Znanych jest też wiele metod wyznaczających liczbę klastrów.[14]

Ważnym krokiem podczas stosowania metod optymalizacyjnych jest wybór liczby skupień. Jeśli ilość grup będzie zbyt mała to będą one składały się z obiektów niepodobnych do siebie. Z drugiej strony, kiedy liczba klastrów będzie za duża powstanie wiele skupisk wewnętrznie bardzo podobnych względem siebie, jednak może wystąpić problem interpretacji oraz jakości wyników klasyfikacji. Wadą metod optymalizacyjnych jest brak możliwości obserwacji sposobu łączenia się obiektów w grupy, co natomiast umożliwiają metody hierarchiczne. Zmiana liczby skupień, zarówno jeśli mowa

[11] DUDEK, A., Klasyfikacja spektralna a tradycyjne metody analizy skupień. *Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu. Ekonometria*, 2009, 23.37 Zastosowania metod ilościowych: 21-34. [21]
[12] WALESIAK, M.; GATNAR, E. (ed.)., *Metody statystycznej analizy wielowymiarowej w badaniach marketingowych*. Wydaw. Akademii Ekonomicznej im. Oskara Langego, 2004. [322]
[13] PIETRZYKOWSKI, R.; KOBUS, P., Zastosowanie modyfikacji metody k-średnich w analizie portfelowej. *Ekonomika i Organizacja Gospodarki Żywnościowej*, 2006, 60: 301-308. [301]
[14] WALESIAK, M.; GATNAR, E. (ed.)., *Metody statystycznej analizy wielowymiarowej w badaniach marketingowych*. Wydaw. Akademii Ekonomicznej im. Oskara Langego, 2004. [330]

o dodaniu lub odjęciu liczby grup, skutkować może zupełnie innym podziałem obiektów. Same klastry mogą być bardzo odmienne od poprzednich skupisk.[15]

Wybór liczby klas sprawia największa trudność w procesie analizy skupień. Decyzja ta zależy głównie od osób wykonujących badanie. Wykorzystuje się tu czasem wiedzę ekspercką, przeprowadzana jest też analiza merytoryczna oraz wykorzystywane bywają indeksy służące do określenia najlepszej liczby skupień. Wiele indeksów zostało opisanych przez Milligana i Coopera [Milligan, Cooper 1985]. W niniejszej pracy zostanie wykorzystany indeks Silhouette, nazywany także indeksem sylwetkowym. Jest on obliczany na podstawie wzoru (1.1).

$$S(u) = \frac{1}{n}\sum_{i=1}^{n} \frac{b(i)-a(i)}{max\{a(i),b(i)\}}, \tag{1.1}$$

gdzie:

- u – liczba skupień,
- n – liczba obiektów, które są grupowane,
- i – numer obiektu (i = 1, 2,…, n),
- $a(i)$ – średnia odległość między i-tym obiektem, a pozostałymi obiektami w skupisku z i-tym obiektem,
- $b(i)$ – średnia odległość i-tego obiektu od obiektów w skupisku najbliższym do skupiska, w którym jest i-ty obiekt.

Indeks ten przyjmuje wartości z przedziału od -1 do 1. Jego wartość najwyższa 1 wskazuje na optymalną liczbę klastrów, zatem obiekty przyporządkowane są do klas w najlepszej ich kombinacji. W przypadku gdy indeks sylwetkowy przyjmuje najmniejszą wartość z przedziału -1 dla konkretnej liczby klastrów, wskazuje na niekorzystny podział obiektów.[16] Interpretacja indeksu sylwetkowego opiera się na wyszukaniu wartości największej indeksu, gdyż wtedy występuje optymalna liczba skupisk.

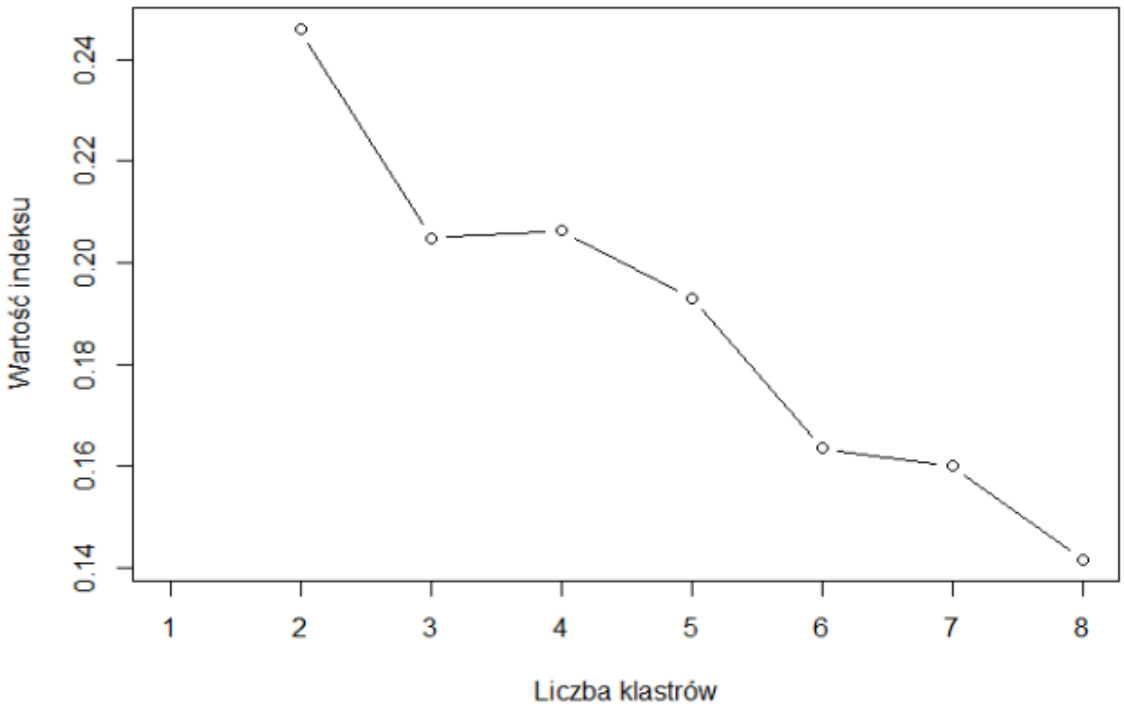
Aby określić na ile grup podzielić obiekty na podstawie indeksu Silhouette, należy wybrać taką liczbę grup dla której podany indeks osiąga maksimum.[17] Przykładowy wykres wyników indeksu sylwetkowego dla różnej liczby klastrów pokazany został na Rysunku 1.1.

---

[15] ROSZKO-WÓJTOWICZ, E., Analiza skupień w ocenie warunków pracy w krajach Unii Europejskiej. *Wiadomości Statystyczne*, 2014, 59.11. [8-9]

[16] Praca magisterska: BIEŃSKOWSKI, D., *Grupowanie konsensusowe danych otrzymanych z mikromacierzy*, Politechnika Warszawska, 2014. [74]

[17] ĆWIĄKAŁA-MAŁYS, A.; MOŚCIBRODZKA, M., Hierarchiczne procedury aglomeracyjne w badaniu poziomu i struktury kosztów publicznych uczelni akademickich. *Metody ilościowe w badaniach ekonomicznych*, 2014, 15.3: 30-41. [31-32]



Rysunek 1.1 Indeks Silhouette dla różnych liczb klastrów

Źródło: opracowanie własne

Na podstawie Rysunku 1.1 zawierającego indeks Silhouette najlepszą liczbą podgrup będzie ta dla której ten wskaźnik jest największy. Największą wartość przyjmuje on podczas podziału obiektów na dwie grupy.

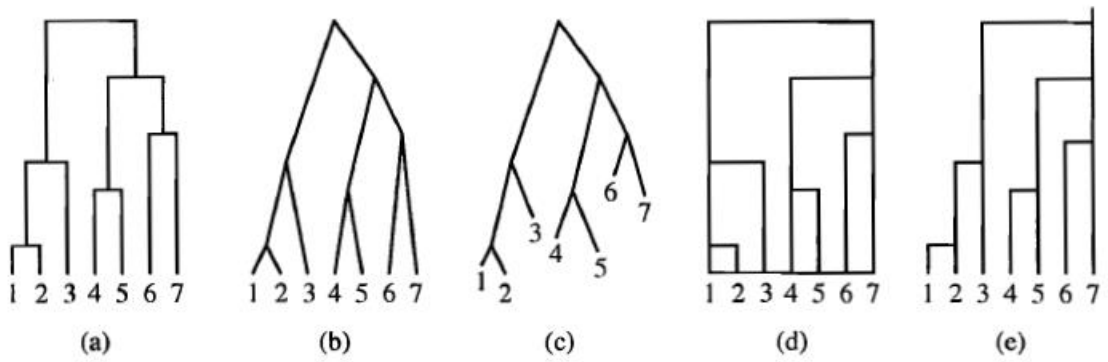
Indeks sylwetkowy zostanie zastosowany w każdej z wykorzystanych metod klasyfikacji. Podczas przeprowadzenia klasyfikacji spektralnej obliczone dodatkowo zostaną indeksy: Calińskiego i Harabasza oraz Daviesa i Bouldina. Wybrano je ze względu na ich częste zastosowanie. Indeks Calińskiego i Harabasza ma podobną interpretację jak indeks sylwetowy. Optymalna liczba klastrów wskazywana jest przez maksimum indeksu. W przypadku indeksu Daviesa i Bouldina wykres czyta się odwrotnie. Najlepsza liczbę grup wskazuje minimalna wartość tego indeksu.

1.3. Grupowanie hierarchiczne

W grupowaniu hierarchicznym każdy z obiektów jest analizowany i porównywany względem pozostałych w procesie iteracyjnym. Wyróżnia się dwa rodzaje tej metod hierarchicznych: aglomeracyjne (obiekty podobne do siebie są łączone do jednego klastra), deglomeracyjne (obiekty z jednego klastra są rozdzielane na dwa

osobne podzbiory). Cały proces podziału kończy się gdy wszystkie elementy zostaną przydzielone do jednego klastra (w przypadku aglomeracji) lub gdy nie ma już możliwości podziału na mniejsze podzbiory (w przypadku deglomeracji).[18]

Wyniki procesu podziału obiektów na poszczególne grupy metodą hierarchiczną są najczęściej przedstawiane za pomocą dendrogramu lub innych diagramów drzew. W pracy została użyta pierwsza wymieniona metoda graficzna. Na Rysunku 1.2 przedstawiono różne formy wizualizacji dendrogramu. Przedstawiają one dokładny przebieg procesu tworzenia grup. Pionowe linie reprezentują odległości między klastrami. Im są dłuższe tym grupy bardziej się od siebie różnią.[19] Dendrogram wizualizuje każdy etap łączenia się obiektów (lub ich rozdzielania). Umożliwia obserwacje, w jakim momencie elementy zostały przyporządkowane do poszczególnych klastrów. Na podstawie takiego wykresu można zdecydować, na ile grup odbędzie się dzielenie obiektów.[20]



Rysunek 1.2 Przykładowe dendrogramy

Źródło: WALESIAK, Marek; GATNAR, Eugeniusz (ed.). *Metody statystycznej analizy wielowymiarowej w badaniach marketingowych*. Wydaw. Akademii Ekonomicznej im. Oskara Langego, 2004, strona 323

[18] ALMEIDA, J. A. S.; BARBOSA, L. M. S.; PAIS, A. A. C. C.; FORMOSINHO, S. J., Improving hierarchical cluster analysis: A new method with outlier detection and automatic clustering. *Chemometrics and Intelligent Laboratory Systems*, 2007, 87.2: 208-217. [210]

[19] WALESIAK, M.; GATNAR, E. (ed.)., *Metody statystycznej analizy wielowymiarowej w badaniach marketingowych*. Wydaw. Akademii Ekonomicznej im. Oskara Langego, 2004. [322-323]

[20] ALMEIDA, J. A. S.; BARBOSA, L. M. S.; PAIS, A. A. C. C.; FORMOSINHO, S. J., Improving hierarchical cluster analysis: A new method with outlier detection and automatic clustering. *Chemometrics and Intelligent Laboratory Systems*, 2007, 87.2: 208-217. [208]

Przed zastosowaniem metody grupowania hierarchicznego należy określić dwie istotne rzeczy. Są to: rodzaj miary podobieństwa jaka zostanie obliczona dla obiektów i klastrów oraz rodzaj techniki hierarchizacji (aglomeracyjna czy deglomeracyjna).[21]

Miary odległości pomiędzy obiektami oddają ich podobieństwo. Im wartość jest większa tym elementy są bardziej różne. Odległość mała świadczy o wysokim podobieństwie. Wartości te są podstawą do grupowania dlatego bardzo ważny jest wybór ich sposobu obliczenia.[22] Najczęściej wybierana jest odległość Euklidesowa lub jej kwadrat.

Drugim ważnym krokiem jest dobór techniki grupowania. W pracy zostanie użyta jedna z metod aglomeracyjnych (ang. *agglomerative*). Technika ta jest także najczęściej stosowana w badaniach mających na celu podział na grupy. Polega ona na tym, że na początku każdy obiekt stanowi jednoelementową grupę, następnie dwa obiekty najbardziej do siebie podobne łączą się w jeden klaster. W kolejnym ruchu następuje ponowne porównanie grup i łączenie. Algorytm ten powtarza się do momentu powstania jednego klastra z wszystkich obiektów. Aby określić podobieństwo grup składających się z kilku elementów, stosowane są różne techniki. W tym miejscu rodzi się nowy problem wyboru jednej z nich. Do hierarchicznych metod połączenia zalicza się metody m.in.: pojedynczego łączenia (ang. *single linkage*), kompletnego łączenia (ang. *complete-linkage*) oraz metodę Warda (ang. *Ward's-linkage*). [23] W pracy zostanie użyta ostatnia z wymienionych tutaj metod i zostanie dokładnie opisana w rozdziale 2.3.

Metoda pojedynczego łączenia, nazywana także metoda najbliższego sąsiada, polega na dołączaniu do grupy takich nowych obiektów, które są najbliżej jakiegoś elementu w danej grupie. Sposób ten jest prosty do obliczenia, jednak niesie ze sobą wady. Jedną z nich jest łańcuchowy przebieg łączenia się obiektów w klasy. Kiedy dwa obiekty utworzą grupę, kolejne elementy są zazwyczaj przyłączane tylko do tej istniejącej. Można wtedy zauważyć brak tworzenia się nowych klas. Kolejną wadą jest

[21] ALMEIDA, J. A. S.; BARBOSA, L. M. S.; PAIS, A. A. C. C.; FORMOSINHO, S. J., Improving hierarchical cluster analysis: A new method with outlier detection and automatic clustering. *Chemometrics and Intelligent Laboratory Systems*, 2007, 87.2: 208-217. [210]

[22] NOWAK-BRZEZIŃSKA, A.; WAKULICZ-DEJA, A. Ś., Wybór miary podobieństwa a efektywność grupowania reguł w złożonych bazach wiedzy. *Studia Informatica, Wyd. Pol. Śląskiej*, 2010, 31.2A: 89. [190]

[23] ALMEIDA, J. A. S.; BARBOSA, L. M. S.; PAIS, A. A. C. C.; FORMOSINHO, S. J., Improving hierarchical cluster analysis: A new method with outlier detection and automatic clustering. *Chemometrics and Intelligent Laboratory Systems*, 2007, 87.2: 208-217. [210]

powstawanie grup o obiektach, między którymi występują większe odległości, co niesie za sobą małe podobieństwo.[24]

Metoda kompletnego łączenia, nazywana także metodą najdalszego sąsiada, działa na zasadzie odwrotnej niż metoda pojedynczego łączenia. Grupy tworzone są na podstawie najdalej położonych od siebie obiektów – największej odległości pomiędzy dwoma obiektami z dwóch różnych skupisk, które mają zostać połączone. W tej metodzie nie występuje efekt łączenia łańcuchowego.[25]

Drugą techniką grupowania jest technika deglomeracyjna (ang. *divisive*). Jej działanie jest odwrotne niż w przypadku metod aglomeracyjnych. Na początku wszystkie obiekty stanowią jedną grupę. Następnie w kolejnych krokach następuje podział na mniejsze grupy – w każdym ruchu pojawia się nowe skupisko. Algorytm zatrzymuje się, gdy jest tyle klas ile elementów liczył pierwotny zbiór obiektów, czyli gdy każdy obiekt tworzy jednoelementową grupę.[26]

[24] WALESIAK, M.; GATNAR, E. (ed.)., *Metody statystycznej analizy wielowymiarowej w badaniach marketingowych*. Wydaw. Akademii Ekonomicznej im. Oskara Langego, 2004. [325]
[25] WALESIAK, M.; GATNAR, E. (ed.)., *Metody statystycznej analizy wielowymiarowej w badaniach marketingowych*. Wydaw. Akademii Ekonomicznej im. Oskara Langego, 2004. [327]
[26] WALESIAK, M.; GATNAR, E. (ed.)., *Metody statystycznej analizy wielowymiarowej w badaniach marketingowych*. Wydaw. Akademii Ekonomicznej im. Oskara Langego, 2004. [328]

2. Wybrane metody analizy skupień do grupowania województw

W literaturze przedmiotu z zakresu grupowania danych wymienia się wiele metod. Dla danych typu o skali przedziałowej i ilorazowej zastosowanie mają metody opisane w rozdziale poprzednim. Wśród nich szczególnie często wymienia się stosowanie metody k-średnich ze względu na jej szybkość w działania oraz prostotę.[27] Jeśli dane są niestandardowo ułożone ciekawe efekty można otrzymać wykorzystując klasyfikację spektralną. Metoda ta daje często lepsze wyniki w porównaniu z metodami tradycyjnymi analizy skupień.[28] Najczęściej stosowaną metodą z grupy metod hierarchicznych jest metoda Warda. Ma ona zastosowanie do danych, które nie mają dużej liczby obserwacji oraz nie posiadają wielu wartości odstających.[29]
W następnych rozdziałach zostaną omówione szczegółowo wybrane trzy metody grupowania.

2.1. Metoda k-średnich

Metoda k-średnich (ang. *k-means*) stosowana jest w przypadku analizowania dużego zbioru danych oraz zalicza się do najczęściej wybieranych niehierarchicznych algorytmów podziałowych podczas przeprowadzania badań. Jej główną cechą jest reorganizacja informacji na temat obiektów w postaci danych, w taki sposób aby tworzyła one ogólniejsze dziedziny. Taka operacja przyczynia się do zmniejszenia czasu i nakładu pracy podczas przeprowadzenia badania. Umożliwia także łatwiejszą interpretację oraz jaśniejszy obraz analizowanego problemu przy jak najmniejszym poziomie utraty informacji.[30]

Przeprowadzenie analizy skupień wymaga ujednolicenia oraz porównywalności zmiennych. Dokonuje się tego po przez operację transformacji normalizacji. Jej głównym

[27] ARTHUR, D.; MANTHEY, B.; RÖGLIN, H., Smoothed analysis of the k-means method. *Journal of the ACM (JACM)*, 2011, 58.5: 1-31. [1]
[28] PEŁKA, M., et al. Wielomodelowa klasyfikacja spektralna danych symbolicznych. *Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu*, 2017, 468: 180-187. [182]
[29] MIGDAŁ-NAJMAN, K.; NAJMAN, K., Analiza porównawcza wybranych metod analizy skupień w grupowaniu jednostek o złożonej strukturze grupowej. *Zarządzanie i finanse*, 2013, 3.2: 179-194. [180]
[30] ROSZKO-WÓJTOWICZ, E., Analiza skupień w ocenie warunków pracy w krajach Unii Europejskiej. *Wiadomości Statystyczne*, 2014, 59.11. [9]

celem jest sprowadzenie zmiennych do jednakowego poziomu wielkości pomiarów. Uwzględnienie różnych miar poszczególnych zmiennych skutkuje tym, że jedne dane mogą mocniej wpływać na wyniki klasyfikacji. Ujenolicenie niweluje ten skutek.

Istnieje wiele formuł normalizacyjnych. W pracy zostanie wykorzystana standaryzacja, która jest realizowana z wykorzystaniem wzoru (2.1).

$$z_{ij} = \frac{(x_{ij} - \bar{x}_j)}{s_j} \tag{2.1}$$

gdzie:

- z_{ij} – znormalizowana wartość i-teg obiektu dla j-tego atrybutu?,
- x_{ij} – wartość i-teg obiektu o j-tego atrybutu,
- $\bar{x}_j$ – średnia j-tego atrybutu,
- s_j – odchylenie standardowe j-tego atrybutu.

Formułę (2.1) stosuje się dla danych mierzonych na skali ilorazowej oraz przedziałowej. Niestety przyczynia się do straty informacji w wyniku transformacji wszystkich zmiennych na skalę przedziałową pomiaru zmiennych. Powyższa standaryzacja niweluje różnice w wielkości zmiennych. W ten sposób średnia arytmetyczna zestandaryzowanych zmiennych wynosi 0 a odchylenie standardowe 1.[31]

Zastosowanie metody k-średnich wymaga na początku wyboru rodzaju odległości, jaka będzie obliczana między obiektami (punktami wielowymiarowymi). Najbardziej popularna i najczęściej stosowana jest odległość euklidesowa (ang. *euclidean distance*). Jej główną zaletą jest prostota interpretacji.[32] Odległość euklidesowa jest to miara odległości między dwoma punktami wielowymiarowymi. Bazuje na współrzędnych geometrycznych. W tym przypadku są to wartości atrybutów, które opisują obiekty.

Aby liczbę $d(x, y)$ można było nazwać odległością, musi ona spełniać następujące warunki:

1) musi być ona nieujemna: $d(x, y) \geq 0$,
2) przyjmuje wartość 0 tylko wtedy, gdy wartości x i y są takie same: $d(x, y) = 0 \Leftrightarrow x = y$,

[31] WALESIAK, M.; GATNAR, E. (ed.)., *Statystyczna analiza danych z wykorzystaniem programu R*. Wydawnictwo Naukowe PWN, 2009. [66-68]
[32] ELMORE, K. L.; RICHMAN, M. B., Euclidean distance as a similarity metric for principal component analysis. *Monthly weather review*, 2001, 129.3: 540-549. [524]

3) musi być zachowana symetryczność, czyli odległość z punktu x do y jest taka sama jak z y do x: $d(x, y) = d(y, x)$,
4) musi być spełniona nierówność trójkąta: $d(x, z) \leq d(x, y) + d(y, z)$.

Odległość euklidesowa spełnia wszystkie powyższe warunki. Jest ona stosowana często w metodzie k-średnich.[33] Odległość euklidesowa w opisywanej metodzie jest obliczana między obiektami, a środkami ciężkości (centroidami). Można ją wyrazić wzorem (2.2).[34]

$$d_{il} = \sqrt{\sum_{j=1}^{m} \left(x_{ij} - \bar{x}_{lj}\right)^2}, \quad (i = 1, .., n), \tag{2.2}$$

gdzie:

- d_{il} – odległość między i-tym obiektem a l-tym środkiem,
- x_{ij} – i-ty obiekt o j-tej zmiennej,
- $\bar{x}_{lj}$ – l-ty środek o j-tej zmiennej,
- i – numer obiektu,
- j – numer zmiennej,
- l – numer klasy (identyfikuje jej środek),
- n – liczba wszystkich obiektów,
- m – liczba wszystkich zmiennych.

W pracy podczas stosowania metody k-średnich zostanie wykorzystana odległość euklidesowa, ponieważ zostaje ona stosowana w wielu badaniach przeprowadzanych przez ekspertów. Jej zaletą jest także wrażliwość na wartości odstające, gdyż różnice są podnoszone do kwadratu. Brakuje tego w np. odległości miejskiej.

Ogólna zasada działania metody k-średnich polega na tym, aby podzielić obiekty na k grup. Pierwsze środki klas są wybierane w sposób losowy – mają znajdować się jak najdalej od siebie. Następnie obiekty zostają przypisywane do najbliższych im środków. W ten sposób powstają pierwsze grupy. W kolejnym kroku obliczany jest na nowo środek klasy (środek ciężkości) jako średnia arytmetyczna współrzędnych obiektów należących do poszczególnych skupisk – według wzoru (2.3). W kolejnych krokach następują migracje obiektów między grupami na podstawie ich odległości od środków, tak aby

[33] HUANG, A., Similarity measures for text document clustering. In: *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand*. 2008. p. 9-56. [51]
[34] PIETRZYKOWSKI, R.; KOBUS, P., Zastosowanie modyfikacji metody k-średnich w analizie portfelowej. *Ekonomika i Organizacja Gospodarki Żywnościowej*, 2006, 60: 301-308. [303]

klasy spełniały zasadę jednorodności wewnątrzgrupowej i różnorodności międzygrupowej. Algorytm ten zatrzymuje się, gdy skład grup nie ulega zmianie.[35]

Kolejne etapy działania metody k-średnich można ująć w następujących krokach:

1) losowy wybór centroid k grup;
2) przypisanie obiektów do centroid na podstawie odległości euklidesowej obliczonej według wzoru (2.2) między obiektami a środkami ciężkości; dołączenie obiektu do danej grupy następuje, jeśli odległość między nimi jest najmniejsza;
3) obliczenie nowych centroid na podstawie przynależących do nich punktów, według wzoru (2.3)

$$b_{lj} = \frac{1}{n_l}\sum_{i=1}^{n_l} x_{ijl}, \tag{2.3}$$

gdzie:

- b_{lj} – współrzędna j-tej zmiennej l-tego środka,
- x_{ijl} – współrzędna i-tego obiektu j-tej zmiennej w klasie o l-tym środku,
- i – numer obiekty,
- j – numer zmiennej,
- l – numer klasy (identyfikuje jej środek),
- n_l – liczba obiektów w l-tej klasie;

4) powtarzanie kroków 2) i 3) do momentu, aż obiekty przestaną zmieniać grupy oraz współrzędne centroid nie będą ulegać zmianie.

Metoda k-średnich jest złożona obliczeniowo. Czas jej wykonania może czasami stanowić problem, gdyż jest długi. W dużej mierze podział uzależniony jest od początkowego ułożenia centroid. Z drugiej strony algorytm ten został przeanalizowany przez wielu ekspertów. Potwierdzono także jego skuteczność, dlatego zostanie on użyty w pracy. Jej głównym celem jest porównanie wyników otrzymanych przy użyciu różnych typów metod analizy skupień – nie koncentruje się na metodach optymalizacyjnych.[36]

[35] ROSZKO-WÓJTOWICZ, E., Analiza skupień w ocenie warunków pracy w krajach Unii Europejskiej. *Wiadomości Statystyczne*, 2014, 59.11. [9]

[36] NAZEER, KA A.; SEBASTIAN, M. P., Improving the Accuracy and Efficiency of the k-means Clustering Algorithm. In: *Proceedings of the world congress on engineering*. London: Association of Engineers, 2009. p. 1-3.

2.2. Klasyfikacja spektralna

Klasyfikacja spektralna odbywa się na podstawie zmienionych danych początkowych, tak że tworzą one macierz o tej samej liczbie obiektów (liczba wierszy), natomiast liczba zmiennych jest równa liczbie klas na jaką nastąpi podział (liczba kolumn).[37] Taka ramka danych jest poddawana grupowaniu przy użyciu metody tradycyjnej. W pracy zostanie użyta metoda k-średnich, w celu zauważenia możliwych różnic w wyniku zastosowania przemian klasyfikacji spektralnej w porównaniu do otrzymanego podziału przy użyciu metody k-średnich bez tych zmian.[38]

Proces klasyfikacji spektralnej można podzielić na kolejne kroki postępowania. Pierwszym z nich jest sporządzenie ramki danych $X = [x_{ij}]$. Macierz ta posiada wymiary $n \times m$, gdzie n to liczba obiektów, a m liczba zmiennych opisująca obiekty. Wartości zmiennych i, j są następujące $i = 1, \dots, n; j = 1, \dots, m$. W przypadku danych metrycznych (przedziałowych oraz ilorazowych), utworzona macierz musi zostać znormalizowana. W pracy proces ten odbędzie się przy zastosowaniu wzoru (2.1). Otrzymana zostanie macierz znormalizowana $Z = [z_{ij}]$.[39]

Krokiem drugim jest utworzenie macierzy podobieństw (ang. *affinity matrix*), inaczej mówiąc macierzy odległości $A = [A_{ik}]$. Ma ona wymiary $n \times n$. Wartości zmiennych i, k są następujące: $i, k = 1, \dots, n$. Macierz składa się z odległości między poszczególnymi obiektami, zatem wartości na głównej przekątnej są równe 0 ($A_{ii} = 0$). Pozostałe wartości obliczane są z wykorzystaniem wzoru (2.4).

$$A_{ik} = e^{(-\sigma * d_{ik})} \tag{2.4}$$

gdzie:

- σ – parametr skali,
- d_{ik} – odległość między i-tym a k-tym obiektem.

Wartość d_{ik} może być obliczona za pomocą różnych metod. W pracy zostanie użyta odległość euklidesowa. Metoda ta stosowana będzie także w pozostałych algorytmach klasyfikacji w celu uniknięcia różnic wynikających z obliczania odległości za pomocą odmiennych metod.

---

[37] WALESIAK, M.; DUDEK, A., *Odległość GDM dla danych porządkowych a klasyfikacja spektralna*. 2009. [11]
[38] DUDEK, A., Klasyfikacja spektralna a tradycyjne metody analizy skupień. *Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu. Ekonometria*, 2009, 23.37 Zastosowania metod ilościowych: 21-34. [23]
[39] WALESIAK, M.; DUDEK, A., *Odległość GDM dla danych porządkowych a klasyfikacja spektralna*. 2009. [11]

Do wyznaczania macierzy podobieństw można użyć także innych estymatorów jądrowych. Stosuje się wtedy inną funkcję, niż zaproponowaną wyżej we wzorze (2.4). Przykładem może być jądro gaussowskie. Macierz podobieństw jest wtedy obliczana wzorem (2.5).

$$A_{ik} = e^{\left(-\frac{d_{ik}^2}{2*\sigma^2}\right)} \tag{2.5}$$

gdzie $i \neq k$. Pozostałe oznaczenia są takie same jak we wzorze (2.4) powyżej.[40]

Trzeci etap obejmuje budowę macierzy wag D. Jej główną przekątną tworzą sumy wartości kolejnych wierszy macierzy A uzyskanej za pomocą wzoru (2.4) lub (2.5). Pozostałe pola w macierzy wag są wypełnione przez zera. Jest to macierz diagonalna.[41]

Kolejnym krokiem jest utworzenie znormalizowanej macierzy Laplace'a L. Tworzy się ją według wzoru (2.6).

$$L = D^{-\frac{1}{2}} * A * D^{-\frac{1}{2}} \tag{2.6}$$

gdzie:

- D – macierz wag,
- A – macierz podobieństw (odległości).

Zostaje pominięty etap odejmowania macierzy L od macierzy diagonalnej (na głównej przekątnej znajdują się jedynki, a pozostałe wartości to zera), w celu ułatwienia interpretacji uzyskanych wyników.[42]

Piąty etap obejmuje wyznaczenie wartości oraz wektorów własnych macierzy Laplace'a L. Są one niezbędne do budowy $E = [e_{ij}]$. Wektory własne są uporządkowane w ten sposób, aby wartości własne na podstawie, których zostały obliczone były ustawione malejąco. Niech u będzie ustaloną liczbą klas, wtedy macierz E przyjmuje wartości u najlepszych (pierwszych) wektorów własnych. W ten sposób macierz E zyskuje wymiary $n \times u$ (n – liczba badanych obiektów).[43]

Ostatnim etapem jest wykonanie operacji normalizacji uzyskanej macierzy E. Poszczególne znormalizowane elementy otrzymane są na podstawie wzoru (2.7).

---

[40] WALESIAK, M., Klasyfikacja spektralna a skale pomiaru zmiennych. *Przegląd Statystyczny*, 2012, 1.59: 13-31. [15-16]
[41] WALESIAK, M.; DUDEK, A., *Odległość GDM dla danych porządkowych a klasyfikacja spektralna*. 2009. [11]
[42] WALESIAK, M., Klasyfikacja spektralna a skale pomiaru zmiennych. *Przegląd Statystyczny*, 2012, 1.59: 13-31. [16]
[43] WALESIAK, M.; DUDEK, A., *Odległość GDM dla danych porządkowych a klasyfikacja spektralna*. 2009. [12]

$$y_{ij} = \frac{e_{ij}}{\sqrt{\sum_{j=1}^{u} e_{ij}^2}} \tag{2.7}$$

gdzie:

- i – numer obiektu ($i = 1, \dots, n$),
- j – numer zmiennej ($j = 1, \dots, u$).

Po obliczeniu wartości dla każdego elementu ze wzoru (2.7) powstaje macierz $Y = [y_{ij}]$. Na podstawie tej macierzy dokonuje się klasyfikacji obiektów za pomocą tradycyjnych metod analizy skupień.[44]

Podczas tworzenia macierzy podobieństw A zgodnie ze wzorem (2.4) pojawia się problem doboru wartości parametru skali σ. Jest on ważnym elementem klasyfikacji spektralnej. Istnieją metody heurystyczne szacujące jak najlepszy parametru. Mają one na celu dobór takiej wartości parametru, aby podobieństwo wewnątrzgrupowe było jak największe, a zatem odbywa się proces minimalizacji odległości w grupach.[45]

Jedną z metod oszacowania wartości parametru skali zaproponował Walesiak w swojej pracy, której kolejne etapy można opisać następująco[46]:

Dobór próby bootstrapowej X' z macierzy odległości X (w przypadku danych ilorazowych z macierzy odległości po jej normalizacji). Próba ta zawiera n' elementów, gdzie $\frac{1}{2}n \leq n' \leq \frac{3}{4}n$ (n – liczba wszystkich obserwacji) scharakteryzowanych przez m zmiennych (m – liczba wszystkich zmiennych). $S_0 = [0; D]$ jest to pierwszy przyjęty przedział, w którym znajduje się optymalna wartość parametru skali. D to suma odległości d_{ik} w macierzy podobieństw. Podane poniżej kolejne kroki są powtarzane do momentu znalezienia najlepszej wartości parametru σ.

1. S_k jest to przedział w k- tej iteracji (w pierwszej $S_k = S_0$). Zostaje on poddany operacji dzielenia na R mniejszych przedziałów o równych rozstępach. Można je zatem określić jako $p_r^k = \left[\underline{p_r^k}; \overline{p_r^k}\right]$, gdzie $r = 1, \dots, R$.

2. Dla nowo otrzymanych przedziałów szacowana jest wartość średnia $\sigma_r^k = \frac{\underline{p_r^k} + \overline{p_r^k}}{2}$. Otrzymane wartości są podstawiane do wzoru (2.3) i zostają przeprowadzone klasyfikacje spektralne próby bootstrapowej X' na określaną liczbę grup u.

[44] WALESIAK, M., Klasyfikacja spektralna a skale pomiaru zmiennych. *Przegląd Statystyczny*, 2012, 1.59: 13-31. [16-17]
[45] WALESIAK, M.; DUDEK, A., *Odległość GDM dla danych porządkowych a klasyfikacja spektralna*. 2009. [12]
[46] WALESIAK, M.; DUDEK, A., *Odległość GDM dla danych porządkowych a klasyfikacja spektralna*. 2009. [12-13]

3. Spośród uzyskanych wartości parametru skali, zostaje wybrany ten środek σ_r^k, dla którego zauważono najmniejszą różnicę w poszczególnych grupach, czyli sumy odległości w klasach były najmniejsze.
4. Algorytm kończy działanie, jeśli dla otrzymanego w poprzednim kroku parametru σ_r^k spełniony jest warunek $p_r^k \leq \emptyset$, gdzie $\emptyset = 10^{-3}$.W innym przypadku algorytm wraca do punktu 2 i powtarza działanie.

2.3. Metoda Ward'a

Metody hierarchiczne aglomeracyjne bazują na macierzy odległości. Ma ona wymiary $n \times n$, gdzie n to liczba klasyfikowanych obiektów. Główna przekątna takiej macierzy składa się z zer, gdyż wartości odległości między tymi samymi grupami są równe zero. Macierz odległości po każdym połączeniu klas jest modyfikowana, gdyż obliczane są odległości między nową grupą a pozostałymi skupiskami. Jeśli nastąpi połącznie klasy A_p z klasą A_q w klasę A_r to jej odległość od pozostałych grup można wyrazić wzorem (2.8).

$$d_{ir} = a_p * d_{ip} + a_q * d_{iq} + b * d_{pq} + c * |d_{ip} - d_{iq}|, \tag{2.8}$$

gdzie:

- d_{ir} – odległość między i-tą klasą (A_i) a klasą A_r,
- d_{ip} – odległość między i-tą klasą (A_i) a klasą A_p,
- d_{iq} – odległość między i-tą klasą (A_i) a klasą A_q,
- d_{pq} – odległość między klasą A_p a klasą A_q,
- a_p, a_q, b, c – parametry.

Parametry a_p, a_q, b, c przyjmują odpowiednie wartości w zależności od wzorów, które są charakterystyczne dla poszczególnych metod. W przypadku metody Ward'a są one dane wzorem (2.9).

$$a_p = \frac{N_i+N_p}{N_i+N_r}, \quad a_q = \frac{N_i+N_q}{N_i+N_r}, \quad b = \frac{-N_i}{N_i+N_r}, \quad c = 0, \tag{2.9}$$

gdzie:

- N_i, N_p, N_q, N_r – określają odpowiednio liczbę obiektów w grupach A_i, A_p, A_q, A_r.

Po wykonaniu takiego połączenia wymiar macierzy odległości zmniejsza się o jeden.[47] Metoda Ward'a jak można zauważyć we wzorach (2.9) uwzględnia liczbę elementów znajdujących się w klasie, podczas określania odległości jednej grupy od drugiej.[48]

Do wyznaczenia odległości pomiędzy obiektami zostanie użyta odległość euklidesowa. Odległość ta obrazuje geometryczny układ badanych obiektów. Zdecydowano się wykorzystać tą odległość, ponieważ została ona wybrana we wszystkich poprzednich metodach, zatem wyniki klasyfikacji będą porównywalne. Pozwala to zniwelować różnice, które mogłyby wyniknąć w przypadku użycia dwóch różnych miar.

Ogólna idea metody Ward'a przy wykorzystaniu odległości euklidesowej ma na celu minimalizację sumy kwadratów odległości obiektów należących do danej grupy od jej środka ciężkości. Uzyskuje się go ze wzoru (2.10)[49].

$$\frac{m_i * A_i + m_j * A_j}{m_i + m_j} \tag{2.10}$$

gdzie:

- m_i – liczba elementów w klasie A_i,
- m_j – liczba elementów w klasie A_j,
- A_i – wektor $(x_{i1}, x_{i2}, \dots, x_{im})$, gdzie m to liczba zmiennych,
- A_j – wektor $(x_{j1}, x_{j2}, \dots, x_{jm})$, gdzie m to liczba zmiennych.

Dzięki minimalizacji wariancji klasy składają się z obiektów jak najbardziej podobnych do siebie. Dwie grupy zostaną połączone, jeśli suma kwadratów odległości elementów nowej grupy od nowopowstałego ważonego środka ciężkości jest najmniejsza. Innymi słowy obliczane są wartości dla każdej kombinacji połączeń, a następnie wybierane jest to, dla którego suma była najmniejsza. Następuje wtedy modyfikacja macierzy podobieństw. Proces kończy się, gdy wszystkie obiekty tworzą jedną klasę. Kolejne etapy połączeń są przedstawiane najczęściej na dendrogramie – o którym więcej powiedziane zostało w rozdziale 1.3 Rysunek 1.2. W wyniku zastosowania metody Ward'a powstają

[47] ĆWIĄKAŁA-MAŁYS, A.; MOŚCIBRODZKA, M., Hierarchiczne procedury aglomeracyjne w badaniu poziomu i struktury kosztów publicznych uczelni akademickich. *Metody ilościowe w badaniach ekonomicznych*, 2014, 15.3: 30-41. [31]

[48] WALESIAK, M.; GATNAR, E. (ed.)., *Metody statystycznej analizy wielowymiarowej w badaniach marketingowych*. Wydaw. Akademii Ekonomicznej im. Oskara Langego, 2004. [327]

[49] RASMUSSEN, E. M., Clustering algorithms. *Information retrieval: data structures & algorithms*, 1992, 419: 442.

grupy jednorodne.[50] Wzory (2.8) oraz (2.9) są uproszczeniem numerycznym opisanej idei metody Ward'a.

W przypadku, gdy do obliczenia odległości między obiektami stosowany jest kwadrat odległości euklidesowej, metoda Ward'a wykorzystuję analizę wariancji. Jej podstawowym celem jest jak najmniejszy wzrost wariancji na skutek połączenia dwóch grup. Wzór na wariancję nowopowstałego klastra po połączeniu klas A_i oraz A_j można wyrazić równaniem (2.11)[51].

$$I_{ij} = \frac{m_i * m_j}{m_i + m_j} * d_{ij}^2 \tag{2.11}$$

gdzie:

- m_i – liczba elementów w klasie A_i,
- m_j – liczba elementów w klasie A_j,
- d_{ij}^2 – kwadrat odległości euklidesowej między i-tymi a j-tymi obiektami.

Odległości euklidesowe obliczone są na podstawie danych zestandaryzowanych zgodnie ze wzorem (2.1), natomiast same wartości odległości wyliczane są na podstawie wzoru (2.2).

W pracy zostanie wykorzystana metoda Ward'a, ponieważ jest to jedna z bardzo efektywnych metod klasyfikacji. W odróżnieniu od innych metod połączenia, scalenie klas nie następuje tylko na podstawie odległości najbliższych lub najdalszych obiektów w sąsiednich skupiskach, lecz oparta jest o analizę wariancji. Na podstawie wzoru (2.11), który daje informację o wariancji nowej grupy, można zauważyć, iż im mniejsze są odległości łączenia, tym mniejsza jest wartość wariancji.

[50] ROSZKO-WÓJTOWICZ, E., Analiza skupień w ocenie warunków pracy w krajach Unii Europejskiej. *Wiadomości Statystyczne*, 2014, 59.11. [11]
[51] RASMUSSEN, E. M., Clustering algorithms. *Information retrieval: data structures & algorithms*, 1992, 419: 442.

3. Przeprowadzenie badania

Celem niniejszego badania jest klasyfikacja województw na grupy o podobnych przyczynach zgonów opisanych za pomocą „garbage codes". Otrzymane wyniki uzyskane przy zastosowaniu trzech metod pozwolą na porównanie końcowych skupisk oraz określenie w nich różnic wynikających z różnorodności sposobów przeprowadzenia analizy skupień.

3.1. Prezentacja danych

Dane wykorzystane w pracy zostały zaczerpnięte ze strony internetowej Głównego Urzędu Statystycznego[52]. Dotyczą one przyczyn zgonów opisanych przy użyciu tzw. „kodów śmieciowych", czyli „garbage codes". Dane pochodzą z 2018 roku. Są one kompletne – nie występują w nich braki. Dla każdego województwa podana została liczba osób, których przyczyną zgonu były odpowiednie kody. W celu zniwelowania problemu różnicy wielkości zaludnienia poszczególnych jednostek administracyjnych najwyższego stopnia, miarą danych jest udział procentowy liczby zgonów poszczególnych kodów dla danego województwa. Wyniki prezentuje Tabela 3.1. Województwa opisane za pomocą zmiennych:

- C76 – nowotwór złośliwy umiejscowień innych i niedokładnie określonych,
- I46 – zatrzymanie krążenia,
- I50 – niewydolność serca,
- I70.9 – uogólniona i nieokreślona miażdżyca,
- R54 – starość,
- R96 – śmierć natychmiastowa,
- R98 – śmierć nieoczekiwana,
- R99 – inne, niedokładnie określone i nieznane przyczyny umieralności,
- Y21 – utonięcie, zanurzenie, o nieokreślonym zamiarze,
- Y34 – nieokreślone zdarzenie, nieokreślony zamiar.

[52] *Zgony według przyczyn określanych jako "garbage codes"*, https://stat.gov.pl/obszary-tematyczne/ludnosc/statystyka-przyczyn-zgonow/zgony-wedlug-przyczyn-okreslanych-jako-garbage-codes,3,3.html, 02.07.2020

Tabela 3.1 Dane w postaci udziałów procentowych

Województwo	C76	I46	I50	I70.9	R54
Dolnośląskie	0.0010	0.0000	0.4970	0.2256	0.0967
Kujawsko-pomorskie	0.0212	0.0003	0.3142	0.2354	0.0716
Lubelskie	0.0274	0.0269	0.4088	0.0385	0.3126
Lubuskie	0.0008	0.0000	0.4401	0.0621	0.2555
Łódzkie	0.0077	0.0003	0.3513	0.2249	0.0914
Małopolskie	0.0046	0.0011	0.0095	**0.7047**	0.1275
Mazowieckie	0.0034	0.0003	0.3323	0.1460	0.2164
Opolskie	0.0058	0.0348	0.1992	0.3482	0.1217
Podkarpackie	0.0126	0.0005	0.4017	0.2562	0.0772
Podlaskie	0.0004	0.0000	0.0138	0.1224	0.1789
Pomorskie	0.0007	0.0637	0.6711	0.0193	0.0156
Śląskie	0.0043	0.0144	0.1469	0.4384	0.0982
Świętokrzyskie	0.0295	0.0168	0.4967	0.0247	0.2716
Warmińsko-mazurskie	0.0003	0.0011	0.2923	0.2901	0.2206
Wielkopolskie	0.0006	0.0003	0.4390	0.1234	0.1721
Zachodniopomorskie	0.0024	0.0000	0.5211	0.2043	0.0231
Województwo	**R96**	**R98**	**R99**	**Y21**	**Y34**
Dolnośląskie	0.0020	0.1316	0.0027	0.0009	0.0000
Kujawsko-pomorskie	0.1788	0.0364	0.0684	0.0058	0.0155
Lubelskie	0.1230	0.0030	0.0442	0.0019	0.0039
Lubuskie	0.0115	0.1849	0.0059	0.0027	0.0011
Łódzkie	0.0585	0.0069	0.1945	0.0002	0.0240
Małopolskie	0.0018	0.0015	0.1062	0.0001	0.0007
Mazowieckie	0.1230	0.0927	0.0177	0.0016	0.0237
Opolskie	0.0398	0.0075	0.1644	0.0037	0.0008
Podkarpackie	0.1225	0.0413	0.0431	0.0022	0.0003
Podlaskie	0.2759	0.0022	0.0056	0.0004	0.0015
Pomorskie	0.0067	0.0037	0.0148	0.0000	0.0015
Śląskie	0.0357	0.0147	0.1613	0.0020	0.0002
Świętokrzyskie	0.1087	0.0019	0.0440	0.0000	0.0010
Warmińsko-mazurskie	0.0988	0.0341	0.0132	0.0005	0.0003
Wielkopolskie	0.0024	0.2193	0.0010	0.0002	0.0002
Zachodniopomorskie	0.1569	0.0091	0.0353	0.0044	0.0000

Źródło: opracowanie własne

Tabela 3.1 przedstawia udział procentowy poszczególnych dziesięciu przyczyn zgonów we wszystkich zgonach określonych jako „garbage codes" w danym województwie. Nazwy zmiennych odpowiadają numerom „kodów śmieciowych".

Zauważono, że liczby są zróżnicowane, więc nie powinien wystąpić problem związany ze zmiennością atrybutów. Największy odsetek zanotowano dla zgonów w województwie małopolskim w wyniku uogólnionej i nieokreślonej miażdżycy.

3.1.1. Przygotowanie zmiennych bazowych

Zmienne wykorzystywane do klasyfikacji województw muszą spełniać pewne warunki. Dane zostaną zweryfikowane poprzez sprawdzenie korelacji oraz współczynnika zmienności. Przeprowadzona zostanie ich standaryzacja oraz test na występowanie wartości odstających.

Współczynnik korelacji Pearsona bada liniową zależność miedzy dwoma zmiennymi. Jeśli przyjmuje on wartość dodatnią to oznacza to, iż wzrost jednej zmiennej powoduje wzrost drugiej, natomiast jeśli przyjmuje wartość ujemną to wzrost jednej zmiennej powoduje spadek drugiej zmiennej. Współczynnik korelacji Pearsona dla badanych zmiennych przedstawia Tabela 3.2. Wartości współczynnika korelacji powinny przyjmować wartości mniejsze od 0.9 oraz i większe od -0.9. W Tabeli 3.2 wszystkie otrzymane wyniki zawierają się w tym przedziale. Pomija się wartości na przekątnej, gdyż korelacja między tą samą zmienną jest zawsze równa 1. Na podstawie współczynnika korelacji Pearsona wszystkie zmienne mogą być użyte w dalszej części badania.

Tabela 3.2 Współczynnik korelacji Pearsona między zmiennymi

	C76	**I46**	**I50**	**I70.9**	**R54**	**R96**	**R98**	**R99**	**Y21**	**Y34**
C76	1.00	0.13	0.14	-0.24	0.37	0.29	-0.36	0.13	0.19	0.14
I460	0.13	1.00	0.34	-0.26	-0.14	-0.27	-0.37	0.11	-0.11	-0.21
I50	0.14	0.34	1.00	**-0.68**	-0.11	-0.26	0.28	-0.41	**-0.01**	-0.02
I70.9	-0.24	-0.26	-0.68	1.00	-0.32	-0.25	-0.25	**0.55**	0.06	-0.10
R54	0.37	-0.14	-0.11	-0.32	1.00	0.11	0.22	-0.28	-0.26	**-0.01**
R96	0.29	-0.27	-0.26	-0.25	0.11	1.00	-0.40	-0.20	0.30	0.20
R98	-0.36	-0.37	0.28	-0.25	0.22	-0.40	1.00	-0.47	-0.05	-0.05
R99	0.13	0.11	-0.41	0.55	-0.28	-0.20	-0.47	1.00	0.14	0.29
Y21	0.19	-0.11	-0.01	0.06	-0.26	0.30	-0.05	0.14	1.00	0.11
Y34	0.14	-0.21	-0.02	-0.10	-0.01	0.20	-0.05	0.29	0.11	1.00

Źródło: opracowanie własne

Największą dodatnią korelację zanotowano między zmiennymi określającymi zgony z przyczyn: uogólniona i nieokreślona miażdżyca (I70.9) oraz inne, niedokładnie określone i nieznane przyczyny umieralności (R99). Wynosi ona 0.55. Największą ujemną korelację (-0.68) zauważono w przypadku atrybutów: uogólniona i nieokreślona miażdżyca (I70.9) oraz niewydolność serca (I50). Współczynnik korelacji Pearsona wskazuje najmniejsza korelację między dwoma parami zmiennych: Y21 (utonięcie, zanurzenie, o nieokreślonym zamiarze) i I50 (niewydolność serca) oraz Y34 (nieokreślone zdarzenie, nieokreślony zamiar) i R54 (starość). Wartość współczynnika jest równa -0.01.

Kolejnym aspektem jest współczynnik zmienności. Informuje on w jakim stopniu są zróżnicowane wartości danej zmiennej. Jeśli jest on większy to występuje duża zmienność. W przeciwnym wypadku jest ona mała. Przyjmuje się, że współczynnik zmienności powinien przekraczać 10%. W Tabeli 3.3 przedstawiono wyniki dla badanych zmiennych. Wszystkie wartości są znacznie większe od 10%. Wskazuje to na duże zróżnicowanie wśród danych. Na podstawie współczynnika zmienności każda zmienna kwalifikuje się do dalszego badania.

Tabela 3.3 Współczynnik zmienności dla poszczególnych zmiennych

C76	I46	I50	I70.9	R54	R96	R98	R99	Y21	Y34
127.90	179.91	52.53	81.43	60.57	93.50	141.86	111.05	105.69	179.82

Źródło: opracowanie własne

Największe wartości dla współczynnika zmienności występują dla zmiennych I46 (zatrzymanie krążenia) oraz Y34 (nieokreślone zdarzenie, nieokreślony zamiar). Atrybuty te przyjmują wartości zarówno niskie jak i wysokie. Najmniejszy współczynnik zmienności zaobserwowano dla zgonów w wyniku niewydolności serca (I50). Wynika to z faktu, iż w większości województw odsetek tej przyczyny zgonu jest wysoki.

Kolejnym etapem jest przeprowadzenie transformacji normalizacyjnej, polegającej na standaryzacji danych według wzoru (2.1). Uzyskane wartości przedstawia Tabela 3.4. Po przeprowadzeniu procesu normalizacji, dane są porównywalne co umożliwia wykrycie wartości szczególnych. W tym celu przeprowadzono test sprawdzający czy występują wartości odstające – test Grubbsa. Jako hipotezę zerową przyjmuje on bark występowania wartości szczególnych. Tabela 3.5 zawiera otrzymane wartości p-value. W przypadku zmiennych I46 (zatrzymanie krążenia) oraz I70.9

(uogólniona i nieokreślona miażdżyca) wynik jest mniejszy od 5%, zatem mogą występować tam wartości odstające. Według reguły trzech sigm należy usunąć te obserwacje, które po standaryzacji są większe od 3 lub mniejsze od -3. Na podstawie Tabeli 3.4 zauważono, że zmienna I46 przyjmuje wartość 2.98 dla województwa pomorskiego, natomiast zmienna I70.9 przyjmuje wartość 2.77 dla województwa Małopolskiego. Obie wartości są bliskie 3, więc można podejrzewać, że to one są wartościami odstającymi. Nie przekraczają one jednak przedziału wyznaczonego przez regułę trzech sigm, dlatego zdecydowano o pozostawieniu zbioru danych standaryzowanych bez zmian.

Tabela 3.4 Dane po procesie standaryzacji

Województwo	C76	I46	I50	I70.9	…	R99	Y21	Y34
Dolnośląskie	-0.68	-0.56	0.83	0.05	…	-0.86	-0.45	-0.56
Kujawsko-pomorskie	1.38	-0.54	-0.17	0.11	…	0.17	2.36	1.30
Lubelskie	2.01	0.94	0.35	-1.01	…	-0.21	0.13	-0.09
Lubuskie	-0.70	-0.56	0.52	-0.88	…	-0.81	0.58	-0.43
Łódzkie	0.00	-0.54	0.03	0.05	…	2.14	-0.85	2.31
Małopolskie	-0.31	-0.50	-1.85	2.77	…	0.76	-0.87	-0.48
Mazowieckie	-0.44	-0.54	-0.08	-0.40	…	-0.62	-0.01	2.27
Opolskie	-0.19	1.37	-0.81	0.75	…	1.67	1.18	-0.46
Podkarpackie	0.50	-0.53	0.31	0.22	…	-0.23	0.28	-0.52
Podlaskie	-0.74	-0.56	-1.83	-0.53	…	-0.81	-0.73	-0.38
Pomorskie	-0.71	2.98	1.79	-1.12	…	-0.67	-0.95	-0.38
Śląskie	-0.34	0.24	-1.10	1.26	…	1.62	0.19	-0.53
Świętokrzyskie	2.23	0.38	0.83	-1.09	…	-0.21	-0.95	-0.44
Warmińsko-mazurskie	-0.75	-0.50	-0.30	0.42	…	-0.70	-0.64	-0.52
Wielkopolskie	-0.72	-0.54	0.51	-0.53	…	-0.88	-0.84	-0.53
Zachodniopomorskie	-0.54	-0.56	0.96	-0.07	…	-0.35	1.56	-0.56

Źródło: opracowanie własne z wykorzystaniem programu R

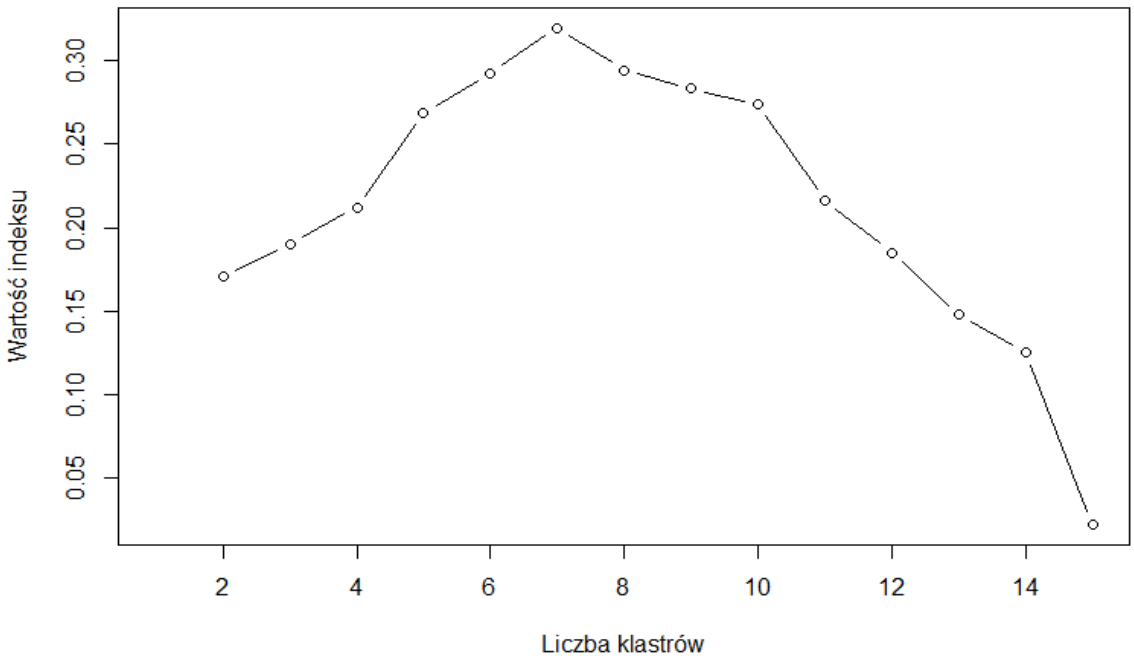
Tabela 3.5 Wartości p-value dla testu Grubbsa

C76	I46	I50	I70.9	R54	R96	R98	R99	Y21	Y34
0.120	**0.002**	0.416	**0.009**	0.404	0.051	0.055	0.169	0.072	0.089

Źródło: opracowanie własne z wykorzystaniem programu R

3.2. Klasyfikacja województw metodą k-średnich

W celu określenia liczby skupisk, na którą zostanie dokonany podział województw za pomocą metody k-średnich, zastosowano m.in. indeks sylwetkowy. Na podstawie Rysunku 3.1, który obrazuje otrzymane wyniki, można podejrzewać, że najlepszy podział danych będzie występował, w przypadku siedmiu klastrów. Dla tej liczby skupisk indeks sylwetkowy przyjmuje największa wartość. W dalszych krokach przeprowadzono analizę merytoryczną, przy podziale na pięć, sześć, siedem orz osiem grup. Są to liczby klastrów, dla których indeks sylwetkowy przyjmuje duże wartości w porównaniu do pozostałych.

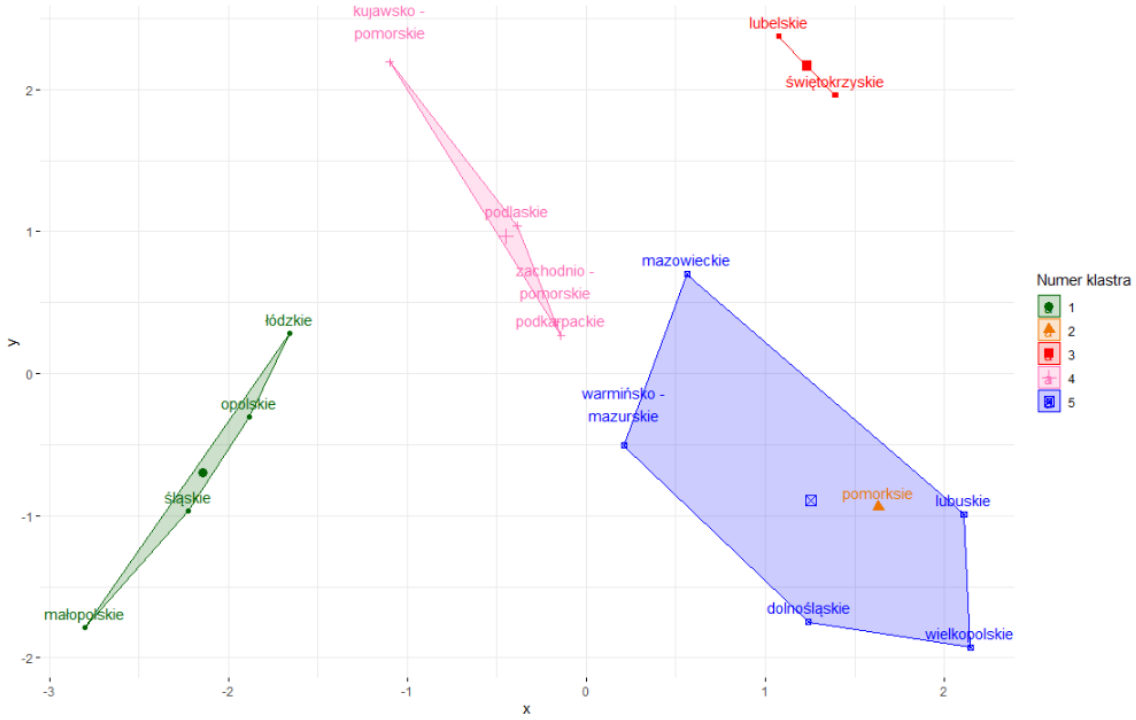


Rysunek 3.1 Indeks sylwetkowy dla metody k-średnich

Źródło: opracowanie własne z wykorzystaniem programu R

Po podziale na pięć grup uzyskano klastry składające się z województw:

1) łódzkie, małopolskie, opolskie, śląskie;
2) pomorskie;
3) lubelskie, świętokrzyskie;
4) kujawsko-pomorskie, podkarpackie, podlaskie, zachodniopomorskie;
5) dolnośląskie, lubuskie, mazowieckie, warmińsko-mazurskie, wielkopolskie.



Rysunek 3.2 Podział na pięć klastrów metodą k-średnich w postaci punktów na wykresie

Źródło: opracowanie własne z wykorzystaniem programu R

Rysunek 3.3 Podział administracyjny Polski z podziałem województw na pięć klastrów uzyskanych metodą k-średnich

Źródło: opracowanie własne z wykorzystaniem programu R

Wyniki zilustrowane w postaci geometrycznej na Rysunku 3.2. umożliwiają zaobserwowanie odległości między klasyfikowanymi obiektami, a zatem ich miarę podobieństwa (im są bliżej tym są do siebie bardziej podobne). Zauważono, że skupisko składające się z województw: lubelskiego oraz świętokrzyskiego znajduje się w znacznej odległości od pozostałych. Podejrzewa się, że w dalszych krokach zwiększając liczbę klastrów nie zostanie on podzielony. W przypadku grupy piątej, która jest najbardziej liczna, można przypuszczać, iż ulegnie ona rozpadowi na mniejsze klastry.

W celu dogodnego zobrazowania ułożenia geograficznego poszczególnych klastrów, wykorzystano mapę administracyjną Polski (Rysunek 3.3). Zaobserwowano, że województwa należące do grupy czwartej (kolor różowy) są najbardziej oddalone od siebie (w sensie geograficznym). Pozostałe klastry składają się z sąsiadujących ze sobą województw.

W celu dalszej oceny powstałych klastrów obliczono odchylenia standardowe dla poszczególnych zmiennych. Została pominięta grupa druga, gdyż tworzy ja jedno województwo pomorskie. Wyniki przedstawiono w Tabeli 3.6. Największą wartość odchylenia standardowego (0.21658) zanotowano w grupie czwartej dla zmiennej I50, która opisuje zgony w wyniku niewydolności serca. W klastrze pierwszym również zauważono wyższe wartości odchyleń standardowych (0.14123 oraz 0.20352). Wszystkie wyższe wartości w Tabeli 3.6 zostały wyróżnione pogrubioną czcionką. Można przypuszczać, że te skupiska zostaną podzielone w kolejnych krokach, aby uległy one zmniejszeniu. Jest to zjawisko pożądane.

Tabela 3.6 Odchylenia standardowe poszczególnych zmiennych w każdym z pięciu klastrów uzyskanych metodą k-średnich; nie uwzględniono grupy jednoelementowej

	I	**III**	**IV**	**V**
C76	0.00152	0.00154	0.00966	0.00124
I46	0.01613	0.00716	0.00024	0.00044
I50	**0.14123**	0.06218	**0.21658**	0.08473
I70.9	**0.20352**	0.00974	0.05881	0.08933
R54	0.01761	0.02895	0.06545	0.06105
R96	0.02360	0.01014	0.06584	0.05860
R98	0.00544	0.00076	0.01946	0.07339
R99	0.03679	0.00010	0.02586	0.00711
Y21	0.00171	0.00134	0.00240	0.00099
Y34	0.01172	0.00207	0.00749	0.01042

Źródło: opracowanie własne z wykorzystaniem programu R

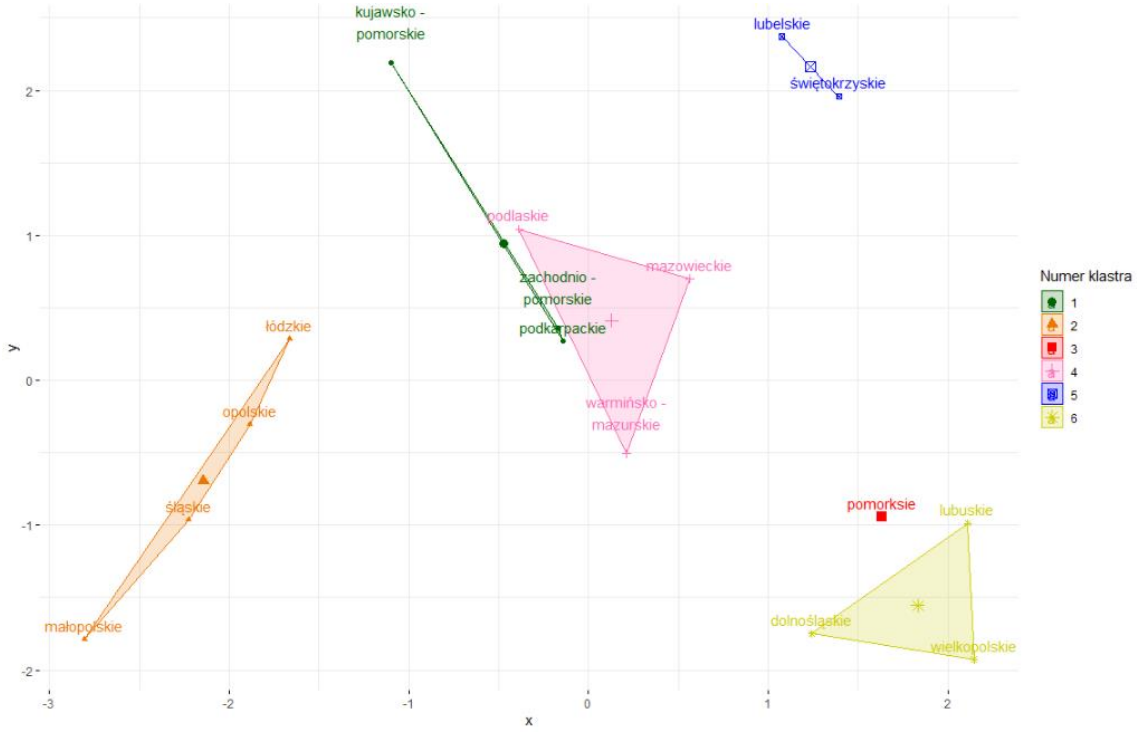
W wyniku podziału na sześć grup powstały klastry składające się z województw:

1) kujawsko-pomorskie, podkarpackie, zachodniopomorskie;
2) łódzkie, małopolskie, opolskie, śląskie;
3) pomorskie;
4) mazowieckie, podlaskie, warmińsko-mazurskie;
5) lubelskie, świętokrzyski;
6) dolnośląskie, lubuskie, wielkopolskie.

Na Rysunku 3.4 przedstawiającym skupiska w postaci geometrycznej można zauważyć, że największa grupa z podziału na pięć klastrów uległa rozpadowi, tak jak przewidywano. Powstały mniejsze skupiska, składające się z obiektów leżących bliżej siebie.

Na podstawie Rysunku 3.5 przedstawiającego mapę administracyjną zauważono, że po podziale na większą liczbę klastrów, grupy składają się z województw jeszcze bliżej położonych względem siebie, niż w poprzednim podziale. Przykładem jest skupisko czwarte, gdzie województwo podlaskie dołączyło się do województw: warmińsko-mazurskie, mazowieckie. W podziale na pięć klastrów łączyło się ono z województwami obecnie należącymi do grupy pierwszej, które są najbardziej rozproszone (nie mają wspólnych granic).

Podobnie jak w podziale na pięć grup obliczono wartości odchyleń standardowych dla każdego klastra z pominięciem grupy trzeciej (województwo pomorskie). Otrzymane wyniki zawiera Tabela 3.7. Zauważono w klastrze pierwszym, który powyżej określono jako najbardziej rozproszony, znaczny spadek wartości odchylenia dla zmiennej I50. Na tej podstawie można stwierdzić, iż odłączenie województwa podlaskiego z tej grupy przynosi dobry efekt – tworzy się skupisko o bardziej podobnych do siebie obiektach. Wysoką wartość odchylenia standardowego (0.20352**)** zanotowano w grupie drugiej dla zmiennej I70.9, która opisuje zgon z przyczyny uogólnionej i nieokreślonej miażdżycy. Klaster ten jest również najbardziej liczny oraz odległość na Rysunku 3.4 między województwem łódzkim i małopolskim jest duża, dlatego można podejrzewać modyfikacje w tym skupisku w kolejnych etapach.



Rysunek 3.4 Podział na sześć klastrów metodą k-średnich w postaci punktów na wykresie

Źródło: opracowanie własne z wykorzystaniem programu R

Rysunek 3.5 Podział administracyjny Polski z podziałem województw na sześć klastrów uzyskanych metodą k-średnich

Źródło: opracowanie własne z wykorzystaniem programu R

Tabela 3.7 Odchylenia standardowe poszczególnych zmiennych w każdym z sześciu klastrów uzyskanych metodą k-średnich; nie uwzględniono grupy jednoelementowej

	I	II	IV	V	VI
C76	0.00942	0.00152	0.00176	0.00154	0.00016
I46	0.00025	0.01613	0.00055	0.00716	0.00016
I50	0.10386	0.14123	**0.17351**	0.06218	0.03315
I70.9	0.02610	**0.20352**	0.09081	0.00974	0.08264
R54	0.02974	0.01761	0.02297	0.02895	0.07939
R96	0.02840	0.02360	0.09605	0.01014	0.00534
R98	0.01733	0.00544	0.04591	0.00076	0.04418
R99	0.01728	0.03679	0.00613	**0.00010**	0.00246
Y21	0.00183	0.00171	0.00069	0.00134	0.00128
Y34	0.00887	0.01172	0.01316	0.00207	0.00057

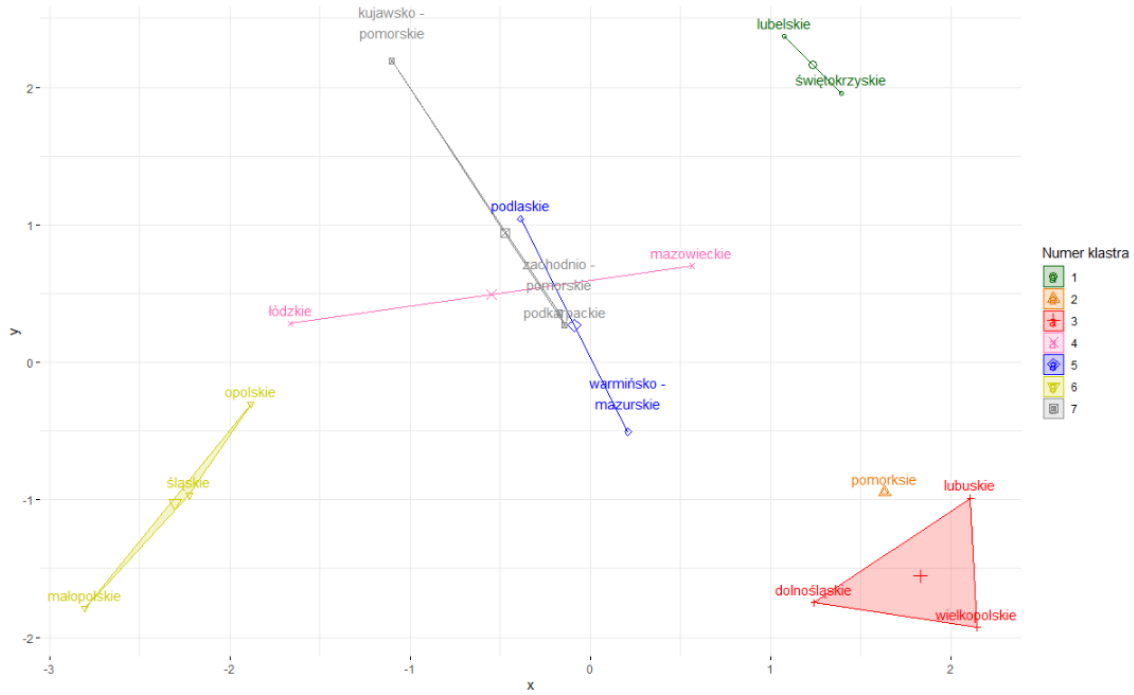
Źródło: opracowanie własne z wykorzystaniem programu R

Podział na siedem grup został uznany za optymalny w wyniku zastosowania indeksu sylwetkowego. Po przeprowadzeniu podziału metodą k-średnich otrzymano klastry zawierające województwa:

1) lubelskie, świętokrzyskie;
2) pomorskie;
3) dolnośląskie, lubuskie, wielkopolskie;
4) łódzkie, mazowieckie;
5) podlaskie, warmińsko-mazurskie;
6) małopolskie, opolskie, śląskie;
7) kujawsko-pomorskie, podkarpackie, zachodniopomorskie.

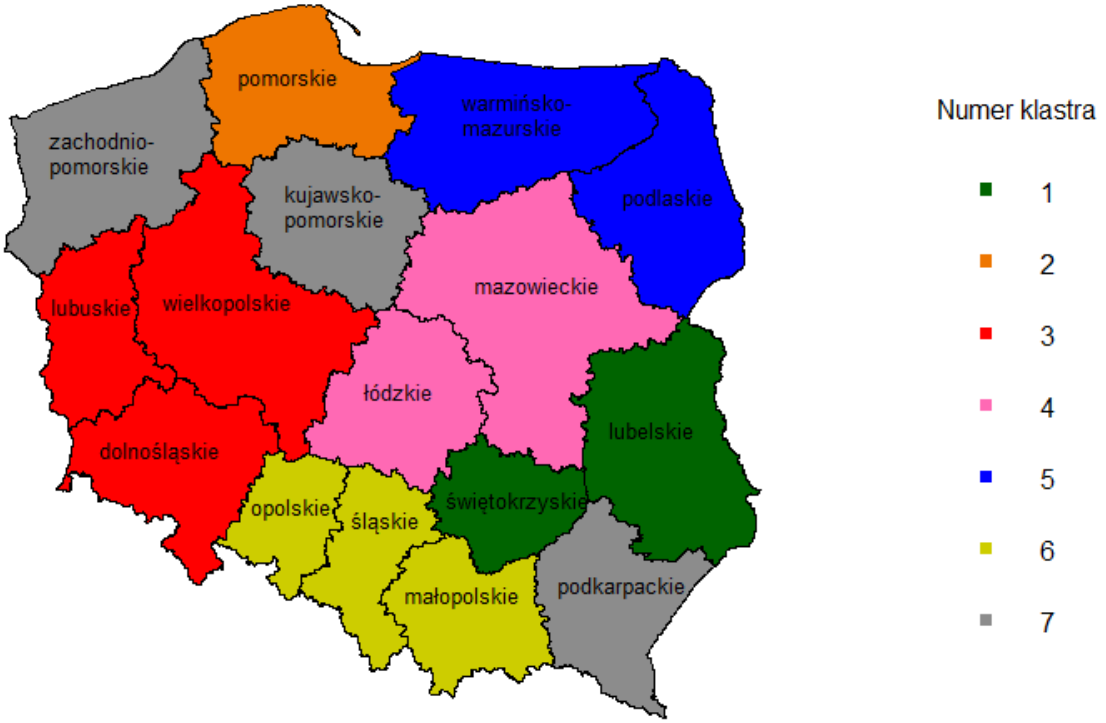
Zwiększenie liczby klas spowodowało powstanie w większości skupisk liczących dwa lub trzy województwa. Klastry znajdujące się z lewej strony Rysunku 3.6 zarówno w górnym jak i dolnym rogu są znacznie oddalone od pozostałych obiektów. Grupa zawierająca województwa: lubelskie oraz świętokrzyskie, według wcześniejszych przewidywań nie uległa zmianie.

Na podstawie Rysunku 3.7 zauważono widoczną zmianę w środkowej części Polski. Powstało nowe skupisko w wyniku odłączenia się województw: łódzkie oraz mazowieckie. Stworzyły one odrębną klasę. Potwierdziły się podejrzenia o modyfikacji klastra zwierającego w podziale na sześć grup województwa: łódzkie, małopolskie, opolskie, śląskie – województwo łódzkie, zostało zakwalifikowane do innego skupiska.



Rysunek 3.6 Podział na siedem klastrów metodą k-średnich w postaci punktów na wykresie

Źródło: opracowanie własne z wykorzystaniem programu R



Rysunek 3.7 Podział administracyjny Polski z podziałem województw na siedem klastrów uzyskanych metodą k-średnich

Źródło: opracowanie własne z wykorzystaniem programu R

Po podziale województw na siedem klastrów, zauważono że odchylenia standardowe dla wszystkich grup są niższe od 0.2. W wyniku odłączenia województwa łódzkiego, w klastrze składającym się z województw: małopolskie, opolskie oraz śląskie zmniejszyła się wartość odchylenia standardowego dla zmiennej I70.9 oraz wartości odchyleń standardowych dla atrybutu I50 nie są tak wysokie jak w poprzednich podziałach.

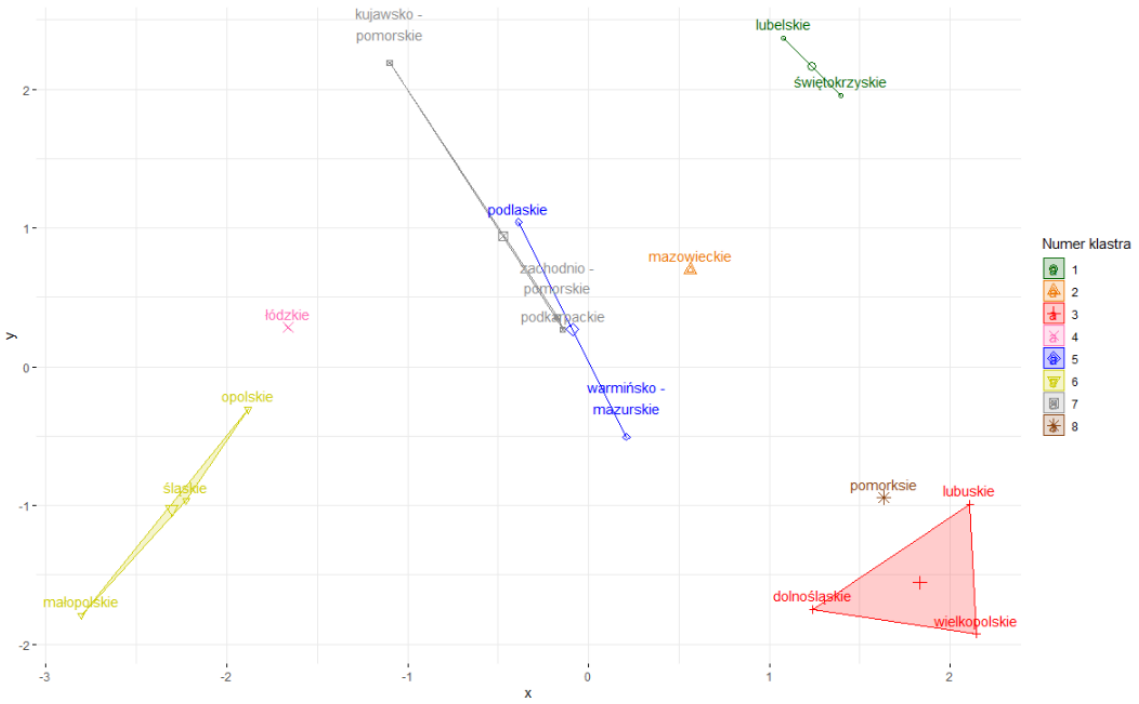
Tabela 3.8 Odchylenia standardowe poszczególnych zmiennych w każdym z siedmiu klastrów uzyskanych metodą k-średnich; nie uwzględniono grup jednoelementowych

	I	**III**	**IV**	**V**	**VI**	**VII**
C76	0.00154	0.00016	0.00304	0.00007	0.00078	0.00942
I46	0.00716	0.00016	0.00003	0.00076	0.01698	0.00025
I50	0.06218	0.03315	0.01343	**0.19694**	0.09797	0.10386
I70.9	0.00974	0.08264	0.05581	0.11863	**0.18534**	0.02610
R54	0.02895	0.07939	0.08836	0.02951	0.01555	0.02974
R96	0.01014	0.00534	0.04562	0.12528	0.02082	0.02840
R98	0.00076	0.04418	0.06070	0.02252	0.00664	0.01733
R99	**0.00010**	0.00246	0.12501	0.00535	0.03274	0.01728
Y21	0.00134	0.00128	0.00103	**0.00012**	0.00180	0.00183
Y34	0.00207	0.00057	0.00023	0.00086	0.00034	0.00887

Źródło: opracowanie własne z wykorzystaniem programu R

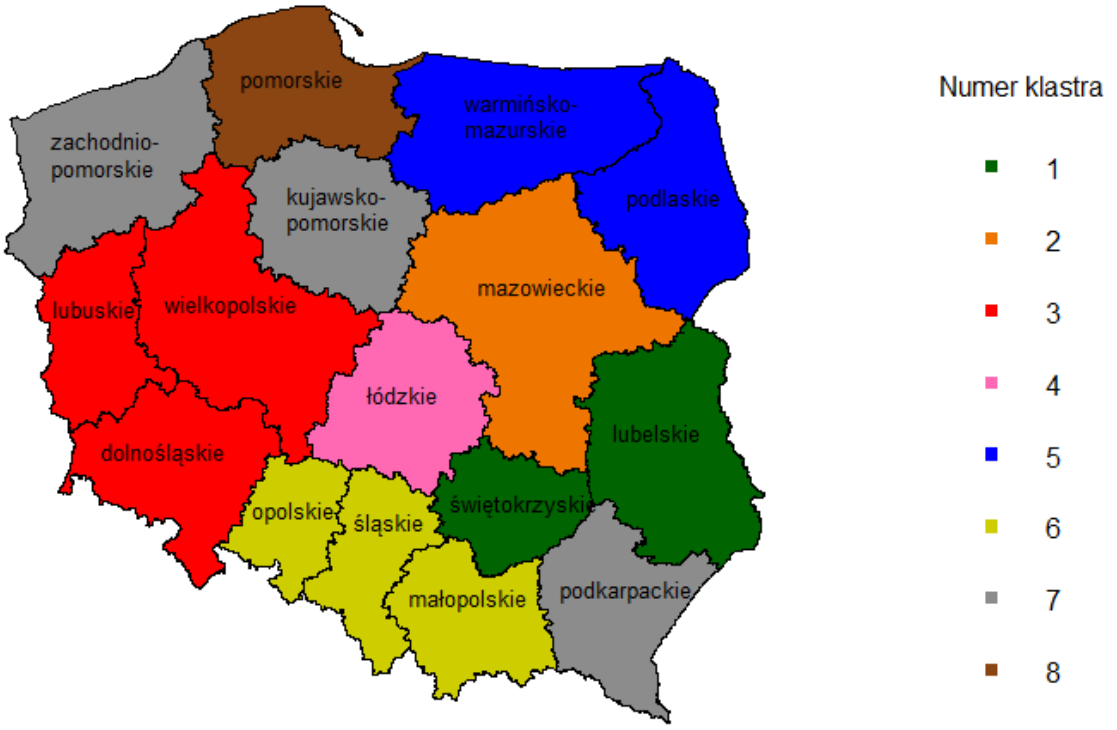
Na skutek podziału na osiem klastrów metodą k-średnich powstały grupy składające się z województw:

1) lubelskie, świętokrzyskie;
2) mazowieckie;
3) dolnośląskie, lubuskie, wielkopolskie;
4) łódzkie;
5) podlaskie, warmińsko-mazurskie;
6) małopolskie, opolskie, śląskie;
7) kujawsko-pomorskie, podkarpackie, zachodniopomorskie;
8) pomorskie.



Rysunek 3.8 Podział na osiem klastrów metodą k-średnich w postaci punktów na wykresie

Źródło: opracowanie własne z wykorzystaniem programu R



Rysunek 3.9 Podział administracyjny Polski z podziałem województw na osiem klastrów uzyskanych metodą k-średnich

Źródło: opracowanie własne z wykorzystaniem programu R

Po klasyfikacji badanych obiektów na osiem klastrów zauważono, iż powstały trzy skupiska jednoelementowe. Rozpadowi uległo skupisko składające się z województw: łódzkie oraz mazowieckie. Na Rysunku 3.8 można zaobserwować, że są to obiekty oddalone od siebie w większym stopniu niż pozostałe. Pozostałe grupy nie uległy zmianom. Na podstawie Rysunku 3.9, który przedstawia podział administracyjny Polski z oznaczonymi skupiskami, zanotowano, że nowopowstałe klastry znajdują się w centralnej części kraju. Wartości odchyleń standardowych są takie same jak w przypadku podziału województw na siedem grup, z wyjątkiem dwóch nowych skupisk jednoelementowych (Tabela 3.9).

Tabela 3.9 Odchylenia standardowe poszczególnych zmiennych w każdym z ośmiu klastrów uzyskanych metodą k-średnich; nie uwzględniono grup jednoelementowych

	I	**III**	**V**	**VI**	**VII**
C76	0.00154	0.00016	0.00007	0.00078	0.00942
I46	0.00716	0.00016	0.00076	0.01698	0.00025
I50	0.06218	0.03315	**0.19694**	0.09797	0.10386
I70.9	0.00974	0.08264	0.11863	**0.18534**	0.02610
R54	0.02895	0.07939	0.02951	0.01555	0.02974
R96	0.01014	0.00534	0.12528	0.02082	0.02840
R98	0.00076	0.04418	0.02252	0.00664	0.01733
R99	**0.00010**	0.00246	0.00535	0.03274	0.01728
Y21	0.00134	0.00128	**0.00012**	0.00180	0.00183
Y34	0.00207	0.00057	0.00086	0.00034	0.00887

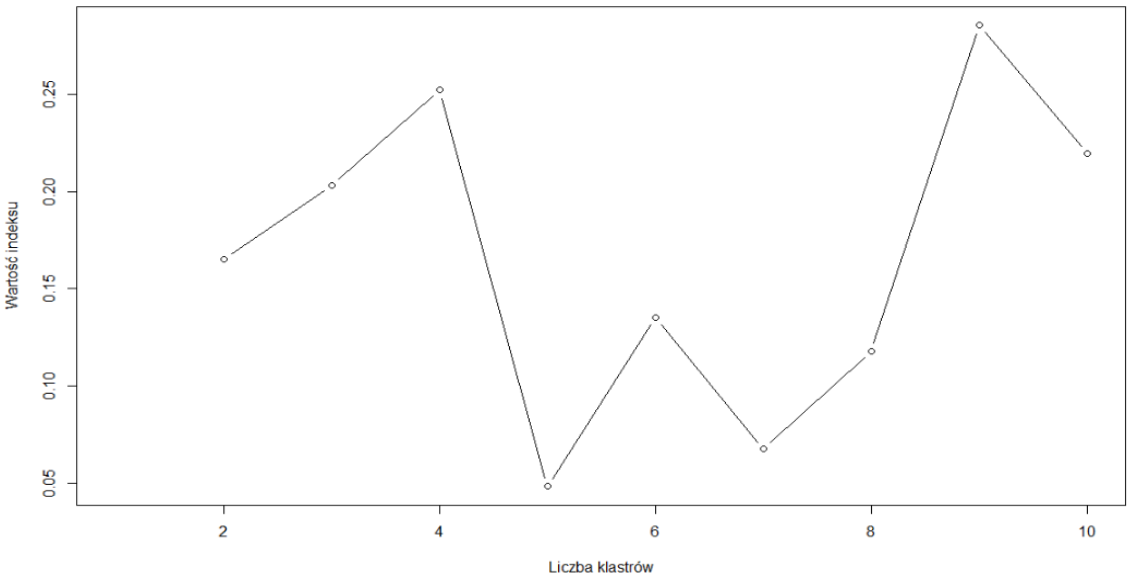
Źródło: opracowanie własne z wykorzystaniem programu R

Podsumowując powyższe obserwacje oraz wyniki otrzymane po poddaniu danych procesowi klasyfikacji metodą k-średnich na pięć, sześć, siedem i osiem klastrów, można zauważyć zmniejszające się wartości odchyleń standardowych w kolejnych podziałach o większej liczebności skupisk. W przypadku analizy klasteryzacji na osiem klas, wystąpiły aż trzy grupy jednoelementowe. Dwie z nich tworzyły jeden klaster w podziale na siedem skupisk. Nie miał on dużych wartości odchyleń standardowych, więc uznano, że siedem skupisk jest maksymalną wielkością podziału. Podczas klasyfikacji metodą k-średnich na pięć oraz sześć klastrów, występowały większe wartości odchyleń standardowych dla zmiennych I50 oraz I70.9 w porównaniu do innych wartości. W przypadku podziału na siedem grup, większe wartości zostały zniwelowane. Taką liczbę klastrów wskazywał

również indeks sylwetkowy. Na podstawie wszystkich rozważań, podjęto decyzję o finalnym podziale województw za pomocą metody k-średnich na siedem klastrów.

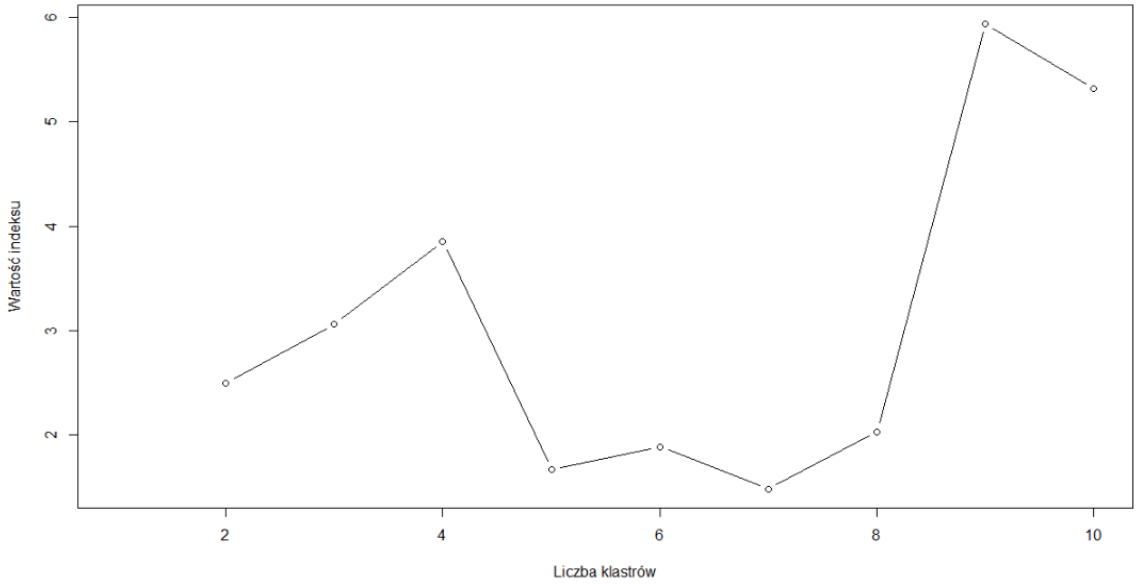
3.3. Klasyfikacja spektralna województw

Klasyfikacja spektralna, podobnie jak klasyfikacja za pomocą metody k-średnich, wymaga określenia na początku liczby klastrów, na jaką zostaną podzielone obiekty – województwa. W tym celu zostały użyte trzy indeksy: sylwetkowy, Calińskiego i Harabasza oraz Daviesa i Bouldina. Pierwszy z nich przedstawia Rysunek 3.10. Największą wartość indeksu zanotowano przy podziale województw na dziewięć klastrów. Wysoką wartości indeksu sylwetkowego można zauważyć także przy występowaniu czterech grup.



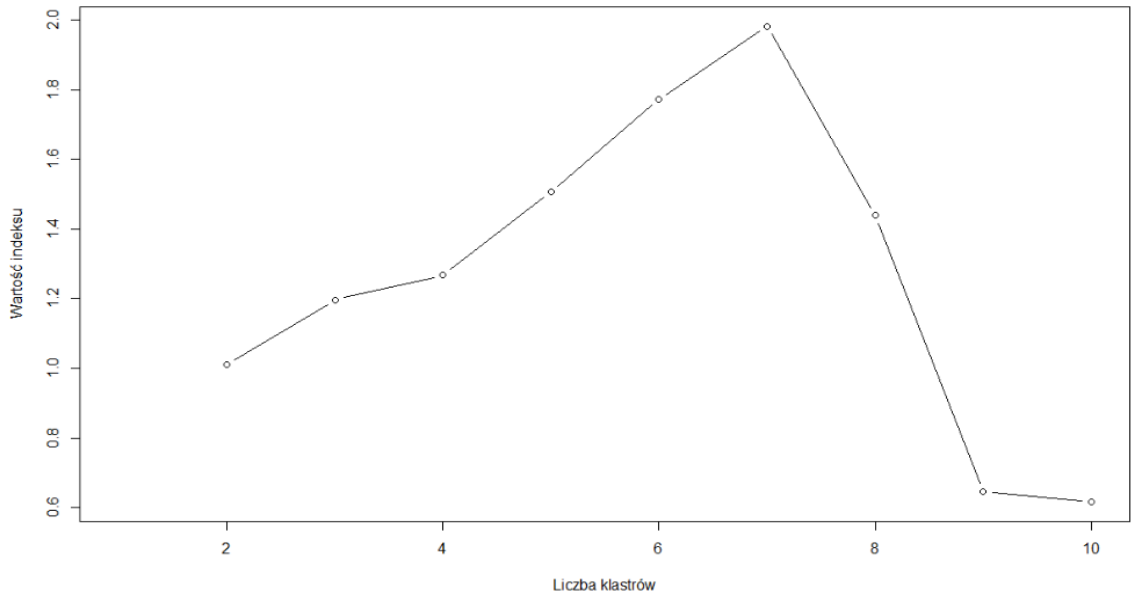
Rysunek 3.10 Indeks sylwetkowy dla klasyfikacji spektralnej
Źródło: opracowanie własne z wykorzystaniem programu R

Otrzymany wykres po obliczeniu indeksu Calińskiego i Harabasza (Rysunek 3.11) przyjmuje największe wartości przy podziale województw na dziewięć, dziesięć oraz cztery klastry. Liczba grup dziewięć oraz cztery pokrywają się z wynikami uzyskanymi na podstawie indeksu sylwetkowego.



Rysunek 3.11 Indeks Calińskiego i Harabasza dla klasyfikacji spektralnej

Źródło: opracowanie własne z wykorzystaniem programu R



Rysunek 3.12 Indeks Daviesa i Bouldina dla klasyfikacji spektralnej

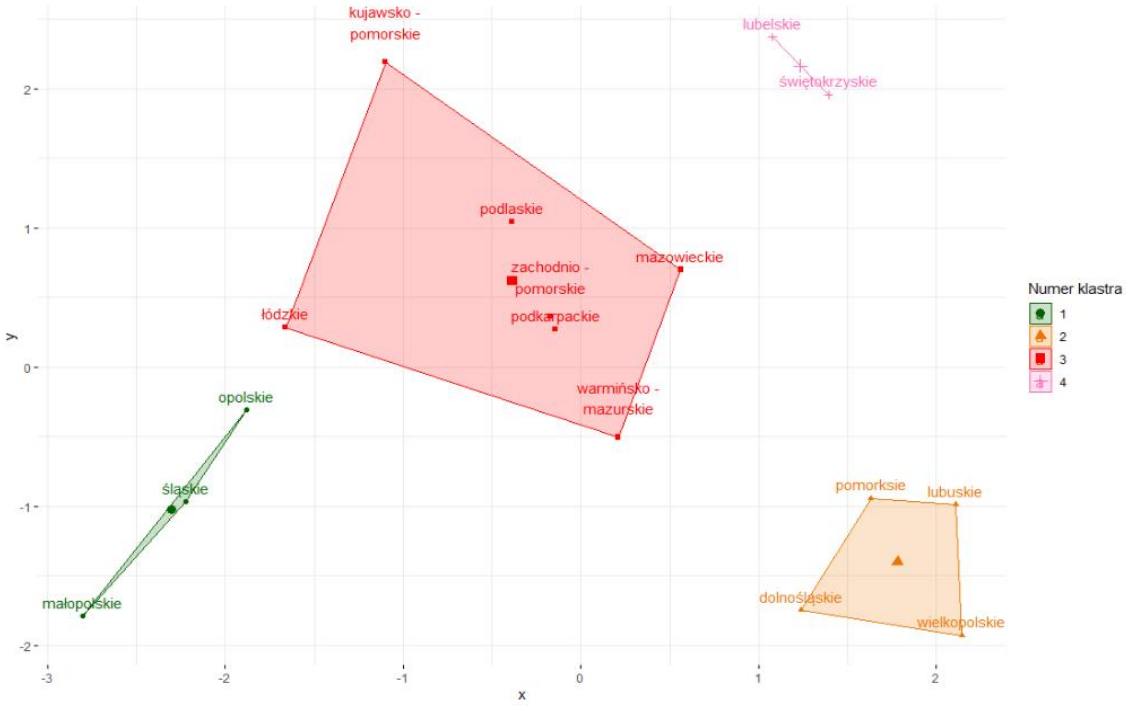
Źródło: opracowanie własne z wykorzystaniem programu R

Indeks Daviesa i Bouldina w przeciwieństwie do dwóch pozostałych indeksów, najlepszą liczbę grup wskazuje, gdy przyjmuje minimum. Rysunek 3.12 przedstawia wykres otrzymany na podstawie wartości otrzymanych dla indeksu Daviesa i Bouldina. Dwie najmniejsze wartości przyjmuje przy podziale województw na dziesięć oraz dziewięć klastrów.

Po przeanalizowaniu otrzymanych wykresów indeksów, zdecydowano o przeprowadzeniu analizy merytorycznej podziału województw na cztery oraz dziewięć klas. W przypadku podziału na cztery skupiska uzyskano klastry składające się z województw:

1) małopolskie, opolskie, śląskie;
2) dolnośląskie, lubelskie, pomorskie, wielkopolskie;
3) kujawsko-pomorskie, łódzkie, mazowieckie, podkarpackie, podlaskie, warmińsko-mazurskie, zachodniopomorskie;
4) lubelskie, świętokrzyskie.

Rysunek 3.13 przedstawia geometryczne odwzorowanie czterech skupisk. Można zauważyć, że klasy są odizolowane od siebie, natomiast obiekty tworzące poszczególne grupy znajdują się blisko siebie. Klaster trzeci (kolor czerwony) składa się z największej liczby województw. Zanotowano, że dwa województwa w tej grupie: kujawsko-pomorskie oraz łódzkie, są najbardziej oddalone od pozostałych obiektów klastra trzeciego.



Rysunek 3.13 Podział na cztery klastry klasyfikacją spektralną w postaci punktów na wykresie
Źródło: opracowanie własne z wykorzystaniem programu R

Rysunek 3.14 Podział administracyjny Polski z podziałem województw na dziewięć klastrów uzyskanych klasyfikacją spektralną

Źródło: opracowanie własne z wykorzystaniem programu R

Na podstawie podziału administracyjnego Polski z uwzględnieniem podziału województw na cztery klastry przy zastosowaniu klasyfikacji spektralnej (Rysunek 3.14) można zauważyć, że obiekty pierwszej grupy znajdują się na południu kraju, natomiast drugiej grupy na zachodzie. Klaster trzeci tworzą województwa z różnych części Polski. Podobnie jak w przypadku podziału metoda k-średnich województwa: lubelskie oraz świętokrzyskie twarzą jedno skupisko.

Tabela 3.10 przedstawia wartości odchyleń standardowych dla każdej zmiennej w poszczególnej grupie. Największą wartość zanotowano w klastrze pierwszym dla atrybutu I70.9 określającego przyczynę zgonu jako uogólniona i nieokreślona miażdżyca. Wartość ta wynosi 0.1853. Wysokie odchylenie standardowe zauważono także w klasie trzeciej – 0.1543, dla zmiennej I50 (niewydolność serca). Większość pozostałych wartości odchyleń standardowych jest niska.

Tabela 3.10 Odchylenia standardowe poszczególnych zmiennych w każdym z dziewięciu klastrów uzyskanych klasyfikacją spektralną

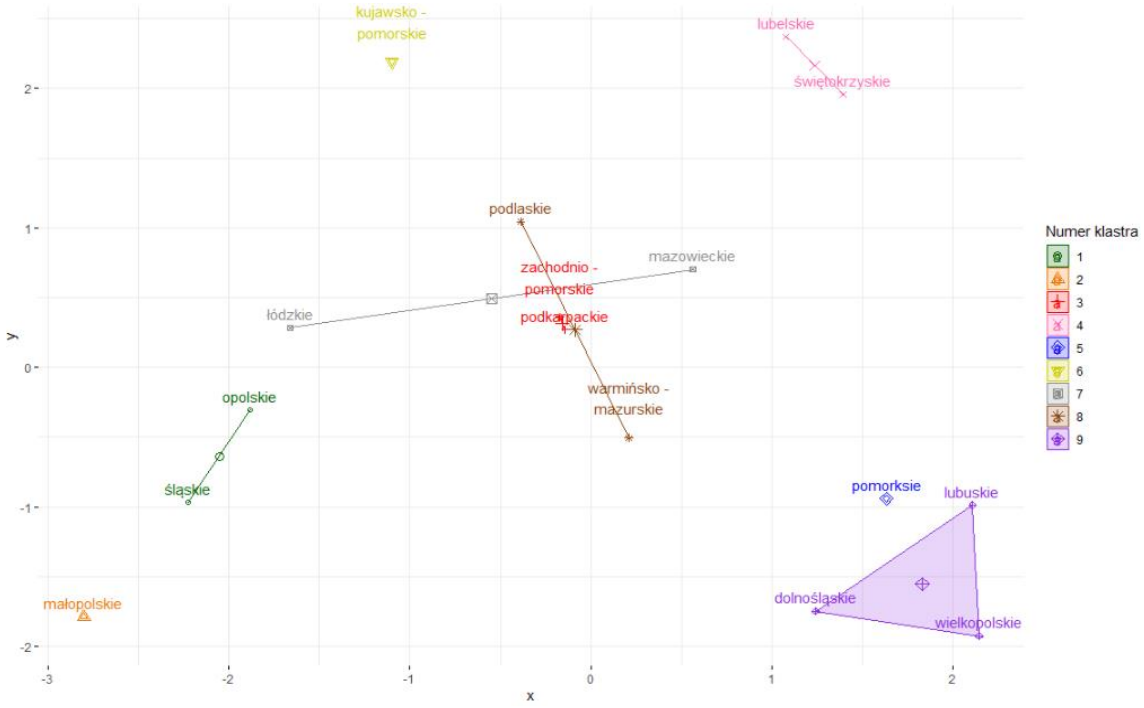
	I	II	III	IV
C76	0.0008	**0.0001**	0.0077	0.0015
I46	0.0170	0.0318	0.0004	0.0072
I50	0.0980	0.1096	**0.1543**	0.0622
I70.9	**0.1853**	0.0896	0.0595	0.0097
R54	0.0156	0.1027	0.0786	0.0290
R96	0.0208	0.0044	0.0696	0.0101
R98	0.0066	0.0946	0.0312	0.0008
R99	0.0327	0.0062	0.0655	**0.0001**
Y21	0.0018	0.0012	0.0022	0.0013
Y34	0.0003	0.0007	0.0113	0.0021

Źródło: opracowanie własne z wykorzystaniem programu R

Podział województw na dziewięć klas został wskazany za najlepszy przez wszystkie trzy zastosowane indeksy. Po klasyfikacji spektralnej na dziewięć grup powstały skupiska złożone z małej liczby obiektów (trzy jednoelementowe). W wyniku podziału uzyskano klastry składające się z województw:

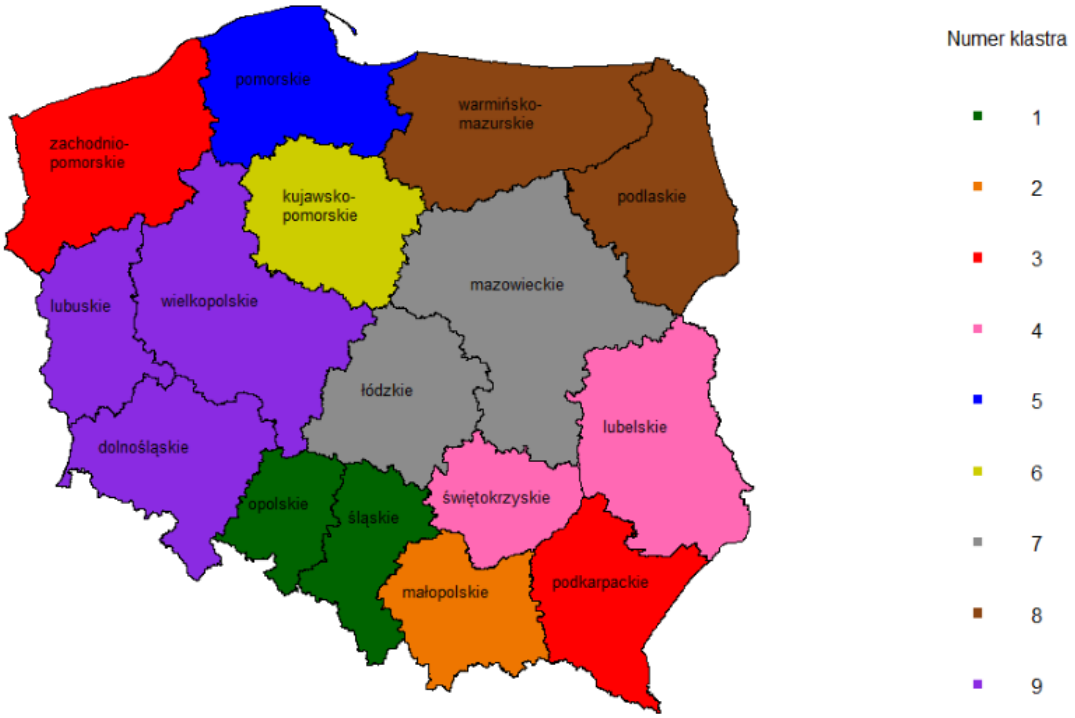
1) śląskie, opolskie;
2) małopolskie;
3) podkarpackie, zachodniopomorskie;
4) lubelskie, świętokrzyskie;
5) pomorskie;
6) kujawsko-pomorskie;
7) łódzkie, mazowieckie;
8) podlaskie, warmińsko-mazurskie;
9) dolnośląskie, lubuskie, wielkopolskie.

Rysunek 3.15 przedstawia geometryczny układ uzyskanych dziewięciu klastrów po klasyfikacji spektralnej województw. Większość skupisk jest od siebie odległa. W centralnej części wykresu grupy trzecia, siódma i ósma pokrywają się ze sobą. Przyczyną tego zjawiska jest dwuwymiarowość wykresu. Obiekty opisane są za pomocą dziesięciu zmiennych czego nie można dokładnie zobrazować na wykresie dwuwymiarowym. Grupy w centralnej części w rzeczywistości są od siebie oddalone.



Rysunek 3.15 Podział na dziewięć klastrów klasyfikacją spektralną w postaci punktów na wykresie

Źródło: opracowanie własne z wykorzystaniem programu R



Rysunek 3.16 Podział administracyjny Polski z podziałem województw na dziewięć klastrów uzyskanych klasyfikacją spektralną

Źródło: opracowanie własne z wykorzystaniem programu R

Na podstawie Rysunku 3.16 przedstawiającym administracyjny podział Polski z podziałem województw na dziewięć grup można zauważyć, że w znacznej większości skupiska tworzą województwa sąsiadujące ze sobą. Wyjątkiem jest klasa składająca się z województw: podkarpackie, zachodniopomorskie. Znajdują się one na przeciwległych końcach kraju.

Tabela 3.11 przedstawia wartości odchyleń standardowych dla każdej zmiennej w poszczególnych grupach. Największe wartości zanotowano w klastrze ósmym dla zmiennych I50 (niewydolność serca) oraz R96 (śmierć natychmiastowa). Wynoszą one odpowiednio: 0.1969 oraz 0.1253. Wyższą wartość odchylenia standardowego zauważono także w klasie siódmej dla zmiennej R99 (inne, niedokładnie określone i nieznane przyczyny umieralności). Jest ona równa 0.1250. Pozostałe wartości odchyleń standardowych są niskie, co świadczy o podobieństwie obiektów poszczególnych klasach. W grupie pierwszej po odłączaniu województwa małopolskiego od województw: opolskie i śląskie, wartość odchylenia standardowego dla zmiennej I70.9 znacznie się zmniejszyła. W podziale na cztery klastry była to największa wartość. Jest to pożądane zjawisko.

Tabela 3.11 Odchylenia standardowe poszczególnych zmiennych w każdym z dziewięciu klastrów uzyskanych klasyfikacją spektralną; nie uwzględniono grup jednoelementowych

	I	**III**	**IV**	**VII**	**VIII**	**IX**
C76	0.0010	0.0072	0.0015	0.0030	**0.0001**	0.0002
I46	0.0144	0.0004	0.0072	0.0000	0.0008	0.0002
I50	0.0370	0.0845	0.0622	0.0134	**0.1969**	0.0332
I70.9	0.0637	0.0367	0.0097	0.0558	0.1186	0.0826
R54	0.0167	0.0382	0.0290	0.0884	0.0295	0.0794
R96	0.0028	0.0243	0.0101	0.0456	**0.1253**	0.0053
R98	0.0051	0.0227	0.0008	0.0607	0.0225	0.0442
R99	0.0022	0.0055	0.0001	**0.1250**	0.0054	0.0025
Y21	0.0012	0.0016	0.0013	0.0010	**0.0001**	0.0013
Y34	0.0005	0.0002	0.0021	0.0002	0.0009	0.0006

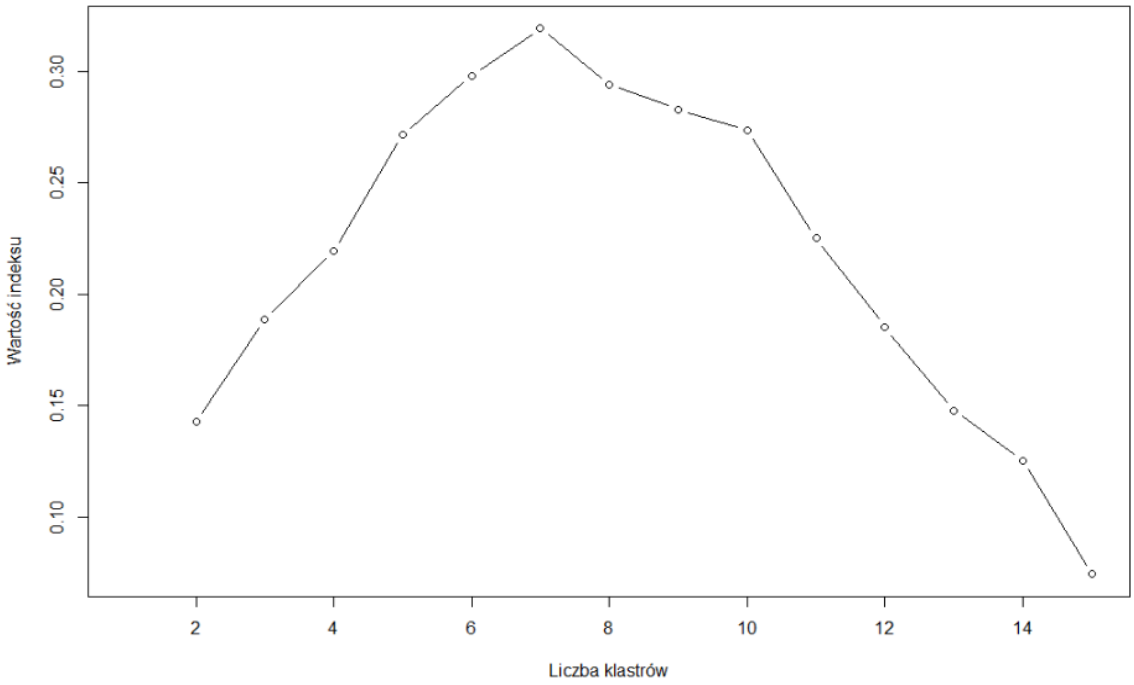
Źródło: opracowanie własne z wykorzystaniem programu R

Podsumowując wyniki otrzymane w procesie klasyfikacji spektralnej na cztery oraz dziewięć klastrów podział na dziewięć klas uznano za lepszy, pomimo wielu skupisk składających się z małej ilości obiektów. Największe wartości odchyleń standardowych znajdują się w jednej z grup. Pozostałe skupiska zawierają województwa o mało

różniących się cechach. W przypadku podziału na cztery grupy w każdym klastrze występowała wartość odchylenia standardowego większa od 0.1.

3.4. Klasyfikacja województw metodą Ward'a

W celu ustalenia liczby klastrów przy zastosowaniu metody Ward'a zostały obliczone wartości indeksu sylwetkowego. Otrzymany wykres (Rysunek 3.17) ma podobny kształt do wykresu indeksu sylwetkowego w przypadku użycia metody k-średnich (Rysunek 3.1). Na rysunku 3.17 zauważono, że maksymalna wartość indeksu sylwetkowego występuje podczas podziału województw na siedem klastrów, zatem jest to optymalna liczba grup. W kolejnej części pracy został przeanalizowany dendrogram utworzony za pomocą metody Ward'a.



Rysunek 3.17 Indeks sylwetkowy dla metody Ward'a

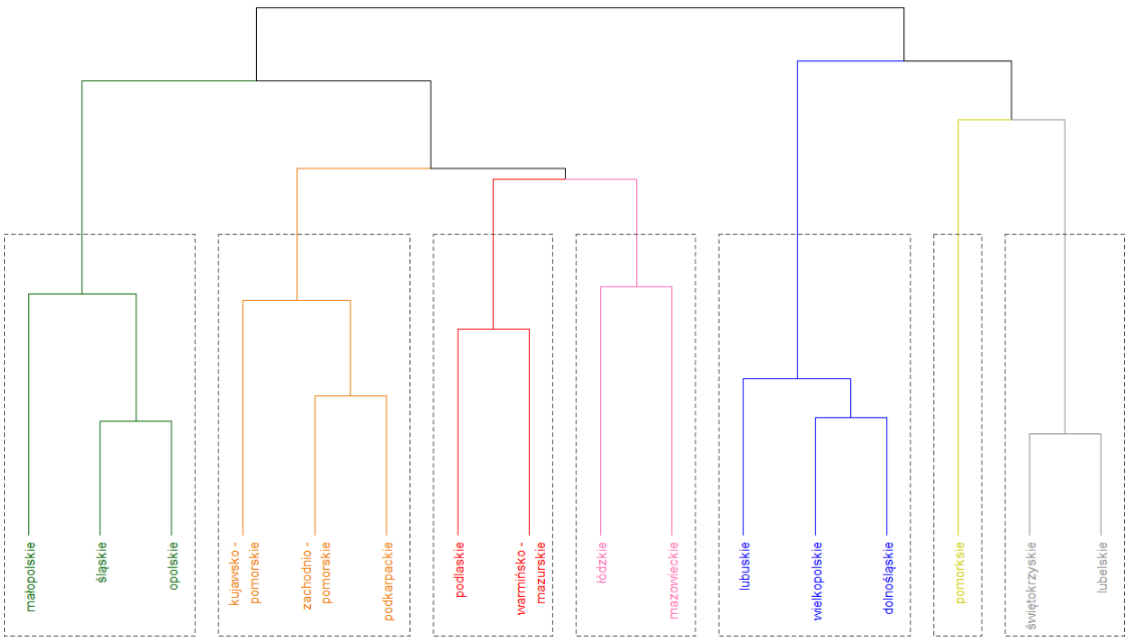
Źródło: opracowanie własne z wykorzystaniem programu R

Po przeprowadzeniu klasyfikacji województw za pomocą metody Ward'a otrzymano dendrogram, który jest zilustrowany na Rysunku 3.18. Został na nim zaznaczony podział województw na siedem grup (optymalna liczba klas według indeksu sylwetkowego). Układ ten został uznany za najlepszy, ponieważ w punkcie przecięcia dendrogramu, linie pionowe są długie, zatem odległości między grupami są duże. Jeśli

podział nastąpił by na mniejszą liczbę klastrów, składały by się one z obiektów mniej podobnych do siebie (krótkie lnie pionowe). W przypadku podziału na większa liczbę klas, na podstawie Rysunku 3.18 można zauważyć, że powstawałoby wiele skupisk jednoelementowych. Przy podziale na siedem klas powstały klastry składające się z województw:

1) małopolskie, opolskie, śląskie;
2) kujawsko-pomorskie, podkarpackie, zachodniopomorskie;
3) podlaskie, warmińsko-mazurskie;
4) łódzkie, mazowieckie;
5) dolnośląskie, lubuskie, wielkopolskie;
6) pomorskie;
7) lubelskie, świętokrzyskie.

Takie same grupy uzyskano podczas podziału województw metodą k-średnich. Mają one zatem takie same wartości odchyleń standardowych jakie są przedstawione w Tabeli 3.8. Wartości te nie są wysokie oraz zostały uznane za lepsze, w porównaniu z wartościami odchyleń standardowych przy podziale na inną liczbę klas. Geograficzna prezentacja na mapie z podziałem administracyjnym przedstawia Rysunek 3.19. Większość województw tworzących poszczególne grupy są sąsiadujące. Układ klas jest identyczny jak w przypadku zastosowania metody k-średnich dla siedmiu klastrów.



Rysunek 3.18 Dendrogram uzyskany metodą Ward'a

Źródło: opracowanie własne z wykorzystaniem programu R

Rysunek 3.19 Podział administracyjny Polski z podziałem województw na siedem klastrów uzyskanych metodą Ward'a

Źródło: opracowanie własne z wykorzystaniem programu R

3.5. Interpretacja

Po przeprowadzeniu klasyfikacji województw za pomocą metody k-średnich, klasyfikacji spektralnej oraz metody Ward'a otrzymano dwa typy podziału. Klastry powstałe na skutek użycia metody optymalizacyjnej oraz hierarchicznej są takie same (Tabela 3.12). Jest ich siedem. W przypadku klasyfikacji spektralnej powstało dziewięć grup. W większości są one takie same jak w klastrach powstałych przy użyciu innych metod. Dwa skupiska dodatkowe w klasyfikacji spektralnej w porównaniu z metodą k-średnich i Ward'a powstały na skutek odłączenia po jednym województwie od innych grup: małopolskiego oraz kujawsko-pomorskiego.

Tabela 3.12 Klastry powstałe w wyniku przeprowadzenia badania

Metoda k-średnich	Metoda Ward'a	Klasyfikacja spektralna
Klaster 1: małopolskie, opolskie, śląskie Klaster 2: kujawsko-pomorskie, podkarpackie, zachodniopomorskie Klaster 3: podlaskie, warmińsko-mazurskie Klaster 4: łódzkie, mazowieckie Klaster 5: dolnośląskie, lubuskie, wielkopolskie Klaster 6: Pomorskie Klaster 7: lubelskie, świętokrzyskie	Klaster 1: małopolskie, opolskie, śląskie Klaster 2: kujawsko-pomorskie, podkarpackie, zachodniopomorskie Klaster 3: podlaskie, warmińsko-mazurskie Klaster 4: łódzkie, mazowieckie Klaster 5: dolnośląskie, lubuskie, wielkopolskie Klaster 6: Pomorskie Klaster 7: lubelskie, świętokrzyskie	Klaster 1: opolskie, śląskie Klaster 2: podkarpackie, zachodniopomorskie Klaster 3: podlaskie, warmińsko-mazurskie Klaster 4: łódzkie, mazowieckie Klaster 5: dolnośląskie, lubuskie, wielkopolskie Klaster 6: Pomorskie Klaster 7: lubelskie, świętokrzyskie Klaster 8: małopolskie Klaster 9: kujawsko-pomorskie

Źródło: opracowanie własne

Średnie wartości zmiennych w klastrach powstałych przy użyciu metody k-średnich i metody Ward'a przedstawione zostały w tabeli 3.13. W klastrze pierwszym złożonym z województw: małopolskie, opolskie oraz śląskie zauważono znacznie wyższą średnią wartość dla zmiennej I70.9, która określa zgon z przyczyny uogólnionej i nieokreślonej miażdżycy – niemal 50%. W pozostałych skupiskach atrybut ten przyjmuje o wiele mniejsze wartości. Jest to jednocześnie największa średnia wartość w klasie drugiej. Najwyższą wartość na tle pozostałych klas zauważono także dla zmiennej określającej zgony jako niedokładnie określone i nieznane przyczyny umieralności (R99).

Wynosi ona ponad 14%. Klaster ten wyróżnia także znacznie mniejszy procent zgonów spowodowanych niewydolnością serca (I50) – niecałe 12%.

Grupę drugą, którą tworzą województwa: kujawsko-pomorskie, podkarpackie, zachodniopomorskie charakteryzuje najwyższy w porównaniu do pozostałych odsetek zgonów spowodowanych utonięciem, zanurzeniem o nieokreślonym zamiarze (Y21). Wynosi on ponad 0.41%. W pozostałych klasach są to wartości poniżej 0.1%.

Województwa: podlaskie, warmińsko-mazurskie tworzą grupę trzecią. Wyróżnia ją najwyższy odsetek zgonów kodowanych jako śmierć natychmiastowa (R96) w porównaniu z pozostałymi skupieniami. Wynosi on ponad 18%. Zauważono także najmniejszą wartość dla zgonów spowodowanych nowotworami złośliwymi umiejscowień innych i niedokładnie określonych (C76).

Do klasy czwartej zaliczone są województwa: łódzkie, mazowieckie. Charakteryzuje ją największy odsetek (ponad 2%) zgonów, których przyczynę określono jako nieokreślone zdarzenie o nieokreślonym zamiarze (Y34). W pozostałych klastrach są to wartości mniejsze od 1%. Pozostałe średnie nie wyróżniają się na tle innych klas.

W klastrze piątym, a zatem w województwach: dolnośląskie, lubuskie, wielkopolskie występuje w znacznym stopniu większy, w porównaniu do pozostałych grup, odsetek zgonów z przyczyny określonej jako śmierć nieoczekiwana (R98). Jest to ponad 17%. W przypadku innych klastrów wartości są poniżej 5%. Województwa te wyróżniają się także najmniejszymi średnimi wartościami dla zmiennych określających przyczyny zgonów ze względu na: zatrzymanie krążenia (I46) – 0.00009 oraz inne, niedokładnie określone i nieznane przyczyny (R99) – 0.00319. W pozostałych grupach przyczyny te występują częściej.

Grupa szósta jednoelementowa, którą tworzy województwo pomorskie znacznie wyróżnia się na tle pozostałych klas. Przyjmuje największe wartości dla zmiennych: I46 (zatrzymanie krążenia – ponad 6%) oraz I50 (niewydolność serca – ponad 67%) w stosunku do wartości tych atrybutów w pozostałych klastrach. Najmniejszą liczbę zgonów zanotowano w tym województwie z przyczyn: uogólnionej i nieokreślonej miażdżycy (I70.9 – niecałe 2%), starość (R54 – ponad 1.5%) oraz co jest zaskakujące, z powodu, iż jest to województwo pomorskie, nie występuje tam ani jeden przypadek utonięcia, zanurzenia o nieokreślonym zamiarze (Y21).

Klaster siódmy tworzony przez dwa województwa: lubelskie, świętokrzyskie wyróżnia się na tle pozostałych klas najczęściej występującymi przyczynami zgonów z przyczyn: nowotworów złośliwych umiejscowień innych i niedokładnie określonych

(C76 – prawie 3%) oraz starość (R54 – niemal 30%). Województwa te charakteryzują się także małymi wartościami w przypadku śmierci nieoczekiwanej w porównaniu z innymi grupami.

Tabela 3.13 Średnie wartości zmiennych dla poszczególnych klastrów uzyskanych metodą k-średnich oraz metodą Ward'a

	I	II	III	IV	V	VI	VII
C76	0.00491	0.01205	0.00032	0.00552	0.00080	0.00074	0.02845
I46	0.01676	0.00026	0.00054	0.00028	0.00009	0.06370	0.02186
I50	0.11852	0.41233	0.15302	0.34180	0.45868	0.67111	0.45275
I70.9	0.49710	0.23197	0.20624	0.18546	0.13704	0.01926	0.03160
R54	0.11582	0.05732	0.19974	0.15388	0.17478	0.01556	0.29210
R96	0.02578	0.15270	0.18735	0.09078	0.00529	0.00667	0.11586
R98	0.00787	0.02894	0.01816	0.04981	0.17859	0.00370	0.00247
R99	0.14394	0.04891	0.00936	0.10610	0.00319	0.01481	0.04409
Y21	0.00195	0.00411	0.00045	0.00091	0.00124	0.00000	0.00095
Y34	0.00056	0.00529	0.00088	0.02382	0.00042	0.00148	0.00243

Źródło: opracowanie własne z wykorzystaniem programu R

Średnie wartości zmiennych dla uzyskanych klas podczas klasyfikacji spektralnej przedstawia Tabela 3.14. Klastry od trzeciego do siódmego są takie same jak klastry o tych samych numerach w Tabeli 3.13, zatem ich interpretacja jest identyczna z wyjątkiem grupy siódmej, gdzie zmienna R98 nie przejmuje już najmniejszej wartości, lecz nadal jest jedną z najniższych. Zostały więc zinterpretowane klasy: pierwsza, druga, ósma i dziewiąta.

Województwa: opolskie oraz śląskie wchodzące w skład pierwszego klastra. Charakteryzują się największą wartością w porównaniu z innymi klasami zgonami z przyczyn określanych jako inne, niedokładnie określone i nieznane przyczyny umieralności (R99 – ponad 16%).

Drga grupa składająca się z województw: podkarpackie oraz zachodniopomorskie wyróżnia się na tle innych klastrów najmniejszym odsetkiem śmierci z powodu nieokreślonego zdarzenia o nieokreślonym zamiarze (Y34 – 0.0002).

Klaster ósmy tworzy województwo małopolskie. Zanotowano, iż występuje tam znacznie mniejszy odsetek osób umierających na niewydolność serca (I50 – niecały 1%) w stosunku do pozostałych grup. Zaniżał on średnią wartość w grupie pierwszej w Tabeli 3.12. Takie samo zjawisko zauważono w przypadku zmiennych: R96 (śmierć

natychmiastowa) oraz R98 (śmierć nieoczekiwana). Największa liczba przyczyn zgonów występuje w przypadku uogólnionej i nieokreślonej miażdżycy (I70.9 – ponad 70%).

Ostatni klaster, czyli województwo kujawsko-pomorskie wyróżnia się na tle pozostałych klas największym odsetkiem osób umierających z powodu utonięcia, zanurzenia o nieokreślonym zamiarze (Y21 – ponad 0.5%). Od grupy drugiej, z którą województwo to jest połączone w przypadku użycia pozostałych metod, województwo kujawsko-pomorskie odróżnia w szczególności większa wartość zmiennych: C76, R99 oraz Y34.

Tabela 3.14 Średnie wartości zmiennych dla poszczególnych klastrów uzyskanych w wyniku klasyfikacji spektralnej

	I	**II**	**III**	**IV**	**V**	**VI**	**VII**	**VIII**	**IX**
C76	0.0051	0.0075	0.0003	0.0055	0.0008	0.0007	0.0285	0.0046	0.0212
I46	0.0246	0.0003	0.0005	0.0003	0.0001	0.0637	0.0219	0.0011	0.0003
I50	0.1730	0.4614	0.1530	0.3418	0.4587	0.6711	0.4528	0.0095	0.3142
I70.9	0.3933	0.2303	0.2062	0.1855	0.1370	0.0193	0.0316	0.7047	0.2354
R54	0.1100	0.0502	0.1997	0.1539	0.1748	0.0156	0.2921	0.1275	0.0716
R96	0.0377	0.1397	0.1874	0.0908	0.0053	0.0067	0.1159	0.0019	0.1788
R98	0.0111	0.0252	0.0182	0.0498	0.1786	0.0037	0.0025	0.0015	0.0364
R99	0.1628	0.0392	0.0094	0.1061	0.0032	0.0148	0.0441	0.1062	0.0684
Y21	0.0029	0.0033	0.0005	0.0009	0.0012	0.0000	0.0010	0.0001	0.0058
Y34	0.0005	0.0002	0.0009	0.0238	0.0004	0.0015	0.0024	0.0007	0.0155

Źródło: opracowanie własne z wykorzystaniem programu R

Podsumowanie

Na skutek klasyfikacji województw ze względu na przyczyny zgonów znanych jako „garbage codes" za pomocą metod: k-średnich, metody Ward'a oraz klasyfikacji spektralnej uzyskano dwa rodzaje podziałów. Za pomocą użycia dwóch pierwszych metod: optymalizacyjnej i hierarchicznej uzyskano siedem takich samych klastrów. W przypadku klasyfikacji spektralnej otrzymane wyniki się różnią, gdyż jest więcej klas. Cztery z nich pokrywają się z grupami uzyskanymi za pomocą metody k-średnich oraz Ward'a. Podczas procesu zwiększania liczby klastrów w metodach „tradycyjnych", zwiększała się liczba jednoelementowych skupisk. Po podziale na osiem grup metodą k-średnich, zaobserwowano wystąpienie trzech klas zawierających po jednym województwie. Jeśli liczba klastrów zwiększona by była do dziewięciu, byłoby jeszcze więcej skupisk jednoelementowych. Identyczne zjawisko zauważono w przypadku metody Ward'a. Klasyfikacja spektralna wydaje się tutaj lepszym zastosowaniem, gdyż przy większej liczbie skupisk, nie występuje tak duża liczba jednoelementowych klastrów jak w przypadku klasyfikacji metodami „tradycyjnymi".

Po obliczeniu średnich wartości w poszczególnych klasach (Rozdział 3.5 Tabela 3.13 oraz 3.14) potwierdziły się przypuszczenia o wysokim odsetku przyczyn zgonów w wyniku niewydolności serca. Wyjątkiem jest grupa ósma (województwo małopolskie) w Tabeli 3.14, gdzie wartość jest mniejsza od 1%. W pozostałych klastrach zmienna I50 przyjmuje znacznie wyższe średnie wartości. Należy pamiętać, iż są to „kody śmieciowe", a zatem wiele przypadków zgonów z przyczyny niewydolności serca była nieokreślona i za słabo zdiagnozowana. Jedynie województwo małopolskie mające najmniejszy odsetek w porównaniu z pozostałymi klasami jest pod tym względem bardziej skrupulatne i dokładne.

Z drugiej strony w województwie małopolskim występuje znacznie większa wartość średnia w przypadku osób zmarłych z powodu uogólnionej i nieokreślonej miażdżycy. Świadczy to o braku szczegółowej diagnozy osób zmarłych na miażdżycę. Podobne zjawisko można zauważyć w przypadku zmiennej R54, określająca przyczynę zgonu jako starość. W klastrze siódmym, w którego skład wchodzą województwa: lubelskie i świętokrzyskie, odsetek zgonów spowodowanych starością jest znacznie większy (niemal 30%) w porównaniu z pozostałymi skupiskami. Świadczy to

o zaniedbaniu w diagnozie przyczyn umieralności osób starszych w tych województwach.

Zaskakującym wynikiem było okrycie, iż w województwie pomorskim nie występuje ani jeden przypadek utonięcia, zanurzenia o nieokreślonym zamiarze. Biorąc pod uwagę to, że jest to województwo nadmorskie oraz wiele osób udaje się tam na wakacje, można było się spodziewać, że utonięcia mogą występować tam liczniej niż w pozostałych województwach. Jak się okazało przyczyny, na skutek których nastąpiły utonięcia w roku 2018 zostały w pełni poprawnie zdiagnozowane.

Bibliografia

1. ALMEIDA, J. A. S.; BARBOSA, L. M. S.; PAIS, A. A. C. C.; FORMOSINHO, S. J., Improving hierarchical cluster analysis: A new method with outlier detection and automatic clustering. *Chemometrics and Intelligent Laboratory Systems*, 2007, 87.2: 208-217.
2. ARTHUR, D.; MANTHEY, B.; RÖGLIN, H., Smoothed analysis of the k-means method. *Journal of the ACM (JACM)*, 2011, 58.5: 1-31.
3. ĆWIĄKAŁA-MAŁYS, A.; MOŚCIBRODZKA, M., Hierarchiczne procedury aglomeracyjne w badaniu poziomu i struktury kosztów publicznych uczelni akademickich. *Metody ilościowe w badaniach ekonomicznych*, 2014, 15.3: 30-41.
4. DUDEK, A., Klasyfikacja spektralna a tradycyjne metody analizy skupień. Prace *Naukowe Uniwersytetu Ekonomicznego we Wrocławiu. Ekonometria*, 2009, 23.37 Zastosowania metod ilościowych: 21-34.
5. ELMORE, K. L.; RICHMAN, M. B., Euclidean distance as a similarity metric for principal component analysis. *Monthly weather review*, 2001, 129.3: 540-549.
6. GACZEK, W. M.; HEŁPA, M.; KASPRZYK, A., Niehierarchiczna analiza skupień—nowa metoda klasyfikacji zjawisk społeczno-gospodarczych. 1980.
7. GNANADESIKAN, R., *Discriminant analysis and clustering*. National Academies, 1988.
8. HUANG, A., Similarity measures for text document clustering. In: *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand*. 2008. p. 9-56.
9. KORZENIEWSKI, J., *Metody selekcji zmiennych w analizie skupień. Nowe Procedury*. Wydawnictwo Uniwersytetu Łódzkiego, 2012.
10. MIGDAŁ-NAJMAN, K.; NAJMAN, K., Analiza porównawcza wybranych metod analizy skupień w grupowaniu jednostek o złożonej strukturze grupowej. *Zarządzanie i finanse*, 2013, 3.2: 179-194.
11. NAZEER, KA A.; SEBASTIAN, M. P., Improving the Accuracy and Efficiency of the k-means Clustering Algorithm. In: *Proceedings of the world congress on engineering*. London: Association of Engineers, 2009. p. 1-3.

12. NOWAK-BRZEZIŃSKA, A.; WAKULICZ-DEJA, A. Ś., Wybór miary podobieństwa a efektywność grupowania reguł w złożonych bazach wiedzy. *Studia Informatica, Wyd. Pol. Śląskiej*, 2010, 31.2A: 89.
13. PEŁKA, M., et al. Wielomodelowa klasyfikacja spektralna danych symbolicznych. *Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu*, 2017, 468: 180-187.
14. PIETRZYKOWSKI, R.; KOBUS, P., Zastosowanie modyfikacji metody k-średnich w analizie portfelowej. *Ekonomika i Organizacja Gospodarki Żywnościowej*, 2006, 60: 301-308.
15. Praca magisterska: BIEŃSKOWSKI, D., *Grupowanie konsensusowe danych otrzymanych z mikromacierzy*, Politechnika Warszawska, 2014.
16. RASMUSSEN, E. M., Clustering algorithms. *Information retrieval: data structures & algorithms*, 1992, 419: 442.
17. ROSZKO-WÓJTOWICZ, E., Analiza skupień w ocenie warunków pracy w krajach Unii Europejskiej. *Wiadomości Statystyczne*, 2014, 59.11.
18. *Umieralność w Polsce. Najczęstsze przyczyny zgonów*, https://www.politykazdrowotna.com/48211,umieralnosc-w-polsce-najczestsze-przyczyny-zgonow, 12.08.2020
19. WALESIAK, M., Klasyfikacja spektralna a skale pomiaru zmiennych. *Przegląd Statystyczny*, 2012, 1.59: 13-31.
20. WALESIAK, M.; DUDEK, A., *Odległość GDM dla danych porządkowych a klasyfikacja spektralna*. 2009.
21. WALESIAK, M.; GATNAR, E. (ed.)., *Metody statystycznej analizy wielowymiarowej w badaniach marketingowych*. Wydaw. Akademii Ekonomicznej im. Oskara Langego, 2004.
22. WALESIAK, M,; GATNAR, E. (ed.)., *Statystyczna analiza danych z wykorzystaniem programu R*. Wydawnictwo Naukowe PWN, 2009.
23. WASILEWSKI, W.; GAŁĘCKI, J.;STAWIŃSKA-WITOSZYŃSKA, B., *Poradnik szkoleniowy dla lekarzy orzekających o przyczynach zgonów i wystawiających kartę zgonów.* Warszawa, 2019 (aktualizacja).
24. *Zgony według przyczyn określanych jako "garbage codes"*, https://stat.gov.pl/obszary-tematyczne/ludnosc/statystyka-przyczyn-zgonow/zgony-wedlug-przyczyn-okreslanych-jako-garbage-codes,3,3.html, 02.07.2020

Spis tabel i rysunków