Laboratorium IX: Analiza skupień

Spis treści

Laboratorium IX: Analiza skupień 1
Wiadomości ogólne 2
1. Wstęp teoretyczny. 2
1.1. Wprowadzenie. 2
1.2. Metody hierarchiczne analizy skupień. 3
1.3. Grupowanie metodą k-średnich. 5
1.4. Założenia analizy skupień. 6
2. Analiza skupień w STATISTICE 6
2.1 Aglomeracja. 6
Interpretacja dendrogramu 8
2.2 Grupowanie metodą k-średnich 9
Założenia analizy skupień 11
Ćwiczenia 12
Część I 12
Część II 12

Wiadomości ogólne

1. Wstęp teoretyczny.

1.1. Wprowadzenie.

Analiza skupień to zbiór metod wielowymiarowej analizy statystycznej, służących wyodrębnianiu jednorodnych podzbiorów obiektów badanej populacji. Metody analizy skupień są stosowane wówczas, gdy nie dysponujemy hipotezami a priori, a badania są w fazie eksploracyjnej. Znajdywanie grup (skupień) obiektów odbywa się w oparciu o zmienne charakteryzujące analizowane obiekty, dlatego też istotnym elementem analizy skupień jest odpowiedni wybór zmiennych służących do wyodrębniania spójnych grup obiektów. Analiza skupień jest również bardzo wrażliwa na przypadki odstające, dlatego przed rozpoczęciem analizy należy usunąć przypadki odstające oraz zmienne słabo różnicujące badane obiekty.

Dzięki analizie skupień można:

- Wykryć czy otrzymane skupienia wskazują na jakąś prawidłowość (np. związek pomiędzy symptomami a faktycznym stanem chorobowym);
- Dokonać redukcji olbrzymiego zbioru danych do średnich poszczególnych grup;
- Potraktować rozdzielenie na grupy jako wstęp do dalszych wielowymiarowych analiz.

Miary odległości

Odległość $d(O_i, O_j)$ to funkcja niepodobieństwa pary obiektów (O_i, O_j), ponieważ im większa jest odległość pomiędzy dwoma obiektami, tym bardziej są one do siebie niepodobne. W grupowaniu łączone są zatem obiekty leżące blisko siebie, równocześnie będące daleko od innych, tworzących inne skupienie. Funkcja odległości określona jest na parach obiektów i przyjmuje wartości w zbiorze liczb rzeczywistych nieujemnych. Jej postać zależy od skali pomiarowej (zmienne ilościowe, porządkowe czy jakościowe) zmiennych charakteryzujących analizowane obiekty. Najczęściej wykorzystywane funkcje odległości dla skal co najmniej przedziałowych to:

- **Odległość Czebyszewa:** $d(x, y) = \max|x_i - y_i|$
- **Odległość euklidesowa:**

$$d(x, y) = \sqrt{\sum_{i=1}^{p}(x_i - y_i)^2}$$

  - Jest to najczęściej wybierana metryka, jako najbardziej „naturalna”
- **Odległość miejska** (Manhattan):

$$d(x, y) = \sum_{i=1}^{p}|x_i - y_i|$$

  - W tej metryce sfera jest powierzchnią kostki
- **Odległość euklidesowa do kwadratu:**

$$d(x, y) = \sum_{i=1}^{p}(x_i - y_i)^2$$

Aby móc korzystać z wyżej wymienionych metryk, różne zmienne muszą być porównywalne. Dlatego też wskazane jest wstępne standaryzowanie zmiennych. W praktyce najczęściej standaryzuje się zmienne według wzoru: $z_i = \frac{x_i - \bar{x}}{s_x}$, gdzie $\bar{x}$ to średnia, zaś s_x to odchylenie standardowe zmiennej w próbie.

Wybranie odpowiedniej metryki umożliwia utworzenie kwadratowej macierzy odległości. Macierz taka jest symetryczna ($d_{ij} = d_{ji}$) oraz na głównej przekątnej ma zera ($d_{ii} = 0$).

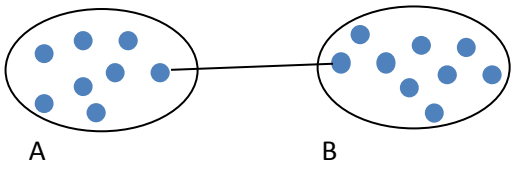
Wśród wielu metod grupowania możemy wyróżnić:

- Metody hierarchiczne
- Grupowanie metodą k-średnich

1.2. Metody hierarchiczne analizy skupień.

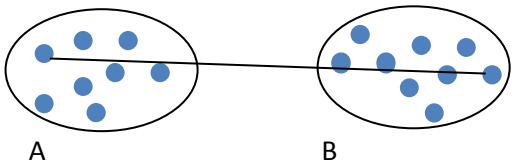
Wśród metod hierarchicznych najczęściej wykorzystywane są techniki aglomeracyjne, w których początkowo każdy obiekt stanowi osobne skupienie, następnie obiekty leżące najbliżej siebie są łączone w nowe skupienie aż do uzyskania jednego skupienia. Problemem jest określenie odległości (czyli zasady wiązania) między nowymi skupieniami, powstającymi z połączonych obiektów. Istnieje szereg różnych zasad wiązania, które między sobą różnią się jedynie sposobami obliczania odległości między skupieniami. W pakiecie STATISTICA można wybrać jedną z następujących metod:

- **Metoda pojedynczego wiązania** (***single linkage metod***), nazywa inaczej **metodą najbliższego sąsiedztwa**. W tej metodzie odległość między dwoma skupieniami określona jest przez odległość między najbliższymi obiektami (sąsiadami) należącymi do różnych skupień. W ten sposób obiekty tworzą skupienia łącząc się w ciągi, a wynikowe skupienia tworzą łańcuchy (rys. 1).



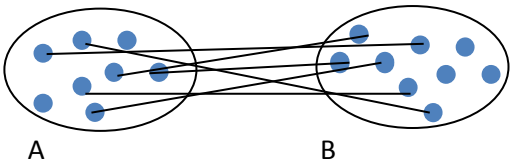
Rys. 1. Metoda najbliższego sąsiedztwa

- **Metoda pełnego wiązania** (***complete linkage***), zwana również **metodą najdalszego sąsiedztwa**. Tutaj odległość między skupieniami określa odległość między najdalszymi sąsiadami, czyli odległość ta jest równa odległości między najdalej położonymi obiektami należącymi do różnych skupień (rys. 2). Stosuje się ją wówczas, gdy obiekty faktycznie formują naturalnie oddzielone „kępki", natomiast metoda ta nie jest odpowiednia w przypadku gdy skupienia są w jakiś sposób wydłużone lub mają charakter łańcucha.



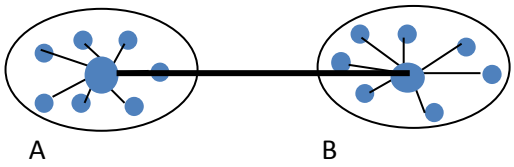
Rys. 2. Metoda najdalszego sąsiedztwa

- **Metoda średnich połączeń** (***UPGMA – unweighted pair-group metod using arithmetic averages***). Odległość między dwoma skupieniami oblicza się za pomocą średniej arytmetycznej wyznaczonej ze wszystkich odległości obiektów należących do dwóch różnych skupień (rys. 3). Można ją stosować zarówno w przypadku skupień wydłużonych, jak i wtedy, gdy obiekty formują naturalne kępki. W tym drugim przypadku metoda ta jest efektywniejsza.



Rys. 3. Metoda średnich połączeń

- **Metoda średnich połączeń ważonych** (***WPGMA – weighted pair-group metod using arithmetic averages***). Jest metodą podobną do UPGMA, z tą różnicą, że tutaj uwzględnione są wielkości skupień (liczby zawartych w nich obiektów). Tą metodę stosuje się wówczas, gdy istnieje podejrzenie wyraźnej różnicy pomiędzy licznościami skupień.
- **Metoda środków ciężkości** (***UPGMC – unweighted pair-group metod using the centroid average***).). Odległość między dwoma skupieniami jest równa odległości pomiędzy ich środkami ciężkości (rys. 4).



Rys. 4. Metoda środków ciężkości

- **Metoda ważonych środków ciężkości** (**mediany**). Na określenie tej metody stosowany jest również skrót ***WPGMC*** (***weighted pair-group metod using the centroid average***). Jest metodą podobną do UPGMC, z tą różnicą, że wprowadza się ważenie uwzględniające różnice między wielkościami skupień. Metoda ta jest lepsza od poprzedniej w przypadku, gdy podejrzewamy istnienie znacznych różnic w rozmiarach skupień.
- **Metoda Warda** – różni się od wszystkich pozostałych, ponieważ do oszacowania odległości między skupieniami wykorzystuje podejście analizy wariancji. Metoda ta zmierza do minimalizacji sumy kwadratów odchyleń wewnątrz skupień. Miarą zróżnicowania skupienia względem wartości średnich jest *ESS* (*Error Sum of Squares*), zwane również błędem sumy kwadratów. ESS jest określone wzorem:

$$ESS = \sum_{i=1}^{k} (x_i - \bar{x})^2$$

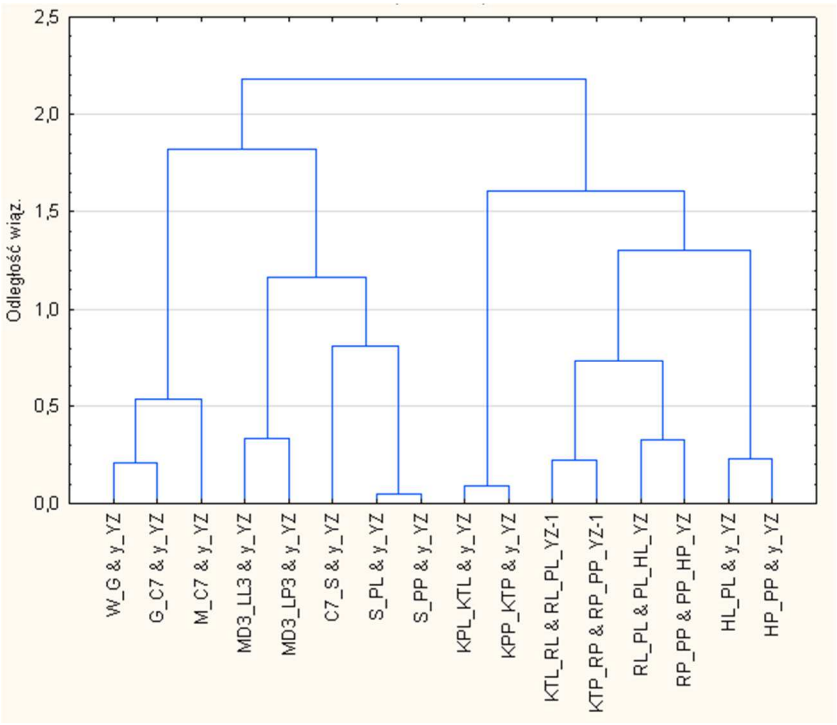
Gdzie:

x_i – wartość zmiennej będącej kryterium segmentacji dla i-tego obiektu

k - liczba obiektów w skupieniu

Metoda Warda jest traktowana jako bardzo efektywna, chociaż zmierza do tworzenia skupień o małej wielkości.

W wyniku zastosowania metod hierarchicznych uzyskujemy dendrogram, który ilustruje hierarchiczną strukturę zbioru obiektów ze względu na zmniejszające się podobieństwo między nimi. Przykładowy dendrogram przedstawia rys. 5.



Rys. 5. Przykładowy dendrogram

1.3. Grupowanie metodą k-średnich.

Grupowanie metodą k-średnich należy do niehierarchicznych metod grupowania. W przeciwieństwie do metod hierarchicznych, w tego typu metodach w efekcie uzyskujemy rozbicie, w którym żadne skupienie nie jest podkupieniem innego. W przypadku tej techniki zakładamy, że znamy a priori liczbę skupień. Ogólnie rzecz biorąc, za pomocą metody k-średnich zostanie utworzonych k skupień, możliwie jak najbardziej różniących się od siebie.

Wyboru początkowych k środków skupień dokonuje się w sposób losowy, arbitralnie lub według pewnego kryterium. W kolejnych krokach dokonuje się korekty tego rozwiązania. Algorytm przenosi obiekty pomiędzy skupieniami tak, aby zminimalizować zmienność wewnątrz skupień i zmaksymalizować zmienność między skupieniami. Kolejne kroki algorytmu charakteryzowane są przez funkcję błędu. Zatrzymanie procedury następuje wówczas, gdy wartość funkcji błędu nie

wykazuje istotnych zmian lub gdy zostanie przekroczona z góry zadana liczba iteracji. W praktyce liczba iteracji potrzebnych do otrzymania końcowego rozwiązania rzadko przekracza 10.

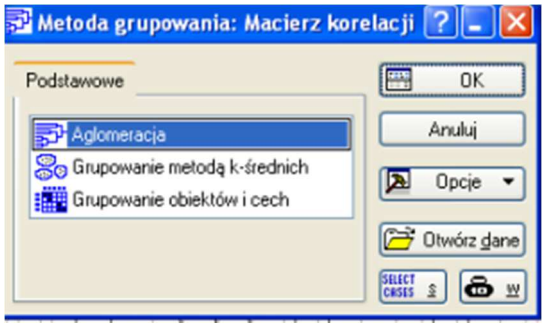
1.4. Założenia analizy skupień.

W przypadku analizy skupień krytycznymi zagadnieniami są reprezentatywność próby i współliniowość. Brak reprezentatywności może prowadzić do zafałszowania struktury skupień. Aby możliwe było uogólnienie na całą populację, próba musi być pobrana w sposób losowy. Współliniowość występuje wówczas, gdy zmienne niezależne są ze sobą mocno skorelowane. W takim przypadku układ skupień może również być nierzeczywisty, ponieważ współliniowe zmienne mogą mieć większy wpływ na miary podobieństwa (odległości).

Analiza skupień jest również wrażliwa na obecność punktów odstających. Do wykrywania obserwacji odstających można wykonać wykres profili przypadków. Na takim wykresie punkty odstające będą miały wyróżniające się profile, które będą najczęściej wyznaczone przez wartości ekstremalne jednej lub kilku zmiennych. W STATISTICE taki wykres można utworzyć wybierając z menu ***Wykresy*** opcję ***Wykresy 2W***, a następnie z rozwijanego podmenu opcje ***Wykresy liniowe*** (***Profile przypadków***).

2. Analiza skupień w STATISTICE

Analizę skupień uruchamiamy wybierając z menu ***Statystyka*** opcję ***Wielowymiarowe techniki eksploracyjne/Analiza skupień***. Po otwarciu modułu pojawia się okno przedstawione na rys. 6.



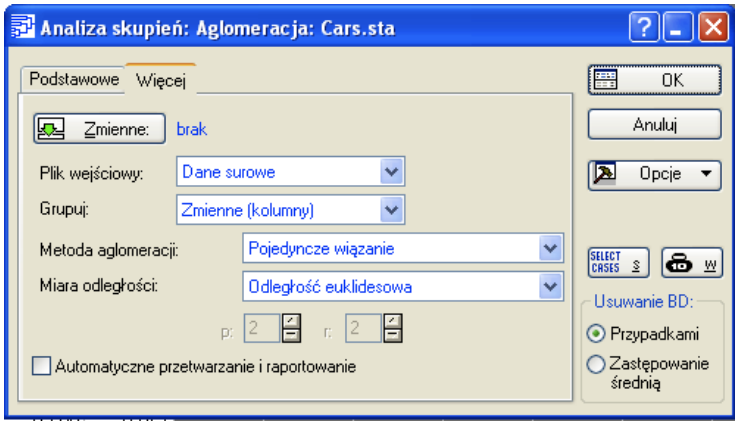
Rys. 6. Okno wyboru metody grupowania.

Na karcie Podstawowe wybieramy metodę grupowania.

2.1 Aglomeracja.

W oknie metody aglomeracji znajdują się dwie karty: ***Podstawowe*** – umożliwiająca szybko określić analizę i uzyskać wyniki oraz ***Więcej***, w której można zdefiniować dokładniejszą analizę (Rys. 7). Jako plik wejściowy można wybrać jedną z dwóch opcji:

- *Dane surowe*
- *Macierz odległości*



Rys. 7. Opcje metod aglomeracji.

W zależności od potrzeb grupowania można dokonywać na podstawie ***Przypadków*** (wierszy) lub ***Zmiennych*** (kolumny). Domyślnie jako opcja grupowania wybrane są ***Zmienne***. Następnie dokonuje się wyboru metody aglomeracji (rozwijana lista zawiera 7 opisanych wyżej metod. Domyślnie wybrana jest metoda ***Pojedyncze wiązanie*** (metoda najbliższego sąsiedztwa), którą w razie potrzeby można zmienić .

Metoda aglomeracji łączy kolejno obiekty o rosnącej odległości. Sposoby jej obliczania dostępne są na liście **Miara odległości**. Są to: **kwadrat odległości euklidesowej**, **odległość euklidesowa**, **odległość miejska (Manhattan)**, **odległość Czebyszewa**, **odległość potęgowa**, **niezgodność procentowa** oraz **1 – r Pearsona** (r oznacza współczynnik korelacji liniowej).

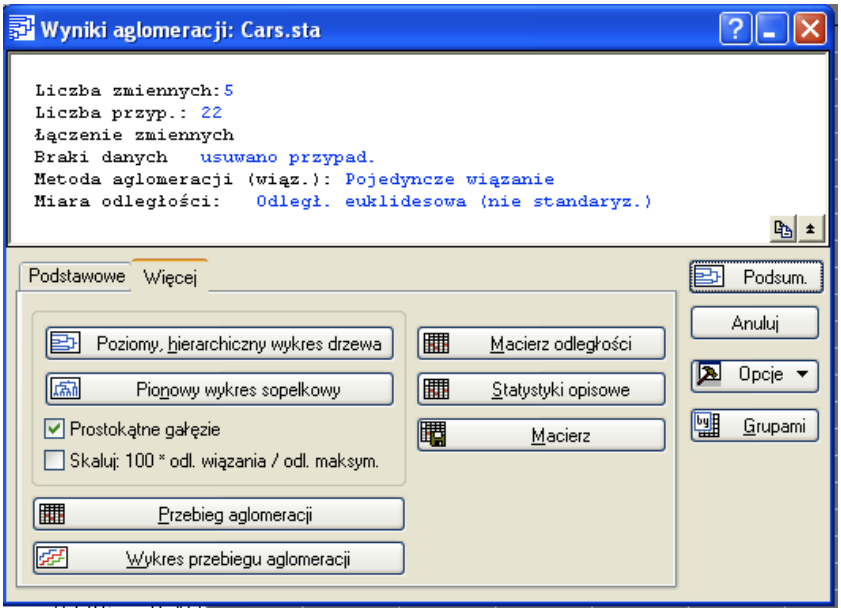
W oknie tym mamy jeszcze możliwość zaznaczenia opcji **Automatyczne przetwarzanie i raportowanie** – wówczas po kliknięciu przycisku OK automatycznie zostanie wykonana cała analiza, a jej wyniki przesłane do skoroszytu, samodzielnych okien lub raportu (w zależności od ustawień). Analogicznie jak przy innych analizach, tu także mamy możliwość usuwania brakujących danych (opcja **Usuwanie BD**): przypadkami lub zastępując je średnią.

Analiza rozpoczyna się po kliknięciu przycisku **OK**. Na ekranie pojawia się wówczas okno **Wyniki aglomeracji** (rys. 8).

W górnej części okna znajduje się ogólne podsumowanie wyników bieżącej analizy. W dolnej części okna znajduje się szereg przycisków otwierających arkusze wynikowe i interpretacje graficzne podsumowujące daną analizę:

- **Statystyki opisowe** – przycisk ten umożliwia wyświetlenie typowego arkusza zawierającego wartości średnie i odchylenia standardowe dla każdego obiektu włączonego do analizy skupień.
- **Macierz odległości** – przywołanie arkusza z macierzą odległości. Macierz ta jest punktem wyjścia do wyodrębniania skupień.
- **Przebieg aglomeracji** – przywołanie arkusza z opisem przebiegu procesu aglomeracji. W pierwszej kolumnie zawarte są odległości wiązań, na których zostały uformowane odpowiednie skupienia, zaś w kolejnych wierszach podane są nazwy obiektów, które formują nowe skupienia.

- **Poziomy hierarchiczny wykres drzewa** oraz **Pionowy wykres sopelkowy** – naciśnięcie jednego z tych przycisków powoduje utworzenie dendrogramu przedstawiającego strukturę powstałych skupień. Pierwszy z nich utworzy poziomy diagram drzewkowy, który przedstawia następstwo grupowania obiektów. Drugi natomiast tworzy pionowy diagram drzewkowy. Na obu typach wykresów istnieje możliwość wyboru wyświetlania prostokątnych lub ukośnych gałęzi (opcja **Prostokątne gałęzie**). Po wyborze opcji **Skaluj: 100*odl.wiązania/odl.maks.** następuje przeskalowanie drzewa do skali standaryzowanej.



Rys. 8. Okno Wyniki aglomeracji.

- **Wykres przebiegu aglomeracji** – opcja ta tworzy wykres liniowy odległości wiązań względem kolejnych etapów procesu wiązania. Jest on przydatny do identyfikowania miejsc, w których formuje się wiele skupień w przybliżeniu w takiej samej odległości wiązania, co może wskazywać na naturalną „nieciągłość" w sensie odległości między obserwowanymi obiektami.
- **Macierz** – kliknięcie tego przycisku powoduje utworzenie arkusza macierzowego dla macierzy odległości. Arkusz ten wyświetlany jest w osobnym oknie.

Interpretacja dendrogramu

Przy interpretacji dendrogramu pod kątem identyfikacji i charakterystyki wyróżnionych skupień pomocne mogą okazać się poniższe wskazówki:

- Analizujemy dendrogram pod względem różnic odległości między kolejnymi węzłami. Duża wartość różnic oznacza, że skupienia są odległe. Zatem dokonujemy podziału w miejscu, gdzie odległość pomiędzy gałęziami drzewa osiąga maksimum.
- Wykorzystujemy wykres przebiegu aglomeracji. Wykres ten pokazuje odległości pomiędzy skupieniami w momencie, gdy były one łączone. Jeśli na wykresie widać wyraźne spłaszczenie (dłuższa linia pionowa), to oznacza, że w tym miejscu skupienia są odległe i jest to najlepszy punkt odcięcia.

- Wykorzystujemy współczynnik RMSSTD (*root-mean-square standard deviation*) dla skupień. Definiujemy go dla k-tego skupienia na podstawie wzoru:

$$\sqrt{\frac{\sum_{i=1}^{p}\sum_{j\in S_k}\left(x_{ij}-\bar{x_i}\right)^2}{p(n_k-1)}}$$

Gdzie: p – liczba zmiennych

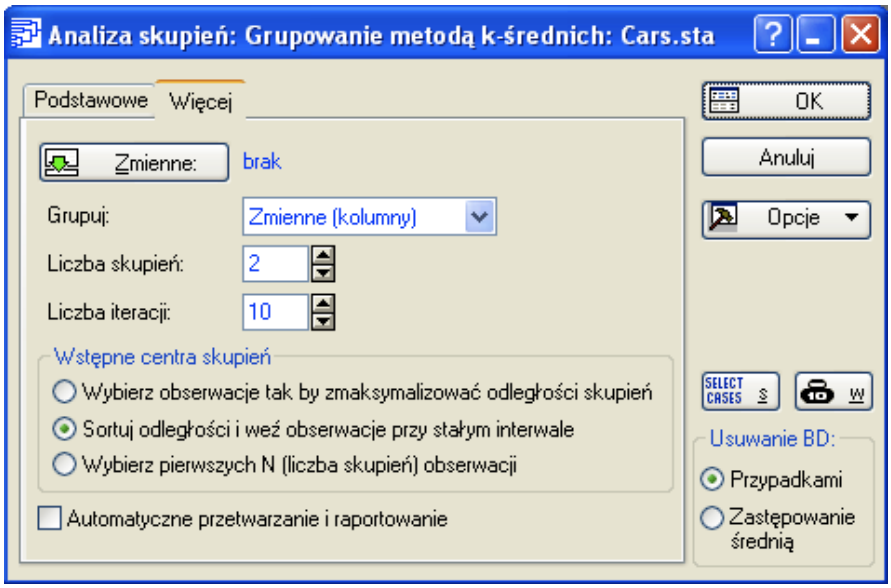
S_k – k-te skupienie

n_k – liczebność k-tego skupienia

Przydatny okazuje się wykres RMSSTD względem kolejnych kroków tworzących drzewo skupień. Jeśli odmienne (odległe) skupienia są łączone na wykresie, to będzie widoczny znaczny wzrost nachylenia linii.

2.2 Grupowanie metodą k-średnich

W oknie wyboru metody grupowania (rys. 6) wybieramy opcję Grupowanie metodą k- średnich. Wówczas otwiera się następujące okno (rys. 9):



Rys. 9. Opcje grupowania metodą k- średnich.

W oknie tym znajdują się dwie karty: **Podstawowe** i **Więcej**. Korzystając z nich możemy szybko określić analizę i uzyskać interesujące nas wyniki:

- Przycisk **Zmienne** pozwala na wybór zmiennych podlegających analizie
- Lista rozwijalna **Grupuj** zawiera dwie pozycje: **Zmienne (kolumny)** i **Przypadki (wiersze)**. Od dokonanego tutaj wyboru zależy, czy wybrane do analizy zmienne będą traktowane jako wymiary (opcja **Przypadki**) czy jako obiekty do tworzenia skupień (opcja **Zmienne**).
- Pole **Liczba skupień** – podajemy w nim liczbę tworzonych skupień, która musi być większa niż 1 i mniejsza niż liczba obiektów. Dla zrealizowania tego zadania program przenosi obiekty do

różnych skupień, zmierzając do minimalizacji zmienności wewnątrz skupień i maksymalizacji zmienności między skupieniami.

- Pole **Liczba iteracji** – podajemy w nim maksymalną liczbę iteracji, które mają zostać wykonane. Domyślne ustawienie (10 iteracji) zwykle jest wystarczające i nie wymaga zmiany.
- Pole wyboru **Automatyczne przetwarzanie i raportowanie** – analogicznie jak przy metodzie aglomeracji, po zaznaczeniu tej opcji cała analiza zostanie wykonana automatycznie.

W dolnej części okna znajduje się również grupa opcji **Wstępne centra skupień**, w której możemy wybrać opcje określające sposób wyznaczenia wstępnych centrów. W pakiecie STATISTICA mamy do wyboru jedną z trzech opcji:

- **Wybierz obserwacje tak, aby zmaksymalizować odległości skupień.** Po wybraniu tej opcji, jako wstępne centra skupień zostaną wzięte obserwacje lub obiekty zgodnie z zasadami maksymalizacji wstępnych odległości między skupieniami. Procedura ta może jednak prowadzić do utworzenia skupień składających się z pojedynczych obserwacji, jeśli w danych występują wyraźne przypadki odstające.
- **Sortuj odległości i weź obserwacje przy stałym interwale.** W przypadku tej opcji w pierwszej kolejności wszystkie odległości między obiektami zostaną posortowane, a następnie na początkowe centra skupień zostaną wybrane obiekty przy stałych interwałach.
- **Wybierz pierwszych N (liczba skupień) obserwacji.** Jeśli wybierzemy tą opcję, na wstępne centra skupień zostanie wziętych N (liczba skupień) pierwszych obserwacji. W ten sposób opcja ta umożliwia pełną kontrolę nad wyborem wstępnej konfiguracji. Jest ona przydatna zwłaszcza wtedy, gdy mamy pewne oczekiwania a priori co do natury analizowanych skupień. W takim przypadku przypadki, które mają stać się wyjściowymi centrami skupień, musimy przenieść na początek pliku.

Po ustawieniu wstępnych opcji i kliknięciu przycisku **OK** ukaże się wstępne okno wynikowe (rys. 10).

Rys. 10. Wyniki grupowania metodą k- średnich.

W górnej części tego okna znajduje się ogólne podsumowanie wyników bieżącej analizy. Przyciski w dolnej części okna otwierają szereg arkuszy wynikowych i interpretacji graficznych podsumowujących analizę:

- **Podsum.: Średnie skupień i odległ. euklid.** – po kliknięciu tego Pu zostaną utworzone dwa arkusze wynikowe: w pierwszym zostały zebrane wartości średnie każdego wymiaru w obrębie każdego skupienia, zaś w drugim – odległości euklidesowe między centrami skupień (pod przekątną) i ich kwadraty (nad przekątną).
- **Analiza wariancji** – celem procedury grupowania metodą k- średnich jest pogrupowanie obiektów w określoną przez użytkownika liczbę skupień. Aby oszacować trafność tej klasyfikacji, porównuje się zmienność wewnątrz skupień (mała, jeśli klasyfikacja jest dobra) ze zmiennością między skupieniami (duża, jeśli klasyfikacja jest dobra). To oznacza, że robimy typową analizę wariancji dla każdego wymiaru.
- **Wykres średnich** – po kliknięciu tego przycisku zostanie wyświetlony wykres liniowy średnich dla poszczególnych skupień. Jest ona przydatny do wizualnego zestawienia różnic w średnich między skupieniami.
- **Statystyki opisowe każdego skupienia** – przycisk ten umożliwia utworzenie arkuszy danych dla każdego skupienia, zawierających statystyki opisowe.
- **Elementy każdego skupienia i odległości** – przycisk ten pozwala wyświetlić elementy każdego skupienia i ich odległości (euklidesowe) od odpowiednich centrów skupienia (sredniej). Pozwala to interpretować potencjalne „złe" elementy skupień, tzn. takie obiekty, które są bardzo odległe od centrum skupienia, choć najwyraźniej nie należą też do żadnego innego skupienia. Każdy wyświetlony arkusz prezentuje wyniki dla jednego skupienia.
- **Zapisz klasyfikacje i odległości** – po kliknięciu tego przycisku zostanie utworzony samodzielny arkusz zawierający informacje na temat klasyfikacji i odległości grupowanych obiektów. Otrzymane wyniki możemy umieścić w skoroszycie lub raporcie za pomocą przycisków **Dodaj do skoroszytu** i **Dodaj do raportu**.

Założenia analizy skupień

W przypadku analizy skupień mamy dwa istotne zagadnienia: reprezentatywność próby oraz współliniowość. Brak reprezentatywności może doprowadzi do zafałszowania struktury skupień. Próba musi być pobrana w sposób losowy, aby otrzymane efekty można było uogólnić na całą populację. Współliniowość występuje wówczas gdy zmienne niezależne są ze sobą mocno skorelowane. Jej wystąpienie utrudnia ocenę prawdziwego wpływu poszczególnych zmiennych. W analizie skupień współliniowość może utworzyć nierzeczywisty układ skupień, ponieważ współliniowe zmienne mogą mieć większy wpływ na miary podobieństwa (odległości). Otrzymany wówczas układ skupień może być mylący.

Analiza skupień jest również wrażliwa na występowanie punktów odstających. Możemy wówczas otrzymać fałszywą strukturę grup. Do wykrywania obserwacji odstających można skorzystać z wykresu profili przypadków. Na takim wykresie punkty odstające będą miały wyróżniające się profile, najczęściej będą one wyznaczone przez ekstremalne wartości jednej lub kilku zmiennych. W pakiecie STATISTICA taki wykres można utworzyć wybierając z menu **Wykresy** opcję **Wykresy 2W**, a następnie z rozwijającego się podmenu opcje **Wykresy liniowe** (**Profile przypadków**).

Instrukcja do laboratorium

Ćwiczenia

Część I

Dane znajdują się w pliku dane9.xls. Zawierają one dane na temat wartości średnich parametrów opisujących postawę ciała człowieka dla różnych jednostek chorobowych.

1. Wczytaj plik z danymi. Przy wczytywaniu zaznacz opcje wczytywania nazw zmiennych i przypadków z pierwszego wiersza i pierwszej kolumny.
2. Przed rozpoczęciem analizy dokonaj standaryzacji zmiennych
   *Wskazówka: Opcja standaryzacji jest dostępna w menu **Dane**.*
3. Uruchom analizę skupień metodą aglomeracji. Jako metodę aglomeracji wybierz metodę średnich połączeń ważonych
4. Wyświetl i przeanalizuj dendrogram. Do określenia punktu odcięcia wykorzystaj wykres przebiegu aglomeracji.
5. Co możemy powiedzieć o analizowanych grupach chorób?

Część II

1. Otwórz plik Cars.sta (z menu Plik → Otwórz przykłady. Plik jest dostępny w katalogu Datasets). Dane zawierają poniższe informacje o różnych modelach samochodów:
   - Przybliżoną cenę (zmienna Cena),
   - Przyśpieszenie (od 0 do 60 mil na godzinę, podane w sekundach; zmienna Przysp),
   - Zdolność hamowania (droga hamowania od 80 mil na godzinę do całkowitego zatrzymania; zmienna Hamowan),
   - Indeks zdolności trzymania się drogi (zmienna Wsk_trzy)
   - Zużycie paliwa samochodu (mile na galon; zmienna Zużycie).
2. Uruchom analizę skupień: grupowanie metodą k- średnich. Jako liczbę skupień wpisz 3.
3. Chcemy skupić w sensowne grupy różne samochody, a więc na **Grupowanie** z domyślnego ustawienia **Zmienne** trzeba zmienić na **Przypadki**.
4. Przeglądnij wyniki analizy wariancji oraz elementy każdego skupienia. Wyciągnij wnioski.
5. Aby potwierdzić wnioski uzyskane w punkcie 4, wyświetl wykresy średnich dla skupień.