**Ćwiczenie 3. Analiza podobieństwa**

1. **Cel ćwiczenia:**

Celem ćwiczenia jest nabycie umiejętności poprawnego wykonania analizy wiązkowej oraz zinterpretowanie uzyskanych wyników.

1. **Zagadnienia do samodzielnego opracowania:**

Przygotowanie danych do analizy podobieństwa: centrowanie, skalowanie i standaryzacja zmiennych. Wykrywanie punktów odbiegających i ich wpływ na wynik analizy. Podobieństwo wielowymiarowe. Miary odległości w przestrzeni cech: odległość euklidesowa i jej kwadrat, odległość miejska (Manhattan), odległość Czebyszewa, odległość potęgowa, miary odległości oparte na współczynniku korelacji. Sposoby wyrażania odległości pomiędzy grupami obiektów: metoda najbliższego sąsiada, metoda najdalszego sąsiada, metoda centroidowa, metoda średniej grupowej. Algorytm analizy wiązkowej oraz interpretacja diagramu drzewa. Metody określania liczby skupień. Zastosowanie analizy wiązkowej do zredukowania liczby wymiarów przestrzeni cech i wyboru reprezentatywnych zmiennych: metoda środka ciężkości i metoda potencjałów. Zastosowania metod analizy podobieństwa w chemii.

1. **Przebieg ćwiczenia:**
   * + 1. W laboratorium analizy chemicznej oznaczono zawartość 11 pierwiastków [mg/kg m.s.] w 15 próbkach gleby (wyniki zebrano w tabeli danych). Tabelę z danymi, proszę pobrać z: <http://www.chem.univ.gda.pl/pchs/dane/cw_3a.ods>. Na podstawie powyższych danych proszę obliczyć macierz odległości (odległość euklidesowa) w R Commander, korzystając z komendy **x=dist(mydata, method="euclidean")**. Następnie proszę wykonać analizę wiązkową metodą najbliższego oraz najdalszego sąsiada **Statistics > Dimensional analysis > Cluster analysis > Hierarchical cluster analysis**. Na podstawie hierarchicznego pogrupowania cech proszę wykonać dendrogramy oraz wskazać różnice dla obu użytych metod wyrażania odległości pomiędzy kolejnymi grupami. Dodatkowo, proszę „ręcznie“ obliczyć macierz odległości (odległość euklidesowa) pomiędzy 15 obiektami (próbkami gleby) w przestrzeni zdefiniowanej przez stężenia dwóch pierwiastków: Mg i Ca.
       2. Instytut Ochrony Środowiska Morskiego przeprowadził badania stężeń ośmiu kongenerów polichlorowanych naftalenów (PCN) w powierzchniowej warstwie osadów dennych pobranych ze 150 stanowisk zlokalizowanych na Morzu Bałtyckim. Pobierz dane z: <http://www.chem.univ.gda.pl/pchs/dane/cw_3b.ods> Dane w pliku zawierają wartości średnie dla każdej ze 150 lokalizacji.
2. Wykorzystując analizę wiązkową proszę zbadać, które kongenery akumulują się w osadach w podobny sposób.
3. Porównać profile kongenerów w próbkach z poszczególnych stanowisk. Które stanowiska różnią się najbardziej pod względem zawartości poszczególnych PCN?
4. Wiedząc, że kongenery podstawione głównie w pozycjach *alpha* (1,4,5,8) pochodzą z tzw. preparatów technicznych wykorzystywanych np. w urządzeniach elektroenergetycznych (np. transformatory) a podstawione głównie w pozycjach *beta* (2,3,6,7) tworzą się w wyniku spalania materii organicznej (np. w spalarniach odpadów), proszę ustalić główne źródło zanieczyszczeń chloronaftalenami w każdym ze stanowisk pomiarowych.
5. Za pomocą metody środka ciężkości, proszę wybrać jeden kongener reprezentatywny dla preparatów technicznych i jeden dla źródeł termicznych. a następnie narysuj wykres rozrzutu dla wybranych reprezentantów: Graphs > Scatterplot
6. **Sprawozdanie:**
   * + 1. Wyniki uzyskane na zajęciach wraz z interpretacją.
7. **Literatura**
   * + 1. A. Łomnicki, „Wprowadzenie do statystyki dla przyrodników”, Wydanie trzecie uzupełnione, Wyd. Naukowe PWN, Warszawa, 2005
       2. J. Mazerski, „Podstawy chemometrii”, Wydawnictwo Politechniki Gdańskiej, Gdańsk, 2000