Rozprawa doktorska

Metody grupowania danych i ich wybrane modyfikacje dedykowane eksploracji danych eksperymentalnych

mgr Klaudia Drab

Promotor pracy:
dr hab. Michał Daszykowski, prof. UŚ

Instytut Chemii
Wydział Matematyki, Fizyki i Chemii
Uniwersytet Śląski
Katowice, 2016

Składam serdeczne podziękowania,

Mojemu Promotorowi ***dr hab. Michałowi Daszykowskiemu, prof. UŚ***
za umożliwienie realizacji badań, cenne uwagi i sugestie oraz poświęcony mi czas,
a także wieloletnią współpracę.

Mojej ***Mamie***
za wsparcie mnie w moich wyborach,
nieocenioną cierpliwość, przekazaną mądrość i nierozerwalną przyjaźń.

Wszystkim ***znajomym i przyjaciołom****,*
którzy wspierali mnie przez cały okres studiów doktoranckich.

Klaudia

Streszczenie rozprawy doktorskiej

Zaawansowana aparatura badawcza umożliwia badanie materiałów różnorodnego pochodzenia. Dlatego znalazła zastosowanie w wielu dziedzinach nauki, gdzie stanowi podstawowe narzędzie w ocenie fizykochemicznych właściwości próbek. Jednak kompleksowa charakterystyka próbki pociąga za sobą pozyskiwanie danych o złożonej strukturze. Opisując każdą analizowaną próbkę za pomocą od kilku do kilku tysięcy zmiennych otrzymuje się tzw. dane wielowymiarowe, co pociąga za sobą potrzebę zastosowania metod matematycznych, pozwalających na analizę, interpretację wyników oraz formułowanie wniosków. W tym celu korzysta się z metod chemometrycznych, w skład których wchodzą metody wstępnego przygotowania danych do dalszej analizy, metody eksploracyjne oraz metody modelowania danych. Szczególnie interesujące są metody eksploracyjne, pozwalające na wgląd w ukrytą strukturę analizowanych danych oraz ujawnienie zależności pomiędzy próbkami i/lub parametrami. Jednym z wariantów metod eksploracyjnych są metody grupowania danych, które są szczególnie przydatne w kontekście wyodrębniania grup podobnych obiektów. Podobieństwo analizowanych próbek, oceniane jest na podstawie ich odległości w przestrzeni eksperymentalnej (te które znajdują się blisko siebie wykazują podobne właściwości fizykochemiczne). W celu określenia podobieństwa pomiędzy próbkami wykorzystuje się tzw. miary podobieństwa, które są matematyczną interpretacją odległości pomiędzy nimi.

Wzrastająca kompleksowość danych pociąga za sobą potrzebę modyfikacji i rozwoju nowych metod eksploracyjnych oraz miar odległości. W związku z czym w niniejszej pracy doktorskiej skupiono się na modyfikacji algorytmu DBSCAN w celu eliminacji problemu błędnego przypisania obiektów brzegowych do odpowiednich grup w przypadku grup sąsiadujących ze sobą w przestrzeni eksperymentalnej. Modyfikacja algorytmu polegała na zmianie sposobu przetwarzania obiektów oraz przypisaniu obiektów brzegowych do grup na podstawie odległości euklidesowej pomiędzy obiektami brzegowymi, a środkami wyodrębnionych grup obiektów. Następnie skupiono się na rozwinięciu koncepcji nowej miary odległości (s_{ij}) pozwalającej porównywać ze sobą dwuwymiarowe chromatograficzne odciski palca, w których występuje problem koelucji substancji i przesunięć pików w czasie. W ostatniej części pracy rozważano problem niepewności pomiarowej towarzyszącej danym eksperymentalnym. Dotychczas, błąd pomiarowy był pomijanym elementem w trakcie analizy danych. Aktualnie rozwój algorytmów uwzględniających niepewności pomiarowe uzyskiwanych danych stanowi nowy trend w pracach naukowych. Korzystając z osiągnieć zaprezentowanych w [122], zaproponowano uwzględnienie niepewności pomiarowych modelowanych dla każdego obiektu np. w algorytmie DBSCAN, poprawiając efektywność metody.

Stosowana w pracy notacja

$\mathbf{A}$	–	macierz wyników o wymiarowości m×k
A	–	podmacierz o wymiarowości I×J
a_{ij}	–	element podmacierzy A, gdzie i ∈ I oraz j ∈ J
a_{iJ}	–	średnia wartości wierszy w podmacierzy A
a_{Ij}	–	średnia wartości kolumn w podmacierzy A
a_{IJ}	–	średnia wartości podmacierzy A
$\mathbf{B}$	–	macierz wag o wymiarowości k×n
$\mathbf{C}$	–	macierz wariancji – kowariancji
CD	–	odległość rdzeniowa stosowana w metodzie OPTICS
d	–	stała rozpadu w metodzie NG
d_i, d_f	–	początkowy i końcowy rozmiar sąsiedztwa w metodzie NG
$d_{\mathbf{xy}}$	–	odległość euklidesowa pomiędzy obiektami **x** i **y**
$d_{\mathbf{xy}}^2$	–	kwadrat odległości euklidesowej pomiędzy obiektami **x** i **y**
$dM_{\mathbf{xy}}$	–	odległość Mahalanobisa pomiędzy obiektami **x** i **y**
DV	–	współczynnik dyspersji
E	–	funkcja kosztów
$\mathbf{E}$	–	macierz reszt o wymiarowości m×n
f	–	liczba czynników głównych
g	–	liczba sąsiadów w otoczeniu i-tego obiektu
H(I, J)	–	średni błąd kwadratowy podgrupy
indeks ^{T}	–	operacja transponowania macierzy
I	–	zbiór wierszy macierzy
Ir_i, Ir_f	–	ilorazy inicjalizacji iteracji i zakończenia iteracji w metodzie NG
J	–	zbiór kolumn macierzy
$\mathbf{K}$	–	macierz podobieństw utworzona na podstawie wprowadzonej miary odległości s_{ij}
k	–	liczba grup
$\mathbf{L}$	–	macierz wag o wymiarowości f×n
m	–	liczba wierszy (obiektów) macierzy
MinPts	–	minimalna liczba sąsiadów w metodzie DBSCAN
n	–	liczba kolumn (zmiennych) macierzy
R	–	współczynnik korelacji Pearsona
RD	–	odległość bezpośrednia stosowana w metodzie OPTICS
r	–	promień sąsiedztwa i-tego obiektu
$\mathbf{S}$	–	macierz wyników o wymiarowości m×f
s_{ij}	–	nowa miara odległości
s_1	–	zwycięski węzeł sieci neuronowej stanowiący środek grupy
s_2	–	najbliższy sąsiad zwycięskiego węzła (s_1) sieci neuronowej
std(x)	–	odchylenie standardowe
t	–	liczba iteracji w metodzie NG
$\mathbf{u}_k$	–	współrzędne środka k-tej grupy

w	–	liczba widm z zakresu UV-VIS w dwuwymiarowym odcisku palca uwzględniana w nowej metodologii opartej na s_{ij}
$\mathbf{w}_j(1,n)$	–	wektor wag dla j-tego węzła sieci
$\mathbf{X}(m\times n)$	–	macierz danych o m wierszach i n kolumnach
$\underline{\mathbf{X}}(m\times n\times p)$	–	tensor danych
x_a	–	element macierzy po autoskalowaniu
x_c	–	element macierzy po centrowaniu
$\mathbf{x}_i$	–	i-ty obiekt w grupie
x_{ij}	–	element macierzy znajdujący się w i-tym wierszu oraz j-tej kolumnie
x i **y**	–	wektory własne (klasyfikacji) macierzy **X**
Y	–	macierz zmiennych zależnych
α	–	kryterium zatrzymania działania algorytmu, zwana również zmienną wieku w metodzie GNG
λ	–	wartość własna macierzy **X**
χ	–	liczba iteracji w metodzie GNG

Stosowane w pracy akronimy

DBSCAN	–	metoda grupowania bazująca na gęstości danych (z ang. Density-Based Spatial Clustering of Application with Noise)
EPR	–	elektronowy rezonans paramagnetyczny (z ang. Electron Paramagnetic Resonance)
GC	–	chromatografia gazowa (z ang. Gas Chromatography)
GK	–	metoda ekspandującego k-średnich (z ang. Growing k-means)
GNG	–	ekspandujący gaz neuronowy (z ang. Growing Neural Gas)
hError	–	metoda grupowania hierarchicznego Warda z uwzględnieniem niepewności pomiarowych
HPLC	–	wysokosprawna chromatografia cieczowa (z ang. High Performance Liquid Chromatography)
kError	–	metoda k-średnich z uwzględnieniem niepewności pomiarowych
LC	–	chromatografia cieczowa (z ang. Liquid Chromatography)
MS	–	spektrometria mas (z ang. Mass Spectrometry)
MLR	–	regresja wieloraka (z ang. Multiple Linear Regression)
NG	–	gaz neuronowy (z ang. Neural Gas)
NMR	–	jądrowy rezonans magnetyczny (z ang. Nuclear Magnetic Resonance)
PC	–	czynnik główny (z ang. Principal Component)
PCA	–	analiza czynników głównych (z ang. Principal Component Analysis)
PCR	–	regresja czynników głównych (z ang. Principal Component Regression)
PI	–	indeks projekcji (z ang. Projection Index)
PLS	–	regresja częściowych najmniejszych kwadratów (z ang. Partial Least Squares Regression)
PP	–	wymuszone projekcje (z ang. Projection Pursuit)
PSO	–	optymalizacja z użyciem roju cząstek (z ang. Particle Swarm Optimization)
SMR	–	regresja macierzy rzadkiej (z ang. Sparse Matrix Regression)
SOM	–	samoorganizujące się mapy Kohonena (z ang. Self-Organizing Maps)
SVD	–	algorytm dekompozycji macierzy na wektory własne i wartości własne (z ang. Singular Value Decomposition)
UV-VIS	–	metoda spektrofotometrii UV-VIS

Spis treści

1. Wprowadzenie

Ustalanie składu chemicznego różnorodnych substancji jest obiektem zainteresowania wszystkich nauk przyrodniczych w tym chemii, geologii, biologii i medycyny. Każda z tych dyscyplin bezpośrednio korzysta z zasobów wiedzy chemii analitycznej. Z tego powodu uległa ona przeistoczeniu z typowej nauki chemicznej w dyscyplinę o charakterze interdyscyplinarnym. W nowoczesnym ujęciu stała się nauką stosowaną, której rola opiera się przede wszystkim na praktycznym zastosowaniu opracowanych metodologii oraz zaawansowanej aparatury w celu analizy materiałów badawczych o złożonym składzie chemicznym. Wykorzystanie nowoczesnych urządzeń pozwala na relatywnie szybką ocenę zarówno składu jakościowego, jak i ilościowego próbek różnorodnego pochodzenia. Ponadto, zadania stawiane przed nowoczesnym laboratorium analitycznym, dysponującym odpowiednio wyposażonym zapleczem naukowym, związane są z oceną zagrożeń środowiska, kontrolą jakości produktów spożywczych oraz diagnostyką medyczną. Na szczególną uwagę zasługują analizy próbek środowiskowych, artykułów spożywczych, roślin, owoców, kosmetyków, leków oraz próbek biologicznych, takich jak: krew, mocz, czy płyn mózgowo-rdzeniowy. Cechą wspólną wymienionych materiałów jest ich złożony skład chemiczny, który najczęściej bada się za pomocą zaawansowanej aparatury pomiarowej. Na szczególną uwagę zasługują takie metody jak: spektrometria mas (MS), chromatografia cieczowa (LC) i/lub gazowa (GC), wysokosprawna chromatografia cieczowa (HPLC), czy jądrowy rezonans magnetyczny (NMR) [1]. Coraz więcej uwagi poświęca się badaniom z wykorzystaniem metod sprzężonych łączących zalety metod separacyjnych oraz metod spektroskopowych. Do tych metod zalicza się m.in. chromatografię cieczową sprzężoną z jądrowym rezonansem magnetycznym (LC-NMR), czy chromatografię gazową łączoną ze spektrometrią mas (GC-MS). Swą rosnącą popularność zawdzięczają m.in. możliwości pozyskiwania kompleksowej informacji o analizowanej próbce [2].

Relatywnie niski koszt i krótki czas prowadzonych analiz przyczynia się do przeprowadzania badań na szeroką skalę podczas których bada się duże ilości próbek charakteryzowanych przez kilka, a nawet kilkanaście tysięcy parametrów. W efekcie gromadzenie danych stało się typowym etapem procesu analitycznego. Mimo to, interpretacja i analiza pozyskanych zestawów danych jest dla chemika analityka wyzwaniem, przede wszystkim ze względu na złożoną strukturę pozyskiwanych danych. Dane tego typu nazywa się wielowymiarowymi, gdzie każdy obiekt (próbka) jest punktem w przestrzeni zdefiniowanej przez liczbę zmierzonych parametrów (zmiennych), a każdy parametr jest punktem w przestrzeni określonej przez liczbę analizowanych próbek [3]. Otrzymane, w procesie badawczym wyniki można zorganizować w macierz danych **X**, o wymiarach m×n, gdzie m reprezentuje liczbę analizowanych próbek, a n liczbę zmierzonych parametrów. W przypadku analizy instrumentalnej, n osiąga wartości od kilku do kilkunastu tysięcy, a liczba analizowanych próbek, w ramach jednego eksperymentu wciąż wzrasta. To też pojawiają się problemy z interpretacją i analizą uzyskanych wyników.

Ograniczenia związane z interpretacją danych wielowymiarowych wynikają z braku możliwości wizualizacji danych o wymiarowości większej niż trzy. Dlatego, coraz częściej podczas analizy danych eksperymentalnych korzysta się z narzędzi chemometrycznych, ułatwiających etap interpretacji i formułowania generalnych wniosków poprzez redukcję wymiarowości danych oraz ich wizualizację. W pierwszym etapie otrzymaną macierz danych poddaje się wstępnemu przygotowaniu danych do dalszej analizy. Jest to kluczowy etap wpływający na jakość otrzymywanych wyników i formułowanie ostatecznych konkluzji. Kolejnym krokiem analizy danych jest ich eksploracja, której nadrzędnym celem jest odkrywanie ukrytej struktury danych. Dostępny pakiet metod eksploracyjnych jest bardzo szeroki, w związku z czym pozwala na dobór metody w zależności od problemu badawczego. Jednym z proponowanych rozwiązań jest zastosowanie metod grupowania danych [4]. Metodologia ta pozwala na wgląd w strukturę danych, dając informację o podobieństwach analizowanych próbek. Jej szczególnym przypadkiem są metody współgrupowania danych [5], umożliwiające równoczesne grupowanie obiektów i parametrów. W efekcie ich działania zostają wyodrębnione podgrupy próbek oraz podgrupy parametrów pozostających ze sobą w ścisłej zależności. Ze względu na liczne zalety, metody grupowania oraz współgrupowania danych znajdują szereg ciekawych zastosowań w wielu dziedzinach nauki. Kluczową rolę odgrywają przede wszystkim w naukach chemicznych oraz biologicznych, takich jak chemia środowiska [6], biochemia, biotechnologia, genomika [7], metabolomika oraz w medycynie [8]. Warto tutaj nadmienić, że wyniki uzyskiwane za pomocą metod eksploracyjnych często wspomagają dobór metody modelowania danych. Informacja ta pozwala na usunięcie obiektów odległych wpływających na konstrukcję modelu. W metodach kalibracji i dyskryminacji umożliwiają utworzenie zbioru testowego i modelowego.

Modelowanie jest ostatnim etapem analizy wielowymiarowych danych. Również i w tym wypadku chemometria proponuje szereg rozwiązań. Wśród nich można znaleźć metody pozwalające na budowę modeli kalibracyjnych, dyskryminacyjnych, czy klasyfikacyjnych, które dobiera się odpowiednio do rozważanego problemu badawczego. Zastosowanie chemometrycznych modeli uzyskanych za pomocą odpowiednio dobranych metod modelowania, pozwala m.in. na redukcję kosztów związanych z przeprowadzaniem rutynowych analiz. Stają się one przydatnym narzędziem rozwiązującym liczne problemy badawcze. Znajdują zastosowanie m.in. podczas oceny zdolności antyoksydacyjnych produktów spożywczych [9], czy badań nad autentycznością artykułów spożywczych oraz leków [10].

Ze względu na wspomniane zalety metod grupowania i współgrupowania danych oraz ich przydatność w różnorodnych dyscyplinach nauki w niniejszej pracy zilustrowano ich działanie w kontekście analizy danych eksperymentalnych. Uwzględniono również modyfikacje wybranych algorytmów grupowania danych, stanowiących niezbędny element ich rozwoju w rozpowszechnieniu ich zastosowania jako narzędzia poznania danych rozmaitego pochodzenia. Poświęcono również uwagę koncepcji miar podobieństwa, jako narzędzi poszukiwania podobieństw (różnic) pomiędzy obiektami i parametrami, wprowadzając nową miarę podobieństwa. Skupiono się na jej

wykorzystaniu w identyfikacji pików chromatograficznych w dwuwymiarowych chromatograficznych odciskach palca.

2. Cele pracy

Badania realizowane w ramach niniejszej pracy doktorskiej obejmowały następujące cele:

- ocenę przydatności wybranych technik grupowania danych w kontekście analizy złożonych danych eksperymentalnych,
- identyfikację kluczowych obszarów zastosowań metod grupowania i współgrupowania wielowymiarowych danych chemicznych,
- identyfikację problemów, które należy uwzględnić w analizie klasterowej oraz eksploracji danych za pomocą metod współgrupowania danych,
- propozycje modyfikacji algorytmów grupowania danych na potrzeby eksploracji danych eksperymentalnych,
- opracowanie nowej miary podobieństwa spełniającej warunki wyznaczone dla miar podobieństwa oraz miar odległości,
- wykorzystanie nowej miary podobieństwa w celu porównania dwuwymiarowych sygnałów analitycznych, w których występują przesunięcia pików oraz problem koelucji,
- uwzględnienie niepewności pomiarowych w eksploracji danych za pomocą metod grupowania danych.

3. Zaawansowane metody instrumentalne

Chemia analityczna jest samodzielną dyscypliną nauki, której zadaniem jest rozwój narzędzi oraz metod umożliwiających poznanie składu chemicznego badanych materiałów. Doświadczenie i wiedza z tego zakresu stają się niezwykle ważne przede wszystkim w kontekście analizy materiałów różnorodnego pochodzenia. Jest ona niezbędna podczas analizy próbek środowiskowych, próbek leków, kosmetyków, artykułów spożywczych, a coraz częściej również podczas analizy próbek pochodzenia biologicznego, takich jak płyny biologiczne np. mocz, krew. Dlatego chemię analityczną można potraktować jako swoiste narzędzie interdyscyplinarne wykorzystywane w różnorodnych dyscyplinach i dziedzinach nauki, wśród których wymienić warto przede wszystkim chemię, biologię, fizykę i medycynę. Interdyscyplinarny charakter chemii analitycznej jest konsekwencją kompleksowości stawianych przed nią zadań oraz różnorodności stosowanych metod badawczych. Niemałe znaczenie odgrywają również ciągłe innowacje w zakresie rozwoju aparatury badawczej i metodologii umożliwiających badanie złożonych materiałów [11].

Zadania stawiane przed nowoczesną chemią analityczną związane są przede wszystkim z określaniem składu jakościowego analizowanych próbek przy równoczesnym określeniu ich składu ilościowego. Dodatkowo analiza może informować o dynamice procesów zachodzących wewnątrz układów, czy strukturze badanej materii, co sprowadza się do uzyskiwania kompleksowej informacji o badanej próbce.

Chemia analityczna stoi również przed wyzwaniem spełnienia wymogów zielonej chemii [12], a więc stosowania jak najmniejszych objętości odczynników, czy zmniejszenia objętości analizowanych próbek. Ważnym aspektem analizy staje się również czas jej trwania, a ograniczenie kosztów prowadzonych procedur staje się nie lada wyzwaniem.

Podstawowym filarem chemii analitycznej jest analiza instrumentalna [13], obejmująca wiele technik badawczych. Metody instrumentalne zyskują coraz większe znaczenie we współczesnym świecie naukowym. Przede wszystkim ze względu na prostotę stosowanych procedur i relatywnie krótki czas oznaczeń przeprowadzanych z wykorzystaniem minimalnych objętości próbek. Warto również podkreślić, że analizy przeprowadzane są z dużą powtarzalnością i odtwarzalnością wyników i mogą odbywać się seriami. Natomiast zestawienie stosowanej aparatury badawczej z komputerem umożliwia uzyskiwanie i wyświetlanie wyników automatycznie, co sprowadza się do pozyskiwania dużej ilości danych, które następnie zostają zarchiwizowane w pamięci komputera.

Istnieje wiele metod instrumentalnych jednak ze względu na tematykę prowadzonych badań w niniejszej pracy skupiono się na scharakteryzowaniu wyłącznie kilku z nich, tj. metody spektrofotometrii UV-VIS oraz metod chromatograficznych, należących do metod separacyjnych. Wspomniano również o metodach sprzężonych będących kombinacją metod separacyjnych i metod detekcji.

3.1 Metody spektroskopowe

Metody spektroskopowe [14] należą do metod analitycznych zajmujących się rejestracją i pomiarem oddziaływań fali elektromagnetycznej z badaną materią. Odziaływanie to rejestruje się w postaci sygnału będącego podstawą do określania właściwości fizykochemicznych materii, co przekłada się na powszechne wykorzystanie technik spektroskopowych w różnorodnego rodzaju badaniach materiałów o złożonym składzie chemicznym. Techniki te wykorzystują właściwości promieniowania elektromagnetycznego, które wykazuje charakter korpuskularno-falowy, a więc przejawia zarówno charakter fali jak i cząstki. Fala ta rozchodzi się z prędkością 300 000 km/s jako periodyczne zmiany pola elektrycznego i magnetycznego. Falową naturę promieniowania charakteryzuje się liczbowo za pomocą dwóch wielkości, jakimi są długość fali (λ) oraz częstości drgań na sekundę (ν). Zakres fal elektromagnetycznych jest bardzo szeroki i mieści się w granicach od 10^{-14} dla fal promieniowania kosmicznego aż do 10^{6} dla fal radiowych. Poza charakterem falowym promieniowanie elektromagnetyczne wykazuje także charakter korpuskularny, a więc wiązka takiego promieniowania może być rozważana jako zbiór kwantów energii, które rozchodzą się w kierunku rozchodzenia się promieniowania. Właśnie, energia promieniowania elektromagnetycznego (1), stanowi podstawę w badaniu właściwości materii:

$$E = h \times \nu = \frac{hc}{\lambda} \tag{1}$$

gdzie:
h – stała Plancka (h = 6,626 069 57(29)· 10^{-34} J×s)
ν – częstość drgań (Hz)
c – prędkość światła $\left(\frac{m}{s}\right)$
λ – długość fali (m)

Powyższa zależność obrazuje że energia promieniowania elektromagnetycznego jest wprost proporcjonalna do częstości drgań i odwrotnie proporcjonalna do jej długości. Wykorzystanie energii z wybranego zakresu promieniowania elektromagnetycznego wywołuje określone zjawisko w cząsteczkach badanej materii. Podczas absorpcji energii promieniowania elektromagnetycznego następuje przejście elektronów z orbitalu o niższej energii na pusty orbital o wyższej energii. Następnie elektron, powraca do stanu podstawowego, co związane jest z emisją energii. Aby zaszła absorpcja energii promieniowania elektromagnetycznego przez badaną materię muszą zostać spełnione tzw. reguły wyboru, które można przedstawić następująco:

1) Absorpcja energii promieniowania elektromagnetycznego następuje wówczas, gdy istnieją dwa takie stany kwantowe cząsteczki Ψ_m oraz Ψ_n, których różnica energii wynosi:

$$E_n - E_m = h\,\nu_{m,n} = \Delta E \qquad (2)$$

2) Podczas absorpcji musi następować zmiana momentu dipolowego cząsteczki (μ).

Zjawisko absorpcji promieniowania elektromagnetycznego, związanego z pochłonięciem energii przez materię, a następnie emisji, czyli oddaniu nadmiaru energii z układu, stanowi podstawę klasyfikacji metod spektroskopowych. W związku z czym wyróżnić można metody spektroskopii absorpcyjnej oraz spektroskopii emisyjnej.
Klasyfikacja metod spektroskopowych jest uzależniona od przyjętego kryterium podziału np. zakresu promieniowania elektromagnetycznego lub formy energii jaka występuje w układach materialnych. W związku z czym wyróżnia się m.in. spektroskopię rentgenowską, radiospektroskopię (w zakresie mikrofalowym i fal radiowych), czy spektroskopię optyczną (w nadfiolecie, w zakresie widzialnym i w podczerwieni), będąca skutkiem podziału metod ze względu na stosowany zakres promieniowania elektromagnetycznego, a także metody spektroskopii elektronowej, oscylacyjnej, rotacyjnej, elektronowego rezonansu paramagnetycznego (EPR), jądrowego rezonansu magnetycznego (NMR), różniące się formą energii układów materialnych.
Istnienie specyficznych układów cząsteczek i występujących w nich wiązań, które wykazują zdolność pochłaniania promieniowania o określonej długości fali, stanowi podstawę badań za pomocą metod spektroskopowych. Wśród nich warto wymienić chromofory, a więc takie cząsteczki które absorbują energię promieniowania elektromagnetycznego w zakresie światła widzialnego i ultrafioletowego. Ta właściwość chromoforów jest podstawą jednej z najstarszych metod spektroskopowych – spektrofotometrii UV-VIS.

3.1.1 Spektrofotometria UV-VIS

Spektrometria w zakresie nadfioletu (z ang. Ultraviolet; UV) i promieniowania widzialnego (z ang. Visible; VIS) [15] jest jedną z najczęściej stosowanych metod instrumentalnych w analizie chemicznej. Oparta jest na zjawisku absorpcji energii promieniowania elektromagnetycznego w zakresie UV, tj. 200–380 nm oraz VIS 380–760 nm przez badane próbki (zob. Rys. 1).

Punktem wyjścia w dokonywaniu oznaczeń jakościowych i ilościowych jest wykorzystanie praw Lamberta, Lamberta-Beera oraz prawa addytywności adsorpcji, opisanych poniżej.

Prawo Lamberta
Absorbancja promieniowania elektromagnetycznego jest proporcjonalna do grubości warstwy absorbującej, jeśli wiązka promieniowania monochromatycznego przechodzi przez jednorodny ośrodek absorbujący:

$$A = \log \frac{I_0}{I} ab \qquad (3)$$

gdzie:
A – absorbancja
I_0 – natężenie światła padającego
I – natężenie światła po przejściu przez ośrodek
a – 0,4343k, gdzie k to współczynnik absorpcji
b – grubość warstwy absorbującej

Prawo Lamberta-Beera
Jeżeli współczynnik absorpcji rozpuszczalnika jest równy zeru, to absorbancja wiązki promieniowania monochromatycznego przechodzącej przez jednorodny roztwór jest wprost proporcjonalna do stężenia roztworu „c" i do grubości warstwy absorbującej „b"

$$A = \log \frac{I_0}{I} = abc \qquad (4)$$

gdzie:
c – stężenie roztworu

Prawo addytywności absorpcji
Absorbancja roztworu wieloskładnikowego równa się sumie absorbancji poszczególnych składników:

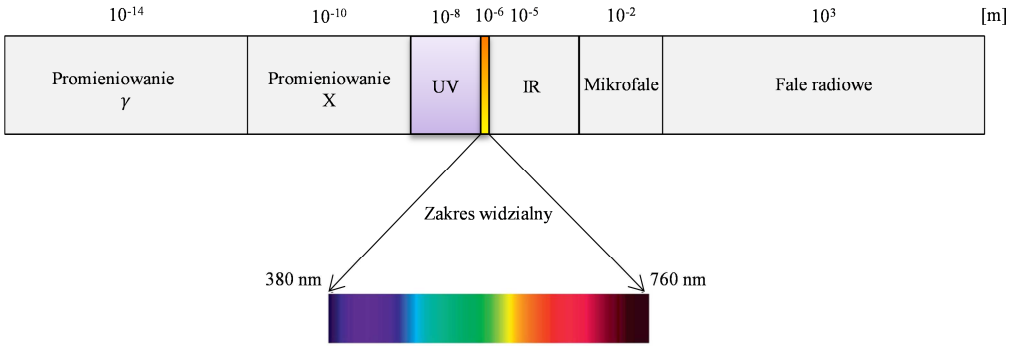
$$A = A_1 + A_2 + A_3 + ... + A_n \qquad (5)$$

gdzie:
A_1,A_2, A_3, A_n – absorbancja poszczególnych składników

Jeżeli stężenie danego składnika wyraża się w mol/L, wówczas prawo to można wyrazić następująco:

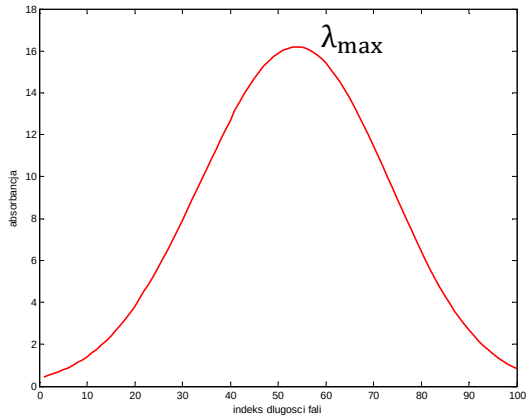
$$A = \varepsilon ab \tag{6}$$

gdzie:
ε – molowy współczynnik absorpcji



Rys. 1 Zakres promieniowania elektromagnetycznego z wyróżnieniem zakresu UV-VIS.

W wyniku absorpcji promieniowania elektromagnetycznego następuje zmiana stanu elektronowego badanej substancji. Rejestrując zachodzące zmiany uzyskuje się elektronowe widma absorpcyjne. Dlatego poprawniej jest mówić o spektrofotometrii UV-VIS. Rejestrowanie zachodzących, pod wpływem promieniowania UV-VIS, zmian w cząsteczce w postaci widma stanowi podstawę analizy jakościowej i ilościowej.
Elektronowe widmo absorpcyjne jest graficznym sposobem zapisu zmian wartości absorbancji w zależności od długości fali przechodzącej przez badany roztwór. Na widmie obserwuje się zazwyczaj jedno maksimum dla konkretnej długości fali, które stanowi podstawę przy oznaczeniach ilościowych substancji w próbce zgodnie z prawem Lamberta-Beera. Na Rys. 2 zaprezentowano teoretyczne widmo z zaznaczonym maksimum absorpcji długości fali.



Rys. 2 Przykładowe widmo otrzymywane za pomocą metody spektrometrii UV-VIS.

Otrzymywane widma są zapisem pochłaniania energii przez chromofor (lub chromofory). A dokładniej tą część cząsteczki, która odpowiada za absorpcję promieniowania, a więc pierścienie aromatyczne i wiązania wielokrotne. Dzięki obecności specyficznych układów wiązań w chromoforach metoda ta pozwala na oznaczenia związków organicznych zawierających w cząsteczce wiązania typu π (np. węglowodory aromatyczne, aldehydy, ketony, kwasy karboksylowe i aminy), a także związków nieorganicznych (np. ozon, tlenek siarki(IV) oraz pierwiastki ziem rzadkich). Oznaczeniom podlegają też związki wykazujące absorpcję w nadfiolecie i absorbujące promieniowanie w zakresie widzialnym, (np. barwne związki organiczne (barwniki), czy barwne sole metali, takie jak manganian(VII) potasu, siarczan(VI) miedzi), a także te substancje, których formy absorbujące promieniowanie w takim zakresie można uzyskać poprzez reakcje kompleksowania. W związku z tym spektrofotometria UV-VIS znajduje praktyczne zastosowanie m.in. podczas oznaczeń kationów metali w formie barwnych związków kompleksowych z ligandami organicznymi [16], w biochemii oraz chemii organicznej. Wykorzystywana jest coraz częściej w badaniach nad DNA [17], [18], [19], pozwala m.in. na monitorowanie procesu denaturacji podwójnej nici DNA pod wpływem temperatury [20]. Okazuje się, że współczynnik absorpcji pojedynczej nici DNA jest znacznie niższy niż dla nici podwójnej. Działając temperaturą na cząsteczkę DNA następuje jej rozpad na dwie pojedyncze nici co skutkuje wzrostem absorbancji. Zastosowanie spektrofotometrii UV-VIS pozwala na wyznaczenie temperatury denaturacji i określenia stopnia przylegania do siebie nici DNA. Badania te są możliwe ponieważ guanina, adenina, tymina i cytozyna ze względu na obecność podwójnych wiązań absorbujących światło w zakresie UV, są dobrymi chromoforami. Technika ta cieszy się również zastosowaniem podczas śledzenia reakcji utleniania i redukcji enzymów (np. NAD^+ do NADH) [12]. W badaniach nad białkami np. przy określaniu zmian w ich konformacji, co z kolei umożliwiają oddziaływania aminokwasów stanowiących swoiste chromofory

z promieniowaniem elektromagnetycznym w zakresie UV. W chemii środowiska wykorzystywana jest m.in. do oznaczenia azotu azotanowego w próbkach wody [21], czy fosforanów w wodzie lub glebie [22]
Jak każda metoda spektrofotometria UV-VIS także posiada ograniczenia, które są związane np. z rejestracją widm. Pochłonięciu ulega jedynie kwant energii o określonej wartości, dlatego otrzymywane widma powinny przedstawiać maksymalnie kilka (w zależności od substancji i obecnych w niej chromoforów) maksimów. Niemniej jednak, widma otrzymywane w praktyce bardzo różnią się od tego teoretycznego obrazu. Relatywnie często pojawia się problem związany z otrzymywaniem 2-3 szerokich pasm nakładających się częściowo na siebie, co uznaje się za wadę metody. Kolejnym problemem mogą być błędy wykonywanych oznaczeń sięgające nawet 30%. Dodatkowo widma uzyskiwane dla podobnych grup molekuł niewiele się różnią od siebie, co skutkuje ograniczeniem metody w oznaczeniach jakościowych. Jednak wedle prawa lamberta Beera absorbancja jest wprost proporcjonalna do stężenia, dlatego metoda spektrofotometrii UV-VIS jest z powodzeniem wykorzystywana w oznaczeniach ilościowych. Warto również wspomnieć o czułości metody, umożliwiającej jej zastosowanie jako sposobu detekcji w metodach separacyjnych [23], dając metody sprzężone np. HPLC-DAD.

3.2 Metody separacyjne

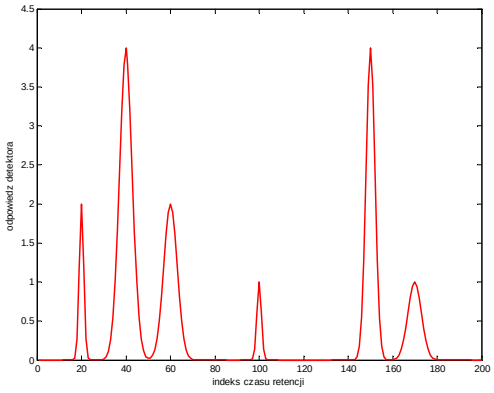
Próbki pochodzenia naturalnego są mieszaninami związków chemicznych. Stąd, sposób ich rozdzielania stał się jednym z podstawowych zagadnień chemii analitycznej. Niezbędne stają się metody umożliwiające selektywne rozdzielenie od siebie wszystkich składników obecnych w złożonych mieszaninach, czy wyizolowanie konkretnego komponentu mieszaniny. Proponuje się tutaj zastosowanie tzw. metod separacyjnych [24], w których rozdział składników mieszaniny oparty jest na wykorzystaniu ich różnych właściwości fizykochemicznych. Metody te, poza wyodrębnieniem poszczególnych komponentów próbki, pozwalają również na ich ocenę ilościową. Równoczesna analiza ilościowa i jakościowa stanowi przewagę nad innymi metodami instrumentalnymi i umożliwia uzyskiwanie dużych ilości informacji o badanym układzie podczas jednej operacji analitycznej.
Metody separacyjne można podzielić na techniki izolacji analitu z matrycy, techniki chromatograficzne i techniki elektromigracyjne.
Do metod izolacji analitu z matrycy próbki zalicza się metody ekstrakcyjne, np. ekstrakcja rozpuszczalnikiem, ekstrakcja do fazy stałej, mikroekstrakcja. Drugi typ metod, a więc metody chromatograficzne, tworzą najobszerniejszą grupę metod separacyjnych. Należą do niej chromatografia planarna, cieczowa, gazowa, wysokosprawna chromatografia cieczowa i ich odmiany chromatografia adsorpcyjna, podziałowa, powinowactwa, wykluczenia i jonowymienna. Do trzeciego typu metod

zaliczamy elektroforezę kapilarną oraz planarną, micelarną chromatografię elektrokinetyczną, kapilarne ogniskowanie izoelektryczne i izotachoforezę kapilarną.
Obecnie, najczęściej stosowanymi metodami są metody chromatograficzne. Przede wszystkim ze względu na możliwość relatywnie łatwego oznaczania składu ilościowego i jakościowego badanych materiałów. Ponadto, są to metody analityczne oraz preparatywne, które można wykorzystać w celu izolacji czystych substancji z mieszaniny.

3.2.1 Chromatografia

Chromatografia jest fizyczną metodą rozdzielania substancji, w której składniki ulegają podziałowi pomiędzy dwie niemieszające się ze sobą fazy: fazę stacjonarną i fazę ruchomą. Podziału metod chromatograficznych dokonuje się ze względu na stan skupienia fazy ruchomej, mechanizmu rozdzielania substancji i sposób prowadzenia procesu chromatograficznego [25]. Ze względu na stan skupienia fazy ruchomej wyróżnia się: chromatografię gazową (z ang. Gas Chromatography, GC) [26], będąca techniką w której jako fazę ruchomą stosuje się gaz. Następnie, chromatografię cieczową (z ang. Liquid Chromatography; LC) [27], gdzie fazą ruchomą jest ciecz. Wymienić tu warto także wysokosprawną chromatografię cieczową (z ang. High Performance Liquide Chromatography, HPLC) [28], chromatografię adsorpcyjną, podziałową, jonową, wykluczania i powinowactwa. W klasyfikacji ze względu na sposób prowadzenia procesu chromatograficznego wyróżnia się chromatografię: kolumnową, planarną, analityczną oraz preparatywną.
Efekt rozdziału badanej próbki w toku procesu chromatograficznego zapisuje się w postaci wykresu przedstawiającego zmiany stężenia substancji w fazie ruchomej opuszczającej kolumnę i przechodzącej przez detektor. Najczęściej zmiany te przedstawia się w funkcji czasu. Wykres ten nazywa się chromatogramem (Rys. 3), gdzie na osi x przedstawia się czas retencji, a na osi y odpowiedź detektora, przedstawianą w postaci pików odpowiadających substancjom występującym w próbce. Chromatogram jest niezwykle ważnym źródłem informacji. Informuje przede wszystkim o składzie jakościowym analizowanej próbki, gdyż identyfikacji związków dokonuje się na podstawie czasu retencji przy którym pojawia się pik danej substancji, ale również o składzie ilościowym rozdzielanych składników próbki, co wynika z dokonywania oceny ilościowej na podstawie pola powierzchni piku.
Metody chromatograficzne ze względu na relatywnie niski koszt analizy oraz krótki czas prowadzonych oznaczeń znajdują zastosowanie w wielu dziedzinach nauki. Właściwe ciężko sobie wyobrazić przeprowadzenie analizy ilościowej i jakościowej bez ich zastosowania. Wykorzystywane są w kryminalistyce [29], [30], laboratoriach kosmetycznych [31], czy chemii przemysłowej np. podczas oznaczenia zawartości polichlorowanych bifenyli (PCB) w różnorodnych materiałach [32]. W monitoringu środowiska przy oznaczeniach pestycydów w żywności [33], w ocenie jakości wód

pitnych [34], czy czystości powietrza [35]. Stają się podstawowym narzędziem w branży farmaceutycznej. Pozwalają m.in. na wyizolowanie czystych związków, czy ocenę zawartości poszczególnych komponentów w próbkach leków. Umożliwiają separację chiralnych komponentów leków różniących się ułożeniem poszczególnych grup atomów w przestrzeni. Jest to niezwykle istotne, ze względu na wykazywanie odmiennej aktywności biologicznej przez enancjomery (np. talidomid, gdzie jeden z enancjomerów jest teratogenny [36]). Chromatografia preparatywna umożliwiająca izolację czystych molekuł wzorcowych np. białek, wykorzystywana jest w biochemii i w biofarmacji. Coraz częściej z zaplecza metod chromatograficznych korzysta się w naukach typu – mika, tj. proteomika [37], czy metabolomika. Ponieważ pozwalają np. na śledzenie zmian w profilach metabolicznych, separacje komponentów leków zawartych płynach biologicznych, wykrywanie metabolitów wskazujących na rozwój jednostki chorobowej [38], itp.
W kontroli jakości produktów np. kosmetycznych [39], czy spożywczych umożliwiają detekcję dodatków do żywności, konserwantów, białek, czy witamin [40].



Rys. 3 Przykładowy chromatogram o sześciu pikach odpowiadających sześciu substancjom występujących w analizowanej próbce.

3.3 Instrumentalne metody sprzężone

W celu uzyskania pełniejszego opisu analizowanej próbki, coraz częściej metody separacyjne zestawia się z zaawansowanymi wielokanałowymi metodami detekcji. Ostatecznie uzyskuje się metody sprzężone, które łączą zalety co najmniej dwóch technik. Najczęściej, metody sprzężone są kombinacją metod chromatograficznych z metodami spektroskopowymi [2]. Do proponowanych rozwiązań zalicza się połączenie wysokosprawnej chromatografii cieczowej ze spektrofotometrią UV-VIS

z wykorzystaniem detektora DAD (HPLC-DAD) [41], chromatografię gazową sprzężoną ze spektrometrią mas (GC-MS) [42], czy chromatografię cieczową sprzężoną z jądrowym rezonansem magnetycznym (LC-NMR) [43] lub spektrometrią mas (LC-MS) [44]. Metody sprzężone nie zawsze stanowią połączenie wyłącznie dwóch metod. Można połączyć ze sobą więcej niż jedną metodę separacyjną i metodę spektroskopową. Wymienić tu warto takie połączenia metod jak LC-MS-MS, czy LC-NMR-MS.

Tak jak już wcześniej wspomniano chromatografia umożliwia separację czystych komponentów próbki, z kolei metody spektroskopowe pozwalają na uzyskanie selektywnej informacji w postaci widm porównywanych z wzorcami lub informacjami zawartymi w bibliotekach. Dzięki temu, techniki łączone są selektywne wobec oznaczanych analitów, czułe w szerokim zakresie stężeń oraz pozwalają na identyfikację poszczególnych komponentów próbki.

Zastosowanie technik sprzężonych w badaniach umożliwia pozyskiwanie kompleksowej informacji o badanych materiałach. Ma to odbicie w rosnącym zainteresowaniu tymi technikami w analizie materiałów różnorodnego pochodzenia. W ostatnich latach techniki sprzężone wykorzystywane są w rozwiązywaniu złożonych problemów analitycznych. Umożliwiają m.in. ocenę jakościową i ilościową oraz identyfikacje komponentów próbek naturalnych ekstraktów roślinnych [45], ziół [46], próbek środowiskowych [47], kontroli jakości produktów. Znajdują zastosowanie w biologii, biochemii, biomedycynie, farmacji [48], fitochemii [49], chemotaksonomii [50], czy metabolomice [51] do identyfikacji metabolitów wtórnych, w wyznaczaniu profili metabolicznych oraz wielu innych dziedzinach naukowych oraz branży przemysłowej. Swą popularność metody sprzężone zawdzięczają m.in. redukcji czasu analizy, minimalizacji objętości próbek, automatyzacji, czy wykluczeniu etapu przygotowania próbek do analizy, dzięki czemu dają one możliwość oznaczania substancji w surowych próbkach naturalnych.

Poza licznymi zaletami, metody sprzężone tak jak wszystkie techniki badawcze posiadają pewne ograniczenia. Jednym z nich jest kompleksowość uzyskiwanych danych, która zdecydowanie utrudnia interpretację oraz wysuwanie generalnych wniosków. Wyniki pozyskiwane przy pomocy technik sprzężonych otrzymuje się w postaci tzw. odcisku palca (z ang. fingerprint), a analiza tego typu danych wymaga zastosowania narzędzi chemometrycznych.

3.4 Ograniczenia metod instrumentalnych

Wszystkie metody instrumentalne poza licznymi zaletami mają także swoje ograniczenia. Pojawiają się one między innymi, dlatego że każdy sygnał analityczny składa się z trzech komponentów tj. szumu, linii podstawowej oraz sygnału właściwego [52]. Pierwsze dwa są niekorzystne i utrudniają odczyt informacji opisywanej przez otrzymywane sygnały. W związku z tym istnieje szereg metod chemometrycznych

poprawiających jakość uzyskanych sygnałów [53] poprzez korektę linii podstawowej np. za pomocą metody asymetrycznych najmniejszych kwadratów z funkcją kary [54], eliminację szumu za pomocą tzw. binningu [55] lub metod odszumiania [56]. Równie często stosuje się metody poprawiające stosunek sygnału do szumu. Kolejnym napotykanym problemem jest wymiarowość danych. Poza zaletami związanymi z otrzymywaniem kompletnej informacji o właściwościach fizykochemicznych wraz z uwzględnieniem składu jakościowego i ilościowego badanych próbek, uzyskanie danych, gdzie każda próbka jest opisana za pomocą kilku tysięcy zmiennych, przyczynia się do utrudnienia interpretacji i formułowania wniosków. Wynika to m.in. z problemu wizualizacji tego typu danych. Należy również nadmienić, iż dane złożone chemicznie zawierają zazwyczaj skorelowane zmienne lub zmienne nieistotne, które nie wnoszą istotnej informacji do analizy danych. Stąd zaproponowano wiele metod pozwalających na selekcje i usuwanie zmiennych, których obecność jest zbędna dla powodzenia analizy danych i wyłącznie utrudnia odczyt istotnej chemicznie informacji [57]. Pomimo stosowania metod selekcji zmiennych ich liczba wciąż jest na tyle wysoka, że konieczne staje się zastosowanie metod eksploracji oraz analizy danych, które zostały opisane w kolejnych rozdziałach niniejszej pracy. Wielowymiarowość danych jest szczególnie problematyczna w przypadku danych uzyskanych metodami sprzężonymi, ponieważ wyniki wzbogacone są o dodatkowy wymiar. Następnym problemem z jakim należy się zmierzyć są przesunięcia pików wywołane nieznacznymi fluktuacjami warunków pomiarowych. W celu poprawienia jakości uzyskanych sygnałów korzysta się z pakietu metod nakładania sygnałów na siebie [58]. Kolejne zagadnienie związane jest z koelucją substancji, o czym wspomniano w podrozdziale *12.2.1*. Jest to dość powszechny problem spotykany w przypadku metod chromatograficznych. Przyczyną tego typu zjawiska są problemy związane z niewystarczającą rozdzielczością stosowanej aparatury. Współwymywanie substancji stanowi istotny problem, zwłaszcza w przypadku danych biologicznych np. z zakresu metabolomiki, gdzie koelucja może wpłynąć na niewyodrębnienie wskaźników biologicznych odpowiedzialnych za rozwój choroby. Również i w tym przypadku wskazane jest skorzystanie z metod matematycznych pozwalających na polepszenie rozdziału i weryfikację, czy dany pik sygnału instrumentalnego reprezentuje jedną czy dwie substancje.

Wśród ograniczeń związanych z metodami instrumentalnymi należy wymienić także ograniczenia typowo aparaturowe, wynikające z konstrukcji urządzeń, takich jak niewystarczająca rozdzielczość, czy zjawisko rozpraszania światła.

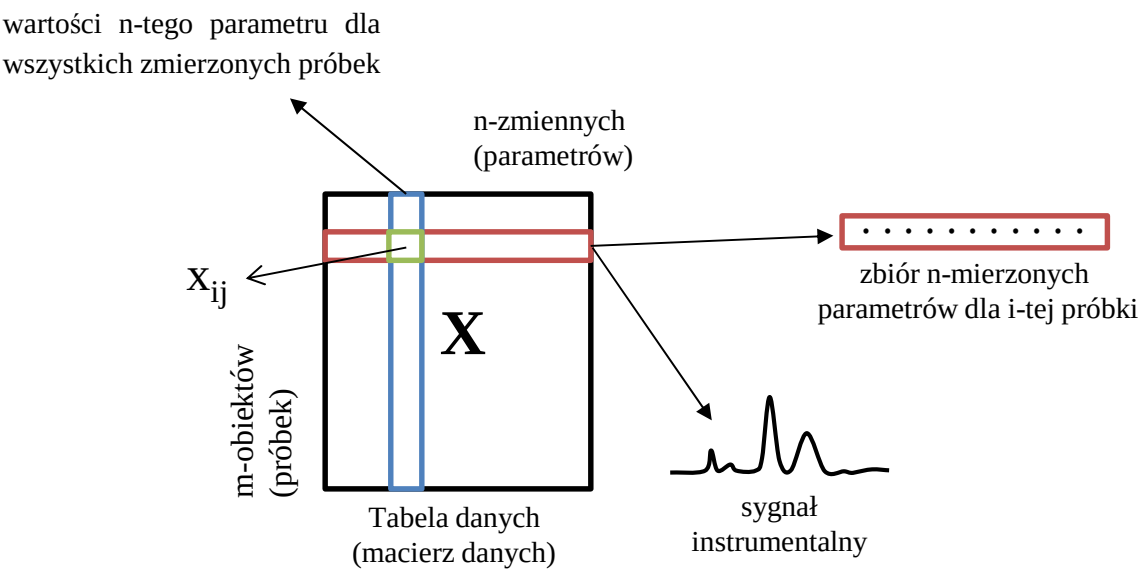
4. Struktura danych eksperymentalnych

Opis analizowanych próbek za pomocą wielu parametrów prowadzi do uzyskania danych wielowymiarowych, określanych również danymi wieloparametrowymi lub złożonymi. Dane takie reprezentuje się zazwyczaj za pomocą macierzy danych (tablicy

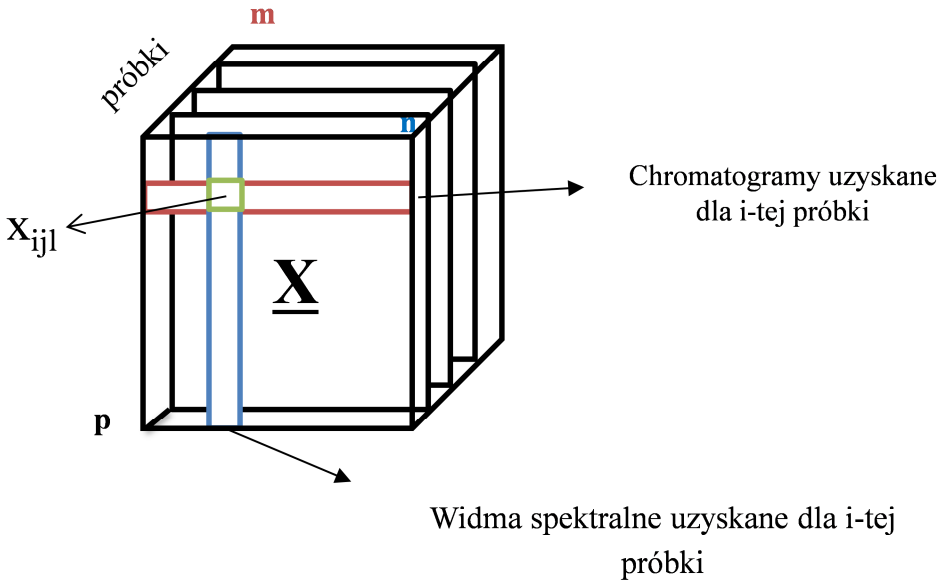
danych) **X** o wymiarowości m×n, gdzie m wierszy macierzy przedstawia obiekty, a n kolumn zmierzone parametry (Rys. 4). Terminy obiekty i próbki to synonimy, niemniej jednak terminu obiekty poprawniej jest używać w przypadku wykonania pomiaru dla tej samej próbki, ale w różnych odstępach czasu, tak aby zarejestrować zachodzące w niej zmiany. Z kolei parametry można również nazywać zmiennymi. W zależności od wykorzystanej techniki, bądź technik analitycznych, do pozyskania informacji o badanych materiałach, wiersze macierzy danych mogą tworzyć albo sygnały instrumentalne, takie jak widma UV-VIS zmierzone w określonym zakresie spektralnym, chromatogramy, czy widma masowe, ale mogą to być również wektory o n elementach, reprezentujące wyniki przeprowadzonych analiz. Z takim przypadkiem można się spotkać podczas określania wybranych właściwości fizycznych lub chemicznych badanych układów np. stężenia poszczególnych komponentów w próbce, pH roztworu czy temperatury. Każdą próbkę reprezentuje się jako punkt w przestrzeni zdefiniowanej przez wartość pewnej liczby zmierzonych parametrów, natomiast każdy parametr jako punkt w przestrzeni zdefiniowanej przez wartości tego parametru dla pewnej liczby analizowanych próbek. Między innymi właśnie stąd wynikają trudności z analizą danych wielowymiarowych. Jeżeli liczba próbek m i/lub parametrów n jest większa niż 3, wówczas zostajemy pozbawieni możliwości ich wizualizacji, będącej najdogodniejszą formą pozyskiwania i odczytu informacji o badanych materiałach i zjawiskach [59].

Poza typową strukturą dwuwymiarową, pozyskiwane dane mogą mieć także trójwymiarową organizację. Dane takie otrzymuje się w przypadku zastosowania metod sprzężonych, gdzie jeden z wymiarów przedstawia np. długość fali, drugi czas retencji a trzeci liczbę analizowanych próbek. W konsekwencji każda próbka scharakteryzowana jest za pomocą tablicy, gdzie w zależności od stosowanych metod w wierszach znajdują się, np. chromatogramy, a w kolumnach widma spektroskopowe. W takim wypadku dane przedstawia się w postaci prostopadłościanu – tensora.

Przy okazji omawiania struktury danych wielowymiarowych warto zwrócić uwagę, iż wszystkie wyniki zawarte w macierzy danych **X** są wypadkową dwóch komponentów: sygnału analitycznego oraz błędu eksperymentalnego.



Rys. 4 Macierz danych eksperymentalnych **X** zawierająca m próbek opisanych przez n parametrów.



Rys. 5 Tensor danych eksperymentalnych **X** o wymiarowości p ×n× m.

5. Wstępne przygotowanie danych do dalszej analizy

Przygotowanie danych do dalszej analizy jest ważnym etapem wpływającym na efektywność ich eksploracji i modelowania [60]. Celem wstępnego przygotowania danych do dalszej analizy jest korekta lub eliminacja niepożądanych efektów fizycznych związanych z obecnością szumu instrumentalnego, błędów pomiarowych, przesunięć pików spowodowanych wpływem nieznacznych zmian zewnętrznych parametrów towarzyszących pomiarom (np. ciśnienie, temperatura), ale również wewnętrznych (np. pH). Jej dobór jest uzależniony od typu analizowanych danych, a zatem czy obiektem wstępnego przygotowania są sygnały instrumentalne, tablica pików, czy elementy reprezentujące wyniki n analiz. Metody te można podzielić na dwie grupy. Do pierwszej z nich należą techniki modyfikacji indywidualnych zmiennych, działające na kolumnach macierzy **X**, takie jak techniki transformacji logarytmicznej, centrowania (7) i skalowania danych np. autoskalowanie (8), będące szczególnym przypadkiem binarnego ważenia zmiennych [59]. Najpowszechniej stosowanym sposobem obróbki danych jest centrowanie [61]. Polega ono na usuwaniu wartości średniej każdej kolumny macierzy danych od poszczególnych elementów tej kolumny. Taki zabieg pozwala na przesunięcie danych do początku układu współrzędnych. Drugim zalecanym sposobem przygotowania danych do dalszej analizy jest autoskalowanie. Sosowane jest przede wszystkim w celu ujednolicenia jednostek.

$$x_c = (x_{ij} - \bar{x}) \tag{7}$$

gdzie:
x_c – element x_{ij} po centrowaniu
x_{ij} – element macierzy **X** występujący w i-tym wierszu i j-tej kolumnie
$\bar{x}$ – wartość średnia j-tej kolumny

$$x_a = \frac{(x_{ij} - \bar{x})}{\mathrm{std}\,(x_j)} \tag{8}$$

gdzie:
x_a – element macierzy o współrzędnych i i j po autoskalowaniu
x_{ij} – element macierzy **X** występujący w i-tym wierszu i j-tej kolumnie
$\bar{x}$ – wartość średnia danej kolumny
$\mathrm{std}(x_j)$ – odchylenie standardowe j-tej kolumny, wyrażane wzorem:

$$\mathrm{std}(x_j) = \sqrt{\frac{1}{m-1}\sum_{i=1}^{m}\left(x_{ij} - \bar{x}_j\right)^2}$$

Zdarza się bowiem, że zmierzone dla próbek zmienne reprezentowane są za pomocą różnych jednostek. Aby uniknąć związanych z tym błędów zaleca się dane poddać autoskalowaniu. Efektem czego, jest możliwość porównania ze sobą próbek reprezentowanych przez poszczególne zmienne. Autoskalowanie polega na centrowaniu, a następnie podzieleniu każdego elementu danej kolumny przez jej odchylenie standardowe. Po autoskalowaniu odchylenie standardowe kolumny wynosi 1. Ze względu na nadanie tej samej wagi wszystkim zmiennym macierzy, autoskalowania nie stosuje się w przypadku sygnałów instrumentalnych zawierających szum.

Drugą grupę metod tworzą algorytmy modyfikacji indywidualnych obiektów macierzy, np. sygnałów instrumentalnych. Są to przede wszystkim techniki eliminacji szumu (np. transformacja Fouriera [62]) oraz korekty linii podstawowej (np. metoda asymetrycznych najmniejszych kwadratów z funkcją kary, (z ang. Asymmetric Least Squares; ALS [54]), procedury normalizacyjne, tworzenia pochodnych oraz metody nakładania sygnałów instrumentalnych (np. metoda nakładania widm maksymalizująca ich wzajemna korelację, z ang. Correlation Optimized Warping, COW [63]). Szczególnie istotne wydają się metody nakładania sygnałów instrumentalnych, ponieważ dopóki sygnały nie mają tej samej długości, a odpowiadające sobie zmienne nie zajmują tych samych miejsc w kolumnach macierzy danych **X**, nie można wykorzystać metod eksploracji danych. Przesunięcia pików, będące następstwem fluktuacji warunków pomiarowych tj. pH roztworów, ciśnienia, temperatury, czy homogeniczności pola magnetycznego, tak jak ma to miejsce w przypadku metody NMR, należą do podstawowych problemów rejestracji sygnałów instrumentalnych. Dlatego też narzędzia umożliwiające ich poprawne nałożenie są niezbędnym elementem wstępnej obróbki danych. Istnieje wiele algorytmów nakładania sygnałów jednak najczęściej stosowana jest wspomniana metoda COW, czy metoda dynamicznego nakładania sygnałów instrumentalnych (z ang. Dynamic Time Warping; DTW) [64] oraz metoda parametrycznego nakładania sygnałów instrumentalnych (z ang. Parametric Time Warping) [65].

6. Określanie podobieństwa występującego w danych eksperymentalnych

Pojęcie podobieństwa jest podstawowym terminem wykorzystywanym w życiu codziennym. Człowiek w naturalny sposób klasyfikuje otaczające go obiekty na podstawie przyjętych kryteriów takich jak np. kolor, smak, zapach czy kształt. Również w przypadku danych eksperymentalnych pojęcie podobieństwa stanowi podstawę w wyodrębnieniu obiektów lub parametrów wykazujących zbliżone właściwości fizyczne, chemiczne, czy biologiczne. Określenie relacji pomiędzy obiektami stanowi podstawę podczas procesu interpretacji danych oraz formułowania ostatecznych konkluzji. Niemniej jednak, w przypadku wielowymiarowych danych, porównywanie

poszczególnych obiektów ze sobą, aby określić ich podobieństwo, wymaga zastosowania odpowiednich narzędzi. Dlatego, w celu wyodrębnienia grup obiektów niezbędne jest wprowadzenie kryterium wyrażającego podobieństwo próbek i/lub parametrów w przestrzeni eksperymentalnej. Sprowadza się to przede wszystkim do określenia odległości pomiędzy obiektami. Zgodnie z zasadą, że obiekty leżące blisko siebie w przestrzeni eksperymentalnej są do siebie bardziej podobne aniżeli te które w tej przestrzeni są od siebie znacznie oddalone. Z tego powodu niezbędne są matematyczne miary pozwalające na określenie dystansu pomiędzy poszczególnymi obiektami lub parametrami. W literaturze opisano ok. 60 miar odległości, które z powodzeniem mogą zostać wykorzystane w celu określenia odległości pomiędzy obiektami reprezentowanymi w przestrzeni parametrów (lub odwrotnie), a jej dobór zależy od formy reprezentacji danych, a więc czy ma się do czynienia z danymi dyskretnymi, ciągłymi, binarnym lub inną formą ich matematycznego zapisu. Uogólniając, termin odległość jest numerycznym sposobem opisu dystansu pomiędzy obiektami w przestrzeni eksperymentalnej.
Na wartość odległości/podobieństwa obiektów wpływają następujące czynniki:

1) sposób opisu obiektów przez zmierzone parametry,
2) schemat ważenia elementów,
3) wybrana miara odległości.

Wśród wymienionych czynników decydujący wpływ ma wybór odpowiedniej miary odległości. Jest to trudne, między innymi ze względu na dużą liczbę dostępnych sposobów jej wyznaczania. Ponadto, każda z dostępnych miar umożliwia odkrycie różnych ukrytych w danych źródeł informacji nie ujawnianych przez pozostałe miary odległości.
Miary odległości wykorzystywane są w każdej metodzie grupowania danych, a także innych licznych metodach chemometrycznych. Ze względu na sposób określania odległości metody te można sklasyfikować na takie, w których odległość określa się pomiędzy:

1) obiektami lub parametrami (np. metody grupowania danych, analiza czynników głównych),
2) obiektami i wybranym punktem odniesienia (np. metody klasyfikacji),
3) dwoma zestawami danych (np. kanoniczna analiza korelacji).

Miary odległości/podobieństwa wykorzystywane są np. w metodzie k-średnich do przypisywania obiektów do poszczególnych grup na podstawie ich odległości mierzonej względem środka grupy lub w metodach hierarchicznych do wyznaczania odległości pomiędzy poszczególnymi obiektami, co stanowi podstawę przy konstrukcji dendrogramu. Kolejnym przykładem zastosowania miar odległości są samoorganizujące się mapy Kohonena, w których odległość wykorzystywana jest do tworzenia sieci neuronowych. Innym przykładem zastosowań jest wykorzystanie koncepcji podobieństwa w metodzie analizy czynników głównych, co stanowi

podstawę podczas eksploracji danych, czy w analizie dyskryminacyjnej w celu określenia kowariancji każdej klasy.
Terminów odległość i podobieństwo często używa się jako synonimów. Jednak ich matematyczna definicja jest różna. Funkcja D: $\mathbf{X}\times\mathbf{X} \rightarrow \mathbb{R}$ uznana jest za *odległość*, gdy dla zbioru danych **X** oraz x, y $\in$ **X** spełnione zostają następujące warunki:

1)	$D_{xy} \geq 0$	warunek pozytywności
2)	$D_{xx} = 0$	warunek refleksji
3)	$D_{xy} = D_{yx}$	warunek symetrii

Zdarza się jednak, że spełnione zostają tylko warunki 1 i 2. Wówczas mówi się o tzw. quasi-odległości.
Rozszerzając warunki jakie musi spełniać *odległość* można zdefiniować **miarę odległości** dla wszystkich x, y, z:

1')	$D_{xy} \geq 0$	warunek pozytywności
2')	$D_{xy} = 0$, jeśli x = y	warunek silnej refleksji
3')	$D_{xy} = D_{yx}$	warunek symetrii
4')	$D_{xy} \leq D_{xz} + D_{zy}$	warunek nierówności trójkąta

Tak jak w przypadku odległości nie wszystkie powyższe warunki muszą być spełnione przez rozważaną funkcję odległości. Z tego powodu można wyróżnić:

- pseudo-miarę odległości, jeśli nie zostaje spełniony warunek silnej refleksji, a jedynie refleksji,
- quasi-miarę odległości, która nie spełnia warunku symetrii funkcji odległości.

Można również rozważyć przypadek ultra-miary, tj. odległości która spełnia warunki 1-3 określone dla miary odległości oraz warunek nierówności ultramiary:

1'')	$D_{xy} \geq 0$	warunek pozytywności
2'')	$D_{xy} = 0$, jeśli x = y	warunek silnej refleksji
3'')	$D_{xy} = D_{yx}$	warunek symetrii
4'')	$D_{xy} \leq \max\{D_{xz}, D_{zy}\}$	warunek nierówności ultramiary

Natomiast terminu *podobieństwo* należy używać gdy funkcja S: $\mathbf{X}\times\mathbf{X} \rightarrow \mathbb{R}$ dla **X**, gdzie x, y $\in$ **X**, spełnia następujące warunki:

1''')	$D_{xy} \geq 0$	warunek pozytywności
2''')	$D_{xx} = 0$	warunek identyfikacji
3''')	$D_{xy} = D_{yx}$	warunek symetrii

Natomiast, **miara podobieństwa** dodatkowo przyjmuje wartości z przedziału [0,1], spełniając tym samym warunek ograniczenia:

$0 \leq S_{xy} \leq 1$ warunek ograniczenia

gdzie, wartość 1 oznacza idealne podobieństwo pomiędzy obiektami, a wartość 0 jej całkowity brak.
Za pomocą odpowiednich transformacji można przejść od miar odległości do miar podobieństwa. Rodzaj transformacji umożliwiającej takie przejście jest uzależniony od tego czy miara odległości z jaką mamy do czynienia należy do ograniczonych, czy nieograniczonych. W niektórych przypadkach wartości odległości są ograniczone do 1 w innych przyjmują dowolne wartości wyznaczone podczas obliczeń. Jednak poprzez zastosowanie normalizacji i metod skalowania danych, każda miara odległości może zostać ograniczona do 1. Transformacje proponowane dla odległości, których wartości są ograniczone do 1 są następujące:

1) $S_{xy} = 1 - D_{xy}$
2) $S_{xy} = 1 - (D_{xy})^2$
3) $S_{xy} = \sqrt{1 - (D_{xy})^2}$

Z kolei transformacje proponowane dla odległości nieograniczonych wyraża się następująco:

1) $S_{xy} = \frac{1}{1+D_{xy}}$

2) $S_{xy} = 1 - \frac{D_{xy}}{D_{max}}$

3) $S_{xy} = e^{-D_{xy}}$

Można również przeprowadzić transformację odwrotną przeprowadzając podobieństwo w odległość poprzez zastosowanie dowolnej transformacji monotonicznej.
Zanim zastosuje się wybraną miarę podobieństwa dane powinno się poddać wstępnemu przygotowaniu do dalszej analizy, co przedstawiono w rozdziale 5 niniejszej pracy. Etap ten ma na celu przede wszystkim ujednolicić dane tak, aby zastosowana miara podobieństwa, reprezentowała najwiarygodniej relacje pomiędzy obiektami, czy parametrami. Przede wszystkim chodzi o ujednolicenie jednostek w jakich wyrażane są mierzone parametry, ponieważ skala za pomocą której reprezentuje się analizowane wyniki ma decydujący wpływ na obserwowane efekty określania odległości. Pominięcie etapu skalowania danych może doprowadzić do uzyskania błędnych wyników.

Tak jak wspominano, wachlarz dostępnych miar podobieństwa jest relatywnie szeroki. Może to utrudniać wybór najodpowiedniejszej miary, pozwalającej określić podobieństwo obiektów i tym samym ujawnić ukrytą strukturę danych. Pomocne może się okazać określenie typu danych z jakimi ma się do czynienia: binarne, realne wartości, częstotliwość, itp. Dane te są rozróżniane na podstawie rodzaju parametrów charakteryzujących obiekty. Przykładowymi miarami odległości dla danych typu rzeczywistych wartości (z ang. Real Data), np. intensywność sygnału, aktywność biologiczna, stężenie, temperatura, itp., są odległość euklidesowa, Mahalanobisa, Manhattan, jako przykłady miar odległości nieograniczonych. Z kolei współczynnik korelacji reprezentuje ograniczoną miarę podobieństwa dla tego typu danych. Dla danych uwzględniających rangę obiektów i parametrów (z ang. Rank Data) przykładem stosowanych w ich eksploracji miar odległości jest odległość Spearmana, a dla danych określających częstotliwość występowania poszczególnych składowych danych (z ang. Frequency Data) np. wydarzeń, najpopularniejszą miarą odległości jest odległość Tanimoto. W przypadku danych binarnych można skorzystać z wielu współczynników podobieństwa wśród których wymienić należy np. współczynnik Jaccarda-Tanimoto [66]. Należy jednak cały czas mieć na uwadze, iż na określenie odległości pomiędzy obiektami wpływają metody wstępnego przygotowania danych do dalszej analizy, gdyż istnieją miary odległości które są bardzo czułe na wszelkie transformacje danych oraz liczbę zmiennych definiujących przestrzeń pomiarową.
W podrozdziale *6.1* bliżej omówiono kilka miar odległości cieszących się największą popularnością w przypadku analizy danych reprezentujących tzw. rzeczywiste wartości. Ponieważ poprzez odpowiednie transformacje można otrzymać z miar odległości odpowiednie miary podobieństwa i odwrotnie to, dla uproszczenia, w niniejszej pracy terminy te będą wykorzystywane zamiennie, zgodnie z ogólnie przyjętą koncepcją.

6.1 Wybrane miary podobieństwa

Jak już wcześniej wspomniano, istnieje wiele miar odległości. Jednak tylko nieliczne są powszechnie stosowane podczas weryfikacji podobieństwa obiektów. Poniżej omówiono te, które stanowią punkt odniesienia dla rozwoju pozostałych miar podobieństwa. Wymienione w kolejnych podrozdziałach odległość euklidesowa, odległość Mahalanobisa, czy współczynnik korelacji Pearsona, umożliwiają ocenę podobieństwa obiektów lub parametrów przy pomocy prostych zależności wynikających z algebry liniowej. Dodatkowo, należą one do miar podobieństwa, w których podobieństwo obiektów oceniane jest zgodnie z intuicją.

6.1.1 Odległość euklidesowa

Najczęściej stosowaną miarą odległości jest odległość euklidesowa [67]. Wynika ona bezpośrednio z twierdzenia Pitagorasa i można ją opisać równaniem (9). Największą zaletą tej miary odległości jest łatwość interpretacji wyników. Dodatkowo, odległość euklidesowa nie zależy od transformacji danych (np. translacja, rotacja). Z kolei jako wadę uznaje się nadawanie dużej wagi tym zmiennym, które mają duże wartości w przestrzeni zdefiniowanej przez wartości parametrów. Odległość pomiędzy dwoma punktami **x** oraz **y** w przestrzeni n parametrów zdefiniowano równaniem (9).
Wyniki obliczeń odległości euklidesowej pomiędzy poszczególnymi próbkami można zestawić w tzw. macierz odległości, która na diagonali zawiera wartości zerowe, co wynika z faktu że odległość pomiędzy tymi samymi próbkami wynosi 0:

$$d_{\mathbf{xy}} = \sqrt{\sum_{j=1}^{n}(x_j - y_j)^2} \tag{9}$$

gdzie:
$d_{\mathbf{xy}}$ – odległość euklidesowa pomiędzy obiektami **x** i **y** w przestrzeni n parametrów
x_j – j-ty element obiektu **x**
y_j – j-ty element obiektu **y**

0	0,9443	0,7454	0,9112	0,6681
0,9443	0	0,7025	0,9447	1,0074
0,7454	0,7025	0	1,1740	0,6160
0,9112	0,9447	1,1740	0	0,9068
0,6681	1,0074	0,6160	0,9068	0

Rys. 6 Macierz odległość reprezentująca odległość euklidesową pomiędzy poszczególnymi próbkami (wierszami) macierzy **X** zawierającej pięć wierszy.

Szczególnym przypadkiem tej miary podobieństwa jest kwadrat odległości euklidesowej [68], w której pomija się pierwiastkowanie:

$$d_{\mathbf{xy}}{}^2 = \sum_{j=1}^{n}(x_j - y_j)^2 \quad (10)$$

6.1.2 Odległość Mahalanobisa

Kolejną, równie często stosowaną, miarą odległości jest odległość Mahalanobisa [69]. Zaletą tej miary jest uwzględnienie korelacji występującej w danych, przez wprowadzenie macierzy wariancji-kowariancji, dzięki czemu uwzględniony zostaje rozkład obiektów przestrzeni pomiarowej. Niemniej jednak, uwzględnienie macierzy wariancji-kowariancji w obliczaniu odległości obiektów prowadzi również do ograniczenia jej zastosowalności. Ponieważ, odległości Mahalanobisa nie można zastosować w przypadku danych o dużej liczbie zmiennych, takich jak np. widma uzyskane za pomocą kamery NIR. Wynika to z obecności skorelowanych zmiennych, które nie wnoszą istotnej informacji w całokształt analizy, czy eksploracji. Obecność tego typu zmiennych, zgodnie z założeniami algebry liniowej, prowadzi do otrzymania macierzy osobliwej, której wyznacznik jest równy zero, co uniemożliwia wyznaczanie macierzy odwrotnej. Oznacza to, że liczba obiektów macierzy musi przewyższać liczbę zmiennych, a zmienne te nie mogą być skorelowane. W konsekwencji sprowadza się to do zastosowania metod wyboru zmiennych zanim przystąpi się do obliczenia odległości pomiędzy obiektami.
Zaletą tej miary podobieństwa jest niezależność jej wyników od skalowania danych, czego nie obserwuje się w przypadku większości miar podobieństwa. Dodatkowo można ją zastosować w celu detekcji obiektów odległych [69].
Punkty oddalone od środka grupy o stałą wartość odległości Mahalanobisa tworzą w przestrzeni hipersferę, a w przestrzeni dwuwymiarowej elipsę. Z kolei w przypadku odległości euklidesowej punkty te utworzą kulę, a w przestrzeni dwuwymiarowej okrąg. Jak łatwo się domyśleć, przyczyną tej różnicy jest uwzględnienie macierzy wariancji-kowariancji przy obliczaniu odległości Mahalanobisa.
Odległość Mahalanobisa pomiędzy dwoma obiektami **x** i **y** można zdefiniować następująco:

$$dM_{\mathbf{xy}} = \sqrt{\sum_{j=1}^{n}(x_j - y_j)^2 \mathbf{C}^{-1}} \quad (11)$$

gdzie:
$dM_{\mathbf{xy}}$ – odległość Mahalanobisa pomiędzy obiektami **x** i **y**
x_j – j-ty element obiektu **x**
y_j – j-ty element obiektu **y**
$\mathbf{C}^{-1}$ – odwrotna macierz macierzy wariancji-kowariancji

W metodach grupowania danych odległość Mahalanobisa jest wykorzystywana w celu wyznaczenia środków grup. Jednak należy pamiętać, że odległość ta nie sprawdza się w przypadku danych wysoce skorelowanych, które uniemożliwiają wyznaczenie macierzy odwrotnej z macierzy wariancji-kowariancji.

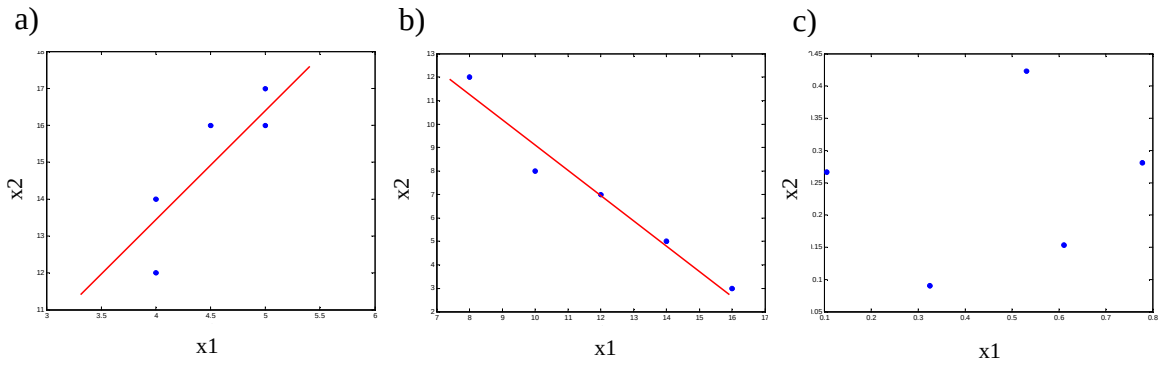
6.1.3 Współczynnik korelacji Pearsona

Podobieństwa można się również doszukiwać pomiędzy parametrami. W tym celu uwzględnia się zależność parametrów od siebie. Stosowaną miarą ich podobieństwa jest współczynnik korelacji Pearsona (R) [59]. Pokazuje on w jakim stopniu dwie zmienne są od siebie liniowo zależne, poprzez wyznaczenie cosinusa kąta pomiędzy wektorami reprezentującymi poszczególne parametry. Wartość współczynnika korelacji waha się od -1 do 1 i oznacza odpowiednio korelację przeciwną oraz korelację dodatnią. Współczynnik korelacji dla pary parametrów definiuje się poniższym wzorem:

$$R = \frac{(x_1 - \bar{x}_1)(x_2 - \bar{x}_2)}{\sqrt{std(x_1)std(x_2)}} \tag{12}$$

gdzie:
x_1, x_2 – zmienna pierwsza i druga
$\bar{x}_1$, $\bar{x}_2$ – średnia wektora odpowiednio pierwszego i drugiego
std(x_1), std(x_2) – odchylenie standardowe odpowiednio zmiennej pierwszej i drugiej

Z korelacją przeciwną można się spotkać w przypadku gdy np. w próbce występują dwa pierwiastki których stężenia zależą od siebie w taki sposób, że zawartość jednego pierwiastka w próbce rośnie, a drugiego maleje. W przypadku korelacji dodatniej obserwuje się sytuację odwrotną, czyli wzrost stężeń obu komponentów próbki. Można rozważyć trzeci wariant, czyli zerową wartość współczynnika Pearsona, co jest równoznaczne z brakiem korelacji dwóch zmiennych (zmienne nie zależą od siebie). Mówi się wówczas, że zmienne te są względem siebie ortogonalne. Przykłady korelacji danych zaprezentowano na Rys. 7. Zgodnie z warunkami wprowadzonymi dla miar podobieństwa, współczynnik korelacji Pearsona nie spełnia warunku pozytywności (warunek 1”’, rozdział 6).
Jako wadę współczynnika korelacji Pearsona uznaje się jego czułość na obecność obiektów odległych.



Rys. 7 Trzy przypadki korelacji dwóch zmiennych x_1 i x_2 a) dodatnia ($R \cong 0,80$), b) ujemna ($R \cong -0,98$) oraz c) brak korelacji ($R \cong 0,20$).

7. Klasyfikacja metod chemometrycznych

Pakiet metod chemometrycznych obejmuje metody uczenia z nadzorem (z ang. Supervised Methods) oraz metody uczenia bez nadzoru (z ang. Unsupervised Methods) [70]. Metody uczenia z nadzorem wykorzystywane są do modelowania danych, a więc tworzenia modeli kalibracyjnych, dyskryminacyjnych lub klasyfikacyjnych w zależności od analizowanego problemu badawczego. W tym celu wykorzystuje się zbiór zmiennych macierzy **X** oraz tzw. macierz zmiennych zależnych **Y**. Ogólne równanie modelu można zapisać następująco:

$$\mathbf{Y}_{(m,k)} = f(\mathbf{X}_{(m,n)}) + \mathbf{E}_{(m,k)} \tag{13}$$

gdzie, m i n to odpowiednio liczba próbek i zmiennych macierzy **X**, k to liczba zmiennych zależnych macierzy **Y**, a **E** to macierz reszt, wyrażająca błąd jaki popełnia się stosując wybrany model, np. model opisany równaniem (13).
Ze względu na złożony charakter wyników (zbyt mała liczba próbek, błąd eksperymentalny) poznanie prawdziwych zależności jest niemożliwe. Dlatego, stosuje się różnorodne modele które są ich aproksymacją. Dzięki, utworzonym modelom możliwe jest przewidywanie zmiennych zależnych, a w związku z tym redukcja nakładów finansowych związanych z analizą chemiczną większej liczby próbek.
W zależności od powodu z jakiego dane poddaje się modelowaniu rozróżnia się metody kalibracji oraz dyskryminacji lub klasyfikacji. Pierwsze z nich są relatywnie często stosowane. Pozwalają na ocenę ilościową wybranych własności, np. przewidywanie stężenia wybranych składników próbki. Najczęściej stosowanymi metodami kalibracji

są metoda regresji wielorakiej (z ang. Multi Linear Regression; MLR), regresja czynników głównych (z ang. Principal Component Regression; PCR) oraz regresja częściowych najmniejszych kwadratów (z ang. Partial Least Squares; PLS).

Z kolei zadaniem technik dyskryminacyjnych i klasyfikacyjnych jest opracowanie takich reguł logicznych, które pozwoliłyby na podstawie tzw. zbioru próbek treningowych, należących do z góry określonych grup, przewidzieć przynależność do tych grup nowych próbek. Modele te wykorzystywane są np. podczas badania autentyczności produktów spożywczych, czy farmaceutycznych na podstawie ich składu chemicznego.

Druga klasa metod, czyli techniki uczenia bez nadzoru wykorzystywane są w celu identyfikacji grup obiektów wykazujących podobne właściwości fizykochemiczne lub ujawnienia próbek odległych, czyli próbek znacząco różniących się od pozostałych. Grupowania dokonuje się w oparciu o zbiór zmiennych **X**, charakteryzujący próbki. Przypisanie próbek do grup odbywa się bez wiedzy na ich temat. W rezultacie, zastosowanie metod uczenia bez nadzoru pozwala na efektywne wyodrębnienie grup obiektów wykazujących zbliżone właściwości fizykochemiczne.

Do typowych technik uczenia bez nadzoru należą: analiza czynników głównych (z ang. Principal Component Analysis; PCA), metoda wymuszania projekcji (z ang. Projection Pursuit; PP), samoorganizujące się mapy Kohonena (z ang. Self-Organizing Map; SOM), czy techniki grupowania danych (z ang. Clustering Methods). Metody te pozwalają na eksplorację analizowanych zestawów wielowymiarowych danych.

Często w obliczu kompleksowej analizy próbek metody uczenia bez nadzoru pełnią rolę efektywnego narzędzia eksploracji danych. Wykorzystywane po wstępnym przygotowaniu danych do dalszej analizy, pozwalają na zapoznanie z ukrytą strukturą zgromadzonych danych oraz wspomagają dobór metody uczenia z nadzorem jako techniki ich modelowania.

Zadania metod eksploracji oraz wynikające z nich następstwa opisano w kolejnych rozdziałach pracy.

8. Metody eksploracji danych

W celu uzyskania szczegółowych informacji o obiekcie badań niezbędne jest przeprowadzenie złożonego procesu badawczego, który jest procesem wieloetapowym. Etapy te stanowią: pobieranie próbek, analiza jakościowa i ilościowa, analiza chemometryczna, interpretacja wyników i formułowanie wniosków. Procedura ta jest procedurą kaskadową, w której wyniki poszczególnych etapów zależą od poprawnego przeprowadzania etapów poprzedzających. Dlatego każdy z nich jest równie ważny i nie może zostać pominięty.

Wykorzystanie zaawansowanych metod analitycznych w trakcie procesu badawczego, pozwala na pomiar nawet do kilkunastu tysięcy zmiennych charakteryzujących próbkę, a tym samym na kompleksową charakterystykę wykazywanych przez nią właściwości fizykochemicznych. Z drugiej jednak strony, taki przyrost informacji przyczynia się do rozrostu danych, które stają się bardzo złożone. W konsekwencji, pojawia się problem wizualizacji danych, a przez to ich interpretacji. Korzysta się wówczas z analizy chemometrycznej, która również jest procesem wieloetapowym. Należy tu wymienić wstępne przygotowanie danych do dalszej analizy, eksplorację oraz ich modelowanie (Rys. 8).

Etap eksploracji jest nieodłączną częścią chemometrycznej analizy. A jej potrzeba wzrasta wraz ze wzrostem zgromadzonych danych oraz koniecznością ekstrakcji i przetwarzania istotnej chemicznie informacji na użyteczną wiedzę. Powodzenie etapu eksploracji zależy w dużej mierze od doboru metody wstępnego przygotowania danych. A samo działanie tych metod można przyrównać do działania szkła powiększającego, umożliwiającego wniknięcie w głąb analizowanych danych oraz poznanie ich struktury, w tym relacji pomiędzy obiektami i/lub zmierzonymi parametrami. Wśród technik eksploracyjnych wyróżnia się metody projekcji [71] oraz metody grupowania danych [72].

Metody projekcji służą przede wszystkim wizualizacji danych oraz redukcji ich wymiarowości. Projekcje otrzymanych wyników umożliwiają subiektywne zdefiniowanie grup. W przypadku metod projekcji o przynależności obiektu do danej grupy decyduje osoba dokonująca eksploracji, a przypisanie obiektów do grup odbywa się na podstawie wizualnej i subiektywnej oceny danych reprezentowanych w nowo zdefiniowanych podprzestrzeniach. Podczas gdy, metody grupowania danych, jak sama nazwa wskazuje, służą przede wszystkim grupowaniu obiektów. W przeciwieństwie do metod projekcji, algorytmy grupowania umożliwiają przypisanie obiektów do grup w sposób automatyczny, a wyniki grupowania reprezentowane są w postaci tzw. listy zawierającej informację o przynależności obiektów do poszczególnych grup.

Metodom grupowania danych, ze względu na tematykę prowadzonych badań, poświęcono rozdział 9, natomiast metody projekcji zostały przedstawione w podrozdziale *8.1*.

Rys. 8 Schemat wieloetapowej procedury analitycznej z uwzględnieniem poszczególnych kroków analizy chemometrycznej uzyskanych danych.

8.1 Metody projekcji danych

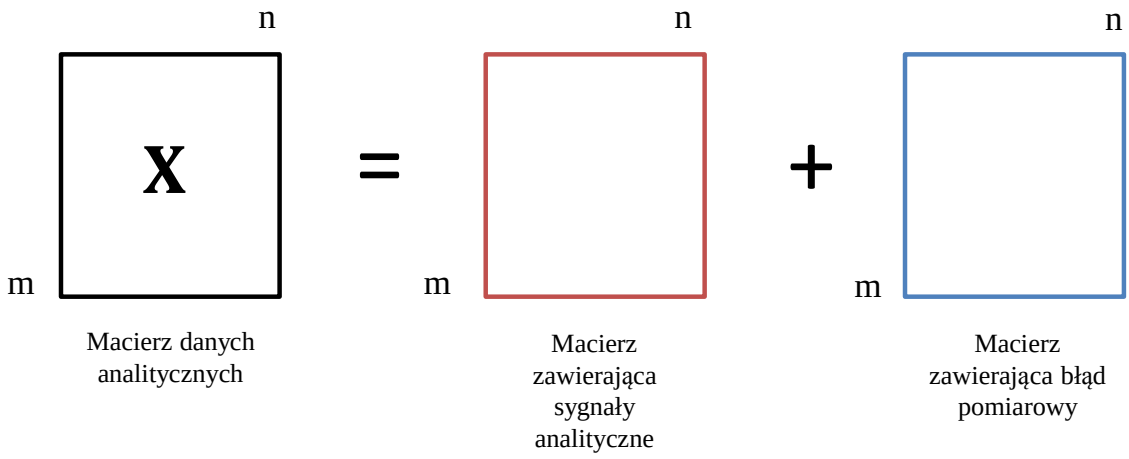
Jak już wcześniej wspomniano przykładem metod eksploracyjnych są metody wymuszania projekcji (z ang. Projection Pursuit; PP). Wprowadzone zostały przez Roya [71], [73] w latach 50. ubiegłego wieku, a następnie opisane przez Kruskala [71]. Natomiast, zasługę praktycznego zastosowania PP przypisuje się Friedmanowi i Tukeyowi. Dzięki wprowadzeniu tzw. indeksu projekcji (z ang. Projection Index, PI) [71] umożliwili ocenę informacji przedstawionych na projekcjach, co znacząco poprawiło efektywność metody.

Zadaniem PP jest poszukiwanie liniowych kombinacji analizowanych danych, czego skutkiem jest redukcja ich wymiarowości. Kombinację liniową otrzymuje się przez optymalizację wspomnianego indeksu projekcji. Można wyróżnić dwa rodzaje PI, parametryczne oraz nieparametryczne [74]. Pierwsze z nich mają za zadanie uchwycić rozkład danych, z kolei nieparametryczne są bardziej ogólne i nie skupiają się na rozkładzie danych w przestrzeni pomiarowej. W literaturze można znaleźć wiele przykładów indeksów m.in. wariancję, entropię [75], czy indeks Yenyukova [76]. Jego wybór jest jednak zawsze związany ze znalezieniem takiego kierunku wektora, który będzie najlepiej opisywał informację zawartą w danych. Wektory te powinny być zarówno jednostkowe jak i ortogonalne, czyli ortonormalne. Ortogonalność wektorów jest czynnikiem zapewniającym maksymalizację wariancji danych. Oznacza to, że informacja opisywana przez jeden wektor jest dopełniana przez kolejne ortogonalne

względem niego wektory. W ten sposób zostaje opisana wyłącznie istotna informacja zawarta w wyjściowych danych.
Jeżeli indeksem projekcji jest wariancja, mówi się wówczas o metodzie PCA, która jest szczególnym przypadkiem metod projekcji.

8.1.1 Analiza czynników głównych

Na szczególną uwagę zasługuje metoda analizy czynników głównych (z ang. Principal Components Analysis; PCA) [77], będąca podstawową metodą PP. Jako prekursora metody uznaje się Pearsona (1901 r.) – praca o prostych i płaszczyznach reprezentujących próbki w przestrzeni pomiarowej. Następnie, pojawiły się algorytm NIPALS (Fisher, MacKenzie), o którym ponownie przypomniał Wold (1966 r.) [78]. Kolejne modyfikacje algorytmu zawdzięcza się Hottelingowi.
Punktem wyjścia zastosowania tej metody jest fakt, że dane analityczne są obarczone błędem i można je przedstawić jako sumę dwóch macierzy, reprezentującej sygnał analityczny oraz macierzy przedstawiającej błąd pomiarowy (Rys. 9). Pozwala ona na przedstawienie wielowymiarowych danych w podprzestrzeniach zdefiniowanych przez nowe zmienne. Nowo utworzone zmienne nazywane są czynnikami głównymi (z ang. Principal Components; PC) i są one względem siebie ortogonalne. Maksymalizują one wariancję danych i są liniową kombinacją oryginalnych zmiennych. W nowym układzie współrzędnych odległości pomiędzy obiektami pozostają niezmienione, a informacja w nich zawarta zostaje zachowana. PC-ty są wektorami własnymi macierzy korelacji lub macierzy kowariancji. A wektory własne i odpowiadające im wartości własne definiują kierunki czynników głównych, w taki sposób aby opisywały one jak największą wariancję danych. Każdy kolejny PC musi być ortogonalny względem poprzedniego, dzięki czemu informacja zawarta w danych nie zostaje utracona. Warto również podkreślić, że tym sposobem pierwszy PC opisuje najwyższy procent całkowitej wariancji danych, a każdy kolejny opisuje jej coraz mniej. Dodatkowo, nowo utworzone osie maksymalizują wariancję danych w taki sposób, aby każda kolejna oś opisywała informację nie opisaną przez poprzednie osie, czego konsekwencją jest częściowa redukcja błędów eksperymentalnych oraz często ujawnienie obecności obiektów odległych. Obiektami odległymi są zazwyczaj próbki obarczone błędem grubym lub reprezentujące ich unikatowe właściwości. Z tych powodów, znalezienie i wskazanie obiektów odległych jest niezwykle istotne dla powodzenia późniejszej analizy. Liczba czynników głównych zależy od chemicznego rzędu macierzy danych **X**. Matematyczny rząd macierzy odpowiada maksymalnej liczbie liniowo niezależnych wektorów (kolumn lub wierszy macierzy), co w praktyce oznacza, że wynosi minimum z wymiarowości macierzy **X**(m, n). Chemiczny rząd macierzy najczęściej jest znacznie niższy od matematycznego.



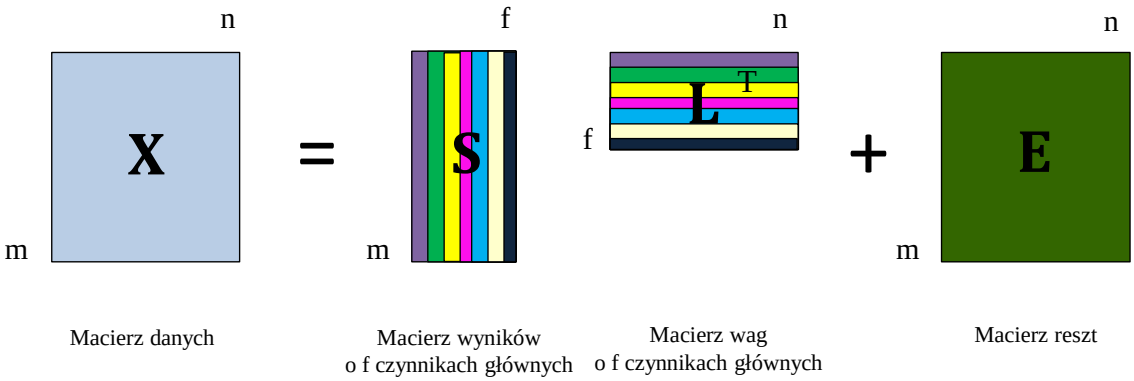
Rys. 9 Graficzne przedstawienie poszczególnych składowych macierzy **X**, tj. sygnał analityczny oraz błąd pomiarowy.

Te wektory własne, którym odpowiadają małe wartości własne zostają uznane za błąd eksperymentalny (np. szum instrumentalny) i opisane są przez tzw. macierz reszt (**E**). W PCA oryginalna macierz zostaje zdekomponowana do macierzy wyników (**S**), macierzy wag (**L**) oraz macierzy reszt (**E**), co wyraża równanie:

$$\mathbf{X}_{(m,n)} = \mathbf{S}_{(m,f)}\ \mathbf{L}^{T}_{(f,n)} + \mathbf{E}_{(m,n)} \qquad (14)$$

gdzie:
X – macierz danych m×n
S – macierz wyników o wymiarowości m×f, gdzie f określa liczbę czynników głównych
$\mathbf{L}^{T}$ – transponowana macierz wag o wymiarowości f×n, gdzie f określa liczbę czynników głównych
E – macierz reszt o wymiarowości m×n

Schematycznie dekompozycję macierzy **X** można przedstawić następująco:



Schemat 1 Dekompozycja macierzy **X** do macierzy wyników **S**, macierzy wag **L**, oraz macierzy reszt **E** w metodzie PCA.

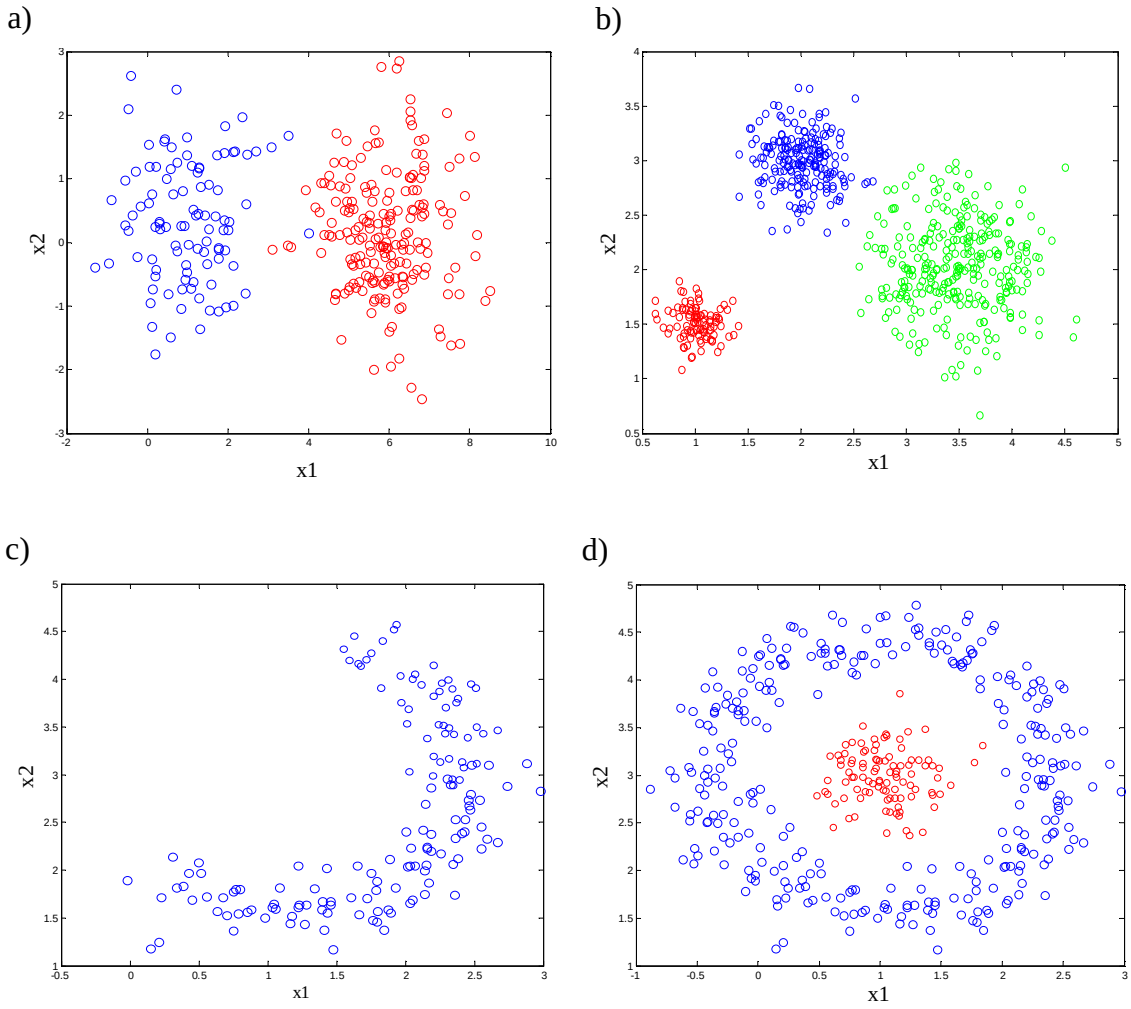
Najczęściej w celu dekompozycji macierzy danych wykorzystuje się algorytm SVD (z ang. Singular Value Decomposition) [59].
PCA wykazuje dwie podstawowe właściwości: umożliwia redukcję wymiarowości danych oraz pozwala na wizualizację ich ukrytej struktury.
Istotne czynniki główne definiują nowy układ współrzędnych, w którym reprezentuje się dane. Wyniki zostają zwizualizowane w postaci projekcji obiektów i parametrów na płaszczyzny zdefiniowane przez wybrane czynniki główne. Najczęściej projekcje wykonuje się na pierwszy i drugi czynnik główny, jako że opisują one największą część wariancji danych. Ich analiza ułatwia poznanie relacji pomiędzy obiektami oraz relacji pomiędzy parametrami, a także uwzględnienie wpływu parametrów na obserwowaną strukturę danych.

9. Metody grupowania danych

Techniki grupowania danych (z ang. Cluster Analysis) [79] należą do metod interpretacyjnych, ułatwiających poznanie ukrytej informacji zawartej w wielowymiarowych danych. Idea grupowania opiera się na identyfikacji naturalnych skupisk obiektów, w których obiekty podobne zostały umieszczone w jednej grupie, podczas gdy obiekty różniące się od siebie w różnych grupach. Przez obiekty podobne rozumie się takie, które w przestrzeni pomiarowej znajdują się blisko siebie, a więc wykazujące zbliżone właściwości fizykochemiczne, o czym wspomniano przy opisie miar odległości (rozdział 6).
Proces grupowania danych można podzielić na trzy kluczowe etapy:

1) wybór algorytmu grupowania oraz dobór odpowiedniej miary podobieństwa,
2) grupowanie właściwe, podczas którego utworzone zostają grupy obiektów i/lub zmiennych,
3) ocena liczby grup i interpretacja wyników.

Grupowanie danych jest dużym wyzwaniem przede wszystkim ze względu na różnorodność spotykanych typów grup obiektów. Istotnym czynnikiem decydującym o doborze metody grupowania jest kształt grup. Naturalnie występujące grupy różnią się rozkładem, rozmiarem, gęstością oraz lokalizacją w przestrzeni pomiarowej. Można je podzielić na grupy o kształcie sferycznym (kompaktowym), będące konsekwencją tworzących się w przestrzeni pomiarowej ugrupowań mających w przybliżeniu rozkład normalny; na grupy elipsoidalne, w których rozkład obiektów jest wzdłuż osi o większej wariancji; grupy bananowe; grupy zawarte w sobie, gdzie mniejsze grupy ulokowane są w obszarze większych, czy grupy wewnątrz siebie, w których obszary o większej gęstości obiektów są zawarte w obszarach o mniejszej gęstości obiektów (zob. Rys. 10) [80]. Wykrywanie grup określonego typu uzależnione jest od zastosowanej miary odległości, np. odległość euklidesowa umożliwia detekcję grup kompaktowych, a odległość Mahalanobisa grup elipsoidalnych [69].
Z racji na dużą liczbę dostępnych algorytmów grupowania, metody te najczęściej klasyfikuje się uwzględniając ich hierarchiczny i niehierarchiczny charakter. Zdarza się, także że klasyfikacja ta zostaje wzbogacona o kolejną grupę, tj. algorytmy grupowania bazujące na gęstości obiektów w przestrzeni pomiarowej [81]. W niniejszej pracy zdecydowano się na właśnie taki podział metod grupowania.
Innym, ze względu na sposób przypisania obiektów do grup, równie często występującym w literaturze podziałem algorytmów grupowania jest podział na tzw. metody jednoznaczne (z ang. Hard Clustering) oraz rozmyte (z ang. Soft Clustering) [81]. W metodach jednoznacznych, dany obiekt może należeć wyłącznie do jednej grupy, podczas gdy w metodach rozmytych każdy obiekt należy do wszystkich utworzonych grup jednak przynależy do nich w różnym stopniu. Przykładami rozmytych metod grupowania danych są: metoda rozmytych c-średnich [82] oraz k-harmonijnych średnich [83].



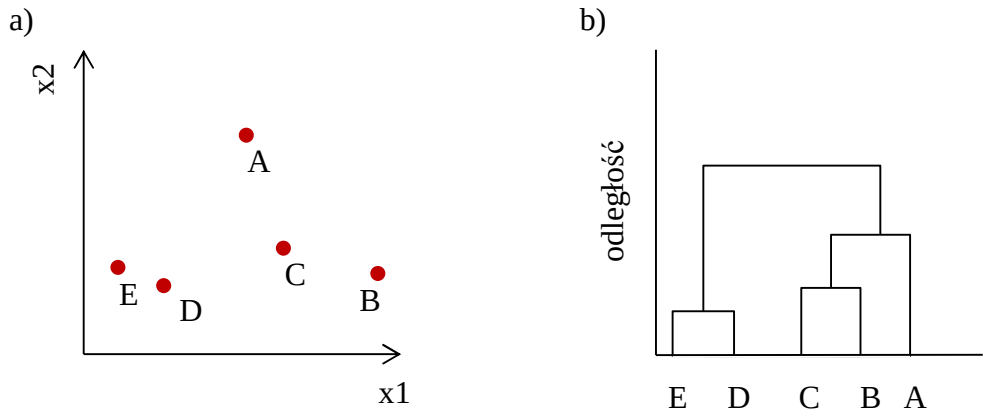
Rys. 10 Najczęściej występujące kształty grup obiektów przedstawione w przestrzeni dwuwymiarowej: a) elipsoidalne, b) sferyczne, c) bananowe oraz d) zawarte w sobie.

9.1 Metody hierarchiczne

Algorytmy grupowania hierarchicznego (z ang. Hierarchical Methods) [84] umożliwiają poznanie struktury analizowanych danych poprzez utworzenie hierarchii obiektów tworzących kolejne grupy. Relacje pomiędzy obiektami lub parametrami przedstawia się w postaci tzw. dendrogramu (Rys. 11), gdzie na osi odciętych umieszcza się obiekty (lub parametry), a oś rzędnych reprezentuje stopień podobieństwa pomiędzy nimi.

Ze względu na sposób tworzenia hierarchii, algorytmy grupowania danych można podzielić na metody aglomeracyjne oraz metody deglomeracyjne (podziałowe). W procedurach aglomeracyjnych rozpoczyna się od liczby grup odpowiadającej liczbie obiektów. Następnie, w celu zredukowania ich liczby, skupiska leżące najbliżej siebie w przestrzeni pomiarowej, łączy się ze sobą dopóki nie utworzą jednej grupy.

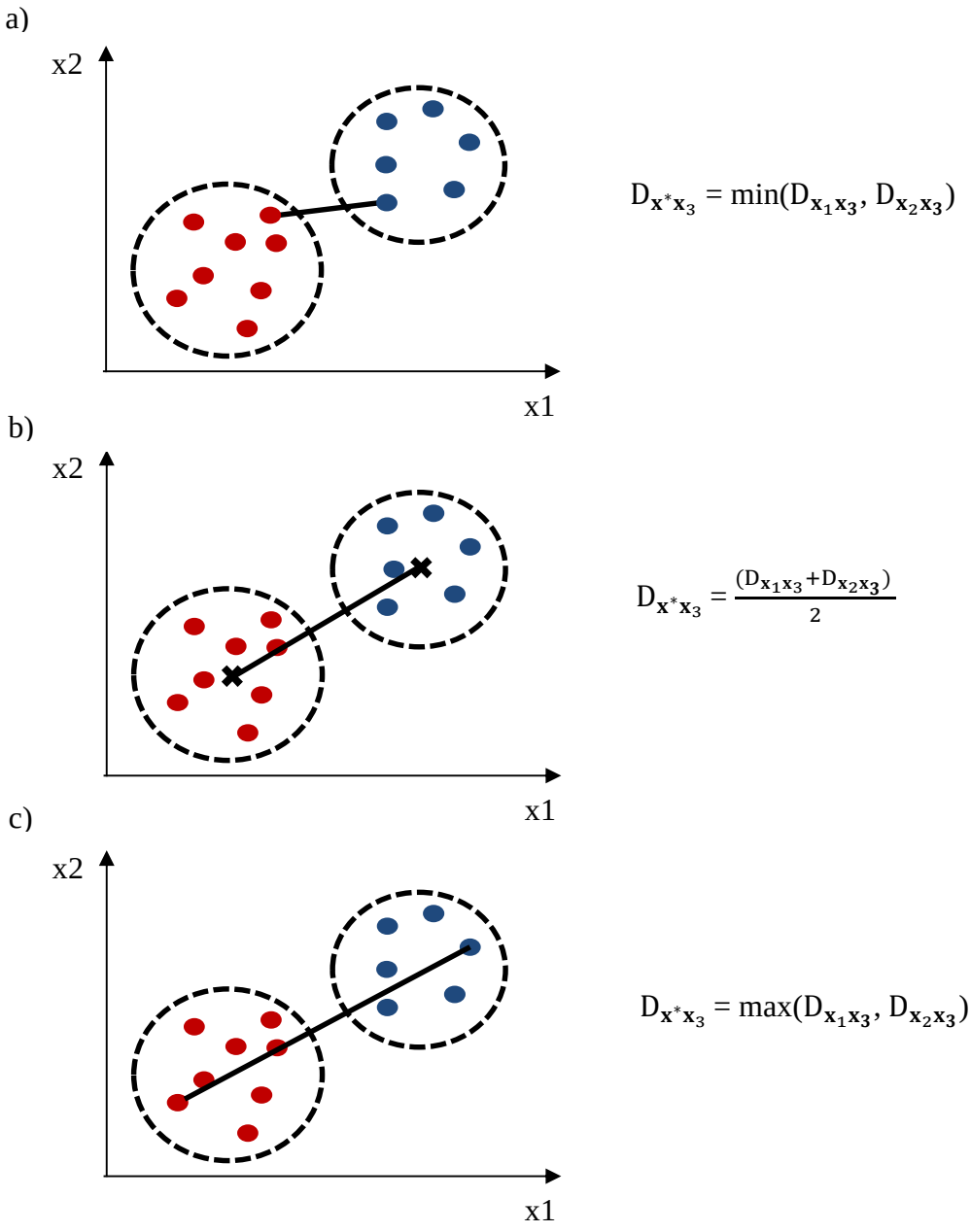
Procedury podziałowe działają przeciwnie do metod aglomeracyjnych. Początkowo, wszystkie obiekty stanowią jedną grupę, którą następnie dzieli się na mniejsze ugrupowania tak długo, aż każda grupa będzie zawierała wyłącznie jeden obiekt.
W grupowaniu hierarchicznym, poza sposobem tworzenia hierarchii ważnym aspektem jest również metodyka łączenia obiektów ze sobą. Do najpopularniejszych algorytmów należą: metoda pojedynczych połączeń, średnich połączeń oraz całkowitych połączeń.



Rys. 11 Dendrogram dla przykładowego rozkładu pięciu obiektów w przestrzeni dwuwymiarowej opisanej przez zmienne x1 i x2.

Algorytmy grupowania hierarchicznego rozpoczynają przeszukiwanie przestrzeni w celu znalezienia dwóch najbliżej położonych względem siebie obiektów np. $\mathbf{x}_1$ oraz $\mathbf{x}_2$. Odległość pomiędzy obiektami oblicza się na podstawie z góry przyjętej miary podobieństwa. Od momentu zidentyfikowania obiektów $\mathbf{x}_1$ oraz $\mathbf{x}_2$ traktowane są one przez algorytm jak jeden obiekt ($\mathbf{x}^*$), do którego przyłącza się kolejne obiekty zgodnie z przyjęta miarą podobieństwa i zasadami działania algorytmu, a także sposobem łączenia obiektów. Przykładowo, w metodzie pojedynczych połączeń poszukuje się obiektu $\mathbf{x}_3$ który jest najbliżej położony względem połączonych, $\mathbf{x}_1$ i $\mathbf{x}_2$. Z kolei w algorytmie średnich połączeń, oblicza się średnią z odległości pomiędzy nowym obiektem $\mathbf{x}_3$, a obiektami $\mathbf{x}_1$ i $\mathbf{x}_2$. W ostatnim z algorytmów, poszukuje się obiektu najbardziej oddalonego względem pozostałych. Różnice w działaniu algorytmów ze względu na sposób łączenia obiektów, przedstawiono schematycznie na Rys. 12. Należy jednak podkreślić, iż procedura łączenia lub rozdziału obiektów przeprowadzona zostaje tylko raz. Oznacza to, że do sekwencji utworzonej hierarchii nie można wprowadzić zmian, co często uznawane jest za wadę metod hierarchicznych. Jednak, sekwencyjne łączenie obiektów i/lub parametrów, ujawniające ich hierarchię oraz wzajemne relacje, a także możliwość wizualizacji struktury danych,

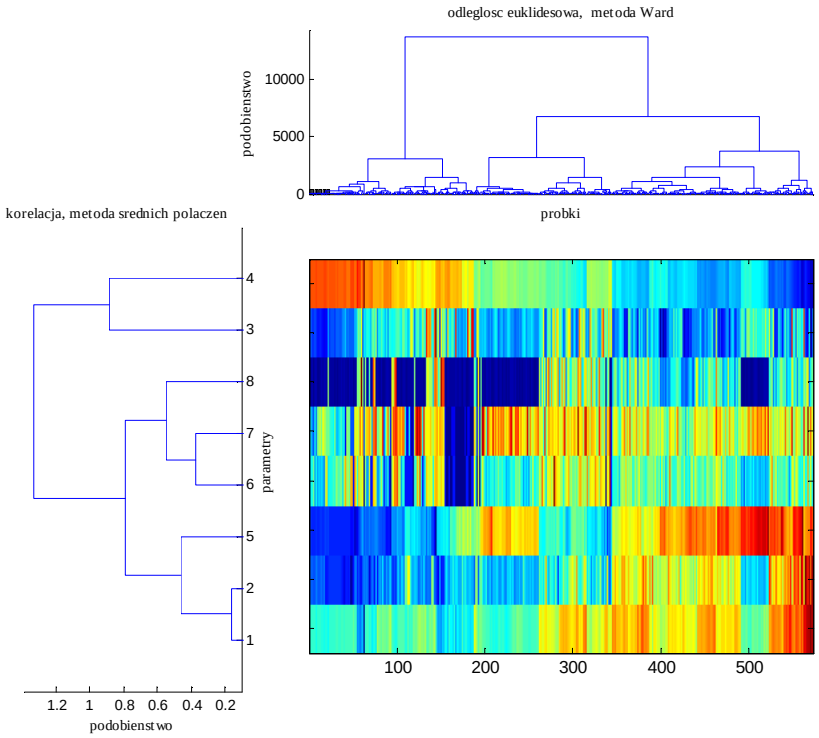
wspomagające interpretację wyników, przyczynia się do powszechnego ich stosowania podczas eksploracji danych. Ze względu na możliwość wizualizacji struktury analizowanych danych metody hierarchiczne mogą stanowić alternatywę dla metody PCA.



Rys. 12 Różne sposoby łączenia ze sobą obiektów w grupy w metodach hierarchicznych: a) metoda pojedynczych połączeń, b) metoda średnich połączeń oraz c) metoda całkowitych połączeń.

9.1.1 Dwukierunkowe grupowanie hierarchiczne

Szczególnym przypadkiem metod grupowania hierarchicznego jest tzw. dwukierunkowe grupowanie hierarchiczne (z ang. Two-way Clustering) [85]. Grupowanie to, analogicznie jak klasyczne grupowanie hierarchiczne, przebiega z utworzeniem dendrogramów, lecz wizualizacji podlegają równocześnie dwa wymiary macierzy danych **X** – próbki i parametry. Aby ułatwić interpretację wyników, zbadanie zależności pomiędzy próbkami i parametrami oraz ocenę wpływu parametrów na grupowanie obiektów, dendrogramy wzbogaca się o tzw. kolorową mapę (z ang. Heatmap). Na tej mapie duże wartości parametrów opisujących próbki reprezentuje się kolorem czerwonym, a małe kolorem niebieskim. Pozostałe wartości parametrów przyjmują kolory pośrednie pomiędzy czerwonym, a niebieskim. Dzięki uzupełnieniu wyników kolorową mapą odczytywanie zależności pomiędzy próbkami i parametrami oraz wpływu parametrów na grupowanie obiektów staje się znacznie prostsze. Z łatwością można także odczytać, które parametry maja największy wpływ na kształtowanie się obserwowanych ugrupowań.



Rys. 13 Przykład wyników otrzymywanych za pomocą dwukierunkowego grupowania hierarchicznego.

9.2 Metody niehierarchiczne

Zadaniem metod niehierarchicznych jest wyodrębnienie z heterogenicznego zestawu danych homogenicznych grup obiektów. Niehierarchiczne metody grupowania danych umożliwiają podział obiektów na z góry określoną liczbę grup tak, aby zminimalizować przyjętą funkcję kosztów, E (15) [79]. W praktyce oznacza to, że np. wariancja w grupie powinna być jak najmniejsza. Dzięki temu, obiekty leżące blisko siebie w przestrzeni pomiarowej zostają przypisane do jednej grupy, z kolei utworzone grupy powinny być jak najbardziej oddalone od siebie w przestrzeni pomiarowej.
W przypadku grupowania niehierarchicznego mniej istotny jest fakt istnienia rzeczywistych ugrupowań, a większe znaczenie odgrywa ich liczba [86]. To prowadzi do wyodrębnienia grup spośród obiektów nie wykazujących naturalnej tendencji do grupowania, lub gdy grupy są niecałkowicie rozdzielone (Rys. 14). Ma to duże znaczenie np. przy grupowaniu cząsteczek ze względu na ich właściwości fizykochemiczne.
Liczbę grup ustala się z góry, albo testuje się kilka możliwości i wybiera najbardziej optymalne rozwiązanie charakteryzujące się minimalną wartością funkcji kosztów.
Wśród niehierarchicznych metod grupowania danych na szczególną uwagę zasługuje metoda k-średnich, będąca podstawową techniką tego typu. Wymienić warto również takie metody jak gaz neuronowy, ekspandujący gaz neuronowy, czy ekspandujący algorytm k-średnich, będący techniką łączącą cechy wszystkich trzech wymienionych tu algorytmów.

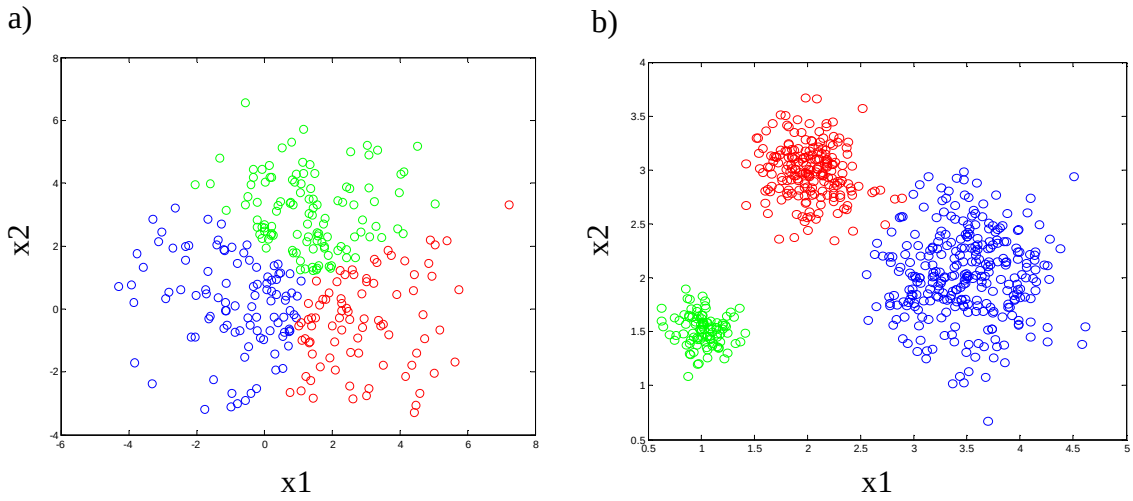
9.2.1 Metoda k-średnich

Metoda k-średnich jest stosowana w celu podziału obiektów na określoną z góry liczbę grup (k). Zgodnie z zasadą, że obiekty należące do jednej grupy są do siebie bardziej podobne niż te należące do różnych grup. Podobieństwo obiektów najczęściej określa się za pomocą odległości euklidesowej (9).
Istnieje wiele wariantów metody k-średnich. Niemniej jednak, najczęściej stosowanym algorytmem jest algorytm MacQueena, który został opracowany w 1967 r. [87]. W algorytmie tym, grupy konstruowane są w trakcie iteracyjnej procedury, zgodnie z następującymi krokami:

1) wybór liczby grup, k,
2) podział obiektów na k grup,
3) obliczenie współrzędnych środków każdej grupy, $\mathbf{u}_k$,
4) przypisanie każdego obiektu do grupy, której środek znajduje się najbliżej aktualnie rozpatrywanego obiektu,
5) obliczenie sumy kwadratów odległości pomiędzy środkiem danej grupy i należącymi do niej obiektami,

6) powtórzenie kroków 3-6, aż do uzbieżnienia się algorytmu.

Metoda k-średnich należy do metod o nieskomplikowanej procedurze obliczeń, a wysokiej efektywności. Ponadto, czas obliczeń odległości uzależniony jest przede wszystkim od ilości próbek, stąd metoda k-średnich nie należy do czasochłonnych. Są to główne atuty tej metody. Dzięki nim może ona zostać wykorzystana w analizie różnorodnych zestawów danych o dużej liczbie próbek. Jak zaznaczono wcześniej, uzyskany podział na grupy może odbiegać od naturalnych ugrupowań. W konsekwencji, również z danych nie wykazujących naturalnej tendencji do grupowania zostają wyodrębnione grupy (Rys. 14).



Rys. 14 Przykładowy podział na grupy dla symulowanego zestawu danych, uzyskany za pomocą algorytmu k-średnich dla k = 3: a) zestaw danych bez wyraźnych grup obiektów, b) z dobrze rozdzielonymi grupami obiektów.

Takie zachowanie algorytmu jest rezultatem poszukiwania lokalnego, a nie globalnego optimum funkcji kosztów (np. wariancji), a także przyjętej definicji funkcji kosztów. W celu uzyskania optymalnego rozwiązania, podział na grupy powtarza się zazwyczaj wielokrotnie i wybiera najbardziej stabilne rozwiązanie, co różnoznaczne jest z minimalizacją funkcji kosztów E, wyrażonej równaniem (15).

$$\min(E) = \sum_{j=1}^{k}(\sum_{i=1}^{m}(\mathbf{x}_i - \mathbf{u}_j)^2) \quad (15)$$

gdzie:
E – funkcja kosztów
k – określona z góry liczba grup
$\mathbf{x}_i$ – i-ty obiekt w grupie
$\mathbf{u}_j$ – środek j-tej grupy

9.2.2 Metoda gazu neuronowego

Algorytm gazu neuronowego (z ang. Neural Gas; NG) został zaproponowany przez Martineza [88]. W algorytmie buduje się sieć neuronową zawierającą k węzłów, odpowiadających środkom grup i równocześnie determinujących liczbę poszukiwanych grup. Każdy węzeł „uczy się" poprzez swobodny ruch w przestrzeni pomiarowej. Stąd też wynika nazwa metody, która odzwierciedla podobieństwo pomiędzy dynamiką poruszania się węzłów w przestrzeni pomiarowej, a cząsteczkami gazu uwięzionymi np. w pojemniku.
Pozycja j-tego węzła sieci jest określona za pomocą wektora wag $\mathbf{w}_j(1, n)$, który określa współrzędne środka grupy w przestrzeni danych. Algorytm rozpoczyna działanie od losowo wybranego obiektu $\mathbf{x}_i(1, n)$ i prezentuje go sieci neuronowej. Następnie wagi są korygowane względem odległości pomiędzy węzłami, a $\mathbf{x}_i$ (jedna iteracja). Największe modyfikacje wag wprowadzone zostają dla węzła, który leży najbliżej $\mathbf{x}_i$. Węzeł ten nazywany jest zwycięzcą. Z kolei wektory wag węzłów najbardziej oddalonych od $\mathbf{x}_i$ ulegają modyfikacjom w mniejszym stopniu.
W każdej kolejnej iteracji wartość funkcji sąsiedztwa ulega zmniejszeniu. Takie podejście zapobiega losowemu (chaotycznemu) podziałowi obiektów na grupy. Podczas etapu uczenia się algorytmu, wagi węzłów określające ich pozycje w sieci, są tak dopasowywane aby zminimalizować funkcję kosztów opisaną równaniem (15). Z kolei sama koncepcja sąsiedztwa pozwala na bardziej elastyczne przypisanie obiektów do węzłów.
Reasumując działanie algorytmu można przedstawić następująco:

Parametry wejścia:
- liczba węzłów k odpowiadająca liczbie poszukiwanych grup w danych;
- wektory wag zawierające losowe wartości przypisane do każdego j-tego węzła (środek grupy);
- zbiór parametrów wpływający na proces uczenia:
  - liczba iteracji (t_{maks}),
  - współczynnik początku i końca stałej uczenia (Ir_i, Ir_f),
  - rozmiar sąsiedztwa początkowy i końcowy (d_i, d_f).

Algorytm:

1) Konstrukcja sieci neuronowej zawierającej k węzłów i losowy wybór jednego obiektu $\mathbf{x}_i(1, n)$ z macierzy danych $\mathbf{X}(m, n)$,
2) Określenie odległości pomiędzy $\mathbf{x}_i$ oraz wektorami wag $\mathbf{w}_j$ każdego węzła j = 1, 2, 3,…,k,
3) Wysortowanie węzłów względem ich wzrastającej odległości do $\mathbf{x}_i$ oraz wybór zwycięskiego węzła charakteryzującego się najmniejszą odległością w odniesieniu do obiektu $\mathbf{x}_i$,
4) W każdej iteracji następuje uaktualnienie wektorów wag zwycięzcy oraz jego sąsiadów zgodnie z równaniem:

$$\mathbf{w}_j(t+1) = \mathbf{w}_j(t)+Ir(t)\times h_d(ind_i)\times(\mathbf{x}_i-\mathbf{w}_j(t)) \quad (16)$$

gdzie:
t – obecna iteracja, t=1, 2, 3,…, t_{maks}
Ir(t) – współczynnik uczenia:

$$Ir(t) = Ir_i \left(\frac{Ir_f}{Ir_i}\right)^{\frac{t}{t_{maks}}} \text{ , } Ir\in(0,1]$$

$h_d(ind_i)$ – funkcja sąsiedztwa:

$$h_d(ins_i)=e^{-\left(\frac{k_i}{d}\right)}$$

d(t) – stała rozpadu:

$$d(t) = d_i \left(\frac{d_f}{d_i}\right)^{\frac{t}{t_{maks}}}, \quad d\in[0,\infty)$$

Jeżeli $t < t_{maks}$ następuje powrót do kroku nr 1,

5) Przypisanie obiektu macierzy **X** do najbliższego węzła.

9.2.3 Metoda ekspandującego gazu neuronowego

Ekspandujący gaz neuronowy (z ang. Growing Neural Gas) [90] można przyrównać do grafu zawierającego k węzłów. Sąsiadujące ze sobą węzły połączone są za pomocą krawędzi. Każdy węzeł opisany jest wektorem wag $\mathbf{w}_j$, opisującym jego położenie

w przestrzeni pomiarowej, tak jak miało to miejsce w przypadku metody gazu neuronowego z poprzedniego podrozdziału (*9.2.2*).
Podczas etapu uczenia się do sieci dodawane są nowe węzły, tak długo aż zostanie osiągnięta maksymalna liczba węzłów k, zdefiniowana na wejściu. Algorytm rozpoczyna z dwoma losowo ułożonymi w przestrzeni węzłami, które łączy się ze sobą krawędzią. Dopasowanie wektorów wag węzłów zachodzi iteracyjnie.
Dla każdego wprowadzonego do sieci obiektu ze zbioru danych **X**, określa się jego najbliższy węzeł zwany zwycięzcą (s1) oraz najbliższego sąsiada zwycięzcy (s2). Następnie węzły s1 oraz s2 łączy się ze sobą przy pomocy krawędzi.
Z każdą krawędzią związana jest tzw. zmienna wieku (α). W każdym etapie procesu uczenia się, polegającym na wprowadzeniu nowego obiektu do sieci, zmienna wieku wszystkich krawędzi wychodzących z zwycięskiego węzła wzrasta o 1. Kiedy krawędź pomiędzy s1 oraz s2 jest tworzona lub już istnieje, zmienna wieku zestawu jest < 0. Śledząc zmiany zmiennej wieku możliwe jest wskazanie węzłów nieaktywnych, czyli tych które nie pełnią roli zwycięzcy. Tym sposobem topologia sieci jest modyfikowana. Te krawędzie które osiągnęły maksymalną wartość α oraz węzły nie posiadające krawędzi zostają usunięte z sieci. Dzięki czemu pojawia się kolejna możliwość ingerencji w strukturę sieci oraz jej modyfikacja.
W algorytmie GNG sąsiedztwo zwycięzcy jest ograniczone do topologicznego sąsiedztwa, czyli węzłów połączonych ze zwycięzcą. Zwycięzca wraz z jego topologicznymi sąsiadami są przesuwani w kierunku obiektów sieci o stałą ustaloną z góry odległość, zdefiniowaną osobno dla zwycięzcy i osobno dla sąsiadów.
W GNG nie występuje funkcja sąsiedztwa, tak jak miało to miejsce w NG oraz koncepcja przetwarzania węzłów jeden po drugim. Wszystkie węzły zostają uaktualniane równocześnie. Następnie oblicza się kwadrat odległości euklidesowej (10) pomiędzy obiektami i zwycięskim węzłem. Informacja o odległości jest wprowadzona do tzw. współczynnika dyspersji, DV, charakteryzującego stopień rozproszenia węzłów w przestrzeni pomiarowej na danym etapie uczenia algorytmu. DV wykorzystywany jest do określania optymalnego położenia węzłów w sieci. Nowy węzeł sieci jest do niej wprowadzany po ustalonej liczbie iteracji (χ).
Liczba iteracji spełnia podobną rolę jak zmienna wieku, określając stopień modyfikacji sieci. Warto również zaznaczyć, że nowe węzły nie mogą zostać wprowadzone do sieci zbyt szybko. Ich pozycja jest średnią arytmetyczną pozycji dwóch węzłów np. q i f, połączonych ze sobą krawędzią oraz charakteryzujących się wysoką wartością DV. Krawędź pomiędzy węzłami q i f zostaje zamieniona na nowe połączenie od q do nowego węzła oraz od nowego węzła do f. Wartość DV po wprowadzeniu nowego węzła zostaje zredukowana o 50%, a wartość tego współczynnika dla nowego węzła jest równa średniej arytmetycznej ze zredukowanego DV węzłów q i f.
Jako kryterium zbieżności (λ) algorytmu wykorzystuje się osiągnięcie maksymalnej liczby węzłów.
Działanie algorytmu można zaprezentować następująco:

1) Utworzenie zbioru A zawierającego dwa węzły a i b, których pozycje $\mathbf{w}_a$ i $\mathbf{w}_b$ w przestrzeni są losowe:

$A = \{a, b\}$

2) Inicjalizowanie zbioru krawędzi C, zawierającego krawędź łączącą węzeł a z b oraz zbioru wieku tych krawędzi równy zero:

$C = \{(a, b)\}$; $wiek_{(a,b)} = 0$

3) Losowy wybór obiektu $\mathbf{x}_i$
4) Określenie węzłów s1 oraz s2 (s1, s2 ∈ A) takich, że:

$\|\mathbf{w}_{s1} - \mathbf{x}_i\| \leq \|\mathbf{w}_c - \mathbf{x}_i\|$, gdzie c∈k

oraz

$\|\mathbf{w}_{s2} - \mathbf{x}_i\| \leq \|\mathbf{w}_c - \mathbf{x}_i\|$, dla wszystkich c z wyjątkiem s1 ∈ k

5) Jeżeli istnieją jeszcze krawędzie – dodać krawędź pomiędzy s1 i s2 do c:

$c=c \cup \{(s1, s2)\}$

W innym przypadku wiek krawędzi pomiędzy s1 oraz s2 = 0

wiek(s1, s2) = 0

6) Dodanie odległości kwadratowej pomiędzy wejściowymi sygnałami i s1 do DV,

$DV_{s1}=\sum\|\mathbf{w}_{s1} - \mathbf{x}_i\|^2$

7) Przesunięcie s1 i s2 w kierunku $\mathbf{x}_i$,

$\mathbf{w}_{s1} = I_{ru}(\mathbf{x}_i - \mathbf{w}_{s1})$

$\mathbf{w}_{s2} = I_{ru}(\mathbf{x}_i - \mathbf{w}_{s2})$

8) Dodanie do wieku s1 jedynkę, dla i oznaczającego zbiór kierunków topologicznych sąsiadów s1

wiek(s1, i) = wiek(s1, i) + 1

9) Usuwanie krawędzi z wieku większego niż a_{maks}. Jeżeli są węzły nie posiadające krawędzi należy je usunąć.
10) Po χ iteracji wstawić nowy węzeł:
    - znaleźć węzeł q z maksymalnym DV
    - interpolować nowy węzeł r pomiędzy węzłami q i f oraz usunąć oryginalną krawędź łączącą węzeł q z f

$c=c\cup\{(x, q), (r, f)\}$

- usunąć C{(q, f)}
- obniżyć wartość DV dla q oraz f

$\Delta DV_a = -\alpha DV_q$ oraz $\Delta DV_f = -\alpha DV_f$

- interpolować DV wszystkich węzłów $c \in k$

$\Delta DV_c = -\beta DV_c$

11) Jeżeli kryterium zatrzymania nie zostało spełnione, przejść do kroku 2.

9.2.4 Metoda ekspandującego k-średnich

Ekspandujący algorytm k-średnich (z ang. Growing k-means; GK) [91] jest relatywnie nowym algorytmem z nieokreśloną topologią sieci neuronowej. Podczas etapu uczenia się węzły są rozproszone w przestrzeni pomiarowej w taki sposób aby zminimalizować funkcję kosztów (15).
Algorytm jest kombinacją algorytmu NG oraz GNG. Tak jak w metodzie gazu neuronowego metoda ta wykorzystuje współczynnik uczenia się, natomiast z metody GNG przejęła rozrostowy charakter sieci, w której pozycje węzłów uaktualniane są po osadzeniu obiektów w sieci. Procedura rozpoczyna się od dwóch węzłów losowo ułożonych w przestrzeni pomiarowej. Podczas etapu uczenia, po wprowadzeniu obiektu do sieci, najbliższy węzeł jest przesuwany w kierunku tego obiektu o stały odcinek zdefiniowany przez użytkownika. Odcinek ten nazywa się współczynnikiem uczenia. Kiedy wybrane obiekty zestawu danych zostaną wprowadzone do sieci następuje przypisanie obiektów do najbliższych węzłów w celu utworzenia grup obiektów podobnych. Następnie, oblicza się kwadrat odległości euklidesowej pomiędzy węzłami i przypisanymi do nich obiektami. W grupie o największej wariancji wprowadza się nowy węzeł umieszczony w połowie odległości pomiędzy węzłem reprezentującym środek grupy a najdalej położonym obiektem tej grupy. Sieć rozrasta się tak długo aż zostanie osiągnięta określona na wejściu liczba k węzłów. Etap uczenia algorytmu zostaje zakończony w momencie uzyskania sieci o k węzłach i wyczerpaniu maksymalnej liczby iteracji.
Poszczególne etapy algorytmu można przedstawić w krokach 1-8, wyszczególnionych poniżej:

1) Sprecyzowanie liczby poszukiwanych grup obiektów podobnych poprzez określenie k poszukiwanych węzłów sieci,
2) Zbudowanie sieci, rozpoczynając z dwoma węzłami losowo umieszczonymi w przestrzeni pomiarowej,
3) Wprowadzenie obiektu $\mathbf{x}_i$ do sieci,
4) Przypisanie obiektu $\mathbf{x}_i$ do najbliższego węzła w sieci. Węzeł ten zostaje określony jako zwycięzca,
5) Przesunięcie zwycięskiego węzła w kierunku obiektu $\mathbf{x}_i$ o określony odcinek,
6) Po wprowadzeniu wybranych obiektów do sieci obliczyć kwadrat odległości euklidesowej pomiędzy węzłami oraz przypisanymi do nich obiektami. Wartości obliczonej odległości wprowadzić do współczynnika rozproszenia węzłów DV,
7) Wprowadzenie nowego węzła pomiędzy węzłem oraz obiektem najbardziej od niego oddalonym w grupie wykazującej największe rozproszenie DV,
8) Jeśli do sieci wprowadzono k węzłów obniżyć współczynnik uczenia zgodnie z pozostałą liczbą iteracji.

9.3 Metody grupowania bazujące na gęstości danych

Odkrywanie naturalnych grup obiektów stanowi wyzwanie analizy chemometrycznej. Istnieje wiele metod eksploracji danych, w tym metod grupowania, pozwalających na wyodrębnienie grup obiektów z wszystkich typów danych eksperymentalnych. Niemniej jednak, każda ze znanych metod eksploracji posiada swoje ograniczenia. Jednym z nich jest poszukiwanie grup reprezentujących konkretny rozkład obiektów w przestrzeni pomiarowej. Innym określanie parametrów wejścia takich jak liczba grup, itp. Problem stanowi również wymiarowość danych, a zwłaszcza wzrastająca liczba analizowanych próbek, powodująca wzrost czasu obliczeń. W konsekwencji pojawiła się potrzeba wprowadzenia algorytmów wychodzących poza wymienione ograniczenia. Algorytmy te z założenia miały znacząco przyspieszyć eksplorację analizowanych zestawów danych oraz nie wymagają od użytkownika wiedzy na temat struktury eksplorowanych danych. Metody te nazywa się metodami grupowania danych bazującymi na gęstości danych. W metodach tych podobieństwo pomiędzy obiektami określa się za pomocą kryterium gęstości danych, a nie miar odległości jak miało to miejsce w przypadku metod grupowania hierarchicznego i niehierarchicznego. Wprowadzenie kryterium gęstości jako miary podobieństwa umożliwia wyodrębnianie grup obiektów o arbitralnych kształtach oraz detekcję obiektów odległych.

Metody te stanowią uzupełnienie klasycznej klasyfikacji metod grupowania. Zazwyczaj klasyfikacja ta obejmuje metody hierarchiczne oraz niehierarchiczne, której podstawą jest sposób łączenia obiektów ze sobą (tj. hierarchicznie lub niehierarchicznie). Ponadto, podobieństwo pomiędzy obiektami wyrażone zostaje za pomocą wybranej miary podobieństwa (np. odległość euklidesowa, Mahalanobisa). Zastosowanie kryterium gęstości pozwala na wyodrębnienie obszarów przestrzeni pomiarowej wykazujących większe zagęszczenie obiektów niż pozostałe obszary tej przestrzeni. Podstawowe metody bazujące na kryterium gęstości to metoda DBSCAN (z ang. Density-Based Spatial Clustering of Application with Noise) [92] oraz metoda OPTICS (z ang. Ordering Points to Identify the Clustering Structure) [93]. Największą zaletą tych algorytmów jest możliwość detekcji grup o arbitralnych kształtach, na co nie pozwalają np. metody niehierarchiczne. Kształt grup jaki tworzą obiekty może być różnorodny, o czym wspominano na początku tego rozdziału.

W przypadku danych eksperymentalnych, najczęściej spotykamy się z grupami sferycznymi (kompaktowymi). Równie często grupy obiektów przyjmują kształt elipsoidalny. Jednakże obiekty mogą utworzyć również skupiska zawarte w sobie, o kształcie banana oraz wiele innych (Rys. 10). Metody hierarchiczne oraz niehierarchiczne ze względu na stosowne miary podobieństwa (np. odległość euklidesowa, Mahalanobisa), umożliwiają wyodrębnienie grup wykazujących kształt sferyczny lub elipsoidalny. W przypadku metod grupowania bazujących na gęstości danych, takich jak DBSCAN, ten problem nie występuje ponieważ algorytm ten przeszukuje przestrzeń obiekt po obiekcie, a sposób ich łączenia prowadzi do efektu łańcucha połączeń (Definicja 2 oraz 3), co umożliwia detekcję grup o arbitralnych kształtach.

Wizualizacja rozkładu obiektów w przestrzeni dwuwymiarowej, umożliwia odróżnienie obszarów stanowiących grupę od tych reprezentujących szum. Związane jest to z rozpoznaniem obszarów wykazujących lokalnie większą liczebność obiektów, a wiec gęstość, niż pozostałe obszary tej przestrzeni. Najczęściej skupiska obiektów rozdzielone są obszarami przestrzeni w której obserwuje się znacznie niższe zagęszczenie obiektów w porównaniu do utworzonych grup. W celu odróżnienia grupy obiektów, od szumu w przestrzeni wieloparametrowej wykorzystuje się matematyczny opis zagadnienia, co zaprezentowano za pomocą Definicji od 1 do 6. Rozpoczęto od wprowadzenia pojęcia sąsiedztwa obiektów, które zostaje wyznaczone przez promień sąsiedztwa r (Definicja 1) oraz minimalnej liczby sąsiadów MinPts. Następnie zdefiniowano bezpośredni łańcuch połączeń obiektów (Definicja 2) oraz łańcuch połączeń obiektów (Definicja 3), które powstają w wyniku przetwarzania poszczególnych obiektów w trakcie działania algorytmu. Ponieważ, nie wszystkie obiekty zawierają w swoim sąsiedztwie MinPts. Z tego powodu wyróżnia się dwa typy obiektów: obiekty rdzeniowe oraz obiekty brzegowe. Obiekty rdzeniowe spełniają warunek minimalnej liczby sąsiadów w sąsiedztwie wyznaczonym przez promień r. Z kolei obiekty brzegowe tego warunku nie spełniają i w swoim sąsiedztwie zawierają mniej niż MinPts obiektów sąsiadujących, ale jeden z sąsiadów musi być obiektem rdzeniowym. Z tego powodu zdefiniowano również pojęcie obiektu brzegowego (Definicja 4), a dodatkowo grupy (Definicja 5) oraz szumu (Definicja 6).

Definicja 1: Promień sąsiedztwa obiektu (r)
Promień sąsiedztwa, r obiektu $\mathbf{x}_i$, określony jako $N_r(\mathbf{x}_i)$ wyraża się następująco:

$$N_r(\mathbf{x}_i)=\{\mathbf{x}_q \in D \mid \mathrm{dist}(\mathbf{x}_i, \mathbf{x}_q) \leq r\}$$

Definicja 2: Bezpośredni łańcuch połączeń obiektów
Obiekt $\mathbf{x}_i$ tworzy bezpośredni łańcuch obiektów z obiektem $\mathbf{x}_q$ przy uwzględnieniu r oraz MinPts, jeśli spełnione zostają następujące warunki:

1) $\mathbf{x}_i \in N_r(\mathbf{x}_q)$
2) $\left|N_r(\mathbf{x}_q)\right| \geq \mathrm{MinPts}$

Łańcuch połączeń obiektów jest symetryczny dla dwóch obiektów rdzeniowych i asymetryczny dla przykładu obiektów brzegowych.

Definicja 3: Łańcuch połączeń obiektów
Obiekt $\mathbf{x}_i$ oraz $\mathbf{x}_q$ należą do łańcucha obiektów dla r oraz MinPts, jeśli istnieje łańcuch obiektów $\mathbf{x}_1, \ldots, \mathbf{x}_m$, gdzie $\mathbf{x}_1 = \mathbf{x}_q$, $\mathbf{x}_m = \mathbf{x}_q$, takich że $\mathbf{x}_{i+1}$ jest w bezpośrednim łańcuchu obiektów z obiektem $\mathbf{x}_i$.

Definicja ta jest kanonicznym rozszerzeniem **Definicji 2**. Symetryczność obiektów obserwuje się wyłącznie w odniesieniu do obiektów rdzeniowych. W większości definicja ta przedstawia asymetryczną naturę obiektów.

Jeżeli dwa obiekty należące do tej samej grupy nie spełniają Definicji 3, ale jeśli w ich sąsiedztwie występuje obiekt który spełnia warunek łańcucha połączeń obiektów to wówczas obiekty określane są jako obiekty brzegowe, co matematycznie można sformułować następująco:

Definicja 4: Obiekt brzegowy
Obiekt $\mathbf{x}_i$ jest obiektem brzegowym z obiektem $\mathbf{x}_q$ względem r i MinPts, jeżeli istnieje obiekt $\mathbf{x}_p$ spełniający definicję łańcucha połączeń obiektów względem ustalonych r oraz MinPts.

Definicja ta przedstawia relację symetryczną.

Zgodnie z intuicją grupę można przedstawić jako skupisko obiektów spełniających definicję 4, co matematycznie wyraża się następująco:

Definicja 5: Grupa
Niech D będzie zbiorem obiektów, a grupa C utworzona względem r i MinPts jest niepustym podzbiorem obiektów ze zbioru D, spełniającym następujące warunki:

1) $\forall \mathbf{x}_i, \mathbf{x}_q$: jeśli C i $\mathbf{x}_q$ są łańcuchem połączeń obiektów zgodnie z ustalonymi r oraz MinPts, wówczas $\mathbf{x}_q \in C$
2) $\forall \mathbf{x}_i, \mathbf{x}_q \in C$: połączone ze względu na gęstość dla $\mathbf{x}_q$, dla r i MinPts.

Definicja 6: Szum
Niech $C_1, \ldots, C_i$ będą grupami zawartymi w zbiorze D, względem parametrów r oraz MinPts, dla i=1,…, k. Wtedy, szumem będą obiekty, które nie należą do żadnej grupy C_i, wówczas szum można wyrazić jako:

$$\text{szum} = \{\mathbf{x}_i \in D | \forall i: \mathbf{x}_i \notin C_i\}$$

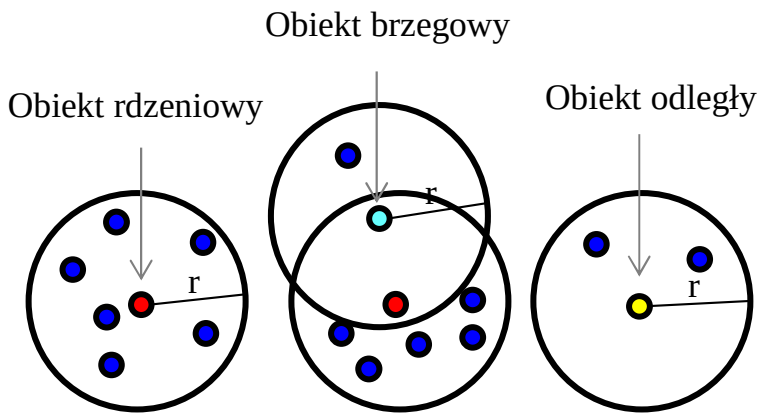
9.3.1 Algorytm DBSCAN

Metody hierarchiczne i niehierarchiczne w eksploracji wielowymiarowych zestawów danych wykazują wysoką efektywność. Jednak, towarzyszące im ograniczenia tj. minimalna wiedza o strukturze eksplorowanych danych, określenie parametrów wejścia, uzależnienie poszukiwanych typów grup obiektów od stosowanej miary odległości oraz problem z analizą danych liczących wysoką liczbę obiektów, np. kilka tysięcy, mogą wpływać na powodzenie przeprowadzanego grupowania. Problemy te rozwiązuje zastosowanie algorytmu DBSCAN, należącego do metod grupowania, wykorzystującego kryterium gęstości. Algorytm ten wymaga wyłącznie określenia promienia sąsiedztwa oraz minimalnej liczby obiektów określanych jako grupa. Dodatkowo, zgodnie z koncepcją kryterium gęstości, pozwala na wyodrębnienie grup obiektów o arbitralnych kształtach. Ponadto, wykazuje się większą efektywnością podczas grupowania bardzo dużych zestawów danych.
Oryginalna metoda DBSCAN opisana w [92] raczej nie jest stosowana w chemometrii. Stosuje się natomiast metodę naturalnych ugrupowań (z ang. Natural Patterns; NP) [94], która powstała w oparciu o bazowy algorytm DBSCAN. Algorytm NP pozwala na oszacowanie liczby naturalnych grup obiektów występujących w zestawie danych oraz wymaga sprecyzowania tylko jednego parametru wejścia [94]. Jednak w dalszej części pracy zdecydowano się nazwę DBSCAN wykorzystywać w odniesieniu do jej zastosowań w kontekście chemometrycznej eksploracji danych.
Algorytm DBSCAN jest metodą umożliwiającą odkrywanie grup obiektów podobnych w środowisku szumu zgodnie z Definicjami 5 oraz 6. Metoda ta idealnie sprawdza się podczas grupowania danych wielowymiarowych zawierających więcej niż kilka tysięcy parametrów. Przeszukiwanie przestrzeni rozpoczyna się od wyznaczenia tzw. obiektu rdzeniowego. Następnie przeszukuje się jego najbliższe otoczenie zataczając okrąg o promieniu r. Wszystkie obiekty znajdujące się w okręgu zostają dopisane do listy członków danej grupy, a następnie przeszukuje się otoczenie o ustalonym promieniu wokół każdego z tych obiektów. Czynności te powtarza się dopóki nie znajdzie się więcej obiektów w przeszukiwanej przestrzeni. Jest to równoznaczne ze znalezieniem kompletnej grupy. Jeżeli pozostały obiekty, których nie przypisano do żadnej z grup, przeszukiwanie rozpoczyna się ponownie. Algorytm działa tak długo, aż wszystkie obiekty zostaną przetworzone. Metoda DBSCAN jest tzw. metodą pojedynczego skanowania danych. Oznacza to, że algorytm przeszukuje przestrzeń pomiarową, w celu utworzenia grup obiektów sąsiadujących i spełniających kryterium gęstości, tylko raz.
Działanie algorytmu DBSCAN można przedstawić w następujących krokach:

1) Zdefiniowanie parametrów wejścia: promienia sąsiedztwa r oraz minimalnej liczby obiektów MinPts w promieniu sąsiedztwa, rozważanych jako grupa,
2) Oznaczenie obiektów rdzeniowych w obszarze wyznaczonym przez promień sąsiedztwa,

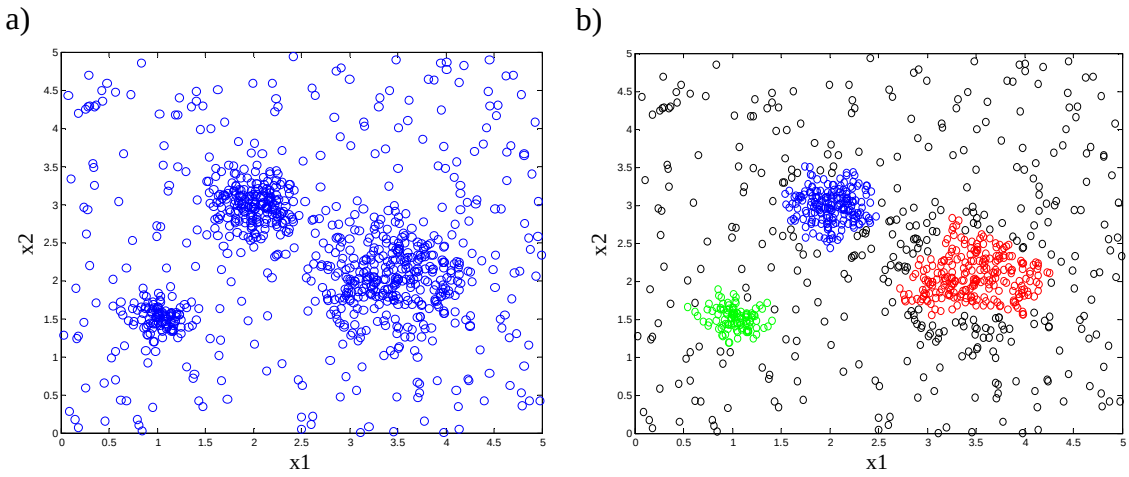
3) Wyszukanie sąsiadów obiektów rdzeniowych w promieniu r i ich dodanie do listy przynależności,
4) Określenie typów obiektów znajdujących się na liście przynależności, poprzez dopisanie do listy wszystkich ich sąsiadów znajdujących się w promieniu r,
5) Usunięcie z listy przynależności przetworzonych obiektów,
6) Jeżeli lista przynależności jest pusta, oznacza to że grupa jest kompletna i należy wrócić do punktu nr 2,
7) Jeżeli na liście pozostają obiekty posiadające w promieniu sąsiedztwa mniej niż MinPts obiektów należy oznaczyć je jako „szum”.

Grupowane obiekty można podzielić na trzy typy. Pierwsze z nich to tzw. obiekty rdzeniowe, tj. posiadające w swoim otoczeniu więcej niż MinPts obiektów, następnie obiekty brzegowe posiadające mniej niż MinPts sąsiadów w otoczeniu, z których przynajmniej jeden jest obiektem rdzeniowym oraz obiekty odległe (szum), które w swoim otoczeniu posiadają mniej niż MinPts obiektów i żaden z nich nie jest obiektem rdzeniowym (Rys. 15).



Rys. 15 Rodzaje obiektów w metodzie DBSCAN, przyjmując liczbę sąsiadów równą 6 oraz promień sąsiedztwa r.

Przeważającym atutem metody DBSCAN nad innymi metodami jest możliwość detekcji grup obiektów w środowisku szumu. Szum pomiarowy traktowany jest zazwyczaj jak obiekty odległe, dzięki czemu możliwe jest wyodrębnienie naturalnych grup obiektów z wyraźnym wyizolowaniem ze środowiska danych szumu przy odpowiednio dobranym promieniu sąsiedztwa i liczbie sąsiadów (Rys. 16).



Rys. 16 Wyodrębnienie grup obiektów obecnych w środowisku szumu pomiarowego za pomocą algorytmu DBSCAN na przykładzie symulowanych danych (900 obiektów w dwuwymiarowej przestrzeni zdefiniowanej przez parametry x1 i x2).

9.3.2 Algorytm OPTICS

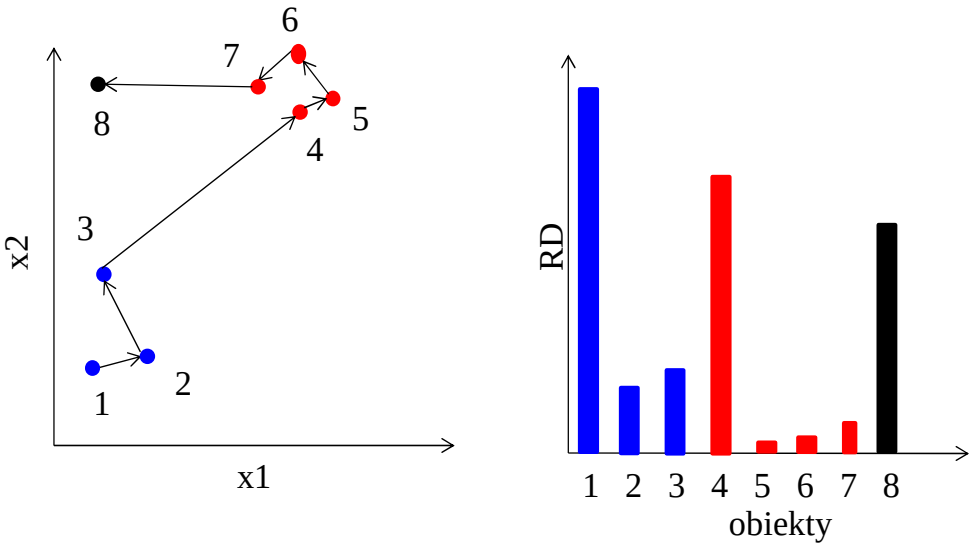
Algorytm OPTICS jest drugim powszechnie stosowanym algorytmem bazującym na analizie gęstości danych. Często określany jest jako rozszerzona wersja algorytmu DBSCAN [96]. Wykorzystywany jest przede wszystkim w celu ujawnienia struktury danych na podstawie przyjętej miary podobieństwa RD. RD dla i-tego obiektu jest maksymalną wartością pomiędzy dwoma odległościami: odległością euklidesową i-tego obiektu i jego najbliższym q-tym sąsiadem oraz odległością euklidesową, zwaną odległością rdzeniową CD, będącą odległością pomiędzy i-tym obiektem oraz jego k-tym sąsiadem.

$$RD_i = \max(d_{iq}, CD_i) \tag{17}$$

W rzeczywistości miara RD_i jest abstrakcyjną miarą podobieństwa, ponieważ wszystkie obiekty dla których q-ty obiekt jest najbliższym sąsiadem mają RD równe CD_q.
Ponieważ algorytm ten jest algorytmem bazującym na kryterium gęstości to wymaga zdefiniowania wielkości promienia sąsiedztwa jako odległości euklidesowej k-tego najbliższego sąsiada. W tym wypadku jest to odległość euklidesowa pomiędzy i-tym oraz k-tym najbliższym sąsiadem, a więc wspomniana odległość CD.
Działanie algorytmu OPTICS można przedstawić następująco:

1) Zdefiniowanie minimalnej liczby obiektów k, rozważanych jako grupa,
2) Wybranie losowo jednego obiektu, od którego nastąpi proces przetwarzania obiektów. Oznaczenie wybranego obiektu jako przetworzonego, poprzez umieszczenie go na początku listy przetworzonych obiektów,
3) Wybranie kolejnego obiektu, którego RD, względem poprzednio przetworzonego jest najmniejsze i dodanie go do listy przetworzonych obiektów. Uznanie nowo przetworzonego obiektu za prekursora w poszukiwaniu kolejnego obiektu,
4) Obliczenie RD kolejnych obiektów, względem poprzednio przetworzonego obiektu i ich dopisanie do listy przetworzonych obiektów z nowo obliczonym RD,
5) Powracanie do kroku nr 2 tak długo aż wszystkie obiekty zostaną przetworzone.

Zasadnicza różnica pomiędzy algorytmem DBSCAN, a OPTICS jest taka, że w metodzie OPTICS ważna jest kolejność przetwarzania obiektów, co reprezentuje się za pomocą tzw. wykresu połączeń obiektów (Rys. 17).



Rys. 17 Tworzenie wykresu połączeń obiektów za pomocą algorytmu OPTICS dla przykładowego rozkładu obiektów w przestrzeni dwuwymiarowej.

Wykres ten na osi y zawiera wartości RD dla kolejno przetwarzanych obiektów przedstawionych na osi x. Interpretacja otrzymanych wyników jest relatywnie prosta i pozwala na wyciagnięcie wniosków dotyczących ilości grup tworzonych przez przetwarzane obiekty. Każda nowa grupa rozpoczyna się wartością RD wyższą od przetworzonego obiektu. Z kolei im gęstsza jest grupa tym niższe wartości RD reprezentują obiekty należące do danej grupy.

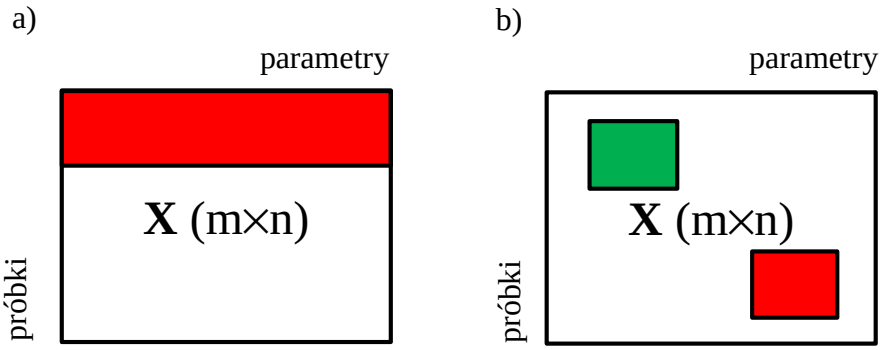
9.4 Grupowanie oparte na modelu statystycznym

Jak wspomniano wcześniej metody grupowania danych należą do metod, w których nie dokonuje się żadnych założeń dotyczących eksplorowanych danych i nie wymagają one wiedzy na ich temat – metody uczenia bez nadzoru. Istnieje jednak metoda, w której wprowadza się założenia dotyczące rozkładu grup w przestrzeni eksperymentalnej w kontekście statystycznym. Metoda ta to metoda grupowania danych oparta na modelu statystycznym (z ang. Model-Based Clustering) [67]. Oparta jest ona na założeniu, że rozkład obiektów w przestrzeni parametrów jest rozkładem normalnym.
Każda grupa reprezentowana jest jako model statystyczny np. rozkład normalny reprezentowany przez krzywą Gaussa. Kształt rozkładu grupy modeluje się na podstawie macierzy wariancji-kowariancji, $\mathbf{C} = \sigma\mathbf{I}$. Następnie dane przedstawia się za pomocą rozkładu normalnego, którego kształt jest uzależniony od macierzy **C**. Może być on sferyczny lub elipsoidalny o odpowiedniej orientacji elipsy w przestrzeni eksperymentalnej. Zastosowany model rozkładu ma jak najlepiej opisać rozkład obiektów w przestrzeni parametrów. Dane modeluje się w ten sposób kilkakrotnie, a następnie wybiera się model statystyczny, który w najbardziej wiarygodny sposób reprezentuje dane, a tym samym pozwala na wyodrębnienie grup obiektów podobnych. Grupowanie tego typu jest zagadnieniem na pograniczu metod grupowania jednoznacznego oraz rozmytego. Gdyż ostatecznie każdy obiekt jest przypisany do każdej grupy z pewnym prawdopodobieństwem.

10. Metody współgrupowania danych

Metody grupowania danych umożliwiają eksplorację wielowymiarowych danych, a ich przydatność dokumentują liczne publikacje naukowe. Znajdują zastosowanie m.in. przy wyborze próbek podczas konstrukcji modeli kalibracyjnych [97], detekcji fluktuacji warunków środowiskowych w monitoringu środowiska [98], czy określaniu liczby związków chemicznych obecnych w próbce [99]. Przykładów zastosowań analizy klasterowej jest wiele, jednak pojawienie nowych nauk typu „-omika” takich jak proteomika, genomika czy metabolomika, położyło nacisk na poszukiwanie nowych rozwiązań pozwalających na równoczesną analizę wierszy i kolumn macierzy. Dyscypliny te wymagają użycia narzędzi ujawniających wpływ parametrów na grupowanie próbek, czyli poszukujących interakcji pomiędzy wierszami, a kolumnami macierzy danych oraz takich, które pozwoliłyby na analizę danych pomimo znacznej niepewności pomiarowej będącej skutkiem np. różnorodności biologicznej, wynoszącej do 30%. Narzędziem takim wydają się metody współgrupowania danych.
Metody współgrupowania danych (z ang. Co-clustering) [100] są technikami o podobnym przeznaczeniu jak metody grupowania danych. Pierwsze metody tego typu powstały w latach 60 ubiegłego wieku, jednak początkowo nie znalazły one praktycznego zastosowania. Termin „współgrupowanie” wykorzystał po raz pierwszy

w 1996 roku Mirkin [101] w celu podkreślenia równoczesnego grupowania wierszy i kolumn macierzy danych.
W przypadku metod grupowania danych, grupy obiektów są opisane przez wszystkie zmierzone parametry. Natomiast metody współgrupowania umożliwiają wyodrębnienie grup obiektów wykazujących podobieństwo wyłącznie w zakresie pewnej grupy parametrów (Rys. 18). Uściślając, celem metod współgrupowania jest znalezienie podmacierzy w wyjściowej macierzy danych **X**. Podmacierz ta, charakteryzuje się spójnymi (koherentnymi) wartościami w wierszach i kolumnach.



Rys. 18 Efekt grupowania za pomocą a) klasycznych metod grupowania danych oraz b) metod współgrupowania.

Atutem metod współgrupowania danych jest możliwość wyodrębnienia nakładających się na siebie grup, czego nie umożliwiają klasyczne metody grupowania danych. Oznacza to, iż jedna próbka może jednocześnie należeć do różnych grup. Cecha ta odgrywa istotną rolę w przypadku analizy danych biologicznych. Szczególnym zainteresowaniem cieszy się metodologia współgrupowania w obszarze badań danych genomicznych, umożliwiając wyodrębnienie grup genów wykazujących zbliżoną koekspresję w określonych warunkach pomiarowych. Z kolei samą ekspresję genów określa się za pomocą mikromacierzy wykorzystując reakcję hybrydyzacji nici DNA [101].
Od 2000 roku trwają intensywne prace nad rozwojem algorytmów współgrupowania danych. Wśród nich można wyróżnić algorytmy poszukujące grup o stałych wartościach w całej podmacierzy, stałych wartościach w wierszach lub kolumnach, albo algorytmy poszukujące konkretnego uporządkowania i koherentności wartości wierszy i/lub kolumn macierzy [103]. Ze względu na specyficzność wyników chemicznych, w których każdy wynik jest wypadkową wartości rzeczywistej oraz błędów (np. błędy aparaturowe, systematyczne, przypadkowe), zwłaszcza ostatnia grupa metod współgrupowania wydaje się być interesująca. Dane chemiczne

charakteryzuje struktura, w której informacja jest ukryta, dlatego techniki umożliwiające przeszukiwanie podprzestrzeni w taki sposób, aby wyodrębnić prawidłowości zawarte w danych wydają się najodpowiedniejsze. Poniżej, w podrozdziale *10.1* omówiono wybrane metody współgrupowania danych.

10.1 Wybrane algorytmy współgrupowania danych

10.1.1 Algorytm CC

Klasycznym algorytmem współgrupowania danych jest algorytm zaproponowany przez Chenga oraz Churcha [104], zwany również algorytmem CC od pierwszych liter nazwisk twórców metody lub tzw. algorytmem grupowania blokowego. Jest to najlepiej poznany algorytm tego typu metod, który służy jako metoda odniesienia dla tworzenia i porównywania nowych algorytmów.
Cheng i Church zdefiniowali pojęcie grupy jako podzbiór wierszy i kolumn macierzy danych o dużym współczynniku podobieństwa, który stanowi średni błąd kwadratowy (H(I, J)) [104]. Średni błąd kwadratowy, dla danego podzbioru wartości macierzy danych, definiowany jest następująco:

$$H(I,J) = \frac{1}{|I||J|}\sum_{i\epsilon I, j\epsilon J}(a_{ij} - a_{iJ} + a_{Ij} + a_{IJ})^2 \quad (18)$$

gdzie:
H(I, J) – średni błąd kwadratowy podgrupy A_{IJ}
I – zbiór wierszy w podgrupie
J – zbiór kolumn w podgrupie
a_{ij} – element podgrupy z i ϵ I oraz j ϵ J
a_{iJ} – średnia wartości wierszy
a_{Ij} – średnia wartości kolumn
a_{IJ} – średnia całej podgrupy

Średni błąd kwadratowy jest wykorzystany jako miara spójności wierszy i kolumn w dowolnej podmacierzy. Z założenia, algorytm poszukuje jak największych podmacierzy, których wartość H(I, J) nie może przekraczać pewnej wartości granicznej δ. Aby warunki te zostały spełnione, algorytm rozpoczyna poszukiwania od jak największej podmacierzy, którą najczęściej stanowi cała macierz danych, a następnie usuwa z niej wiersze i kolumny tak, aby zminimalizować średni błąd kwadratowy. Każdorazowo otrzymuje się jedna grupę, którą następnie „usuwa się” z danych poprzez

zastąpienie odpowiadających jej wartości elementów wartościami losowymi. Nowo utworzona macierz staje się zbiorem danych poddanych działaniu algorytmu. Czynności te powtarza się tak długo, aż wszystkie podgrupy zostaną znalezione.

10.1.2 Algorytm k-spectral

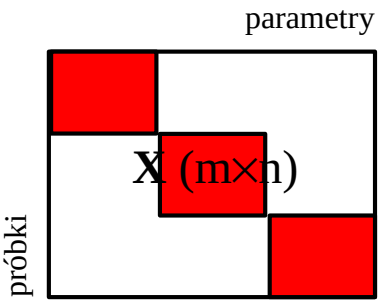
Kolejnym powszechnie stosowanym algorytmem współgrupowania danych jest tzw. algorytm k-spectral. Został on zaproponowany przez Klugera [105] i był pierwotnie dedykowany grupowaniu profili nowotworowych charakteryzowanych za pomocą mikromacierzy RNA. Równoczesne grupowanie wierszy i kolumn na podstawie korelacji profili umożliwia utworzenie podgrup.
W omawianej metodzie współgrupowanie poprzedzone jest odpowiednio dobraną normalizacją macierzy danych. Etap normalizacji umożliwia usunięcie niepożądanych efektów będących przyczyną błędów eksperymentalnych, np. fluktuacji warunków pomiarowych. Specjalnie w tym celu Kluger wprowadził tzw. normalizację bistochastyczną, polegająca na usuwaniu wartości średniej kolumn i wierszy, aż do uzbieżnienia wyników. Jednak równie dobrze można zastosować normalizację logarytmiczną. Należy jednak pamiętać, iż dobór sposobu normalizacji danych będzie wpływał na powodzenie współgrupowania danych w kolejnych etapach eksploracji. Metoda k-spectral jest metodą umożliwiającą ujawnienie w danych tzw. struktury szachownicy (Rys. 19). W celu identyfikacji tej struktury wykorzystuje się techniki z zakresu algebry liniowej, zakładając że strukturę szachownicy opisują tzw. wektory klasyfikacji **x** i **y** odpowiednio dla wierszy oraz kolumn, co matematycznie można przedstawić następująco:

$$\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{x} = \lambda^2\mathbf{x} \quad \text{oraz} \quad \mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{y} = \lambda^2\mathbf{y} \tag{19}$$

gdzie:
X – macierz danych
x, **y** – wektory własne
λ – wartość własna

W rzeczywistości równanie (19) jest równoznaczne z dekompozycją macierzy danych uzyskiwaną za pomocą algorytmu dekompozycji macierz danych **X** na wektory własne i wartości własne (SVD). Organizacja wierszy oraz kolumn macierzy **X** zgodnie z wektorami klasyfikacji ujawnia poszukiwaną w danych strukturę szachownicy. Aby uzyskać informację o przynależności kolumn i wierszy macierzy do poszczególnych grup, zmodyfikowane macierze poddaje się grupowaniu algorytmem k-średnich. Metoda k-spectral, w porównaniu do innych metod współgrupowania

danych, wykazuje charakter metody globalnej, poprzez uwzględnienie w obliczeniach wszystkich kolumn i wierszy podczas poszukiwania podmacierzy.



Rys. 19 Macierz danych **X** o strukturze szachownicy.

10.1.3 Algorytm regresji macierzy rzadkiej

Algorytm tzw. regresji macierzy rzadkiej (z ang. Sparse Matrix Regression; SMR) jest metodą wprowadzoną przez Bro [106]. Pozwala na wyodrębnienie podgrup obiektów oraz parametrów za pomocą zmodyfikowanego sposobu dekompozycji macierzy **X**, polegającego na specyficznym ważeniu elementów macierzy danych poprzez nałożenie tzw. funkcji rzadkości (z ang. Sparse) na każdy czynnik modelu dwuliniowego. Funkcja rzadkości pozwala na rozróżnienie obiektów i parametrów istotnych, w kontekście przynależności do danej podgrupy, od tych które nie wnoszą wkładu do jej utworzenia. W efekcie, poszczególne czynniki macierzy wyników oraz wag zawierają informację o przynależności obiektów i parametrów do danej podgrupy. Zastosowana koncepcja rzadkości, pozwala na przypisanie odpowiednich wierszy i kolumn do konkretnej podgrupy, przedstawiając wszystkie inne wartości współczynników nie należących do grupy jako równe 0. Im wyższą wartość z przedziału od 0 do 1 przyjmuje dany obiekt (parametr) tym większe jest prawdopodobieństwo, że należy on do określonej grupy. Niektóre wiersze i kolumny macierzy **X** mogą nie należeć do żadnej z podgrup lub należeć do tzw. grup nakładających się na siebie.
Matematycznie dekompozycja macierzy **X** za pomocą metody SMR polega na minimalizacji funkcji kosztów, co przedstawia równanie (20).
Tym sposobem metoda SMR łączy cechy wspomnianej, w podrozdziale *8.1.1*, metody PCA oraz metod grupowania danych, umożliwiając tym samym wyodrębnienie podgrup obiektów i parametrów. Równocześnie, algorytm SMR jest metodą współgrupowania danych, w której każdy obiekt należy do wszystkich wyodrębnionych grup, i można ją porównać do metod grupowania rozmytego, o których wspominano w rozdziale 9.

$$\left\|\mathbf{X}-\mathbf{A}\mathbf{B}^{\mathrm{T}}\right\|_{\mathrm{F}}^{2}+\lambda\sum_{\mathrm{i,k}}\|\mathbf{A}_{\mathrm{ik}}\|+\lambda\sum_{\mathrm{j,k}}\left\|\mathbf{B}_{\mathrm{jk}}\right\| \quad (20)$$

gdzie:
X – macierz danych
A – macierz wyników (m×k)
B – macierz wag (k×n)
k – liczba wyodrębnionych podgrup
$\sum_{i,k}, \sum_{j,k}$ – sumy absolutnych wartości będące substytutem nałożonej funkcji rzadkości dla elementów niezerowych
λ – współczynnik rzadkości

10.1.4 Metody wyboru zmiennych

Zastosowanie metod eksploracyjnych w celu ekstrakcji chemicznie istotnej informacji często sprowadza się do wyodrębnienia tzw. głównych efektów. A więc informacji jaka dominuje w danych i determinuje obserwowany podział obiektów na grupy. Zdarza się jednak, że dane zwierają znacznie więcej informacji niż ujawniają główne efekty. Informacja ta najczęściej ukryta jest w podprzestrzeniach macierzy danych **X**.
Znanych jest wiele metod umożliwiających przeszukiwanie podprzestrzeni danych poprzez np. wyodrębnienie podmacierzy danych. Należą do nich metody współgrupowania danych opisane w tym rozdziale. Innym przykładem są metody wyboru istotnych zmiennych (z ang. Feature Selection).
Metody te należą do algorytmów ewolucyjnych, naśladujących swym działaniem zachowania społeczne organizmów żywych egzystujących w środowisku naturalnym w zorganizowanych populacjach np. ławice ryb, klucze ptaków. Osobniki żyjące w danej zbiorowości oddziałują na siebie i równocześnie oddziałuje na nie środowisko w którym żyją. Osobniki te, określane mianem cząstek, posiadają zdolność zapamiętywania swojego położenia i przystosowują się do środowiska w jakim żyją, wybierając przy tym najkorzystniejsze warunki. Równocześnie cząstki cały czas poszukują lepszej lokalizacji (rozwiązań). W związku z czym, posiadają one położenie i prędkość z jaką poruszają się w przestrzeni pomiarowej w celu znalezienia optymalnego rozwiązania.
Jedną z metod wyboru zmiennych jest metoda optymalizacji z użyciem roju cząstek (z ang. Particle Swarm Optimization; PSO). Po raz pierwszy zaproponowana w 1995 r. przez Kennedy'ego i Eberharta [107]. PSO należy do algorytmów optymalizujących w kontekście globalnym. W tej metodzie wszystkie możliwe rozwiązania traktuje się jako rój, natomiast każde rozwiązanie z osobna jako położenie danej cząstki w przestrzeni pomiarowej.
W kolejnych iteracjach cząstki przemieszczają się poszukując najlepszego położenia, a więc rozwiązania stanowiącego optimum funkcji wielu zmiennych. Każda cząstka ma

swoich sąsiadów, których określa się na początku obliczeń. Liderem sąsiadów zostaje ta cząstka w sąsiedztwie, która wykazuje najlepsze położenie. Następnie spośród wszystkich znalezionych rozwiązań wskazuje się lidera, który charakteryzuje się najlepszym położeniem z wszystkich znalezionych rozwiązań. Na początku cząstki reprezentują rój (populację) możliwych rozwiązań. Przyjmują one losowe położenie i prędkość w przestrzeni parametrów. Zasady komunikacji pozwalają na kontaktowanie się poszczególnych cząstek ze sobą i przepływ informacji. Następnie, kiedy cząstki zaczynają się poruszać, poszukiwana zostaje trajektoria ruchu, wyznaczająca optymalne rozwiązanie. Trajektoria ta jest stochastyczna nie deterministyczna, dlatego bierze pod uwagę informacje o każdej cząstce (dopasowanie poszczególnych cząstek) oraz o innych cząstkach (globalne dopasowanie). Informacje te zostają zaktualizowane za każdym razem gdy cząstki znajdują lepsze rozwiązanie.
Matematycznie, trajektorię ruchu dla cząstki i o współrzędnych $\mathbf{x}_i$ i prędkości $\mathbf{v}_i$ w przestrzeni n-parametrowej, wyraża się następująco:

$$\mathbf{x}_i(t) = \mathbf{x}_i(t-1) + \mathbf{v}_i(t) \tag{21}$$

gdzie:
$\mathbf{x}_i$ – i-ta cząstka
$\mathbf{v}_i$ – prędkość i-tej cząstki
t – numer kolejnej iteracji

Po każdej iteracji sprawdza się wydajność każdej cząstki oraz ocenę jej prędkości zgodnie z równaniem (22).

$$\mathbf{v}_i(t) = \mathbf{v}_i(t-1) + c_1(\mathbf{p}_i - \mathbf{x}_i)\mathbf{R}_1 + c_2(\mathbf{p}_g - \mathbf{x}_i)\mathbf{R}_2 \tag{22}$$

gdzie:
c_1, c_2 – współczynniki przyspieszenia
$\mathbf{p}_i$ – dopasowanie i-tej cząstki
$\mathbf{p}_g$ – globalne dopasowanie cząstek
$\mathbf{R}_1$, $\mathbf{R}_2$ – macierze z losowymi elementami z rozkładu równomiernego mieszczącego się w przedziale [0,1]

Aby algorytm PSO mógł zostać wykorzystany w celu wyboru istotnych zmiennych należy wprowadzić do niego modyfikację [108]. Współrzędne cząstek w algorytmie PSO przyjmują wartości 1 lub 0. W ten sposób oznacza się zmienne istotne i nieistotne.

Zmienne które otrzymały wartość 1 zawierają informację istotną, podczas gdy te którym nadano wartość 0 nie odgrywają znaczącej roli i mogą zostać usunięte. Modyfikacja ta sprawia że algorytm realizuje binarny wybór zmiennych, w związku z czym wiąże się to z przeszukiwaniem binarnej n wymiarowej przestrzeni pomiarowej. Prędkości z równania (22) w binarnej wersji algorytmu PSO przedstawia się jako prawdopodobieństwo, zmiany statusu. Jeżeli dla $\mathbf{v}_{id}$, gdzie d jest d-tym elementem wektora prędkości cząstki i, jest równe 0,75 to można powiedzieć, że z prawdopodobieństwem 75% nastąpi zmiana położenia. Matematycznie prędkość wyraża się za pomocą następującego wzoru:

$$\mathbf{v'}_{\mathrm{id}} = \frac{1}{1+ \mathrm{e}^{-\mathrm{v}_{\mathrm{id}}}} \tag{23}$$

Metodę PSO można wykorzystać w celu identyfikacji grup obiektów. Obiekty prezentuje się w przestrzeni zdefiniowanej przez wybrane parametry. Następnie relacje pomiędzy obiektami przedstawia się za pomocą metod grupowania hierarchicznego. Ponieważ metody łączenia obiektów wpływają na jakość obserwowanych wyników w celu ich oceny wprowadzono następujące kryterium dopasowania:

$$\text{dopasowanie} = \frac{\mathrm{d}_{\mathrm{ost}}}{\mathrm{d}_{\mathrm{ost}-1}} \tag{24}$$

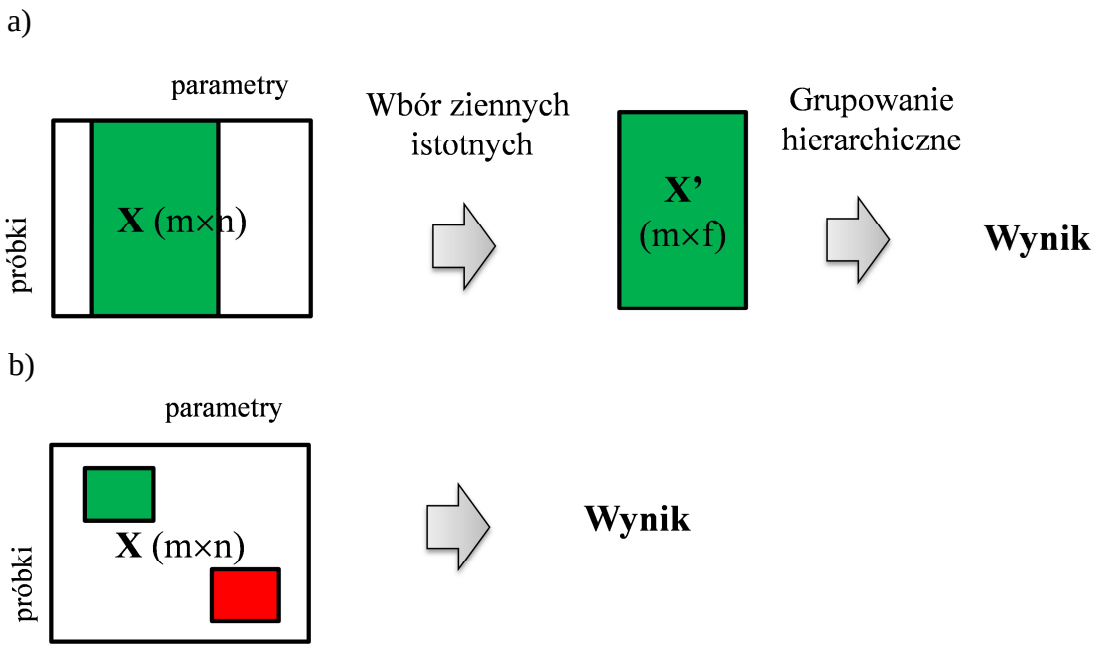
gdzie:
d_{ost} – ostatnie połączenie obiektów w dendrogramie
$d_{ost} – 1$ – przedostatnie połączenie obiektów w dendrogramie

Im wyższa wartość kryterium dopasowania tym lepszy podział na grupy. W celu zwiększenia wiarygodności otrzymywanych wyników oraz ograniczenia ryzyka wpływu obiektów odległych na obserwowane wyniki, liczbę obiektów tworzących grupę należy określić z góry.
Podsumowując, działanie algorytmu PSO można przedstawić w kilku krokach:

1) Inicjalizacja populacji cząstek z losowych wektorów binarnych,
2) Przedstawienie wyników grupowania każdej cząstki w przestrzeni zdefiniowanej przez wybrane zmienne. Dokonanie oceny dopasowania zgodnie z równaniem (24),
3) Aktualizacja wyników dla nowo wybranych zmiennych w przypadku otrzymania w kolejnej iteracji lepszych rezultatów dla kryterium dopasowania dla poszczególnych cząstek,

4) Aktualizacja wyników dla nowo wybranych zmiennych w przypadku otrzymania w kolejnej iteracji lepszych rezultatów dla kryterium dopasowania globalnego
5) Aktualizacja prędkości cząstek zgodnie z równaniami (22 i 23),
6) Wygenerowanie dla każdego parametru losowej wartości r_{id} z przedziału [0,1]. Jeśli $v_{id} > r_{id}$ odwrócić odpowiadający fragment, w przeciwnym wypadku nie wprowadzać zmian,
7) Powracanie do kroku nr 2 do momentu osiągnięcia maksymalnej liczby iteracji i uzbieżnienia się algorytmu.

Metody wyboru zmiennych przeznaczone są przede wszystkim dla danych wykazujących dużą korelację. Bardzo dobrze sprawdzają się w eksploracji danych typu „–mika" np. metabolomika, gdzie pozwalają na ujawnienie istotnych z punktu widzenia biologii, szlaków metabolomicznych lub w identyfikacji biomarkerów.
Z pozoru metody wyboru zmiennych oraz metody współgrupowania danych wydają się podobne. Różnica w działaniu algorytmów jest jednak zasadnicza. Metody wyboru zmiennych przedstawiają grupy obiektów w przestrzeni wybranych parametrów. Jednak w dalszym ciągu otrzymuje się informacje o wszystkich obiektach. Dodatkowo obiekty zostają przypisane wyłącznie do jednej grupy. W przypadku metod współgrupowania danych, jeden obiekt może należeć do kliku grup lub do żadnej z nich. A obiekty należące do danej grupy wykazują podobieństwo ze względu na konkretną podgrupę parametrów.
Na Rys. 19 schematycznie przedstawiono różnice w działaniu algorytmów wyboru zmiennych oraz współgrupowania danych.
Jak łatwo zauważyć w przypadku metod wyboru zmiennych definiuje się nową macierz danych **X**' o wymiarowości m×f, gdzie m oznacza liczbę obiektów, a f liczbę wybranych zmiennych. Następnie relacje pomiędzy obiektami, w przestrzeni zmiennych istotnych, wizualizowane są za pomocą dendrogramu. Dopiero interpretacja wyników uzyskanych za pomocą grupowania hierarchicznego oraz określenie kryterium dopasowania umożliwia interpretację wyników oraz formułowanie wniosków. W przypadku metod współgrupowania danych otrzymuje się podmacierze danych reprezentujące obiekty wykazujące podobieństwo determinowane przez podgrupy parametrów. Znalezienie podmacierzy w głównej macierzy danych **X**, pozwala na interpretację oraz ostateczne wnioski.



Rys. 19 Schematyczne porównanie działania algorytmów a) wyboru zmiennych oraz b) współgrupowania danych [97].

11. Obszary zastosowań metod eksploracji danych

Metody grupowania i współgrupowania danych znajdują zastosowanie w wielu dziedzinach nauki. Przede wszystkim w chemii, monitoringu środowiska, metabolomice, genomice, proteomice, czy medycynie, stanowiąc kluczowe narzędzie eksploracyjne. Prowadzą do interpretacji wyników i ostatecznych konkluzji.

Liczba publikacji naukowych potwierdzająca użyteczność opisanych metod jest ogromna i nie sposób wymienić ich wszystkich. Z tego względu wybrano kilka przykładów zastosowań metod eksploracyjnych.

Większość zastosowań metod grupowania ma na celu wyłonienie kompaktowej reprezentacji danych jakościowych. Cel ten osiąga się poprzez grupowanie obiektów i/lub parametrów na podstawie przyjętej miary podobieństwa. Przedstawienie wielowymiarowych danych w postaci kilku grup obiektów lub parametrów, znacznie upraszcza interpretację problemu badawczego. Przykładowo, metody grupowania danych mogą zostać wykorzystane do wyboru reprezentatywnych próbek do konstrukcji modelu kalibracyjnego [97].

Liczba grup może charakteryzować liczbę związków chemicznych występujących w analizowanych próbkach. Z sytuacją taką spotykamy się np. podczas eksploracji obrazów hiperspektralnych, gdzie metody grupowania wykorzystywane są w celu

detekcji obszarów obrazu reprezentujących odmienny skład chemiczny [99], [109]. W naukach środowiskowych, metody grupowania znajdują zastosowanie podczas detekcji lokalnych zmian klimatu, składu powietrza, wody lub gleby [98], [110]. Techniki grupowania odgrywają również znaczącą rolę w naukach biologicznych, ze szczególnym uwzględnieniem biologii systemowej. Służą m.in. do klasyfikacji gatunków roślin na postawie uzyskanych profili chemicznych [111], np. kwasów tłuszczowych [111], steroli [113] – tzw. chemotaksonomia. W genomice są podstawowym narzędziem grupowania genów wykazujących zbliżoną koekspresję [114]. W proteomice umożliwią identyfikację protein oraz zmodyfikowanych peptydów [115]. W kolejnej dyscyplinie naukowej z tego zakresu, jaką jest metabolomika, narzędzia grupowania umożliwiają detekcję poszczególnych jednostek chorobowych poprzez uwzględnienie metabolitów zawartych w próbkach płynów ustrojowych (mocz, krew, osocze, płyn mózgowo-rdzeniowy) [116]. Metody grupowania można również stosować jako narzędzia diagnostyczne pozwalające przypisać nieznane próbki do istniejących już grup odpowiadających określonym jednostkom chorobowym np. choroby Parkinsona, czy Huntingtona [117].

Wymienione przykłady zastosowań technik grupowania różnią się znacząco od siebie. Jednak zawsze ich rola polega na wsparciu etapu poznawania struktury danych i wyodrębnieniu ukrytej informacji chemicznej.

12. Badania własne

W części teoretycznej niniejszej pracy omówiono większość podstawowych metod grupowania oraz współgrupowania wielowymiarowych danych. W części obejmującej badania własne, szerzej przedyskutowano wyniki badań uzyskane przez zastosowanie wybranych algorytmów grupowania oraz współgrupowania danych. Skupiono się przede wszystkim na przedstawieniu i omówieniu tych modyfikacji, które korelują z celami pracy, tj. działanie zmodyfikowanego algorytmu DBSCAN, wprowadzenie nowej miary podobieństwa próbek, omówienie problemów związanych z zastosowaniem metod współgrupowania danych w kontekście analizy danych chemicznych, takich jak sygnały instrumentalne. Uwzględniono również problem występowania niepewności pomiarowych obserwowanych w danych analitycznych w kontekście ich późniejszego grupowania.

12.1 Modyfikacja metody DBSCAN

Metoda DBSCAN jest metodą bazującą na gęstości danych o czym pisano szerzej w podrozdziale *9.3.1* tej pracy. Metoda ta w porównaniu do metod grupowania hierarchicznego oraz niehierarchicznego wykazuje unikalne właściwości. Przede wszystkim umożliwia wykrycie naturalnych grup obiektów o dowolnym kształcie i rozmiarze. Dodatkową zaletą tego algorytmu jest relatywnie krótki czas grupowania danych. Najbardziej czasochłonnym etapem jest obliczenie odległości pomiędzy obiektami. Jednakże proces ten może zostać przyspieszony poprzez zastosowanie komputera o wyższej mocy obliczeniowej.
Największa wadą algorytmu DBSCAN jest błędne przypisanie do grup obiektów brzegowych w przypadku tzw. grup sąsiadujących ze sobą. Przez grupy sąsiadujące rozumie się takie grupy obiektów, które znajdują się w przestrzeni zmiennych bardzo blisko siebie. W przypadku większości danych wielowymiarowych grupy są dobrze rozdzielone i ich rozróżnienie za pomocą DBSCAN jest łatwe. Jednak w przypadku danych charakteryzujących się dużą gęstością jak np. dane z trójwymiarowej tomografii tkanki nerwowej [117], grupy występują w przestrzeni eksperymentalnej bardzo blisko siebie. Wówczas o przynależności obiektów brzegowych do poszczególnych grup decyduje kolejność przetwarzania obiektów następująca w trakcie ich grupowania. Aby usprawnić działanie algorytmu i rozwiązać problem błędnego przypisywania obiektów brzegowych do odpowiednich grup, zaproponowano wprowadzenie poprawki do algorytmu, w celu polepszenia stabilności rozwiązania [118].
W metodzie DBSCAN, gęstość danych określona jest przez dwa parametry: promień sąsiedztwa, r, oraz minimalną liczbę obiektów w sąsiedztwie, MinPts. Aby poprawnie przypisać obiekty do grup w pierwszej kolejności należy zidentyfikować, które obiekty są obiektami brzegowymi, a które obiektami rdzeniowymi. Następuje to poprzez zdefiniowanie tzw. łańcucha gęstości obiektów (z ang. Density-Reachable Objects).

Matematycznie można przedstawić ten warunek następująco: dwa obiekty $\mathbf{x}_i$ i $\mathbf{x}_m$ spełniają warunek osiągnięcia gęstości, jeśli istnieje łańcuch obiektów od $\mathbf{x}_1,\ldots, \mathbf{x}_i, \mathbf{x}_{i+1},\ldots, \mathbf{x}_m$, taki że $i \geq 1$ i $m \geq 2$ dla wszystkich $i < m$, wtedy $\mathbf{x}_i$ jest obiektem rdzeniowym, jeżeli gęstość($\mathbf{x}_1$) $\geq$ min(MinPts) oraz $\mathbf{x}_{i+1}$ jest sąsiadem $\mathbf{x}_i$, $\mathbf{x}_{i+1} \in$ MinPts($\mathbf{x}_i$).

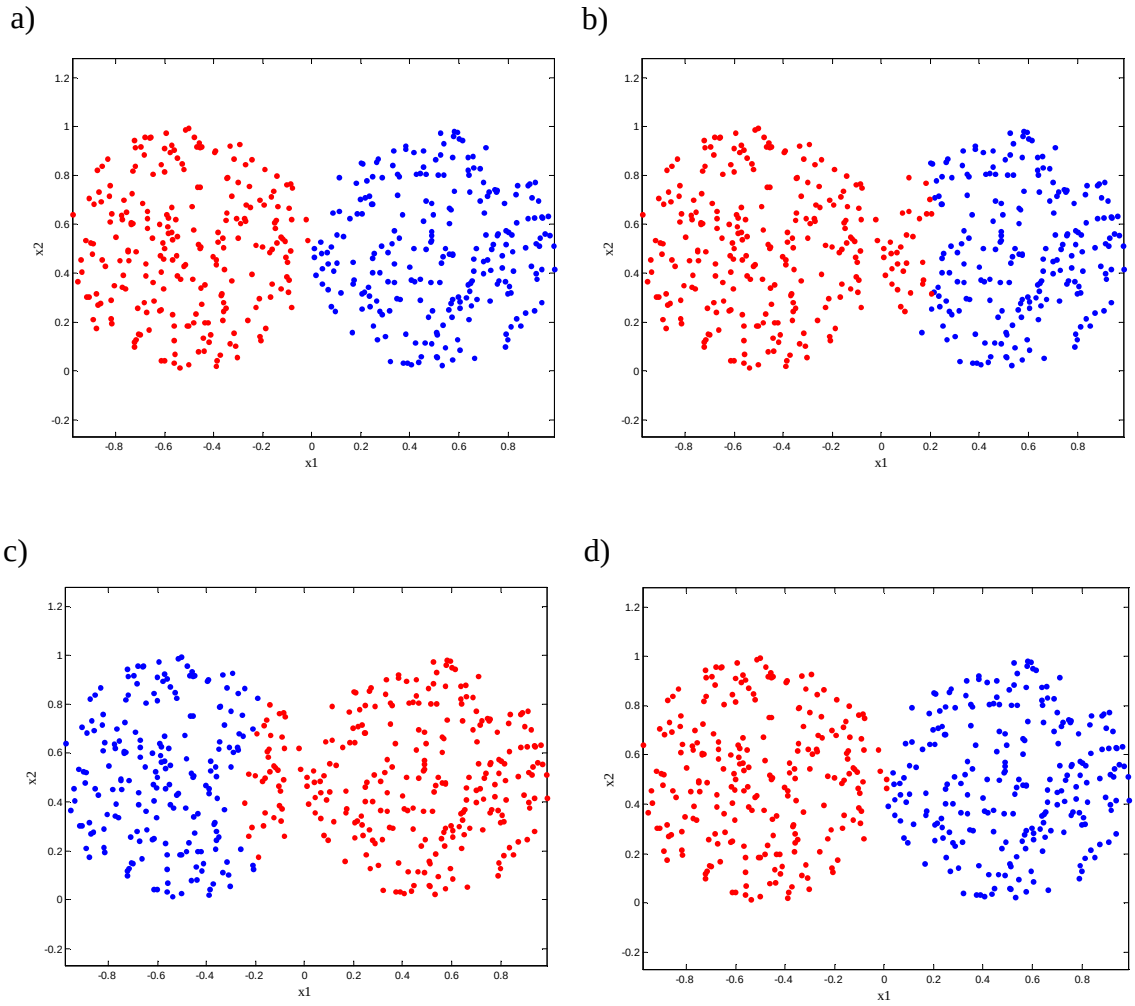
Jeśli powyższe warunki zostały spełnione ze względu na kryterium gęstości otoczenia obiektów to ciąg obiektów $\mathbf{x}_1,\ldots, \mathbf{x}_{i-1}$ nazywamy obiektami rdzeniowymi lub obiektami brzegowymi.

Modyfikacja algorytmu DBSCAN polega na zmianie w sposobie budowy łańcucha gęstości, który w konsekwencji przyjmuje następującą postać: [$\mathbf{x}_{\text{rdzeniowy}},\ldots, \mathbf{x}_{\text{rdzeniowy(n-1)}}, \mathbf{x}_{\text{rdzeniowy(n)}}$] dla obiektów rdzeniowych lub [$\mathbf{x}_{\text{rdzeniowy(1)}},\ldots, \mathbf{x}_{\text{rdzeniowy(n-1)}}, \mathbf{x}_{\text{brzegowy(n)}}$] obiektów rdzeniowymi z wyjątkiem ostatniego, będącego obiektem brzegowym. Wówczas, obiekt brzegowy nie bierze udziału w najważniejszym etapie działania algorytmu, odpowiedzialnym za rozbudowę łańcucha obiektów. Celem zmodyfikowanego algorytmu DBSCAN jest odłączenie ostatniego obiektu, którym jest obiekt brzegowy, od łańcucha obiektów spełniających kryterium gęstości obiektów. Modyfikacja definicji polega na stworzeniu łańcucha obiektów rdzeniowych, co można przedstawić następująco: $\mathbf{x}_1,\ldots, \mathbf{x}_n$, gdzie $\mathbf{x}_i$ jest obiektem rdzeniowym dla wszystkich $i \leq n$.

Praktycznie rzecz ujmując, w zmodyfikowanej wersji algorytmu DBSCAN, przypisanie obiektów do poszczególnych grup jest determinowane przez obiekty rdzeniowe z pominięciem obiektów brzegowych. Modyfikacji zasadniczo ulega sposób przeszukiwania przestrzeni pomiarowej oraz sposób tworzenia listy przynależności poszczególnych obiektów do grup. W pierwotnej wersji algorytmu DBSCAN, lista przynależności próbek zwiera wszystkie typy obiektów wyznaczone przez promień sąsiedztwa, r, natomiast w zmodyfikowanej wersji algorytmu przypisuje się do niej wyłącznie obiekty rdzeniowe. W zmodyfikowanej wersji DBSCAN krok drugi, następujący po określeniu parametrów wejścia, polega na znalezieniu wszystkich obiektów rdzeniowych odpowiednio dla każdej grupy. Dopiero wówczas, rozpoczyna się etap przypisania obiektów brzegowych do odpowiednich grup na podstawie wartości odległości euklidesowej. Sprowadza się to do obliczenia odległości pomiędzy obiektami brzegowymi, a obiektami rdzeniowymi. Ostatecznie, obiekt brzegowy zostaje przypisany do tej grupy, w której znajduje się najbliżej położony obiekt rdzeniowy. Właśnie takie rozwiązanie pozwala uniknąć problemu błędnego przypisania obiektów brzegowych występujących na granicy sąsiadujących ze sobą grup.

Porównanie efektywności algorytmów DBSCAN i zmodyfikowanego algorytmu DBSCAN zilustrowano wykorzystując symulowane dane zawierające dwie sąsiadujące ze sobą grupy w przestrzeni dwuwymiarowej (łącznie 471 próbek). Ich rozkład obrazuje Rys. 20. Na Rys. 20a przedstawiono dwie wysymulowane grupy obiektów oznaczone odpowiednio kolorem granatowym i czerwonym. Symulowanie grup przebiegało w taki sposób aby granica pomiędzy grupami była niewyraźna. Na Rys. 20b oraz 20c zaprezentowano wyniki grupowania obiektów za pomocą bazowego algorytmu DBSCAN. Porównując uzyskane grupy do rzeczywistego rozkładu (Rys. 20a) łatwo zauważyć, iż w obu przypadkach obiekty brzegowe zostały

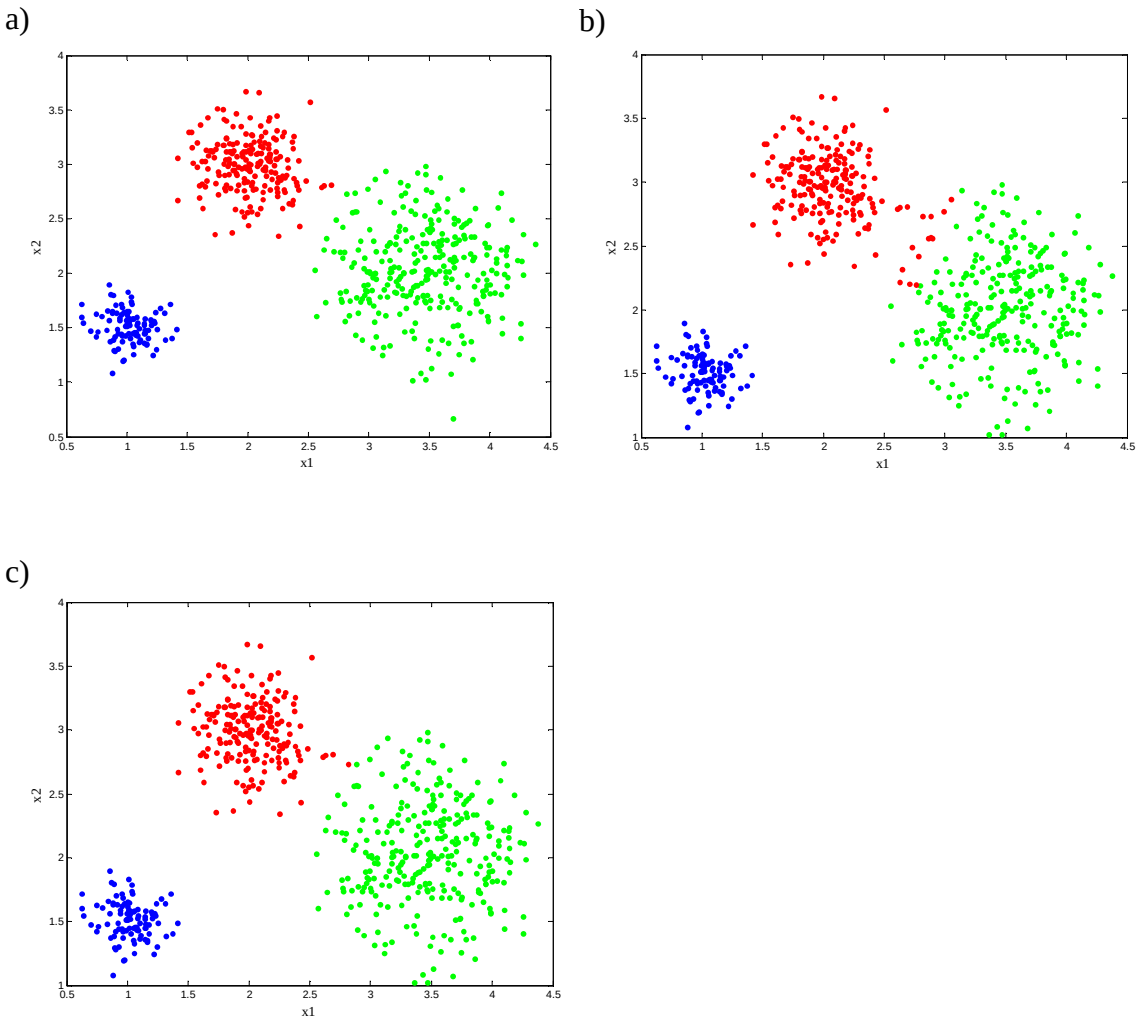
błędnie przyporządkowane. Dodatkowo, Rys. 20b i 20c ilustruje problem niestabilności rozwiązania – wpływ kolejności przetwarzania próbek przez oryginalny algorytm na końcowy wynik grupowania. Za każdym razem efekt końcowy grupowania znacząco różni się od siebie pomimo tych samych parametrów wejścia (MinPts = 255 oraz r = 0,6). Następnie, te same symulowane dane poddano grupowaniu wykorzystując zmodyfikowany algorytm DBSCAN. Wprowadzenie modyfikacji znacząco poprawiło jakość grupowania obiektów i umożliwiło odtworzenie wyjściowych grup. Wszystkie obiekty brzegowe zostały poprawnie przypisane do odpowiadających im grup, czego dowodzi Rys. 20d.



Rys. 20 Symulowany zestaw danych zwierający 471 obiektów tworzących dwie sąsiadujące ze sobą grupy w przestrzeni dwuwymiarowej zdefiniowanej przez parametry x1 i x2: a) oryginalnie wysymulowane grupy, b) oraz c) efekt grupowania bazową metodą DBSCAN oraz d) efekt grupowania zmodyfikowanym algorytmem DBSCAN. Ustalone parametry wejściowe to MinPts = 255 oraz r = 0,6.

Na Rys. 21 przedstawiono trzy grupy obiektów w przestrzeni zdefiniowanej przez parametry x1 oraz x2 zwierające łącznie 600 próbek. W tym przypadku wykazano działanie algorytmu, gdy w przestrzeni eksperymentalnej występują dwie grupy sąsiadujące ze sobą, oznaczone za pomocą kolorów czerwonego i zielonego oraz trzecia grupa oznaczona kolorem niebieskim, która jest dobrze odseparowana od pozostałych. Tak jak w poprzednim przykładzie zilustrowanym na Rys. 20b oraz 20c, podstawowy algorytm DBSCAN zawodzi, gdy grupy obiektów w przestrzeni eksperymentalnej sąsiadują ze sobą. Następuje wówczas błędne przypisanie obiektów brzegowych do odpowiednich grup. Sytuacja prezentuje się odmiennie, gdy te same dane zostaną poddane grupowaniu za pomocą zmodyfikowanej wersji algorytmu DBSCAN (Rys. 21c). Należy stanowczo podkreślić, iż dla danych zawierających grupy obiektów dobrze od siebie odseparowanych w przestrzeni pomiarowej, zmodyfikowana wersja DBSCAN prowadzi do tych samych wyników jak te uzyskane za pomocą bazowego algorytmu (zobacz Rys. 21b i c dla grup odseparowanych).

Wprowadzone do algorytmu zmiany, zwiększyły zakres zastosowalności metody w kontekście eksploracji danych zawierających sąsiadujące ze sobą grupy obiektów, poprzez pokonanie trudności z przypisaniem obiektów brzegowych do rozważanych grup obiektów dając zarazem stabilne rozwiązanie grupowania i niezależne od kolejności przetwarzania danych. Jako ograniczenie metody można potraktować wpływ zastosowanej miary odległości na kształt wyodrębnianych grup obiektów. W omówionym przypadku wyodrębnione grupy maja kształt sferyczny, co jest zgodne z ideą zastosowania odległości euklidesowej.

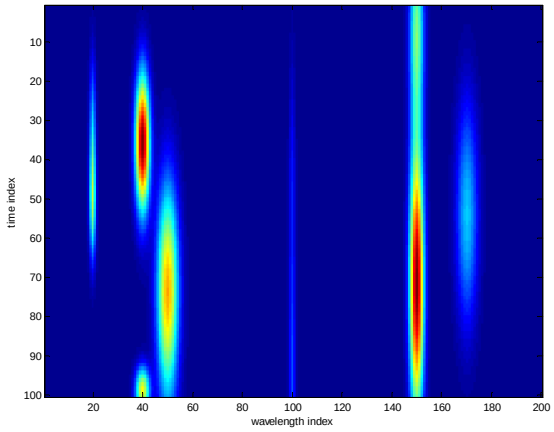


Rys. 21 Symulowany zestaw danych zwierający łącznie 600 obiektów tworzących trzy grupy obiektów zaprezentowanych w przestrzeni dwóch parametrów x1 i x2;
a) oryginalnie wysymulowane grupy obiektów, b) efekt grupowania uzyskany bazową metodą DBSCAN oraz c) efekt grupowania zmodyfikowanym algorytmem DBSCAN. Ustalone parametry wejścia to MinPts = 90 oraz r = 0,6.

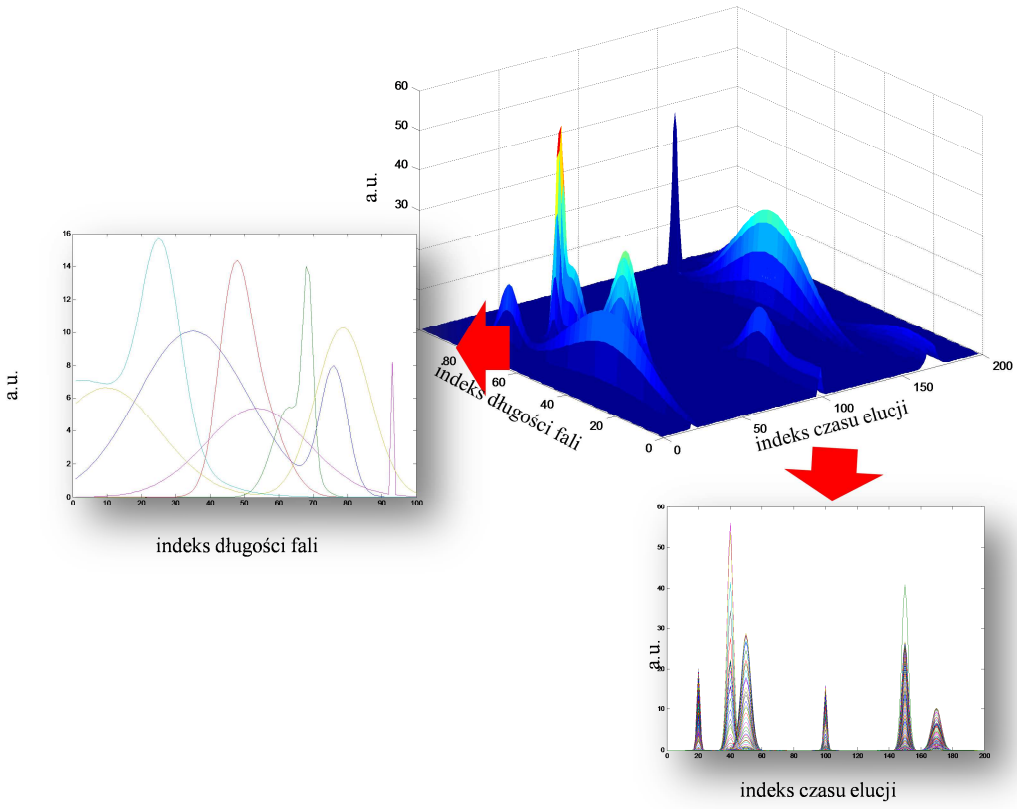
12.2 Nowa metodologia porównywania dwuwymiarowych chromatograficznych odcisków palca

Wykorzystanie zawansowanej aparatury badawczej pozwala na analizę próbek różnorodnego pochodzenia, pozwalając tym samym na ich analizę jakościową, ilościową i ocenę właściwości fizykochemicznych. Próbki charakteryzowane są za pomocą sygnałów instrumentalnych, których złożoność zależy od zastosowanej techniki badawczej – chromatografia, metody spektroskopowe. Cechą wspólną próbek, takich jak: próbki środowiskowe, próbki leków, kosmetyków, artykułów spożywczych,

a przede wszystkim próbek biologicznych jest bardzo złożony skład chemiczny, co przekłada się m.in. na skomplikowaną matrycę próbki. Dlatego w celu ich kompleksowej analizy i charakterystyki, coraz częściej korzysta się z metod sprzężonych będących kombinacją przynajmniej dwóch metod instrumentalnych, najczęściej chromatografii wspieranej metodami spektroskopowymi, o czym pisano w podrozdziale *3.3* niniejszej pracy. Otrzymywana wówczas informacja o analizowanej próbce pochodzi z co najmniej dwóch niezależnych (ortogonalnych) źródeł, co pozwala na jej kompleksową charakterystykę. Stosując metody sprzężone uzyskuje się tzw. wielowymiarowe odciski palca (z ang. Multi-Dimensional Fingerprint) [119]. W przypadku metod typu HPLC-DAD, czy też LC-MS, itd., każda próbka zostaje opisana za pomocą tzw. dwuwymiarowego odcisku palca. Zastosowanie detektora typu DAD lub MS pozwala na uzyskanie dodatkowej informacji w postaci widm UV-VIS lub widm masowych dla wszystkich porcji eluatów próbki. Przykładowy odcisk palca otrzymany poprzez wysymulowanie danych HPLC-DAD zaprezentowano na Rys. 22 oraz 23.



Rys. 22 Dwuwymiarowy chromatograficzny odcisk palca uzyskany za pomocą symulowanych danych HPLC-DAD.

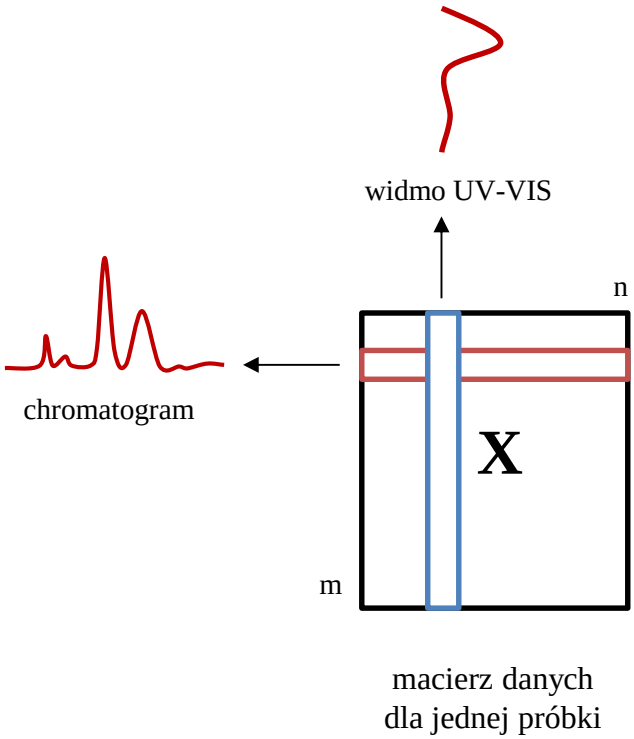


Rys. 23 Oryginalny trójwymiarowy obraz reprezentujący dane chromatograficzne otrzymywane za pomocą metody HPLC-DAD z uwzględnieniem widm UV-VIS oraz chromatogramów czystych składników mieszaniny.

Uzyskane odciski palca można również przedstawić w postaci macierzy danych (Rys. 24). Każda macierz charakteryzuje jedną próbkę. W wierszach macierzy znajdują się chromatogramy uzyskane dla danej długości fali, natomiast w kolumnach widma spektroskopowe zarejestrowane w każdym punkcie czasu elucji. Ponieważ w procesie badawczym analizuje się więcej niż jedną próbkę to w efekcie uzyskuje się tensor złożony z tablic danych opisujących poszczególne próbki. Wymiarowość uzyskanych danych stanowi pierwszy problem ich analizy.

Kolejny problem z jakim można się spotkać w przypadku dwuwymiarowych chromatograficznych odcisków palca jest to problem współwymywania (koelucji) substancji. Niejednokrotnie analizując uzyskane chromatogramy można zaobserwować piki których kształt znacząco odbiega od typowego kształtu krzywej Gaussa. W takim przypadku konieczne jest potwierdzenie, czy dany pik lub piki charakteryzują jedną, czy więcej substancji. Następnym problemem z jakim można się spotkać podczas analizy tego typu zestawów danych jest tzw. problem przesunięć pików. Dotyczy on

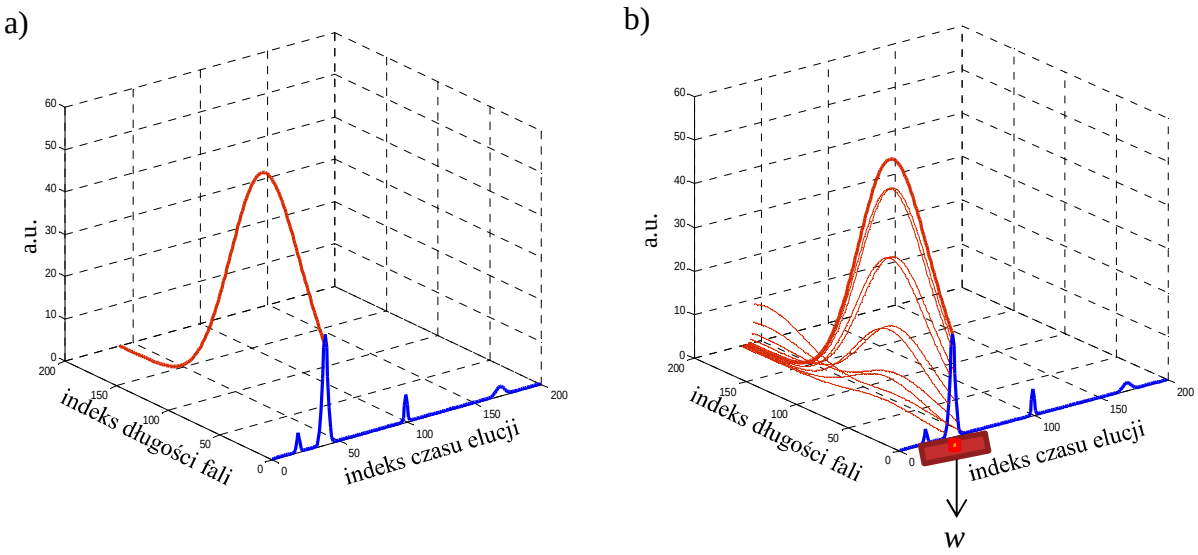
nie tylko danych uzyskanych na drodze analizy chromatograficznej, ale pojawia się również w przypadku innych metod instrumentalnych. Jak wspominano w podrozdziale *3.4*, przesunięcia pików są efektem nawet nieznacznych wahań warunków pomiarowych, wpływających bezpośrednio na rejestrację sygnałów instrumentalnych i ich jakość.



Rys. 24 Macierz danych uzyskana dla jednej próbki za pomocą sprzężonej metody typu HPLC-DAD, zawierająca w wierszach chromatogramy (rejestrowane przy danej długości fali), a w kolumnach widma UV-VIS dla kolejnych porcji eluatu.

Zazwyczaj, proponowanym rozwiązaniem problemu przesunięć sygnałów jest ich nałożenie za pomocą metod nakładania sygnałów instrumentalnych, stanowiących pakiet metod wstępnego przygotowania danych do dalszej analizy. Jedną z najczęściej stosowanych jest metoda COW, ułatwiająca efektywne nakładanie sygnałów na siebie. Pomimo jej powszechności, ma ona także pewne ograniczenia. Zalicza się do nich m.in. wybór wzorca, względem którego dokonuje się nakładanie poszczególnych sygnałów na siebie oraz czas obliczeń. Wybór wzorca jest bardzo ważnym elementem działania algorytmu, gdyż wpływa on na powodzenie późniejszej analizy oraz interpretację danych. W niniejszej pracy w celu rozwiązania problemów koelucji substancji oraz przesunięć pików wykorzystano koncepcję podobieństwa,

wprowadzając nową miarę podobieństwa, s_{ij} (24). Miara ta została opracowana w oparciu o współczynnik korelacji Pearsona, dzięki czemu uwzględnia korelację zarejestrowanych widm obserwowanych przy określonych czasach elucji (tj. porcji eluatu). Wyniki reprezentowane są jako mapa odpowiedzi, przedstawiająca na diagonali podobieństwo analizowanych próbek. Procedura analizy dwuwymiarowych chromatograficznych odcisków palca polega na obliczeniu współczynnika korelacji Pearsona pomiędzy widmami jednej próbki, a widmami w pewnym zakresie punktów pomiarowych osi czasu elucji *w* dla drugiej próbki, co zaprezentowano na Rys. 25.



Rys. 25 Sposób obliczania podobieństwa pomiędzy widmami dwóch próbek, gdzie: a) chromatogram reprezentujący próbkę pierwszą zrejestrowany przy indeksie długości fali równym 3 oraz odpowiadające mu widmo UV-VIS dla indeksu czasu elucji wynoszącego 40, b) chromatogram dla próbki drugiej uzyskany przy indeksie długości fali równym 3 oraz widma UV-VIS dla indeksu czasu elucji wynoszącego 40 oraz przedziału $w = 10$.

Następnie, wzajemne obszary odpowiedniości wizualizuje się za pomocą mapy odpowiedzi, która umożliwia relatywnie łatwą ocenę podobieństwa próbek, identyfikację odpowiadających sobie pików w przypadku sygnałów z ich przesunięciami, czy detekcję substancji które w wyniku procesu chromatograficznego uległy współwymyciu. Nowa miara podobieństwa pozwala na obliczenie stopnia podobieństwa dwóch rozważanych odcisków palca, przyjmując wartość z przedziału [0,1]. Wartość 1 reprezentuje idealne podobieństwo, a 0 oznacza brak podobieństwa analizowanych próbek. Mowa tu o wariancie nowej miary podobieństwa, w którym

uwzględnia się wartość bezwzględną ze współczynnika korelacji. W innym przypadku należałoby, wyrażoną równaniem (24) miarę podobieństwa odjąć od 1, aby został spełniony warunek pozytywności.
Wprowadzona miara podobieństwa spełnia wszystkie warunki od 1''' do 3''' wprowadzone w rozdziale 6 pracy w odniesieniu do miar podobieństwa.

$$s_{ij} = \left[\max\left(\left(\frac{\sum \max(\mathbf{C})}{n}\right), \left(\frac{(\sum \max(\mathbf{C}^{\mathbf{T}}))}{n}\right)\right)\right] \tag{24}$$

gdzie:
n – liczba punktów pomiarowych, odpowiadająca poszczególnym indeksom czasu retencji zarejestrowanych dla obu próbek
C – macierz współczynników korelacji Pearsona
c_{ij} – współczynnik korelacji pomiędzy i-tym widmem próbki 1 oraz j-tym widmem próbki 2

12.2.1 Problem koelucji substancji występujący w dwuwymiarowych chromatograficznych odciskach palca

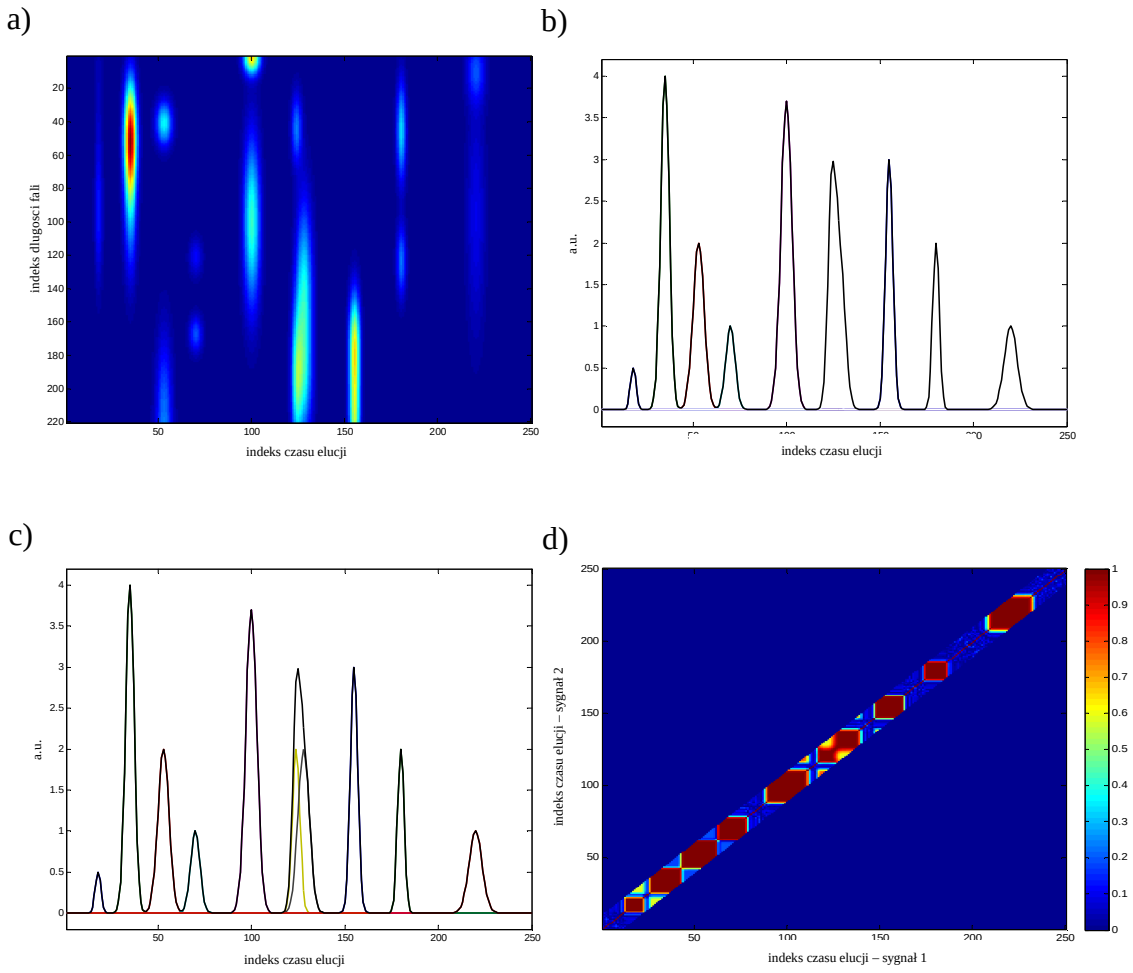
Mimo licznych zalet jakie wykazują metody instrumentalne, w tym metody sprzężone, posiadają one również swoje ograniczenia. Jednym z nich jest możliwy brak pełnego rozdziału substancji. Spowodowany jest przede wszystkim złożonością składu analizowanych próbek i/lub zbyt szybkim tempem prowadzonego rozdziału chromatograficznego. Dlatego mimo doboru warunków rozdziału, prawdopodobieństwo nakładania się pików jest duże. Problem współwymywania substancji skutkujący nakładaniem się pików chromatograficznych może prowadzić do niewłaściwego zidentyfikowania substancji, co skutkuje błędnymi wnioskami. Wymywanie z kolumny chromatograficznej dwóch lub więcej substancji w tym samym czasie elucji (lub w niewielkich odstępach czasu) skutkuje pojawieniem się na chromatogramie pików, które nakładają się na siebie w wyniku czego może pojawić się pik o nieregularnym kształcie odbiegającym od kształtu krzywej Gaussa. Dlatego, niezbędne są narzędzia pozwalające na identyfikację liczby substancji jakie ten pik reprezentuje. Może się również zdarzyć, że niespecyficzny kształt piku jest skutkiem sposobu rejestracji chromatogramu, a nie złym rozdziałem substancji na kolumnie chromatograficznej.
W niniejszej pracy zaproponowano nowatorskie rozwiązanie badania podobieństw pomiędzy próbkami, w których występuje problem koelucji substancji, poprzez zastosowanie procedury opartej na wprowadzonej mierze podobieństwa s_{ij}. W tym celu wykorzystano symulowane dwuwymiarowe chromatograficzne odciski palca, które

wysymulowano w taki sposób, aby odzwierciedlały charakterystykę sygnałów na ogół otrzymywanych metodą HPLC-DAD. Każdą próbkę opisano jako macierz zawierającą w wierszach chromatogramy, a w kolumnach widma. W pierwszym etapie analizy obliczono wartość bezwzględną ze współczynników korelacji Pearsona pomiędzy parami widm próbek, tworząc macierz korelacji zawierającą niezbędną dla dalszych rozważań informację o podobieństwie poszczególnych widm. Jak wcześniej opisano współczynniki korelacji obliczano pomiędzy każdym widmem próbki pierwszej oraz odpowiadającymi mu widmami w pewnym zakresie *w* próbki drugiej. Następnie, opierając się na otrzymanych współczynnikach korelacji wyznaczono wartość s_{ij}, definiującą podobieństwo próbek. Wyniki zwizualizowano za pomocą mapy odpowiedzi, gdzie podobieństwo pomiędzy próbkami przedstawione jest na diagonali. Taki sposób reprezentacji wyników znacząco ułatwia etap interpretacji ponieważ, wizualna ocena wyników reprezentowanych przez mapę odpowiedzi ukazuje liczbę substancji reprezentowanych przez poszczególne piki chromatograficzne. Jeżeli pik obserwowany na chromatogramie przedstawia nałożone na siebie piki różnych substancji, to w przypadku mapy odpowiedzi piki te są dobrze rozdzielone.

Przykład 1

Poniżej omówiono wyniki analizy symulowanych dwuwymiarowych chromatograficznych odcisków palca HPLC-DAD (Rys. 26a). Wymiarowość otrzymanej macierzy **X** wynosiła 250×220, a zatem uzyskane dane zawierają wartości absorbcji zarejestrowane przy 220 długościach fal. Indeks czasu elucji odpowiada kolejnym 250 punktom pomiarowym, tj. 250 porcjom eluatu dla których rejestrowano widma UV-VIS. Podczas symulacji uwzględniono rozważany problem koelucji substancji. Dlatego, jeden z pików widocznych na sumarycznym chromatogramie (pik nr 6), został utworzony poprzez nałożenie na siebie dwóch pików różnych substancji, co pokazano na Rys. 26c. Następnie, obliczono współczynniki korelacji dla tej samej próbki i zaprezentowano wyniki na mapie odpowiedzi (Rys. 26d). Mapa odpowiedzi konstruowana jest w taki sposób, aby każdy jej piksel reprezentował wartość współczynnika korelacji z macierzy korelacji. Stąd wartości R = 1, oznaczone są odpowiednio kolorem czerwonym, a wartości R = 0, kolorem niebieskim. Wartości pośrednie przyjmują kolor będący kolorem pośrednim dwóch granicznych barw, co przedstawia dołączona do mapy legenda. Relatywnie łatwo dostrzec, że na diagonali mapy odpowiedzi (Rys. 26d), w miejscu odpowiadającym pikom, pojawiają się charakterystyczne obszary złożone wyłącznie z ciemnoczerwonych pikseli, co związane jest z relatywnie dużą wartością współczynnika korelacji. Ponieważ w tym przypadku rozważa się podobieństwo tej samej próbki względem siebie, to reprezentują one dokładnie wartość równą 1. W pozycji od 110 do 140 indeksu czasu elucji, zarówno dla sygnału pierwszego jak i drugiego, pojawią się charakterystyczne dwa czerwone obszary w kształcie prostokątów odpowiadające dwóm substancjom. W ten dość łatwy sposób udało się zidentyfikować liczbę substancji reprezentowanych przez

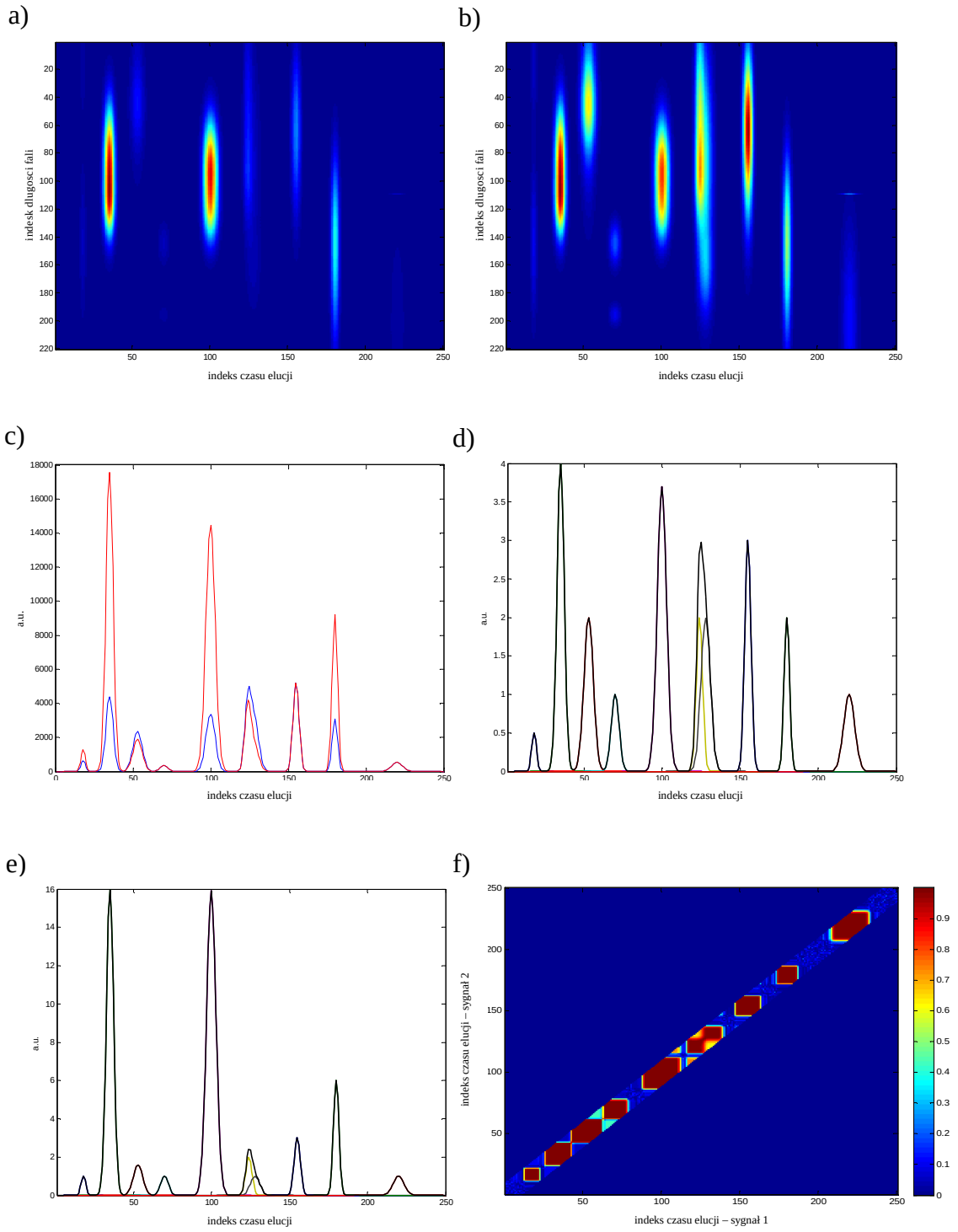
poszczególne piki na chromatogramie. Należy również podkreślić, że w tym przypadku dwuwymiarowy chromatograficzny odcisk palca nie ujawnia informacji o ewentualnej koelucji substancji. Opierając się na informacji o składzie jakościowym próbek reprezentowanym przez otrzymane odciski palców lub chromatogramy, można by stwierdzić że liczba substancji uzyskana w rozdziale chromatograficznym wynosi 9, mimo iż w rzeczywistości próbka zawiera 10 różnych substancji.
Wartość miary podobieństwa s_{ij}, w tym przypadku, wynosi 1.



Rys. 26 Symulowane sygnały HPLC-DAD reprezentujące jedną próbkę
a) dwuwymiarowy chromatograficzny odcisk palca, b) sumaryczny chromatogram (sumaryczna intensywność widma dla danej porcji eluatu), c) sumaryczny chromatogram z zaznaczonymi pikami, które uległy koelucji, d) mapa odpowiedzi, uzyskana na podstawie współczynników korelacji dla w = 10.

Przykład 2

W kolejnym przykładzie rozważano problem koelucji w kontekście analizy dwóch próbek charakteryzujących się takim samym składem chemicznym, jednak różniących się profilami stężeniowym (Rys. 27c). Uwzględniano tutaj wariant, w którym założono, że analiza próbek przebiegała w tych samych warunkach rozdziału. Również w tym przypadku wysymulowano dane zbliżone do sygnałów typu HPLC-DAD o wymiarowości 220×250 dla obu macierzy reprezentujących próbki. Sygnały charakteryzujące próbki wysymulowano w taki sposób, aby miały ten sam skład chemiczny (10 substancji) jednak różniły się stężeniami poszczególnych komponentów, a jeden z pików powstał w wyniku nałożenia pików dwóch substancji (pik 6).
Odciski palców reprezentujące skład chemiczny próbek zaprezentowano na Rys. 27a oraz 27b. Jak łatwo zauważyć różnią się one jedynie poziomem zarejestrowanej intensywności sygnałów dla poszczególnych składników występujących w próbkach, jednak nie wykazują problemu koelucji substancji. Tak jak miało to miejsce w poprzednim przykładzie, wykorzystanie proponowanego rozwiązania, umożliwiło detekcję piku powstałego w wyniku współwymywania substancji. Na mapie odpowiedzi wyraźnie widać, że obszar odpowiadający substancji szóstej powstał na skutek równoczesnego wymycia dwóch komponentów w zbliżonym czasie elucji. W celu wyraźnego zilustrowania badanego problemu oraz skuteczności wykorzystanej metodologii zastosowano szerokość okna $w = 10$. Tak jak w poprzednim przypadku (Przykład 1) obliczono współczynniki korelacji pomiędzy próbkami tworząc macierz korelacji, którą następnie wykorzystano do utworzenia mapy odpowiedzi. Wyniki analizy zaprezentowano na Rys. 27f. Podobieństwo próbek 1 i 2, określone na postawie miary s_{ij}, wynosi 0,7106.
Porównanie map uzyskanych w Przykładzie 1 oraz 2 ujawnia, że na mapach z Przykładu 2 nie występują piksele odpowiadające współczynnikowi korelacji o wartości 1, oznaczające idealne podobieństwo. Jest to skutkiem obliczeń współczynnika korelacji pomiędzy dwoma różnymi próbkami, więc wartość ta musi odbiegać od jedynki ($R \neq 1$).
Proponowane rozwiązanie daje zadowalające wyniki pomimo różnicy stężeń poszczególnych komponentów próbek. Umożliwia to rozwiązanie problemu koelucji w relatywnie łatwy i szybki sposób. Dodatkowo, zmiana wartości s_{ij} sugeruje, że próbki są do siebie znacząco podobne jednak nie identyczne. Oczywiście w przypadku kiedy próbki charakteryzowałby się takim samym składem oraz stężeniem poszczególnych komponentów, a jedyna różnica pomiędzy próbkami wynikałaby z różnicy związanej z szumem eksperymentalnym, wartość s_{ij} wynosiłaby $\cong 1$.



Rys. 27 Symulowane dane HPLC-DAD dla dwóch próbek a) dwuwymiarowy chromatograficzny odcisk palca próbki 1, b) dwuwymiarowy chromatograficzny odcisk palca próbki 2, c) zestawienie sumarycznych chromatogramów dla obu próbek (linia niebieska próbka 1 oraz linia czerwona próbka 2), d) sumaryczny chromatogram próbki 1 z zaznaczonym problemem koelucji, e) sumaryczny chromatogram próbki 2 z zaznaczonym problemem koelucji, f) mapa odpowiedzi uzyskana na podstawie współczynników korelacji obu próbek dla $w = 10$.

12.2.2 Ocena podobieństw bez wstępnego nakładania sygnałów

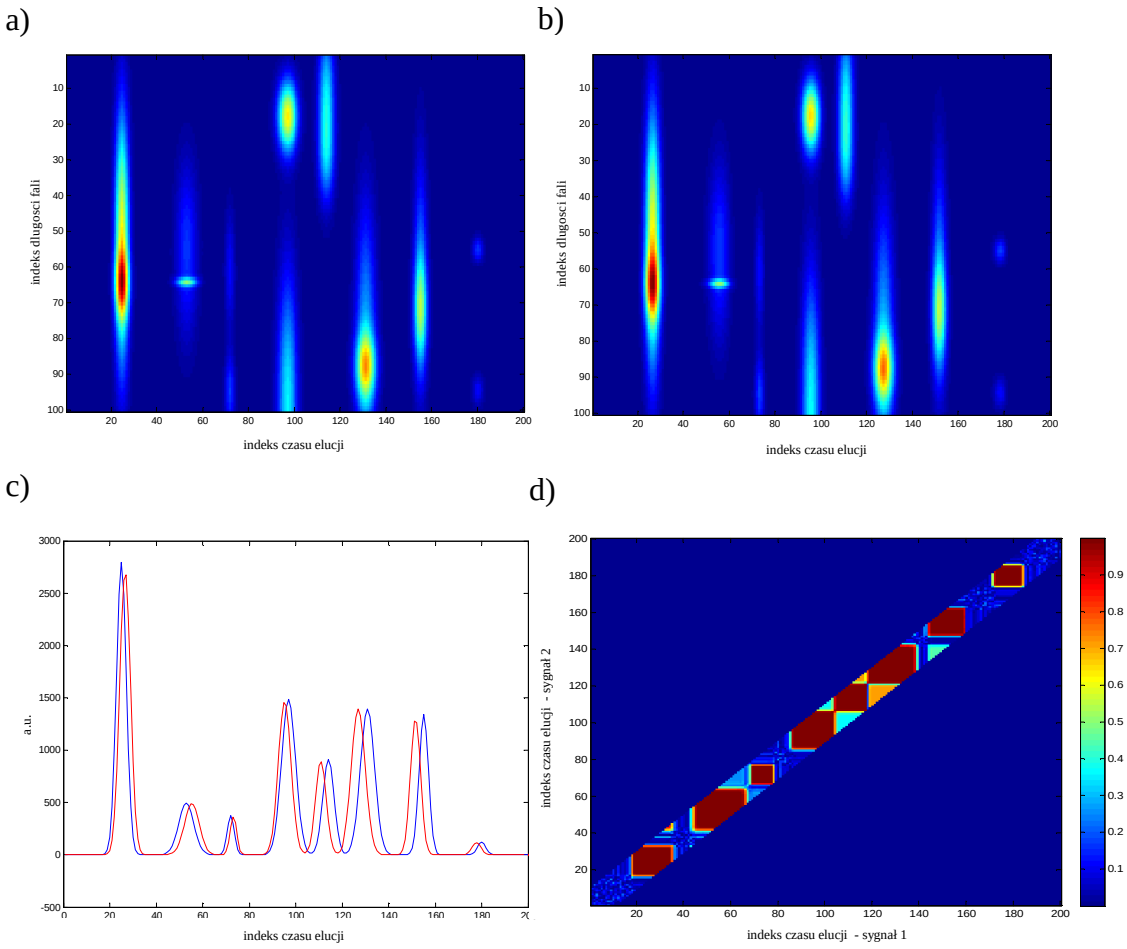
Kolejnym problemem jaki występuje w przypadku sygnałów rejestrowanych za pomocą metod instrumentalnych są tzw. przesunięcia sygnałów. O przesunięciach mówi się gdy te same zmienne występujące w różnych widmach lub chromatogramach, znajdują się w innej pozycji na osi opisującej odpowiednio częstotliwość lub czas elucji. Przesunięcia te są następstwem nieznacznych wahań warunków pomiarowych takich jak temperatura, pH, czy ciśnienie. W przypadku chromatogramów przesunięcia pików są skutkiem fluktuacji temperatury, zmian w składzie fazy ruchomej lub wynikają ze starzenia się kolumny chromatograficznej. Problem nakładania przesuniętych względem siebie sygnałów instrumentalnych jest zagadnieniem związanym ze wstępnym przygotowaniem danych do dalszej analizy. Skuteczność nałożenia widm na siebie jest ważnym elementem wpływającym na powodzenie późniejszej eksploracji i analizy danych. Pominięcie tego etapu uniemożliwia dalszą analizę. W literaturze można znaleźć pakiet metod umożliwiających nałożenie sygnałów na siebie takich jak COW, czy parametryczne nakładanie sygnałów instrumentlanych, o czym wspominano uprzednio w rozdziale 5. Jednak w niniejszej pracy w celu identyfikacji odpowiadających sobie pików, wykorzystano wprowadzoną miarę podobieństwa s_{ij}. Takie podejście do problemu umożliwia identyfikację odpowiadających sobie w kolejnych sygnałach instrumentalnych zmiennych, bez potrzeby nakładania ich na siebie. Pozawala to na skrócenie czasu analizy danych, pominięcie etapu poszukiwania wzorca, względem którego następuje nałożenie sygnałów oraz eliminację ryzyka jego błędnego wyboru.

Przykład 3

Niniejszy przykład obrazuje użyteczność wprowadzonej metodologii w kontekście analizy danych, w których występują przesunięcia pików wzdłuż osi reprezentującej indeks czasu elucji. Wyniki zaprezentowano na symulowanych sygnałach HPLC-DAD, w których wygenerowano przesunięcia chromatogramów. Wymiarowość otrzymanych danych wynosiła dla każdej próbki 200×100. Na Rys. 28a oraz b przedstawiono odpowiadające wysymulowanym próbkom odciski palca. Łatwo zauważyć, że reprezentują próbki o tym samym składzie, jednak obszary odpowiadające intensywności substancji na odcisku palca na Rys. 28b są przesunięte względem odcisku palca na Rys. 28a, o kilka jednostek indeksu czasu elucji. Na Rys. 28c przedstawiono sumaryczne chromatogramy obu próbek, na których zobrazowano przesunięcia pików.

Wyniki identyfikacji pików za pomocą wprowadzonej metodologii zaprezentowano jako mapę odpowiedzi na Rys. 28d. Na diagonali mapy pojawiły się obszary składające się z pikseli o barwie czerwonej (R = 1) odpowiadające porcjom eluatu o tym samym składzie. Wartość nowej miary podobieństwa dla rozważanego przykładu wynosiła

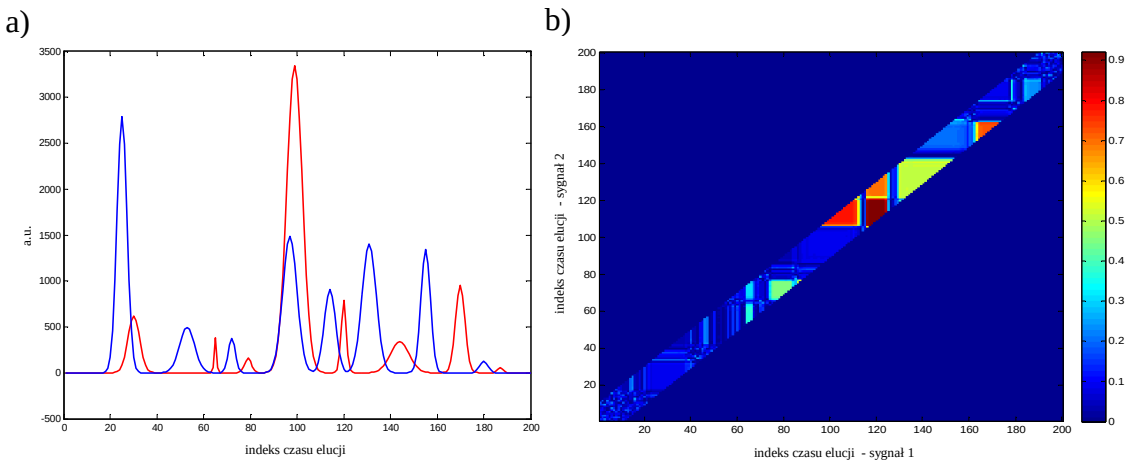
s_{ij} = 0,7167, co wskazuje na relatywnie duże podobieństwo próbek. Zastosowano okno obejmujące łącznie 10 widm (w = 10).
Dzięki zastosowaniu wprowadzonej metodologii identyfikacja odpowiadających sobie pików substancji w chromatogramach z wyraźnym problemem ich przesunięć, stała się relatywnie prosta.



Rys. 28 Symulowane dane HPLC-DAD, uwzględniające problem przesunięć pików w czasie a) odcisk palca uzyskany dla próbki 1, b) odcisk palca uzyskany dla próbki 2, c) zestawienie sumarycznych chromatogramów dla próbek 1 i 2, oznaczonych odpowiednio niebieską i czerwoną linią, d) mapa odpowiedzi utworzona na podstawie współczynników korelacji dla w = 10.

Aby zweryfikować skuteczność zaproponowanej metody zastosowano ją również w celu analizy próbek różniących się składem chemicznym i porównano otrzymane wyniki z tymi otrzymanymi dla poprzednio rozważanego przykładu związanego

z problemem przesunięć pików w czasie. Porównano nowo wysymulowaną próbkę 3 z próbką 1, a wyniki zilustrowano na Rys. 29. Porównanie sumarycznych chromatogramów, reprezentujących odpowiednio próbki 1 oraz 3, ujawnia że próbki te różnią się składem jakościowym. O odmienności próbek świadczy otrzymana wartość s_{ij} = 0,3606. Wyniki wzbogacono, jak zawsze, mapą odpowiedzi utworzoną na podstawie współczynników korelacji, które obliczono pomiędzy widmami analizowanych próbek dla *w* = 10. Wizualna ocena otrzymanej mapy potwierdza, że próbki te znacząco różnią się od siebie. W przeciwieństwie do poprzedniego przykładu na diagonali mapy odpowiedzi nie pojawiają się charakterystyczne obszary o kształcie prostokątów złożonych z czerwonych pikseli, oznaczające R = 1 i odpowiadające pikom tych samych substancji. Zamiast tego diagonala zawiera piksele o barwie niebieskiej, oznaczającej względnie małą korelację pomiędzy widmami porównywanych próbek.



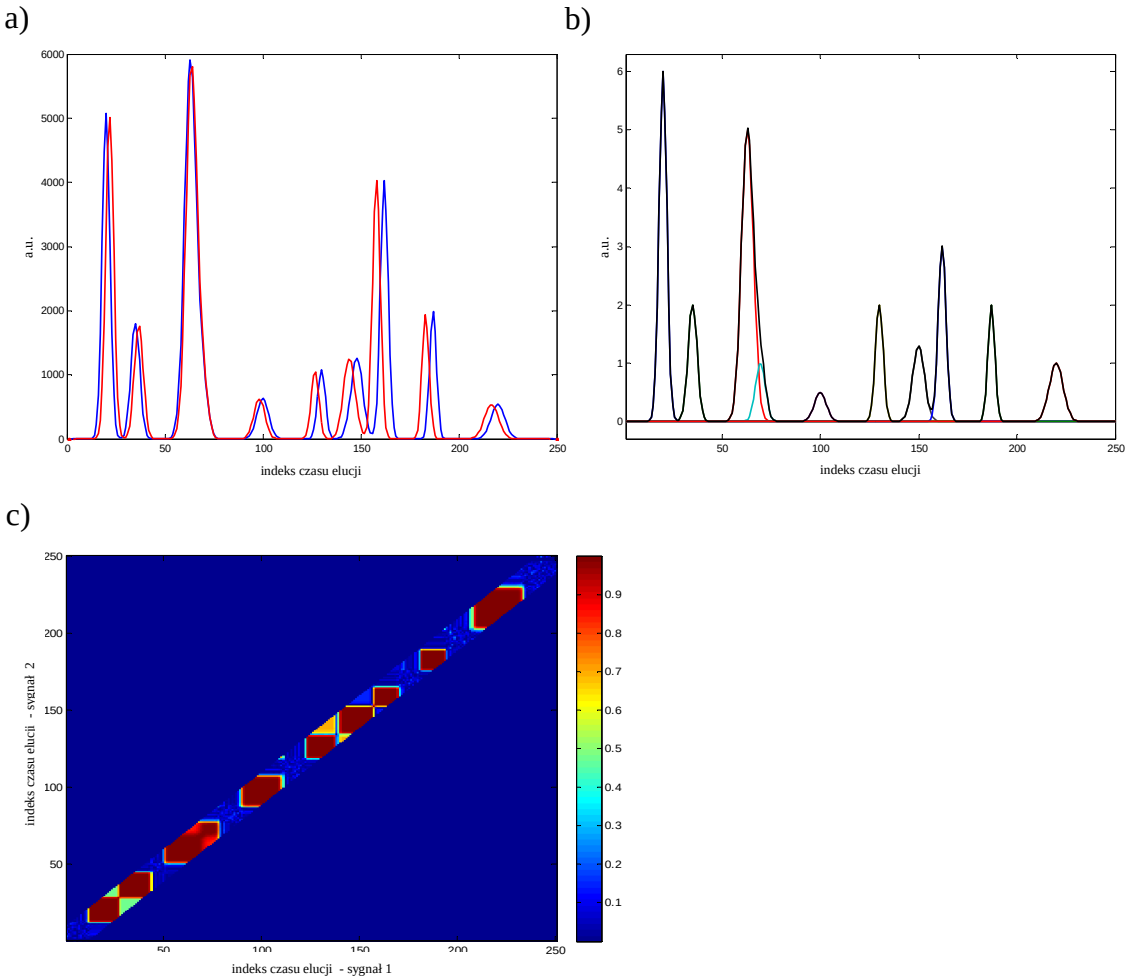
Rys. 29 Symulowane dane HPLC-DAD dla próbek różniących się składem a) zestawienie sumarycznych chromatogramów oznaczonych linią niebieską dla próbki 1 oraz linią czerwoną dla próbki 3, b) mapa odpowiedzi utworzona dla próbek 1 i 3, na podstawie ich współczynników korelacji, dla *w* = 10.

12.2.3 Ocena podobieństw sygnałów przy równoczesnej koelucji substancji i przesunięciach pików

Podczas analizy sygnałów instrumentalnych można się spotkać z równoczesnym wystąpieniem problemu koelucji substancji i przesunięć pików w czasie. Zastosowanie zaproponowanej metodologii analizy porównawczej opartej na nowej mierze odległości, umożliwia równoczesną identyfikację odpowiadających sobie pików oraz liczby reprezentowanych przez nie substancji. Przypadek ten omówiono w Przykładzie 4.

Przykład 4

W celu zwizualizowania efektywności zaproponowanej procedury analizy danych użyto symulowanych sygnałów HPLC-DAD. Wymiarowość macierzy danych opisującej każdą próbkę wynosiła 180×250. Zestawienie sumarycznych chromatogramów, reprezentujących obie porównywane próbki przedstawiono na Rys. 30a, który uwidacznia również problem przesunięć pików chromatograficznych w czasie. Problem koelucji substancji dla próbki 1 zilustrowano na Rys. 30b.
Wyniki analizy danych przedstawiono jako mapę odpowiedzi, która zawiera obszary wysokiej korelacji, odpowiadające pikom tej substancji w porównywanych próbkach. Dodatkowo obszar piku nr 3 jest „podwójny”, co wskazuje na koelucję substancji przy tym indeksie czasu elucji. Uzyskana wartość s_{ij} w tym przypadku wyniosła 0,7103.



Rys. 30 Symulowane dane HPLC-DAD, w których uwzględniono problem koelucji oraz przesunięć pików chromatograficznych: a) sumaryczny chromatogram oznaczony linią niebieską dla próbki 1 i linią czerwoną dla próbki 2, b) sumaryczny chromatogram dla próbki 1 obrazujący koelucję substancji, c) mapa odpowiedzi utworzona dla próbek 1 i 2, na podstawie współczynników korelacji odpowiadających sobie widm w oknie o szerokości $w = 10$.

12.2.4 Analiza tensora danych oparta na nowej mierze odległości

Wprowadzoną metodologię opartą na nowej mierze odległości s_{ij}, można z powodzeniem zastosować w celu analizy danych w formie tensora $\underline{\mathbf{X}}(p \times n \times m)$. Dla danych typu HPLC-DAD, p odpowiada długościom fali przy których zarejestrowano poszczególne chromatogramy, n to liczba punktów odpowiadająca indeksowi czasu elucji, a m odpowiada liczbie analizowanych próbek. W związku z czym, każda poszczególna tablica w tensorze jest macierzą chrakteryzującą określoną

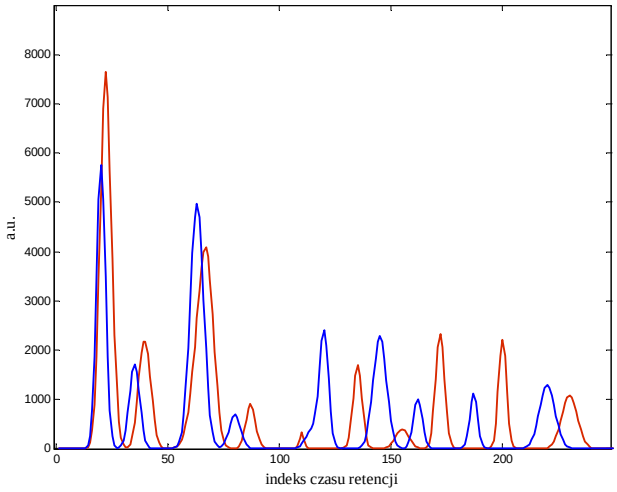
próbkę. Analiza tego typu danych nie należy do najprostszych przede wszystkim ze względu na skomplikowaną formę ich prezentacji. Wymiarowość tego typu zestawów danych znacząco wydłuża czas prowadzonych analiz, utrudnia interpretację, wyciąganie wniosków oraz detekcję zależności pomiędzy poszczególnymi próbkami i/lub parametrami. Zastosowanie zaproponowanej metodologii badania podobieństw pomiędzy próbkami opiera się na obliczeniu współczynników korelacji pomiędzy wszystkimi widmami jednej z próbek i odpowiadającymi im widmami w oknie *w*. Obliczenia powtarza się dla m próbek w tensorze, tak aby porównać ze sobą wszystkie próbki, a ich podobieństwo wyraża się za pomocą miary podobieństwa s_{ij}. Wartości s_{ij}, otrzymane dla par próbek, reprezentuje się w postaci macierzy podobieństwa, **K**, o wymiarowości m×m. Otrzymana w ten sposób macierz zawiera informację o wzajemnym podobieństwie próbek, czyli informację o strukturze danych, co stanowi punkt odniesienia dla dalszej analizy lub eksploracji. Otrzymaną macierz można poddać działaniu metod eksploracyjnych, takich jak metody hierarchiczne, uzyskując w ten sposób informację o istnieniu grup obiektów podobnych lub metody PCA, pozwalającej na wizualizację wyników.

Aby zaprezentować efektywność proponowanej metody wysymulowano dane HPLC-DAD zawierające 26 próbek o wymiarowości 180×250. Ostatecznie otrzymano tensor danych. Dane symulowano w taki sposób aby odzwierciedlały realne problemy występujące np. podczas kontroli jakości produktów. W zawiązku z czym rozważano kilka problemów. W trakcie symulacji uwzględniono problemy koelucji oraz przesunięć pików względem osi indeksu czasu elucji. Zabiegi te miały na celu jak najwierniejsze odzwierciedlenie realnych danych, otrzymywanych za pomocą metody HPLC sprzężonej z metodą spektrofotometrii UV-VIS.

Po obliczeniu współczynników korelacji, określono podobieństwo próbek za pomocą miary podobieństwa s_{ij}. Wynik zebrano w macierz o wymiarowości m×m, gdzie m uzależnione jest od liczby próbek analizowanych w poszczególnych rozważaniach. Następnie na otrzymanej macierzy podobieństwa, **K**, przeprowadzono eksplorację za pomocą metod hierarchicznych z wykorzystaniem metody średnich połączeń oraz odległości euklidesowej. Dla porównania zastosowano również metodę PCA. Zastosowanie metod hierarchicznych oraz metody PCA pozwala na wyodrębnienie grup próbek podobnych oraz detekcję obiektów odległych. Wprowadzona metodologia oparta na nowej mierze odległości w połączeniu z metodami eksploracyjnymi, wydaje się być dobrym narzędziem w analizie danych o złożonej strukturze. Dodatkowym atutem zastosowanego rozwiązania jest relatywnie krótki czas analizy, który uzależniony jest od stopnia złożoności danych. Ponadto, wyniki można wzbogacić o mapy odpowiedzi uzyskane każdorazowo dla porównywanych próbek w celu identyfikacji substancji, które uległy koelucji.

Przykład 5

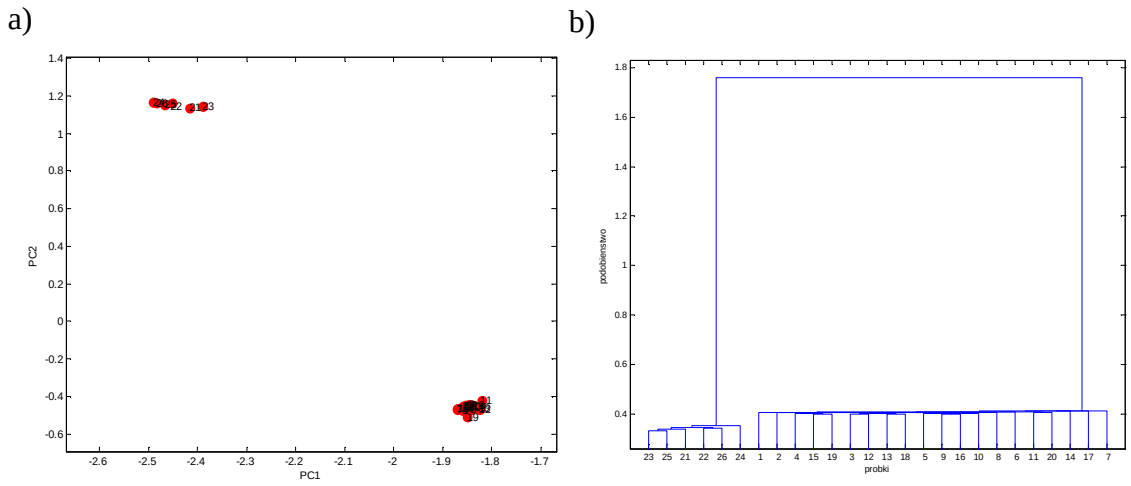
Aby potwierdzić użyteczność nowej miary podobieństwa w kontekście eksploracji wielowymiarowych danych, wysymulowano zestaw danych HPLC-DAD, zawierający dwie grupy próbek różniące się składem. Utworzone grupy zawierały odpowiednio; grupa pierwsza 20 i grupa druga 6 próbek. Każda próbka zawierała 180 długości fali przy których zarejestrowano poszczególne chromatogramy (Rys. 31), zawierające 250 punktów indeksu czasu elucji. Ostatecznie otrzymano tensor o wymiarowości 180×250×26.



Rys. 31 Dwa wybrane chromatogramy reprezentujące różnice w składzie próbek należących do grupy 1 – chromatogram oznaczony czerwoną linią oraz grupy 2 – chromatogram przedstawiony linią niebieską.

Celem przeprowadzonej analizy było określenie przydatności proponowanej metodologii w wyodrębnianiu grup próbek na podstawie ich właściwości fizykochemicznych, a uściślając składu chemicznego opisanego przez widma, jaki reprezentowały poszczególne próbki. Podczas symulacji uwzględniono koelucję substancji oraz możliwość występowania przesunięć chromatogramów w czasie, a także do każdej próbki dodano szum instrumentalny. Dzięki czemu uzyskano dane przypominające eksperymentalne zastawy danych. Procedura analizy polegała na obliczeniu współczynnika korelacji pomiędzy widmami wszystkich próbek, zgodnie z założeniem metody, w której współczynnik korelacji oblicza się pomiędzy widmem jednej próbki i odpowiadającymi mu widmami w oknie o szerokości *w* drugiej próbki.

Następnie, wykorzystując miarę podobieństwa s_{ij}, określono podobieństwo pomiędzy próbkami, a wyniki zestawiono w macierzy podobieństwa, **K**. Następnie, macierz ta została poddana analizie PCA i eksploracji z wykorzystaniem grupowania hierarchicznego, gdzie jako miarę odległości zastosowano odległość euklidesową, a jako metodę łączenia obiektów metodę średnich połączeń. Grupowanie hierarchiczne można wykonać bezpośrednio na macierzy **K** z pominięciem wprowadzania dodatkowej miary odległości jaką jest odległość euklidesowa, gdyż macierz ta zawiera już informację o podobieństwie obiektów. Jednak bez względu na to czy grupowaniu hierarchicznemu będzie podległa macierz **K**, czy macierz odległości euklidesowych otrzymana z macierzy **K**, uzyskany dendrogram będzie reprezentował taki sam podział obiektów na grupy.
W obu przypadkach otrzymano podział na dwie grupy próbek zgodny z podziałem utworzonym w trakcie symulacji danych. Wyniki eksploracji przedstawiono na Rys. 32.



Rys. 32 Wyniki eksploracji symulowanych danych HPLC-DAD o wymiarowości **X**(180×250×26): a) projekcja obiektów na płaszczyznę zdefiniowaną przez PC1 i PC2, b) dendrogram otrzymany w wyniku grupowania hierarchicznego z zastosowaniem metody średnich połączeń i odległości euklidesowej.

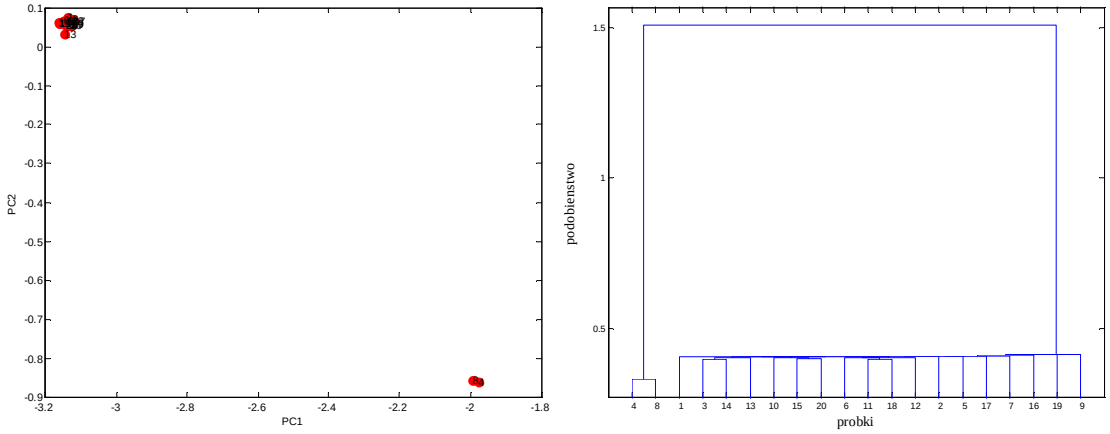
Wizualna ocena otrzymanych wyników ukazuje wyraźnie rozdzielone dwie grupy o małej wariancji w obrębie każdej z nich. Niewielkie rozproszenie próbek na projekcji obiektów na płaszczyznę zdefiniowaną przez pierwszy i drugi czynnik główny oraz zbliżone podobieństwo, czyli zbliżona odległość euklidesowa pomiędzy próbkami, na dendrogramie są wynikiem wyrażenia podobieństwa w macierzy **K** za pomocą liczb z przedziału [0,1]. Uściślając, te próbki które są do siebie bardziej podobne (należące

do jednej grupy) przyjmują wartość s_{ij} około 0,7, a te pochodzące z różnych grup przyjmują wartość s_{ij} rzędu 0,4.

Przykład 6

Analizując wyniki eksploracji dla Przykładu 5, można zauważyć, że nie ma przeszkód aby zaproponowaną metodologię z powodzeniem wykorzystać podczas rutynowej kontroli jakości produktu. Jednym z możliwych obszarów zastosowań jest kontrola jakości preparatów farmaceutycznych pochodzenia syntetycznego i naturalnego. Jeżeli w trakcie procesu produkcji zostanie wytworzony produkt niespełniający obowiązujących norm, to porównując jego skład ze składem innych próbek (lub składem wzorca) można taką próbkę relatywie łatwo zidentyfikować. Takie rozwiązanie problemu eliminuje etap nakładania chromatogramów na siebie, co znacząco skraca czas analizy. Metodologia ta pomija również etap sumowania danych po jednym z wymiarów tensora, co zazwyczaj ma miejsce w przypadku korzystania z innych metod eksploracji czy analizy danych. Dzięki takiemu rozwiązaniu, zachowuje się kompletną informację o badanych obiektach. Wykorzystując macierz podobieństw, **K**, podczas eksploracji danych uzyskuje się informację o obiekcie odbiegającym od ustalonych norm. Dodatkowo, znajomość pozycji obiektów w macierzy **K**, pozwala na utworzenie mapy odpowiedzi dla próbki odległej oraz wzorca lub wybranej próbki, która spełnia ustalone normy. Uzyskana w ten sposób wizualizacja danych może okazać się cennym źródłem informacji na temat przyczyn odmienności składu próbki badanej od wymaganych norm. Aby wykazać przydatność proponowanej metody wykorzystanej do kontroli jakości produktów użyto danych z poprzedniego przykładu. Jednakże dane te poddano modyfikacji polegającej na ograniczeniu liczby próbek do 20. Losowo wybrano 18 próbek tworzących grupę pierwszą i wprowadzono do niej dwa obiekty charakteryzujące się odmiennym składem chemicznym. Następnie, zastosowano zaproponowaną w tej pracy metodologię oceny podobieństw. Na podstawie podobieństwa próbek określonego za pomocą miary s_{ij} utworzono macierz podobieństw, **K** i wykorzystano ją na etapie eksploracji danych. Analogicznie jak w poprzednim przykładzie wykorzystano metodę PCA oraz metodę grupowania hierarchicznego, decydując się na metodę średnich połączeń jako metodę łączenia obiektów oraz określenia podobieństwa za pomocą odległości euklidesowej. Otrzymane wyniki potwierdzają, skuteczność zastosowanego podejścia eksploracji danych. Wyniki otrzymane za pomocą obu metod ujawniają próbki różniące się składem chemicznym. W rozważanym przypadku obie próbki odległe wykazują zbliżony skład, co obrazuje poziom ich podobieństwa na dendrogramie oraz ich położenie na projekcji obiektów zdefiniowanej przez pierwsze dwa czynniki główne. Oczywiście może się zdarzyć sytuacja, że obiekty odległe będą wykazywały odmienny skład chemiczny, co ujawni wizualizacja wyników.

a) b)



Rys. 33 Efekt eksploracji macierzy podobieństwa, **K**, podczas kontroli jakości produktu a) projekcja obiektów na płaszczyznę zdefiniowaną przez PC1 i PC2, b) dendrogram uzyskany za pomocą grupowania hierarchicznego przy użyciu metody średnich połączeń i odległości euklidesowej

12.2.5 Wykorzystanie nowej miary podobieństwa do określania autentyczności próbek leku Viagra na podstawie ich składu chemicznego

Zaproponowaną metodologię porównywania próbek zastosowano do weryfikacji autentyczności leku Viagra. Porównywania dokonano na podstawie składu chemicznego analizowanych próbek leku. Podczas wyboru problemu badawczego kierowano się wzrastającym zainteresowaniem tematem fałszowania różnorodnych produktów, w tym produktów spożywczych, kosmetyków oraz leków. Fałszowanie produktów farmaceutycznych stanowi poważny problem w skali światowej. Zwłaszcza problematyczne staje się podrabianie produktów farmaceutycznych, produkowanych w nielegalnych laboratoriach bez zachowania wymaganych środków ostrożności oraz nadzoru jakości składu farmaceutyków. Prowadzi to do syntezy lekarstw zawierających znaczne ilości zanieczyszczeń i/lub nieodpowiednie stężenia substancji aktywnych. W ostatnich latach odnotowuje się coraz to większy procent przypadków śmierci, której przyczyną było przyjmowanie leków z nielegalnego źródła. Dlatego odróżnienie leków autentycznych od zafałszowanych jest kluczową kwestią w ograniczeniu rozprowadzania nielegalnych środków farmaceutycznych.

Oceny jakościowej i ilościowej leków dokonuje się za pomocą zaawansowanych metod instrumentalnych, w tym m.in. metody HPLC-DAD. Otrzymuje się wówczas, potencjalnie, bogatą informację o składzie chemicznym leku, stężeniu substancji aktywnej i obecnych zanieczyszczeniach. Pojawia się jednak problem, o którym

niejednokrotnie wspominano w niniejszej pracy, jakim jest wymiarowość uzyskiwanych danych. Opis każdej próbki za pomocą chromatograficznego odcisku palca pociąga za sobą konieczność redukcji jednego z wymiarów, a w następstwie prowadzi do częściowej utraty informacji zawartej w danych. Dodatkowo, zastosowanie metod przygotowania danych do dalszej analizy, mających na celu eliminację szumu instrumentalnego, korektę linii podstawowej, czy nałożenie sygnałów na siebie pociąga za sobą konieczność zastosowania odpowiednich metod chemometrycznych, co znacząco wydłuża czas analizy. Najbardziej uciążliwym etapem wstępnego przygotowania sygnałów chromatograficznych jest ich nakładanie. W badaniach starano się wypracować takie podejście, które pozwoliłoby, przynajmniej w określonych sytuacjach, wyeliminować ten etap. Zaproponowano alternatywne podejście eksploracji tego typu danych w oparciu o wprowadzoną metodologię porównywania próbek opisanych przez dwuwymiarowe chromatograficzne odciski palca.

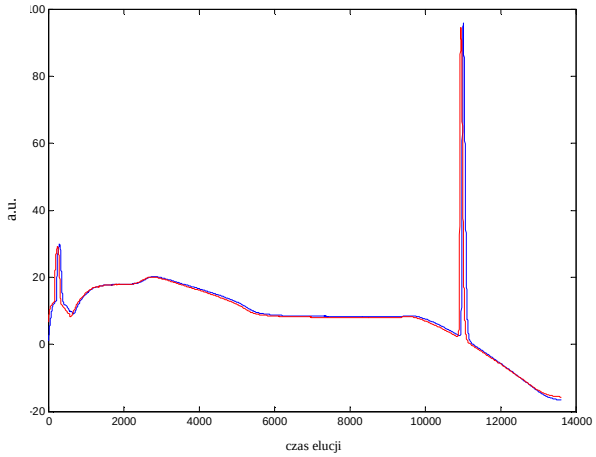
Porównanie próbek leku Viagra na podstawie ich składu chemicznego przy użyciu nowej miary podobieństwa s_{ij} oraz eksploracja otrzymanej macierzy podobieństwa, **K**, metodami grupowania hierarchicznego lub PCA, wydaje się dobrym rozwiązaniem pozwalającym wyodrębnić poszczególne grupy leków.

Próbki leku analizowano metodą HPLC-DAD, otrzymując tensor danych o wymiarowości 162×13620×143. Ze 143 próbek 46 stanowiły próbki autentyczne leku, a pozostałe 97 to próbki Viagry zafałszowanej. Dane przedstawiono bliżej w Przykładzie 7.

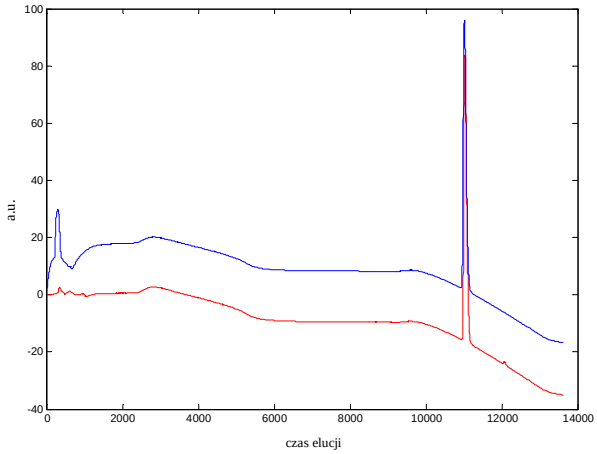
Przykład 7

Spośród 143 analizowanych próbek losowo wybrano 5 autentycznych próbek Viagry oraz 5 próbek, będących próbkami zafałszowanymi. Wizualna analiza wybranych dwuwymiarowych chromatograficznych odcisków palca ujawnia obecność przesunięć pików substancji względem osi czasu retencji (Rys. 34). Ten fakt pozwala zweryfikować skuteczności działania metody porównawczej, podczas analizy realnych danych, w których występuje co najmniej jeden z wymienionych problemów chromatograficznych. Autentyczne próbki Viagry oraz zafałszowane różnią się składem chemicznym (Rys. 35). Leki zafałszowane zawsze zawierają pewne ilości zanieczyszczeń, a ich stężenie jest różne w poszczególnych próbkach. Takie zjawisko można przyrównać do problemu obecności lub braku piku substancji. Jest to czynnik różnicujący próbki i stanowiący podstawę analizy wykorzystującej miarę podobieństwa, s_{ij}. Występowanie dodatkowych substancji w próbkach jest związane z występowaniem widm tych substancji rejestrowanych w zakresie UV-VIS. Dlatego wydaje się, że wprowadzona metoda badania podobieństw, która bazuje na porównywaniu widm UV-VIS ze sobą w zakresie *w*, dobrze sprawdzi się w rozważanym przypadku. Dodatkowym atutem, o którym już wspominano, jest fakt, że metodę porównywania można zastosować bezpośrednio dla surowych danych,

pomijając etap przygotowania danych do dalszej analizy, zwierający korektę linii podstawowej, eliminację szumu, wybór zmiennych, czy nakładanie sygnałów na siebie względem osi czasu.

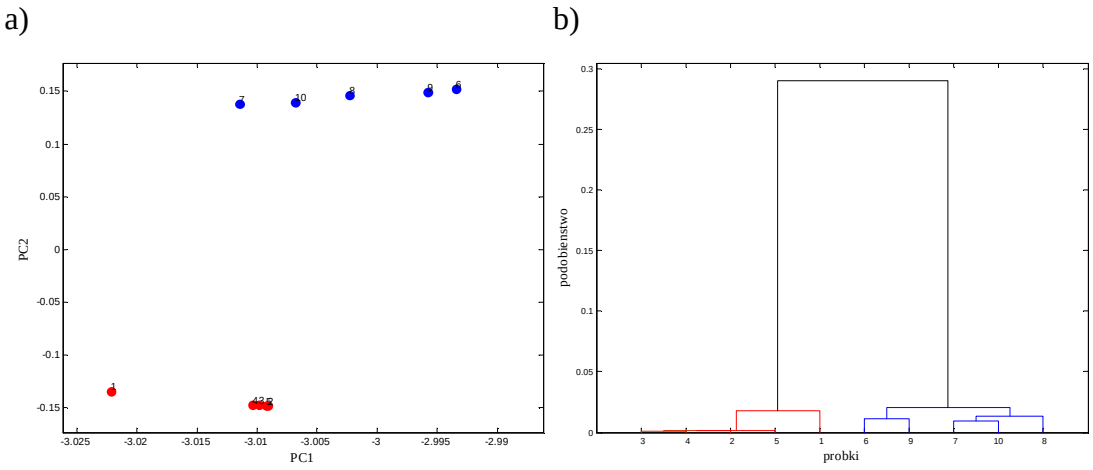


Rys. 34 Sumaryczne chromatogramy reprezentujące problem przesunięć pików chromatograficznych występujący w próbkach leku Viagra.



Rys. 35 Sumaryczne chromatogramy ilustrujące różnice w składzie próbek leku Viagra. Kolorem niebieskim przedstawiono chromatogram próbki autentycznej, a czerwonym próbki zafałszowanej.

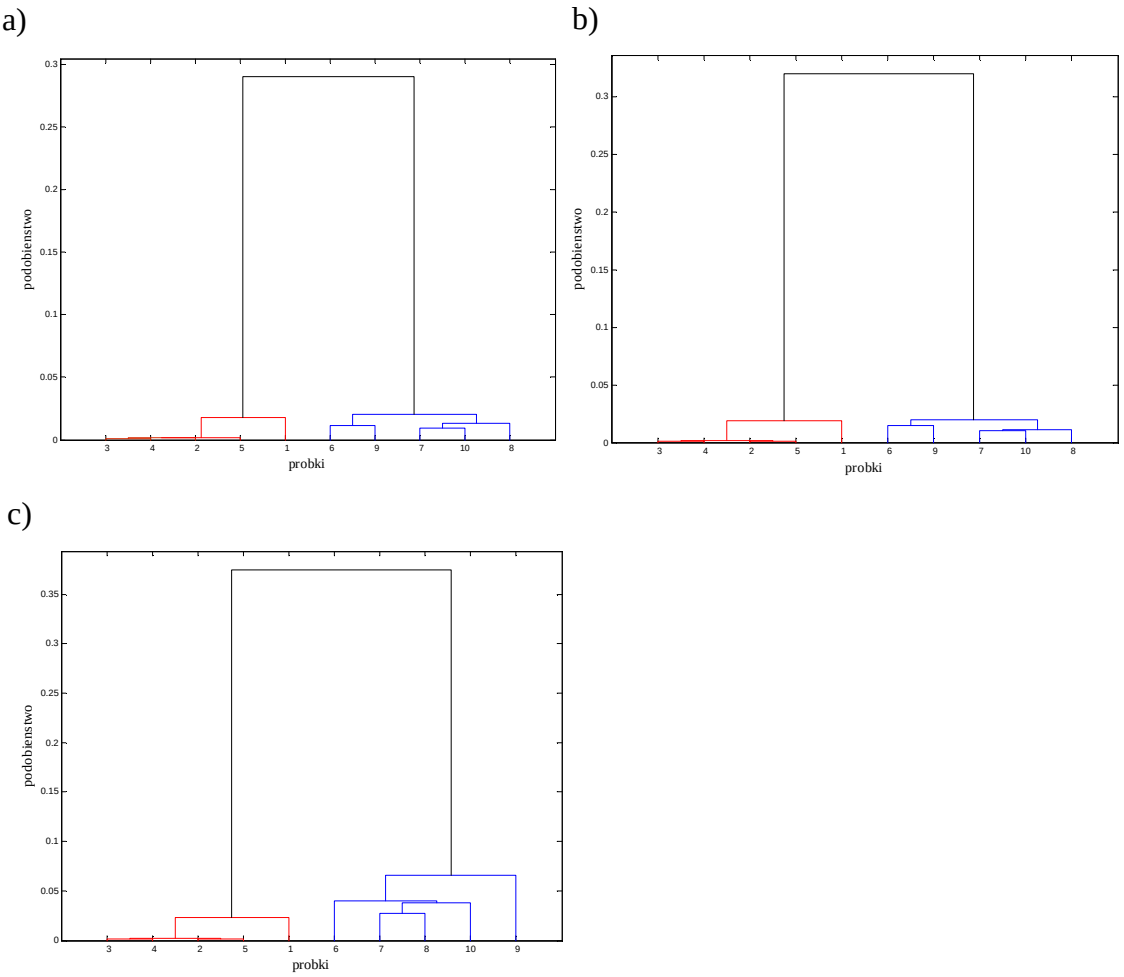
Ze względu na dużą objętość danych (ok. 200 Mb), przeskalowano je, zmniejszając liczbę punktów pomiarowych na osi czasu elucji z 13620 do 6810. Aby zbadać efektywność podejścia, danych celowo nie poddano wstępnemu przygotowaniu do analizy. Przeskalowane dane analizowano za pomocą wprowadzonej metodologii (dla w = 100), otrzymaną macierz podobieństwa, **K**, eksplorowano za pomocą metod hierarchicznych oraz metody PCA (Rys. 36). Wizualna ocena wyników otrzymanych za pomocą obu metod eksploracyjnych ujawniała obecność dwóch grup próbek, jednej utworzonej przez próbki autentyczne i drugiej przez próbki zafałszowane.



Rys. 36 Wyniki eksploracji danych przeprowadzonej w celu weryfikacji autentyczności leku Viagra. Kolorem czerwonym oznaczono próbki Viagry autentycznej, a kolorem niebieskim próbki Viagry nieautentycznej: a) projekcja obiektów na płaszczyzny zdefiniowane przez pierwszy i drugi czynnik główny, b) dendrogram otrzymany za pomocą grupowania hierarchicznego wykorzystując metodę średnich połączeń jako metodę łączenia obiektów i odległość euklidesową jako miarę podobieństwa.

Następnie analizę przeprowadzono dla w = 100, w = 30 oraz w = 0. Doboru szerokości okna w dokonano na podstawie zakresu szerokości przesunięć pików substancji obserwowanych na chromatogramach. W pierwszym przypadku szerokość okna ustalono na podstawie szerokości piku substancji aktywnej, w drugim brano pod uwagę różnicę w czasie elucji występowania pików substancji aktywnej oraz ewentualnych zanieczyszczeń. Trzeci przypadek, w którym nie zastosowano okna, miał ukazać różnice w efektywności działania algorytmu w przypadku jego zastosowania oraz po jego wyeliminowaniu. Wyniki grupowania hierarchicznego macierzy **K** otrzymanej dla wszystkich trzech przypadków zaprezentowano na Rys. 37. Wyniki grupowania otrzymane dla w = 100 oraz w = 30, odpowiadają oczekiwanym różnicom pomiędzy próbkami. Jedyna różnica związana jest z poziomem podobieństwa utworzonych grup

względem siebie. W przypadku gdy, zastosowane okno jest węższe grupy próbek leku są od siebie lepiej odseparowane. Dlatego, w tym przypadku wartość okna *w* = 30 wydaje się korzystniejsza. Porównując odpowiadające sobie widma w poszczególnych próbkach zakładając *a priori* brak przesunięć pików z pominięciem okna *w*, również otrzymano dwie grupy próbek leku oryginalnego oraz zafałszowanego. Jednak podobieństwo próbek w obrębie grupy leku zafałszowanego jest odmienne od tego uzyskanego przy zastosowaniu okna *w* o szerokości 100 i 30.



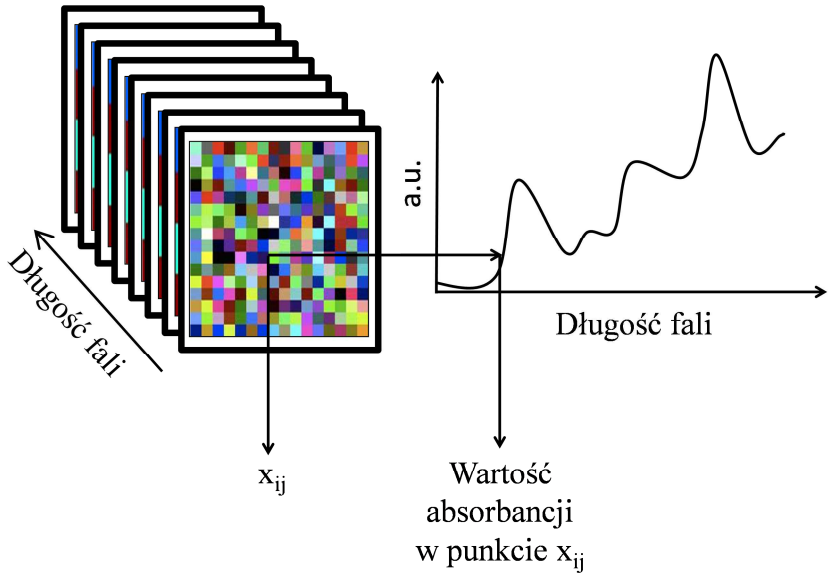
Rys. 37 Zestawienie wyników grupowania hierarchicznego (miara podobieństwa odległość euklidesowa, metoda średnich połączeń) macierzy podobieństwa, **K**, otrzymanej za pomocą nowej metodologii porównywania próbek na podstawie ich składu chemicznego odpowiednio dla a) *w* = 100, b) *w* = 30 oraz c) *w* = 0.

Grupa próbek Viagry, która jest zafałszowana charakteryzuje się bardziej różnorodnym składem aniżeli próbki leku autentycznego i wyniki analizy otrzymane bez zastosowania okna zdecydowanie lepiej to reprezentują. We wszystkich przypadkach próbka 1 w grupie leku autentycznego różni się nieco od pozostałych. Ta różnica wynika z występujących w niej przesunięć pików w czasie w porównaniu do pozostałych sygnałów próbek. Należy jednak pamiętać, że pomimo obecności przesunięć pików w czasie próbka ta została dobrze przypisana do grupy.
Odróżnienie próbek Viagry autentycznej od nieautentycznej na podstawie ich składu chemicznego przy użyciu wprowadzonej metodologii potwierdza jej efektywność. Umożliwia ona eksplorację wyników z pominięciem etapu przygotowania próbek do dalszej analizy. Metodę tę można zastosować dla surowych danych, a w przypadku dużej liczby punktów odpowiadających czasom elucji, dane można przeskalować wykorzystując metody bazujące na interpolacji sygnałów instrumentalnych.

12.3 Zastosowanie metod grupowania danych w segmentacji obrazów hiperspektralnych

W określaniu właściwości fizykochemicznych materiałów różnorodnego pochodzenia coraz częściej wykorzystuje się obrazowanie hiperspektralne [98]. Stosowane jest ono głównie w celu analizy jakościowej i ilościowej powierzchni badanych próbek materiałów. Otrzymane w ten sposób dane mają postać tensora. W każdym pikselu otrzymanego obrazu zapisane jest odpowiadające danemu fragmentowi przedmiotu widmo spektralne, chrakteryzujące jego skład chemiczny oraz właściwości fizykochemiczne. Dzięki takiej reprezentacji materii możliwe jest odróżnienie od siebie poszczególnych obszarów na podstawie różnic w ich oddziaływaniu z falą elektromagnetyczną. Cechy obrazowania hiperspektralnego są wykorzystywane m.in. w medycynie podczas obrazowania tkanek objętych procesem chorobowym oraz obrazowaniu powierzchni ziemi, przy określaniu topografii wybranego terenu.
W procesie obrazowania hiperspektralnego rejestruje się dużą liczbę wysoce skorelowanych widm. Analiza tego typu danych wymaga zastosowania metod chemometrycznych umożliwiających tzw. segmentację obrazów, a więc podział obrazu na jednorodne obszary wykazujące zbliżone właściwości fizyczne i/lub chemiczne. Detekcja tych obszarów bazuje na analizie podobieństw pomiędzy widmami. A zatem, segmentacja obrazu polega na grupowaniu pikseli wykazujących wysoki stopień podobieństwa zgodnie z przyjętym kryterium podobieństwa. Najczęściej w celu ich segmentacji redukuje się jeden z wymiarów obrazu poprzez zsumowanie wartości dla wymiaru zawierającego widma. Do najczęściej stosowanych metod należą metoda PCA oraz metoda k-średnich, umożliwiające segmentację obrazu oraz jego rekonstrukcję na podstawie odpowiednio wybranej liczby czynników głównych oraz zdefiniowanej na wejściu liczby grup. Zaproponowana metodologia oparta na wprowadzonej mierze podobieństwa, s_{ij}, jest narzędziem dzięki któremu segmentacji obrazu można dokonać

pomijając etap redukcji trzeciego wymiaru i przeprowadzeniu eksploracji na surowych danych. Porównując widma określające właściwości fizykochemiczne możliwe jest wyodrębnienie podobnych obszarów obrazu. Ponieważ w przypadku obrazów hiperspektralnych nie odnotowuje się problemów przesunięć pików, czy koelucji substancji, to można pominąć zastosowanie okna *w*.



Rys. 38 Schematyczne przedstawienie obrazu hiperspektralnego.

W niniejszym rozdziale posłużono się ogólnodostępnymi obrazami hiperspektralnymi reprezentującymi tzw. sceny, którymi są np. krajobrazy zarejestrowane w Minho w Portugalii, przedstawiające roślinność (np. liście, drzewa, trawę, glebę) oraz takie które zarejestrowano w miastach Porto i Braga w Portugalii, ilustrujące krajobraz miejski, tj. fragmenty budynków, ulic, itp. [120], [121]. W obrazowaniu hiperspektralnym stosowano długości fali z zakresu widzialnego fali elektromagnetycznej, obejmującego zakres od 410 nm do 710 nm, z przedziałem co 10 nm. Tym samym otrzymano 31 długości fali przy której zarejestrowano obraz reprezentujący wybrany fragment krajobrazu.
Do analizy wybranych obrazów zastosowano typową segmentację za pomocą metody k-średnich oraz metody PCA. Następnie posłużono się wprowadzoną metodologią w celu porównania efektywności metody w rozważanym przypadku. Procedurę przeprowadzonej analizy opisano w Przykładzie 8.

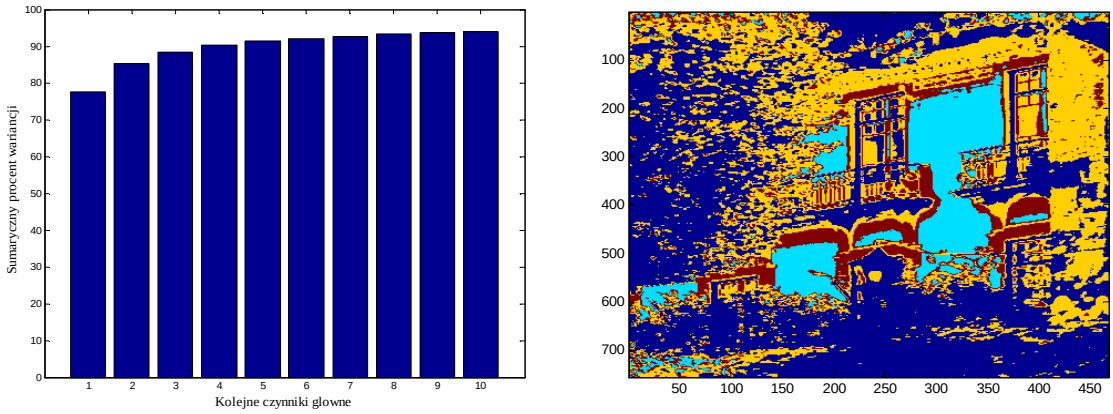
Przykład 8

Analizie poddano obraz hiperspektralny przedstawiający budynek usytuowany wśród roślinności, co ukazuje Rys. 39. Wymiarowość tensora, który zawierał dane dla obrazu hiperspektralnego wynosiła 755×748×31. Aby przeprowadzić segmentację obrazu dane zsumowano, tak aby jego wymiarowość wynosiła 755×748. Następnie dane poddano eksploracji za pomocą metody PCA oraz metody k-średnich. W przypadku metody k-średnich stosowano różne warianty liczby grup, rozpoczynając od k = 2, a na k = 10 skończywszy. Następnie obraz rekonstruowano na podstawie, odpowiednio, liczby czynników głównych oraz liczby grup, oceniając wizualnie poprawność przeprowadzonej segmentacji obrazu. Uzyskane efekty grupowania zaprezentowano na Rys. 40 oraz 41.

Rys. 39 Fragment krajobrazu, który rejestrowano za pomocą kamery hiperspektralnej.

Następnie, surowe dane poddano eksploracji za pomocą wprowadzonej metodologii. Uzyskaną macierz podobieństwa, **K**, poddano grupowaniu hierarchicznemu, uzyskując dendrogram przedstawiający podobieństwo pomiędzy obrazami rejestrowanymi przy 31 długościach fali (Rys. 42). Na podstawie utworzonego dendrogramu 31 obrazów tworzących obraz hiperspektralny można podzielić na 4 grupy obrazów wykazujących odmienne właściwości fizykochemiczne ujawniane ze względu na stosowaną długość fali elektromagnetycznej.

a) b)



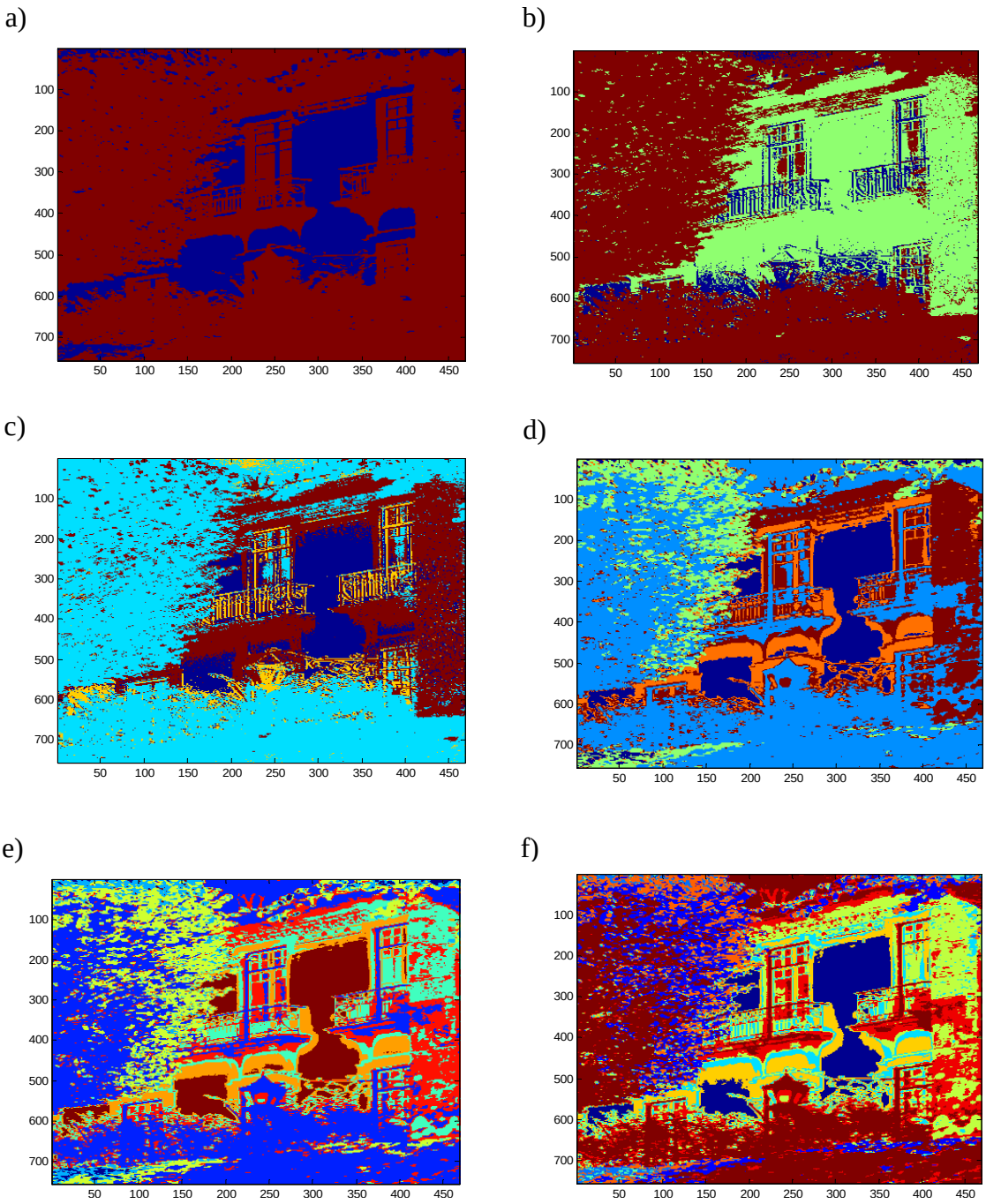
Rys. 40 Efekt rekonstrukcji obrazu hiperspektralnego z Rys. 39 za pomocą metody PCA a) procent wariancji opisany przez 10 czynników głównych, b) obraz zrekonstruowany na podstawie czterech pierwszych czynników głównych.

Takie podejście ułatwia określenie liczby grup podobnych obszarów, a rekonstrukcja obrazu staje się znacznie łatwiejsza. Wprowadzone podejście analizy danych oparte na nowej mierze podobieństwa może być alternatywą dla wykorzystywanych dotychczas metod segmentacji obrazów hiperspektralnych. Rekonstrukcja obrazu hiperspektralnego na podstawie liczby grup określonej przy użyciu otrzymanego dendrogramu (Rys. 42) pociąga za sobą konieczność sumowania jednego z wymiarów obrazu (zawierającego widma), jeżeli obraz ten ma zostać przedstawiony w formie dwuwymiarowej. W przypadku, gdy istnieje możliwość przedstawienia obrazu w formie trójwymiarowej etap sumowania widm może zostać pominięty. Wykorzystując skonstruowany dendrogram obraz można odtworzyć zgodnie z liczbą grup reprezentowaną na dendrogramie. Otrzymuje się wówczas obraz zrekonstruowany na podstawie długości fali, w których obrazowana materia wykazuje zbliżone właściwości fizykochemiczne. Oznacza to, że dendrogram pozwala na segmentację widma na fragmenty odpowiadające substancjom wykazującym swoje właściwości fizyczne i/lub chemiczne przy zastosowanej długości fali (Rys. 43).

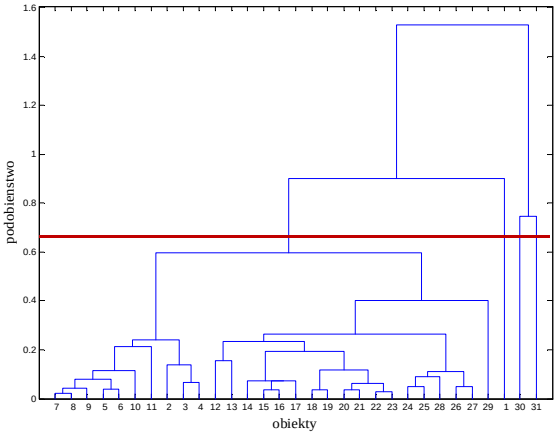
Resumując, zaproponowany sposób analizy obrazów hiperspektralnych stanowi wyłącznie alternatywne podejście dla typowej segmentacji obrazów stosowanej do tej pory. Wykazuje ono dwie zalety, po pierwsze daje możliwość analizy surowych danych bez uprzedniego wstępnego przygotowania, tj. sumowania widm lub zastosowania innych metod. Po drugie, liczba grup zostaje odczytana z dendrogramu, dzięki czemu wielokrotne odtwarzanie obrazu na podstawie zmiennej liczy czynników głównych lub liczby grup w metodach PCA oraz k-średnich, zostaje ograniczone do minimum. Jednak najczęściej wizualizacja obrazu możliwa jest poprzez jego przedstawienie w postaci dwuwymiarowej, a nie trójwymiarowej co pociąga za sobą konieczność

redukcji wymiaru przedstawiającego widma tak jak w przypadku pozostałych dwóch metod. Przeprowadzona segmentacja obrazów hiperspektralnych miała na celu zaprezentowanie możliwości jakie daje wprowadzona metodologia oparta na nowej mierze odległości, s_{ij}, jednak zalety eksploracji obrazów za pomocą zaproponowanego podejścia nie umniejszają segmentacji obrazów pozostałymi metodami eksploracji danych, dlatego wybór sposobu jego analizy uzależniony jest wyłącznie od indywidualnych potrzeb użytkownika.

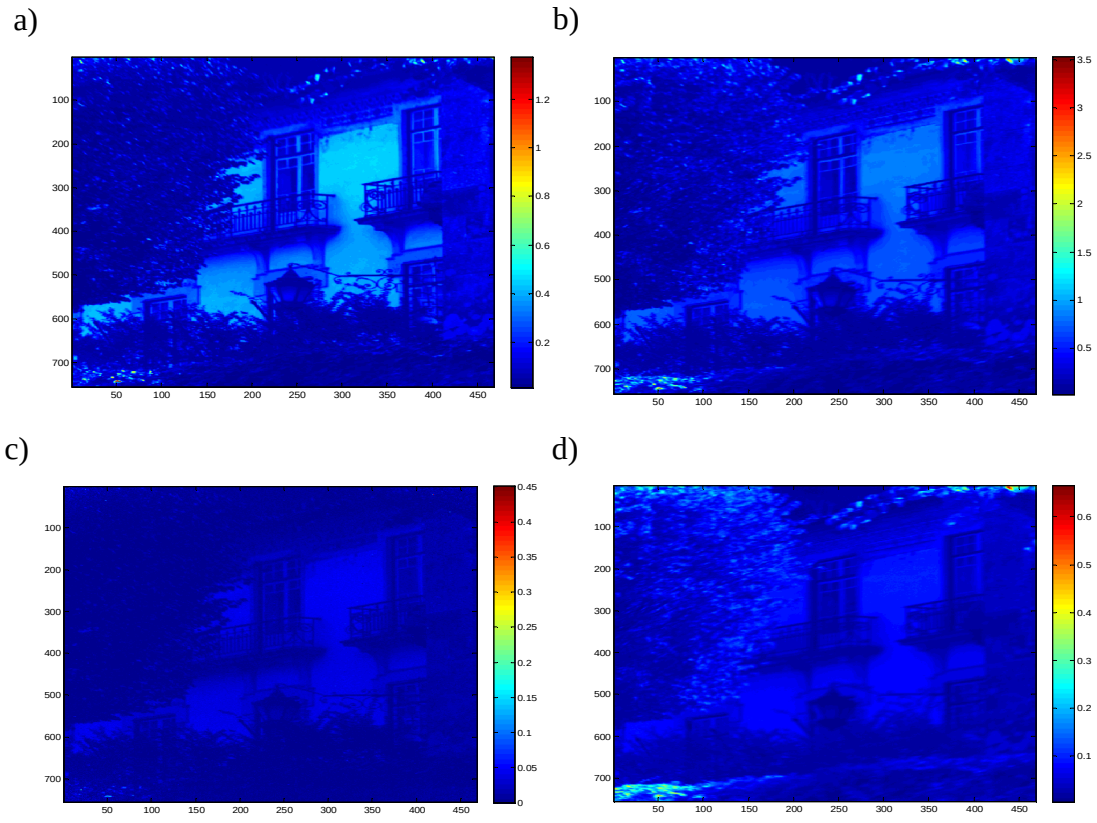
Z kolei zestawienie wyników otrzymanych za pomocą metody PCA oraz metody k-średnich ujawniło, że segmentacja obrazu oraz jego późniejsza rekonstrukcja jest efektywniejsza w przypadku metody PCA. Obraz odtworzony na podstawie czterech czynników głównych lepiej reprezentuje zarejestrowane za pomocą obrazowania hiperspektralnego właściwości fizykochemiczne badanej materii.



Rys. 41 Efekt segmentacji obrazu za pomocą metody k-średnich dla różnej liczby rozważanych grup a) k = 2, b) k = 3, c) k = 4, d) k = 5, e) k = 7 oraz f) k = 10.



Rys. 42 Dendrogram, obrazujący podobieństwo pomiędzy kolejnymi obrazami tworzącymi obraz hiperspektralny, skonstruowany za pomocą metody średnich połączeń, gdzie jako miarę podobieństwa zastosowano odległość euklidesową.



Rys. 43 Rekonstrukcja obrazu hiperspektralnego przedstawionego na Rys. 39, względem grup utworzonych na podstawie długości fali reprezentowanych przez dendrogram a) dla obrazów od 2 do 11, b) dla obrazów od 12 do 29, c) dla 1-szego obrazu, d) dla obrazów 30 oraz 31.

Przykład 9

W ramach analizy obrazów hiperspektralnych sprawdzono wpływ stosowanej miary podobieństwa na obserwowane efekty segmentacji. W tym celu wybrano obraz hiperspektralny przedstawiający fragment krzewu (Rys. 44). Grupując obraz za pomocą metody k-średnich zastosowano różne miary podobieństwa w tym odległość euklidesową, kwadrat odległości euklidesowej, czy współczynnik korelacji. Na Rys. 45 porównano efekt rekonstrukcji obrazu dla współczynnika korelacji oraz kwadratu odległości euklidesowej dla k = 3. Zdecydowano się na przedstawienie wyników dla tych dwóch miar ponieważ efekt rekonstrukcji obrazu w tych dwóch przypadkach był najłatwiej dostrzegalny.
Dodatkowo efekty segmentacji porównano z rekonstrukcją obrazu uzyskaną za pomocą metody PCA (Rys. 46).

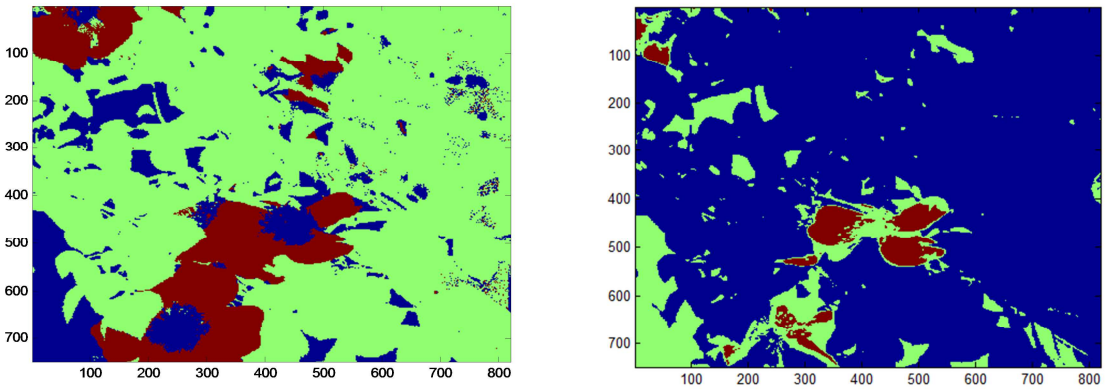
Rys. 44 Zdjęcie przedstawiające fragment krzewu, który poddano obrazowaniu hiperspektralnemu.

Wizualna ocena obrazu zrekonstruowanego za pomocą metody k-średnich z zastosowaniem jako miary podobieństwa współczynnika korelacji oraz kwadratu odległości euklidesowej ujawnia, że dobór miary odległości jest kluczowym elementem wpływającym na skuteczność przeprowadzanej segmentacji obrazu. Ciężko jednak ocenić, która z przeprowadzonych segmentacji przebiegła dokładniej. Zastosowanie współczynnika korelacji pozwoliło na wyodrębnienie jednorodnych obszarów wykazujących zbliżone właściwości fizykochemiczne oraz funkcje biologiczne np. płatki kwiatów, liście i łodygi oraz tło. Z kolei zastosowanie kwadratu odległości euklidesowej pozwala na wyizolowanie poszczególnych obszarów obrazu z większą dokładnością, np. kwiaty reprezentowane są za pomocą dwóch barw, co świadczy o ich odmiennym właściwościach. Informacja ta pokrywa się z tą

zaobserwowaną na oryginalnym obrazie (Rys. 44), płatki kwiatów wykazują zabarwienie od jasnoróżowego do ciemnoróżowego. Niepokojące jest jednak, że podczas rekonstrukcji tło obrazu oraz części kwiatów zostały przypisane do tej samej grupy, co najprawdopodobniej ma związek z barwą obu części obrazu oraz związaną z tym faktem absorpcją promieniowania elektromagnetycznego w podobnym zakresie. Drugi problem związany jest z prawie całkowitą utratą informacji o obecności trzeciego kwiatu (zob. lewy górny róg Rys. 45 b). W kontekście analizy obrazów medycznych lub odwzorowujących topografię terenu lepszym rozwiązaniem wydaje się zastosowanie współczynnika korelacji jako miary podobieństwa w procesie ich grupowania.

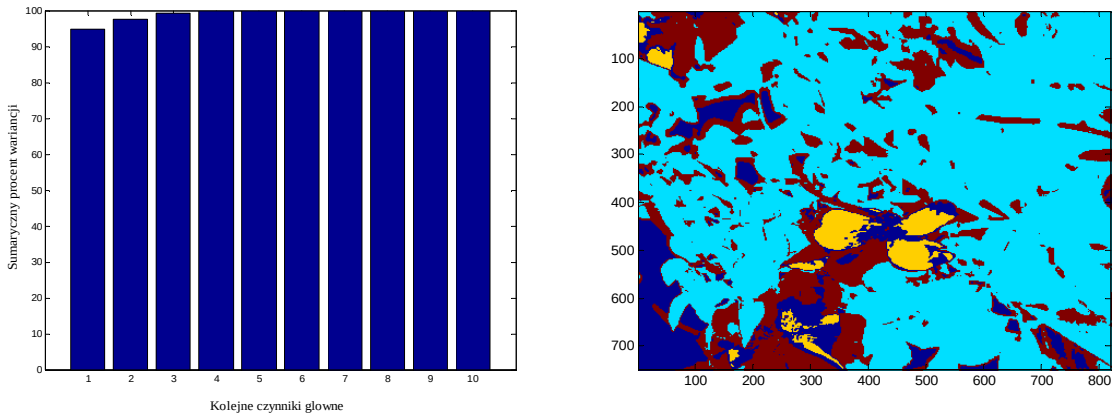
Zestawiając otrzymane metodą k-średnich wyniki w obu wariantach segmentacji z tymi uzyskanymi za pomocą metody PCA tak jak w poprzednim przykładzie (Przykład 8), wizualna ocena zrekonstruowanych obrazów ukazuje, że metoda PCA jest lepszym narzędziem segmentacji obrazów hiperspektralnych, gdyż wszystkie kluczowe obszary obrazu zostały wyizolowane poprawnie.

a) b)



Rys. 45 Rekonstrukcja wybranego obrazu hiperspektralnego po przeprowadzonej segmentacji za pomocą metody k-średnich dla k = 3 oraz dwóch różnych kryteriów podobieństwa odpowiednio: a) współczynnik korelacji Pearsona i b) kwadrat odległości euklidesowej. Na rysunku poszczególne obszary charakteryzujące się podobnymi widmami mają ten sam kolor (granatowy, zielony lub brązowy).

a) b)



Rys. 46 Rekonstrukcja obrazu hiperspektralnego za pomocą metody PCA
a) sumaryczny procent wariancji opisany przez 10 czynników głównych,
b) zrekonstruowany obraz na podstawie czterech czynników głównych.

12.4 Metody współgrupowania danych w eksploracji danych chemicznych

Zastosowanie metod współgrupowania danych w eksploracji danych pochodzących z mikromacierzy, stało się punktem odniesienia do wprowadzenia ich jako narzędzia eksploracji danych chemicznych. Literatura nie zawiera zbyt wielu doniesień informujących o ich wykorzystaniu w tym kontekście. Z tego powodu podjęto próbę oceny ich przydatności w analizie sygnałów instrumentalnych.

Ponieważ metody współgrupowania danych umożliwiają wyodrębnienie podmacierzy z macierzy danych **X**, to wydają się dobrym narzędziem umożliwiającym wyizolowanie podgrup próbek opisanych określoną podgrupą parametrów. Wydaje się to szczególnie obiecujące w przypadku eksploracji danych pozyskiwanych w badaniach z zakresu nauk biologii systemowej takich jak metabolomika, czy proteomika, gdzie analizy próbek dokonuje się za pomocą zaawansowanych metod instrumentalnych. Metody współgrupowania danych wykorzystane jako narzędzie pozwalające na wyodrębnienie wskaźników biologicznych (biomarkerów) w próbkach biologicznych (tj. mocz, krew, czy inne płyny biologiczne) usprawniłoby i znacząco ułatwiło interpretację gromadzonych wyników.

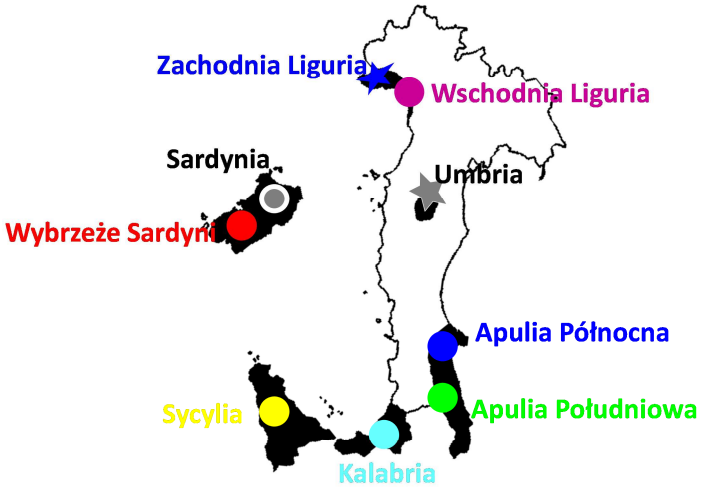
W niniejszej pracy wybrano odpowiednie algorytmy współgrupowania danych w kontekście eksploracji zestawów danych otrzymanych metodami instrumentalnymi. Aplikowano je również w eksploracji symulowanych zestawów danych, tak aby sprawdzić ich użyteczność. Jednak w wielu przypadkach, nie udało się zoptymalizować parametrów wejścia lub algorytm nie uzbieżniał się. W innych przypadkach

otrzymywane wyniki nie miały sensu z chemicznego punktu widzenia. Można przypuszczać, że jedną z przyczyn napotykanych problemów, była wariancja danych chemicznych, różniąca się od wariancji obecnej w danych mikromacierzowych. W przypadku danych uzyskiwanych w badaniach genomicznych, podczas analizy których porównywaniu poddaje się ekspresję genów, należy brać pod uwagę różnorodność biologiczną oraz fakt że ekspresja ulega zasadniczym zmianom w czasie. W przypadku danych chemicznych, np. tych otrzymywanych metodami instrumentalnymi, reprezentują one skład analizowanych próbek, a obserwowane różnice związane są ze stężeniem poszczególnych substancji lub ewentualnie występowaniem lub brakiem dodatkowych substancji w porównaniu z pozostałymi próbkami. Wariancja pomiędzy próbkami wynosi kilka procent, w porównaniu z wariancją biologiczną wynoszącą nawet 60% reprezentowanej informacji jest ona niewielka.
Zadaniem algorytmów współgrupowania danych jest wyizolowanie obszarów (podgrup obiektów i zmiennych) znacząco różniących się od pozostałych obszarów danych. Obszary te wykazują spójność lub inaczej mówiąc homogeniczność zawartej w nich informacji. Dlatego, zdecydowano się zastosować dwa algorytmy współgrupowania danych w celu eksploracji dwóch dobrze poznanych zestawów danych. Znajomość danych pozwala na obiektywną ocenę wyników uzyskanych za pomocą technik współgrupowania, które nie były wcześniej używane w kontekście eksploracji danych chemicznych.
Jako pierwszy wykorzystano algorytm SMR przedstawiony w podrozdziale *10.1.3* w celu eksploracji danych charakteryzujących oliwę z oliwek pochodzącą z różnych rejonów Włoch. Wyniki uzyskane metodą SMR zestawiono z wynikami uzyskanymi metodami PCA oraz dwukierunkowym grupowaniem hierarchicznym. Następnie zastosowano metodę k-spectral w zestawieniu z dwukierunkowym grupowaniem hierarchicznym w eksploracji danych opisujących próbki opium z trzech rejonów Indii. Metoda k-spectral została opisana w podrozdziale *10.1.2*.

Przykład 10

W celu określenia efektywności metody SMR w kontekście analizy danych chemicznych, eksploracji poddano zestaw danych zawierający 572 próbki oliwy z oliwek, pochodzące z 9 regionów uprawnych Włoch, tj. Północna Apulia, Zachodnia i Wschodnia Liguria, Umbria, Wybrzeże Sardynii oraz Sardynia Śródlądowa (nazywana dla uproszczenia Sardynią), Sycylia, Południowa i Północna Apulia oraz Kalabria (Rys. 47).
W próbkach, za pomocą techniki chromatografii gazowej, oznaczono zawartość następujących ośmiu kwasów tłuszczowych: kwas palmitynowy, oleopalmitynowy, stearynowy, oleinowy, linolowy, arachidowy, α-linolenowy, eikozanowy. Następnie, dla uzyskanych chromatogramów utworzono tablicę pików o wymiarowości 572×8.

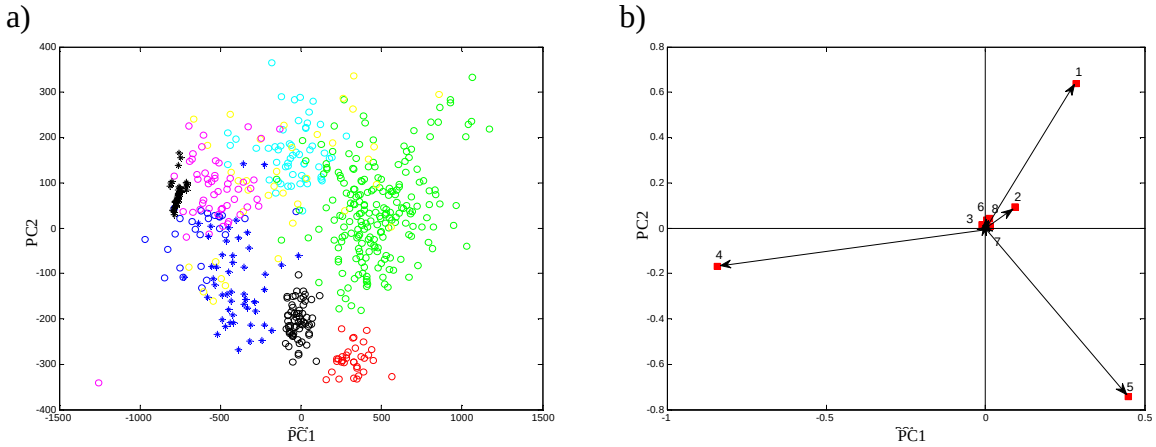
Otrzymaną tablicę pików poddano autoskalowaniu i centrowaniu. Jednak wyniki eksploracji ujawniły, że w rozważanym przypadku metoda centrowania danych wydaje się lepszym rozwiązaniem aniżeli ich autoskalowanie. Po przygotowaniu danych do dalszej analizy, dane poddano eksploracji za pomocą trzech metod: PCA, SMR oraz dwukierunkowego grupowania hierarchicznego. Dendrogramy otrzymano odpowiednio dla próbek, poprzez ich łączenie za pomocą metody Warda oraz odległości euklidesowej, zastosowanej jako miarę oceniającą ich podobieństwo oraz dla parametrów poprzez zastosowanie metody średnich połączeń i współczynnika korelacji jako miary podobieństwa.



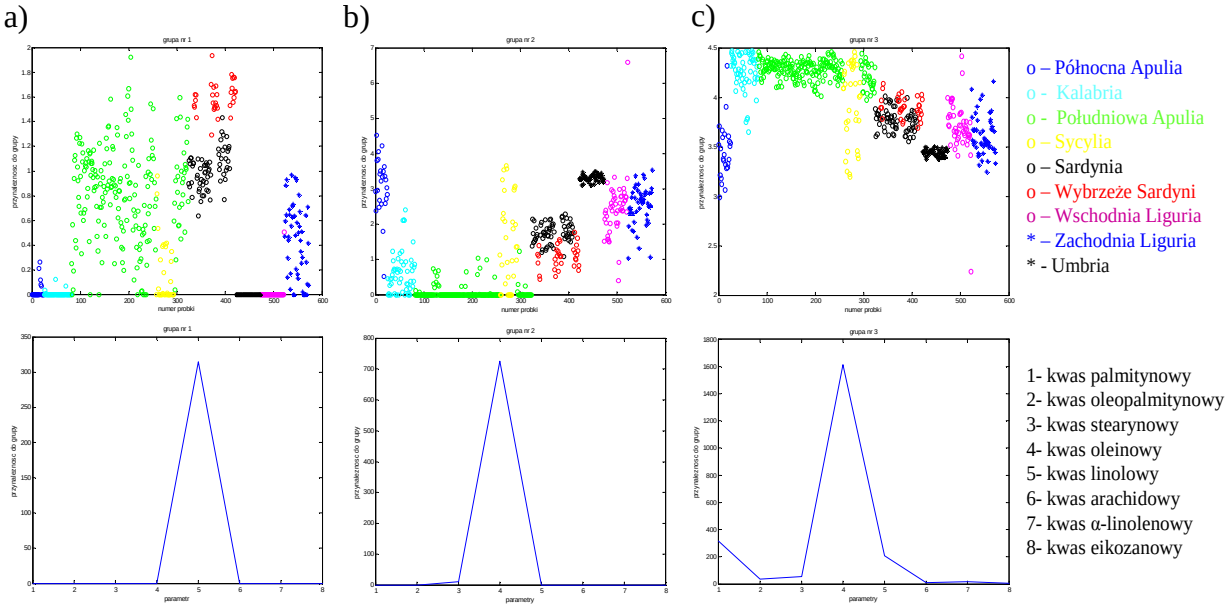
Rys. 47 Mapa Włoch z zaznaczonymi rejonami pobierania próbek.

Wyniki uzyskane za pomocą tychże metod umożliwiły ich porównanie i konfrontację z rzeczywistym pochodzeniem geograficznym próbek. W przypadku metody SMR liczbę grup próbek ustalono każdorazowo jako równą trzy, kierując się geograficznym położeniem regionów, z których pochodziły.
Zastosowanie omawianego podejścia umożliwiło wyodrębnienie parametrów determinujących podział obiektów na grupy i określnie podobieństwa próbek na podstawie zawartych w nich poszczególnych kwasów tłuszczowych. Otrzymane wyniki zilustrowano na Rys. 48, 49 i 50. Interpretacja projekcji wyników w przypadku metody PCA, jak wcześniej wspomniano, jest subiektywna i zależy od osoby interpretującej wyniki. Ponieważ grupy próbek nakładają się na siebie w przestrzeni czynników głównych to nie można jednoznacznie przypisać obiektów do odpowiednich grup.

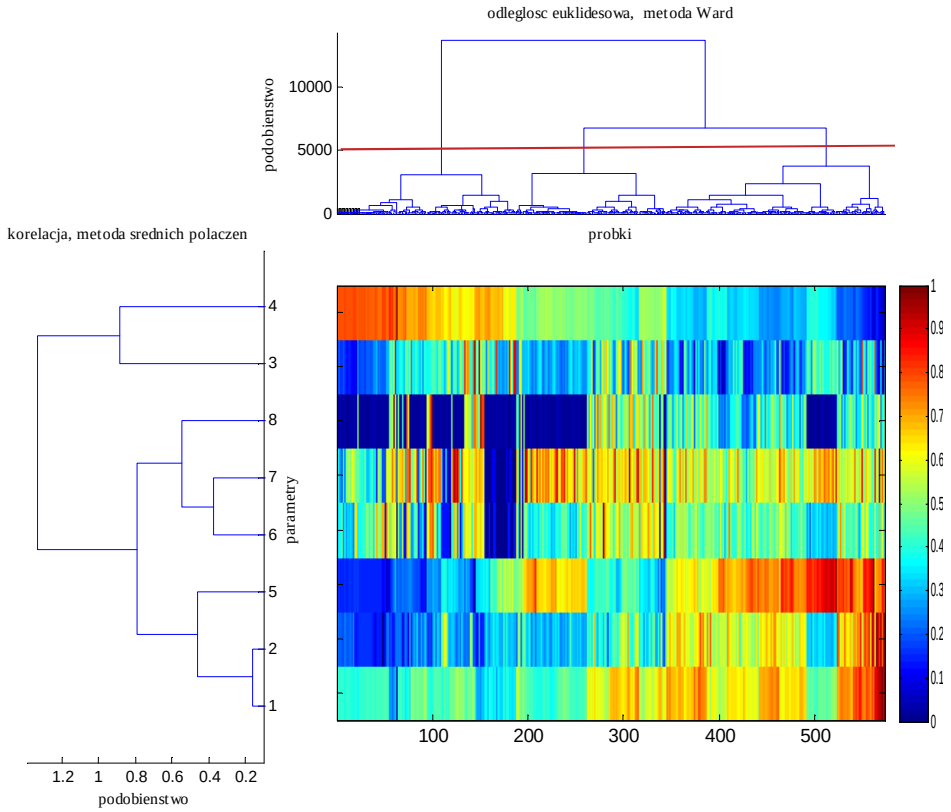
Z kolei w przypadku metody SMR, przynależność obiektów i parametrów do danej grupy odczytuje się z osi rzędnych. Im wartość na osi jest wyższa dla obiektów oraz parametrów tym większe prawdopodobieństwo przynależności do określonej grupy. W metodzie dwukierunkowego grupowania hierarchicznego, struktura danych reprezentowana jest przez dwa dendrogramy skonstruowane osobno dla próbek i osobno dla parametrów. Następnie, w celu ułatwienia interpretacji, dendrogramy wzbogaca się o tzw. kolorową mapę, przedstawiającą wzajemne relacje pomiędzy próbkami i parametrami. Relatywnie wysokie stężenia kwasów tłuszczowych w próbkach reprezentują czerwone wartości, a małą zawartość przedstawiono kolorem ciemnoniebieskim. Stężenia zawarte pomiędzy niskimi i wysokimi wartościami przyjmują barwy pośrednie zgodnie ze skalą kolorów.
Wizualizacja danych za pomocą metody PCA ujawniła, iż dane nie wykazują wyraźnej tendencji do grupowania. Jest to zrozumiałe ponieważ próbki pochodzą z regionów o zbliżonych warunkach klimatycznych (temperatura powietrza, nawodnienie terenu, nasłonecznienie, itp.) wpływających na wzrost i rozwój roślinności. Projekcje parametrów na płaszczyzny zdefiniowane przez pierwszy (PC1) i drugi czynnik (PC2) główny ujawniły, że próbki można pogrupować na podstawie zawartości kwasu oleinowego (4), kwasu linolowego (5) oraz palmitynowego (1) względem PC1 oraz kwasu palmitynowego (1) i kwasu linolowego (5) względem PC2 (Rys. 48). Wynika z tego, że próbki z Północnej Apulii, Zachodniej i Wschodniej Ligurii oraz Umbrii zawierają wyższe stężenia kwasu oleinowego w porównaniu do próbek pochodzących z pozostałych regionów. Co więcej, za pomocą metody SMR oraz dwukierunkowego grupowania hierarchicznego, tak jak w przypadku metody PCA, otrzymano grupy próbek pochodzące z tych samych, regionów o których podobieństwie zadecydował głównie parametr 4. W przypadku metody SMR jest to grupa druga (Rys. 49b), a w przypadku metody dwukierunkowego grupowania hierarchicznego grupa pierwsza (Rys. 50). Kontynuując, projekcja kwasów tłuszczowych z Rys. 48b, ujawniła również że parametr 4 jest przeciwnie skorelowany z parametrami 1 oraz 5. Wskazuje to na niską zawartość kwasów palmitynowego oraz linolowego we wspomnianych próbkach z Ligurii, Umbrii oraz Północnej Apulii w przeciwieństwie do próbek z Wybrzeża Sardynii, Południowej Apulii. Charakteryzują się one relatywnie wysoką zawartością tych kwasów tłuszczowych. Również i w tym przypadku, wyniki pokrywają się z tymi otrzymanymi metodami SMR – grupa pierwsza (Rys. 49a) oraz grupowaniem hierarchicznym w dwóch kierunkach – grupa trzecia (Rys. 50). Reasumując, zastosowane metody eksploracji danych pozwoliły na wyodrębnienie grup próbek różniących się zawartością odpowiednich kwasów tłuszczowych.



Rys. 48 Wyniki eksploracji próbek włoskiej oliwy z oliwek uzyskane za pomocą metody PCA: a) projekcja obiektów na płaszczyznę zdefiniowaną przez pierwsze dwa czynniki główne (PC1 i PC2), b) projekcja parametrów na płaszczyznę zdefiniowaną przez PC1 oraz PC2.



Rys. 49 Wyniki eksploracji danych próbek oliwy z oliwek uzyskanych metodą SMR dla trzech grup. Rysunki górne reprezentują przynależność obiektów do poszczególnych grup, a dolne przynależność parametrów do grup: a) pierwszej, b) drugiej oraz c) trzeciej grupy.

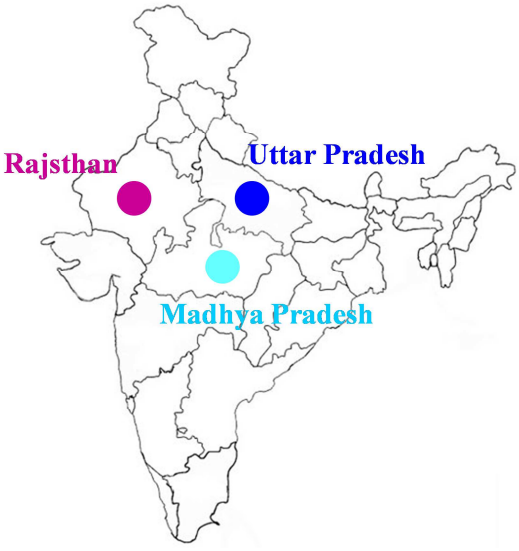


Rys. 50 Wyniki eksploracji uzyskane metodą dwukierunkowego grupowania hierarchicznego wzbogaconego kolorową mapą dla danych opisujących próbki oliwy z oliwek. Dendrogramy dla próbek i parametrów otrzymano odpowiednio poprzez zastosowanie metody Warda i odległości euklidesowej oraz metody średnich połączeń i współczynnika korelacji.

Przykład 11

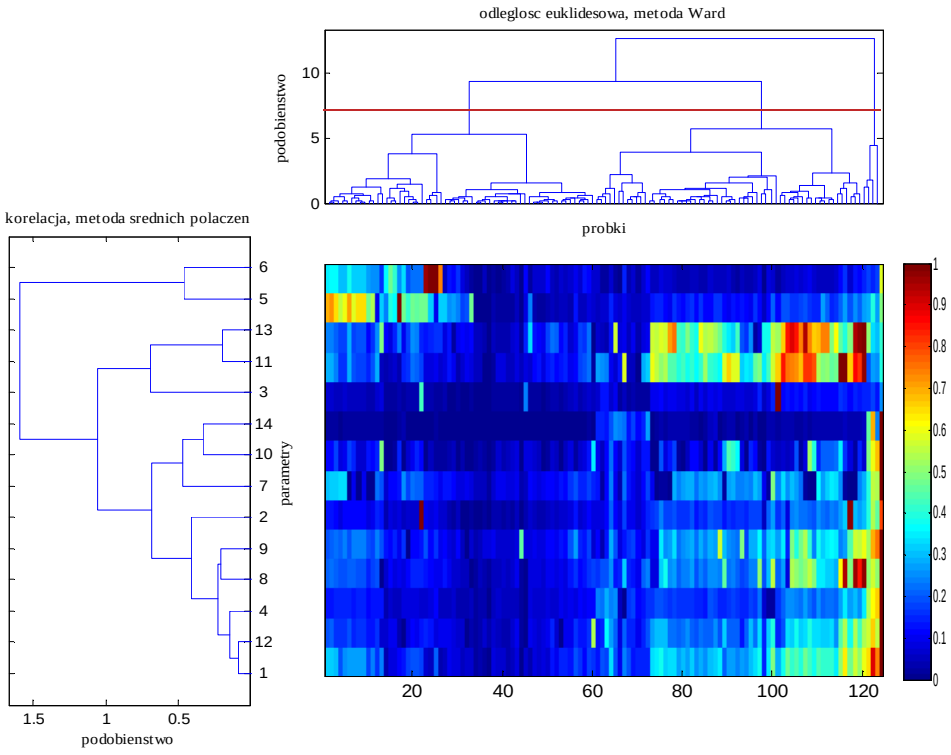
Drugą metodą współgrupowania danych jaką testowano w kontekście eksploracji danych chemicznych była metoda k-spectral. Działanie algorytmu sprawdzono na zestawie danych charakteryzujących próbki opium pobrane z trzech rejonów Indii Rajsthan, Uttar Pradesh oraz Madhya Pradesh (Rys. 51). W próbkach tych, za pomocą chromatografii cieczowej, oznaczono 14 następujących aminokwasów: kwas asparaginowy, treonina, seryna, kwas glutaminowy, glicyna, alanina, walina, izoleucyna, leucyna, tyrozyna, fenyloalanina, histydyna, lizyna oraz arginina.
Z danych chromatograficznych utworzono tablicę pików o wymiarowości 124×14, którą poddano eksploracji za pomocą metody dwukierunkowego grupowania hierarchicznego (Rys. 52) oraz metody k-spectral (Rys. 53). Ze względu na

pochodzenie geograficzne próbek, w metodzie k-spectral liczbę grup na wejściu ustalono jako równą trzy.
Wizualna analiza wyników (Rys. 52 i 53) wykazała że w przypadku metody dwukierunkowego grupowania hierarchicznego próbki również utworzyły trzy grupy. Zestawienie wyników z obu metod ujawniło że pierwszą grupę stanowią próbki z Madhya Pradesh oraz Uttar Pradesh. Grupę drugą tworzą próbki z Rajsthan. Z kolei trzecia grupa została utworzona przez 4 próbki z rejonu Uttar Pradesh. Grupa ta na Rys. 53 stanowi grupę drugą. O ile w przypadku podziału próbek na grupy obserwuje się zgodność wyników otrzymanych zaproponowanymi metodami, o tyle w przypadku parametrów obserwuje się całkowitą rozbieżność. Za utworzenie pierwszej grupy w dwukierunkowym grupowaniu hierarchicznym odpowiadały głównie glicyna (5) oraz alanina (6), a grupy drugiej fenyloalanina (11) i lizyna (13). Trzecia grupa charakteryzuje się relatywnie wysokimi stężeniami argininy (14), tyrozyny (10), waliny (7), treoniny (2), izoleucyny (8), leucyny (9), kwas glutaminowego (4), histydyny (12) oraz kwas asparginowego (1). W przypadku metody k-spectral parametry dominujące, wspierające utworzenie poszczególnych grup to kwas asparginowy (1), fenyloalanina (11) oraz lizyna (13) w przypadku pierwszej grupy. Grupa trzecia wykazuje relatywnie wysokie stężenia kwasu glutaminowego (4) oraz argininy (13). Druga grupa wykazuje wysokie stężenie pozostałych aminokwasów.

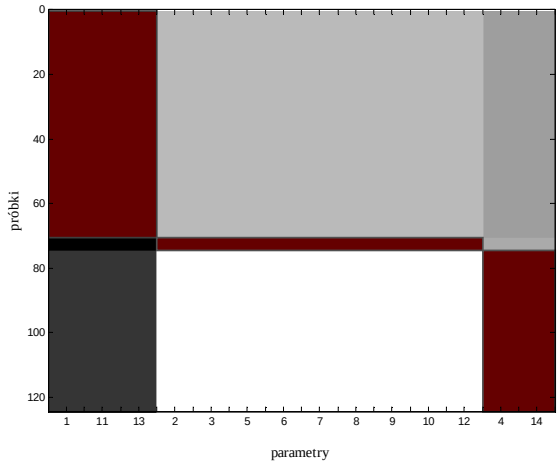


Rys. 51 Mapa Indii z zaznaczonymi rejonami, z których pochodziły próbki opium.

Otrzymane wyniki potwierdzają, że w danych można wyodrębnić trzy grupy próbek jednak ciężko stwierdzić które parametry tak naprawdę determinują obserwowaną tendencję do grupowania.
Przeprowadzona eksploracja z użyciem metody k-spectral ujawniła, że może ona być zastosowana w eksploracji danych chemicznych. Otrzymane w ten sposób wyniki są relatywnie proste w interpretacji, a czas prowadzonych obliczeń jest uzależniony od wymiarowości danych. Jednak w celu weryfikacji otrzymywanych wyników należy zestawić je z wynikami otrzymywanymi za pomocą innych metod eksploracyjnych zaliczanych do metod współgrupowania lub metody PCA i/lub metody dwukierunkowego grupowania hierarchicznego. Podczas doboru metody porównawczej należy uwzględnić fakt że metody współgrupowania ujawniają informację o podgrupach i ich utworzenie może być zdeterminowane przez inne parametry niż utworzenie grup opisanych za pomocą wszystkich zmierzonych zmiennych.



Rys. 52 Wyniki eksploracji uzyskanych metodą dwukierunkowego grupowania hierarchicznego wzbogaconego kolorową mapą dla danych opisujących próbki opium. Dendrogram dla próbek powstał przez ich grupowanie metodą Warda przy zastosowaniu odległości euklidesowej, a dendrogram dla parametrów w wyniku ich grupowania metodą średnich połączeń i uwzględnieniu ich korelacji.



Rys. 53 Wizualizacja wyników współgrupowania danych metodą k-spectral dla k = 3. Podgrupy próbek opium wyodrębnione na podstawie podgrup parametrów oznaczono obszarami w kolorze brązowym.

12.5 Uwzględnienie niepewności pomiarowych w eksploracji danych

Stosując metody grupowania danych w celu eksploracji wielowymiarowych danych, bazuje się na założeniu, iż w danych nie występuje problem niepewności pomiarowych. Uwzględnienie błędów pomiarowych podczas grupowania obiektów jest aktualnie uznawane za nowy kierunek rozwoju algorytmów grupowania. Problem niepewności pomiarowej danych eksperymentalnych towarzyszy wszystkim typom danych ze szczególnym uwzględnieniem danych biologicznych, a zwłaszcza danych otrzymywanych metodą mikromacierzy. Nieuwzględnienie niepewności otrzymanego wyniku może przyczynić się do nieprawidłowego wyodrębniania grup podobnych obiektów, a w konsekwencji błędnej interpretacji wyników oraz wyciągnięcia generalnych wniosków.

Włączenie informacji o niepewności pomiarowej danych podczas ich eksploracji za pomocą metody k-średnich oraz metod grupowania hierarchicznego zostało przedstawione przez M. Kumara i N.T. Patela w [122]. Proponują oni alternatywne podejście grupowania wprowadzając zmodyfikowane algorytmy k-średnich oraz metod hierarchicznych metodą Warda, które nazwano odpowiednio kError i hError, a całą metodologię nazwano grupowaniem danych z uwzględnieniem błędów pomiarowych (z ang. Error-Based Clustering). Z racji na nieadekwatność nazw zaproponowanych algorytmów w polskiej wersji językowej, w niniejszych rozważaniach będą

prezentowane jako algorytm k-średnich uwzględniający niepewności pomiarowe oraz analogicznie metody hierarchiczne uwzględniające niepewności pomiarowe eksplorowanych danych lub za pomocą nazw w angielskiej wersji językowej. Polepszenie otrzymywanych wyników grupowania autorzy otrzymali poprzez uwzględnianie macierzy wariancji-kowariancji reprezentującej niepewność pomiarową dla każdego obiektu. Metoda ta oparta jest na założeniu, że dane są obarczone błędem, który można opisać za pomocą rozkładu Gaussa. Błędy te są modelowane dla każdego obiektu z osobna, a nie dla całego zbioru jak ma to miejsce w metodzie grupowania danych opartym na modelu statystycznym (z ang. Model-Based Clustering) [67]. Dodatkowo, uwzględniono ewentualną korelację błędów odpowiadających poszczególnym obiektom danych. W metodzie grupowania uwzględniającej błędy pomiarowe wykorzystuje się wieloparametrowy rozkład normalny, który opisują takie parametry jak średnia oraz macierz wariancji-kowariancji, **C**, dla każdego obiektu. Można wyróżnić trzy warianty, w których modelowany błąd zależy od formy uwzględnianej macierzy wariancji-kowariancji (Rys. 54). W najprostszej formie, błędy pomiarowe obiektów, reprezentowane są przez rozkłady normalne o danej wartości średniej i macierzy wariancji-kowariancji $\mathbf{C} = \sigma^2\mathbf{I}$. Oznacza to, że wszystkie obiekty są oszacowane z błędem pomiarowym, który przyjmuje sferyczny rozkład o takim samym promieniu. W bardziej skomplikowanym wariancie modelowany błąd wyrażony macierzą wariancji-kowariancji wciąż reprezentuje kształt sferyczny jednak jego zakres wokół obiektu jest inny dla każdego z nich. W ostatnim wariancie modelowany dla obiektu błąd może przyjmować kształt elipsoidalny. Dzięki takiemu podejściu wokół każdego obiektu w przestrzeni eksperymentalnej można wyznaczyć obszar w formie elipsy lub sfery (okręgu), w którym to otrzymany wynik występuje z określonym prawdopodobieństwem. Otwiera to możliwość włączenia niepewności pomiarowych w trakcie grupowania. Celem jest znalezienie grup, w których obiekty mają podobne wartości średnie. Wiąże się to z określeniem niepewności pomiarowej dla każdego obiektu, a następnie dla każdego rozkładu oblicza się średnią i oblicza się odległość pomiędzy nimi (Rys. 55). W kolejnym kroku oblicza się średnią dla całej grupy dzięki czemu możliwe jest wyodrębnienie grup obiektów podobnych.

Matematyczny opis wprowadzonej metody można przedstawić następująco. Dane które mają zostać grupowane zawierają m obiektów od $\mathbf{x}_1$, …, $\mathbf{x}_m$ oraz m pozytywnych macierzy wariancji-kowariancji $\mathbf{C}_i$ reprezentujących błąd obiektów uzyskany dla n parametrów. Wykorzystując założenie, że każdy obiekt pochodzi z n-parametrowego rozkładu Gaussa z jedną z możliwych średnich $\boldsymbol{\theta}_1$,…, $\boldsymbol{\theta}_G$ ze zbioru G, gdzie $G \leq m$, co można wyrazić jako $\mathbf{x}_i \sim N_p(\mathbf{u}_i, \mathbf{C}_i)$, gdzie $\mathbf{u}_i \in \{\boldsymbol{\theta}_1,\ldots, \boldsymbol{\theta}_G\}$, dla i = 1, …, m. Zadaniem algorytmu jest znalezienie grup k_1, …, k_G, dla obiektów o takiej samej średniej, $\mathbf{u}_i$ należących do tych samych grup z $\mathbf{u}_i = \boldsymbol{\theta}_k$.

Niech grupa $S_k = \{i|\mathbf{x}_i \in k_k\}$, dla $\boldsymbol{\mu}_i = \boldsymbol{\theta}_k$ dla $\forall i \in S_k$, k = 1, …, G. Dysponując obiektami $\mathbf{x}_1$,…, $\mathbf{x}_m$ oraz błędami w postaci macierzy wariancji-kowariancji, $\mathbf{C}_1$, …, $\mathbf{C}_m$, maksymalizację prawdopodobieństwa można wyrazić następująco:

$$L(\mathbf{x}_i|S,\boldsymbol{\theta}) = \prod_{k=1}^{G}\prod_{i\epsilon S_k}\frac{1}{(2\pi)^{\frac{p}{2}}}|\mathbf{C}_i|^{-\frac{1}{2}}e^{-\frac{1}{2}(\mathbf{x}_i-\boldsymbol{\theta}_k)^T\mathbf{C}_i^{-1}(\mathbf{x}_i-\boldsymbol{\theta}_k)} \quad (25)$$

gdzie:
$|\mathbf{C}_i|$ – wyznacznik macierzy wariancji-kowariancji $\mathbf{C}_i$ dla i = 1,…, m

Prawdopodobieństwo wyrażone równaniem (25) osiąga maksymalną wartość, wówczas gdy zostanie spełnione wprowadzone kryterium:

$$\min_{S_1,\ldots,S_G}\sum_{k=1}^{G}\sum_{i\epsilon S_k}(\mathbf{x}_i-\widehat{\boldsymbol{\theta}}_k)^T\mathbf{C}_i^{-1}(\mathbf{x}_i-\widehat{\boldsymbol{\theta}}_k) \quad (26)$$

gdzie:
$\widehat{\boldsymbol{\theta}}_k$ – najbardziej wiarygodna średnia dla grupy k, czyli maksymalne oszacowane prawdopodobieństwo wyrażone jako:

$$\widehat{\boldsymbol{\theta}}_k = \left(\sum_{i\epsilon S_k}\mathbf{C}_i^{-1}\right)^{-1}\left(\sum_{i\epsilon S_k}\mathbf{C}_i^{-1}\mathbf{x}_i\right), \text{ k = 1, …, G}$$

Innymi słowy $\widehat{\boldsymbol{\theta}}_k$ jest ważoną średnią obiektów w grupie k_k, czyli tzw. średnią Mahalanobisa dla grupy k_k. jeżeli błąd, czyli kowariancje dla $\widehat{\boldsymbol{\theta}}_k$ oznaczy się jako $\boldsymbol{\psi}_k$ to ostatecznie można ja wyrazić jako:

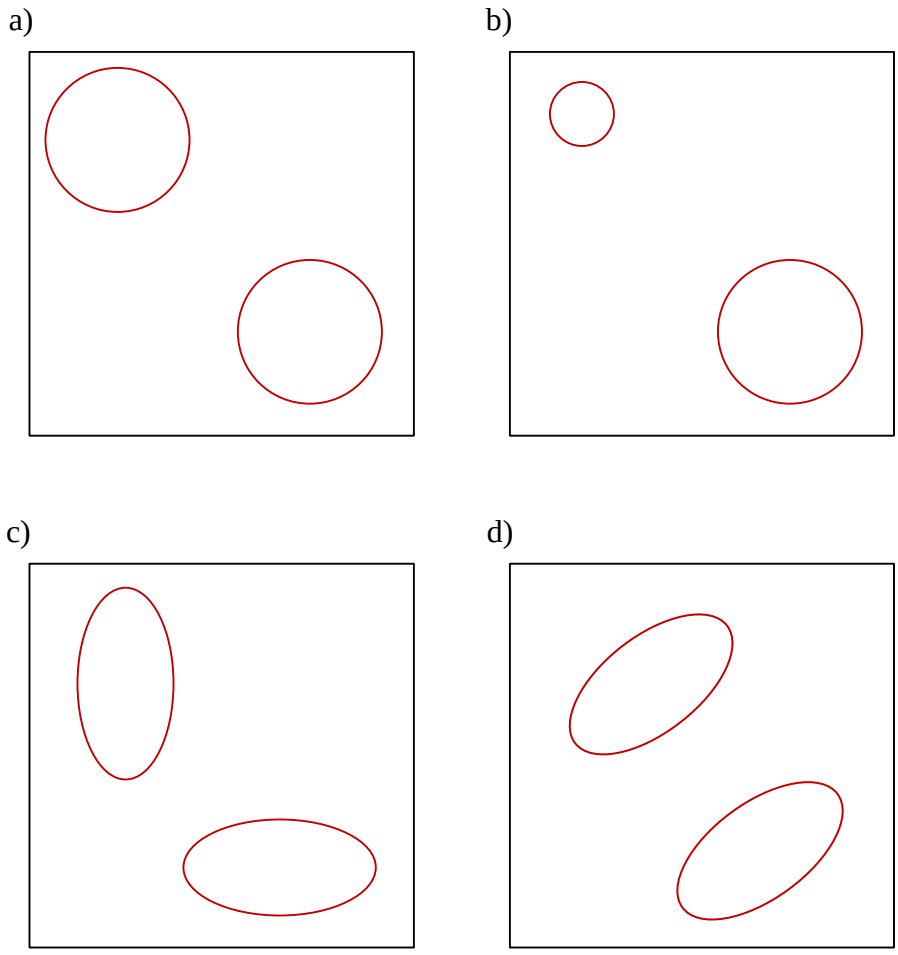
$$\boldsymbol{\psi}_k = \left(\sum_{i\epsilon S_k}\mathbf{C}_i^{-1}\right)^{-1} \quad (27)$$

Z wprowadzonego, na użytek metody grupowania danych z uwzględnieniem błędów pomiarowych równania (26) wynikają dwie właściwości [122]. Po pierwsze, jeżeli przyjmujemy że wszystkie błędy są sferyczne $\mathbf{C}_i = \sigma^2\mathbf{I}$, gdzie $\mathbf{I}$ jest macierzą jednostkową, wówczas wykorzystane w tej metodzie kryterium jest tożsame z kryterium minimalizacji odległości euklidesowej tak jak w przypadku metody k-średnich. W zawiązku z czym metoda ta staje się generalizacją podstawowej miary stosowanej w algorytmie grupowania niehierarchicznego – k-średnich. Po drugie, dzięki zastosowaniu kryterium (26) na wyniki nie wpływają transformacje w przestrzeni eksperymentalnej, ze względu na podobieństwo poszczególnych członów równania do odległości Mahalanobisa.
Na podstawie kryterium wyrażonego równaniem (26), autorzy utworzyli miary podobieństwa wyrażone poniżej za pomocą równań (28) oraz (29), które posłużyły do

grupowania obiektów za pomocą odpowiednio algorytmów Warda oraz k-średnich. Matematyczne uzasadnienie wprowadzonych do algorytmów modyfikacji zostało szczegółowo omówione w [122].

$$d_{ij} = (\hat{\boldsymbol{\theta}}_i - \hat{\boldsymbol{\theta}}_j)^T(\boldsymbol{\psi}_i + \boldsymbol{\psi}_j)(\hat{\boldsymbol{\theta}}_i - \hat{\boldsymbol{\theta}}_j) \tag{28}$$

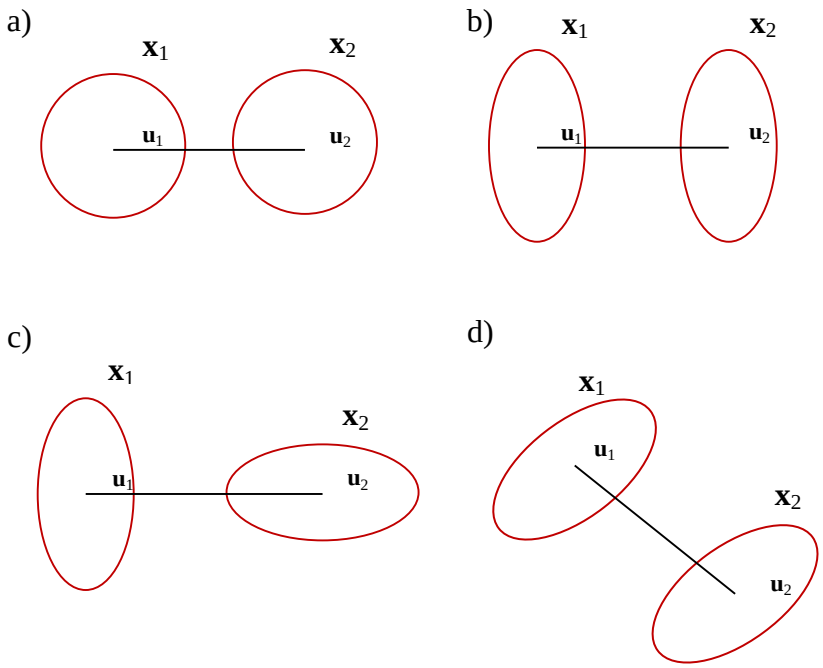
$$d_{ik} = (\mathbf{x}_i - \hat{\boldsymbol{\theta}}_k)^T\mathbf{C}_i^{-1}(\mathbf{x}_i - \hat{\boldsymbol{\theta}}_k) \tag{29}$$



Rys. 54 Przykładowe obszary niepewności pomiarowych dla każdego obiektu charakteryzowane przez różne macierze wariancji-kowariancji a) sferyczne o takim samym promieniu, b) sferyczne o zmiennym promieniu, c) elipsoidalne o różnej orientacji w przestrzeni, d) elipsoidalne o takiej samej orientacji w przestrzeni.

Nowe podejście umożliwiło efektywne grupowanie obiektów, uwzględniające niepewności pomiarowe, dając znacznie lepsze wyniki aniżeli to przeprowadzone bazowymi algorytmami grupowania. W swoim artykule autorzy dowodzą, na przykładzie symulowanych danych, że uwzględnienie niepewności pomiarowych w algorytmach k-średnich i grupowaniu hierarchicznym Warda umożliwia poprawne wyodrębnienie grup obiektów z ok. 80% efektywnością. Szczegółowe porównanie wyników otrzymywanych algorytmami kError, hError, k-średnich oraz metodą Warda, zaprezentowano w [122]. Mając na względzie udowodnioną poprawę w działaniu algorytmów w ramach pracy doktorskiej zaproponowano kolejne modyfikacje metod grupowania poprzez włączenie niepewności pomiarowych w sposób analogiczny do wcześniej proponowanego. Rozszerzenie obejmowałoby wprowadzenie modyfikacji na większą liczbę algorytmów grupowania, np. metod grupowania bazujących na gęstości danych. Zastosowanie nowym miar odległości zawierających informację o błędach pomiarowych w postaci macierzy wariancji-kowariancji np. w algorytmie DBSCAN stało by się nowym narzędziem ich eksploracji i pozwoliło na detekcję grup obiektów o arbitralnych kształtach. Podejście to można również wprowadzić bezpośrednio do algorytmów grupowania niehierarchicznego, których działanie oparte jest na minimalizacji funkcji kosztów E (15), czyli: NP, GNG, czy GK. Inne możliwości wynikające z eksploracji danych z wykorzystaniem miary podobieństwa do konstrukcji tzw. jądra i użycie ich jako dane wejściowe, tak jak ma to miejsce w metodzie wielowymiarowego skalowania danych (z ang. Multidimensional Scaling), a następnie wykorzystanie otrzymanej macierzy do dalszej eksploracji.

Pomimo ciekawych własności, ten sposób grupowania danych ma także ograniczenia. Mają one podłoże natury eksperymentalnej i obliczeniowej. Warunkiem koniecznym do zastosowania opisanej metodologii jest wielokrotny pomiar próbek. Tylko wtedy możliwa jest estymacja błędu pomiarowego wyznaczająca parametry rozkładu normalnego – wartość średnią i macierz wariancji-kowariancji. Natomiast, ograniczeniem proponowanego podejścia jest macierz danych zwierająca skorelowane zmienne. Jeżeli błąd jest skorelowany wówczas niemożliwie jest utworzenie macierzy odwrotnej z macierzy wariancji-kowariancji, a w konsekwencji wyznaczenie niepewności pomiarowej zgodnie z zaprezentowaną ideologią. Jednak, uwzględniając niepewności pomiarowe za pomocą opisanej metody każdy obiekt jest opisany macierzą, w związku z czym każdą macierz można poddać analizie metodami pozwalającymi na eliminację zmiennych zależnych np. PCA.



Rys. 55 Zilustrowanie obliczania odległości pomiędzy rozkładami normalnymi otrzymanymi dla obiektów $\mathbf{x}_1$ oraz $\mathbf{x}_2$ a) dla rozkładów niepewności pomiarowych o kształcie sferycznym o takim samym promieniu, b) elipsy o tej samej orientacji w przestrzeni, c) elipsy o różnej orientacji w przestrzeni eksperymentalnej i d) dwóch elips które są ze sobą skorelowane.

Kody algorytmów hError oraz kError, DBSCAN oraz odległości uwzględniającej niepewności pomiarowe, które zapisano przy użyciu programu Matalab R2009b zaprezentowano w rozdziale Załączniki.

13. Podsumowanie

Metody grupowania danych są cennym narzędziem wykorzystywanym do badań podobieństw dla danych różnorodnego pochodzenia, przez co znajdują zastosowanie w wielu dziedzinach nauki. Coraz częściej korzysta się z ich zalet w eksploracji danych biologicznych, w celu wizualizacji ukrytej struktury danych. Na przykład w badaniach genomicznych ułatwiają wyodrębnienie grup genów wykazujących zbliżoną ekspresję w czasie, w medycynie stosowane są w celu odróżnienia pacjentów chorych od zdrowych, czy wskazania tzw. znaczników biologicznych w postaci metabolitów, odpowiadających za rozwój jednostki chorobowej, co znajduje zastosowanie w naukach typu metabolomika.

Pozyskiwanie danych za pomocą nowoczesnej aparatury badawczej wpływa nie tylko na zasób informacji, ale także na wzrost złożoności pozyskiwanych danych. Ich reprezentacja w postaci macierzy lub tensora wymaga zaawansowanych narzędzi ich analizy. To z kolei wymaga modyfikacji istniejących algorytmów w taki sposób, aby umożliwiały ujawnienie ukrytej struktury danych w jak najkrótszym czasie.

Wprowadzenie innowacyjnych rozwiązań usprawniających działanie algorytmów nie może się odbyć bez uwzględnienia kluczowych obszarów ich zastosowań oraz problemów jakie należy rozważyć w trakcie eksploracji z zastosowaniem tego typu narzędzi. Zestawy danych pozyskiwane w naukach przyrodniczych, chociaż są gromadzone podobnie (bo przy użyciu metod instrumentalnych, zarówno metod spektroskopowych jak i chromatograficznych), charakteryzują się odmienną strukturą i specyfiką. Zupełnie inaczej interpretuje się dane pochodzenia chemicznego niż te pochodzenia biologicznego, nawet jeżeli próbki analizowano z zastosowaniem tej samej metody. W przypadku danych biologicznych grupy obiektów znacznie częściej tworzą tzw. grupy sąsiadujące ze sobą w przestrzeni wymiarowej, a dodatkowo mogą przybierać różnorodne kształty. Z tego względu niezbędne jest narzędzie pozwalające na detekcję grup obiektów o arbitralnych kształtach. Algorytmem takim jest algorytm DBSCAN, który bazując na gęstości danych, pozwala na poprawne ich wyodrębnienie. Ograniczeniem algorytmu jest błędne przypisanie obiektów do grup w przypadku, gdy graniczą one ze sobą w przestrzeni parametrów. Dlatego w ramach niniejszej pracy wprowadzono modyfikację pozwalającą na poprawne przypisanie tzw. obiektów brzegowych do właściwych grup. Wyszukanie w pierwszej kolejności obiektów rdzeniowych, a następnie brzegowych, które zostają przypisane do grupy na podstawie minimalnej odległości pomiędzy nimi, a środkiem utworzonej wcześniej grupy, znacząco poprawiło otrzymywane wyniki. Wspominając o dystansie pomiędzy próbkami należy wyraźnie zaznaczyć, że od doboru miary odległości zależą wyniki przeprowadzonego grupowania. Zupełnie inne efekty będziemy obserwować przy zastosowaniu odległości euklidesowej aniżeli odległości Mahalanobisa. Pomimo przydatności miar odległości w detekcji grup obiektów należy zapoznać się z ich ograniczeniami i obszarami ewentualnych zastosowań. Z powodu ograniczeń jakie niesie za sobą zastosowanie już istniejących miar one również podlegają modyfikacjom. Dlatego w niniejszej pracy na potrzeby m.in. eksploracji

dwuwymiarowych chromatograficznych odcisków palca, zdefiniowano nową miarę odległości. Jej zastosowanie w kontekście omówionych danych otrzymywanych np. za pomocą metody HPLC-DAD, eliminuje etap ich wstępnego przygotowania do dalszej analizy. Szczególnie ważne jest pominięcie czasochłonnego etapu nakładania sygnałów instrumentalnych na siebie. Miara ta pozwala także na detekcje pików substancji w przypadku których pojawia się problem ich koelucji, czy też braku substancji wynikającego z różnicy w składzie analizowanych próbek. Na podstawie wartości nowej miary odległości możliwa jest detekcja wspomnianych problemów tj. przesunięć pików w czasie, czy koelucji substancji oraz wizualizacja wyników za pomocą mapy odpowiedzi.

Nowa miara podobieństwa może zostać wykorzystana w szerszym aspekcie aniżeli eksploracja dwuwymiarowych odcisków palca. Z powodzeniem może zostać zastosowana w analizie danych, które zostały zorganizowane w tensor, np. obrazy hiperspektralne. Pozwala to na wyodrębnienie podobnych obrazów względem zastosowanej długości fali.

Kolejnym ważnym zagadnieniem są niepewności pomiarowe otrzymywanych wyników, które najczęściej w proponowanych dotychczas rozwiązaniach z wykorzystaniem dostępnych algorytmów zostają pominięte. Bazując jednak na zaprezentowanych przez M. Kumara i N.T. Patela nowych miarach podobieństwa uwzględniających niepewność pomiarową każdego obiektu w metodach kError oraz hError, możliwe jest rozwinięcie pozostałych algorytmów otrzymując lepsze wyniki grupowania. W ramach badań prowadzonych w tej pracy doktorskiej zastosowano niepewności pomiarowe w algorytmie DBSCAN polepszając efektywność metody.

Reasumując dotychczasowe rozważania przedstawione w tej pracy doktorskiej można z całą pewnością stwierdzić, że algorytmy grupowania danych są przydatnym narzędziem eksploracji danych różnorodnego pochodzenia, a wprowadzenie modyfikacji, w tym również modyfikacji lub rozwinięcie nowych miar podobieństwa, znacząco poszerza zakres ich zastosowań.

Wprowadzenie nowych miar odległości pozawala na modyfikacje algorytmów grupowania, poprawiając ich efektywność. Dlatego w przyszłości koncepcja nowych miar odległości może zostać wykorzystana w innych metodach. Jedną z możliwości jest wprowadzenie utworzonych macierzy podobieństw do tzw. metod opartych o funkcję jądra np. metoda kernel PCA lub metod modelowania danych takich jak metoda częściowych najmniejszych kwadratów (PLS). Zarówno macierz podobieństw, **K**, otrzymana na podstawie wprowadzonej w tej pracy doktorskiej miary odległości s_{ij}, jak również miary odległości włączające niepewności pomiarowe, mogą posłużyć w efektywniejszej analizie danych. Kolejnym rozszerzeniem zastosowania miar odległości z uwzględnieniem ich niepewności pomiarowych jest fuzja danych. Niejednokrotnie można się spotkać z potrzebą łączenia ze sobą zestawów danych. Na przykład w przypadku badań nad rozwojem jednostki chorobowej, takiej jak stwardnienie rozsiane, badaniom podlegają różne płyny biologiczne np. osocze krwi i płyn mózgowo rdzeniowy. W celu kompleksowej analizy danych i detekcji czynników wpływających na rozwój choroby, dane otrzymane z analizy obu płynów biologicznych łączy się ze sobą. W przypadku tego typu danych z pogranicza biologii

i medycyny poprawienie otrzymywanych wyników można otrzymać poprzez wprowadzenie ideologii niepewności pomiarowej włączonej na etapie fuzji danych. Trzecim wariantem zastosowania niepewności pomiarowych jako narzędzia poprawiającego otrzymywane wyniki jest ich uwzględnienie w trakcie analizy dwuwymiarowych chromatograficznych odcisków palca.

14. Załączniki

```
function d = edist(x1,x2,u1,u2)

% Obliczanie odległości pomiędzy obiektami x1 i x2 z uwzględnieniem ich błędów
% pomiarowych wyrażonych, jako kowariancja u1 i u2. Ad. 1 dla niepewności
% pomiarowej z Rys. 55d oraz ad. 2 dla niepewności pomiarowej zilustrowanej na
% Rys. 55b i c.
% Program utworzony w oparciu o model uwzgledniający niepewności pomiarowe
% [122].
% Funkcja:
% d = edist(x1,x2,u1,u2)
% Parametry wejścia:
% x1 - wektor (1, n) lub macierz (m, n), gdzie m - obiekty i n - zmienne
% x2 - wektor (1,n)
% u1 - macierz (n, n) lub 3-D macierz (n, n, m), będąca macierzą kowariancji dla
% obiektu x1
% u2 - macierz (n, n), reprezentująca kowariancje obiektu x2
```

Ad. 1

```
[m1,n] = size(x1);
m2 = size(x2,1);
iu1 = zeros(n);
iu2 = zeros(n);
b = zeros(n,1);
c = zeros(n,1);

for i = 1:m1

   %sumowanie odwróconej macierzy kowariancji u1
   in = inv(u1(:,:,i));
   iu1 = iu1 + in;
   b = b + in*x1(i,:)';


end

for i = 1:m2

   % sumowanie odwróconej macierzy kowariancji u2
   in = inv(u2(:,:,i));
   iu2 = iu2 + in;
   c = c + in*x2(i,:)';

end

psi1 = inv(iu1);
psi2 = inv(iu2);
```

```
t1 = psi1*(b);
t2 = psi2*(c);
d = sqrt((t1-t2)'*(inv(psi1+psi2))*(t1-t2));
```

Ad. 2

```
function d = edistd(x1,x2,u1,u2)


[m1,n] = size(x1);
m2 = size(x2,1);

iu1 = zeros(1,n);
iu2 = zeros(1,n);
b = 0;
c = 0;

for i = 1:m1

   %sumowanie odwróconej macierzy kowariancji u1
   in = 1./(u1(i,:));
   iu1 = iu1 + in;
   b = b + in.*x1(i,:);
end

for i = 1:m2

      %sumowanie odwróconej macierzy kowariancji u1
   in = 1./(u2(i,:));
   iu2 = iu2 + in;
   c = c + in.*x2(i,:);

end

psi1 = 1./(iu1);
psi2 = 1./(iu2);
t1 = psi1.*(b);
t2 = psi2.*(c);

q = (1./(psi1+psi2));
d = sqrt((t1-t2).^2*q');
```

```
function [class,type]=dbscan (x,u,k,Eps)

% Grupowanie danych za pomocą algorytmu DBSCAN uwzględniającym niepewności
% pomiarowe
% Funkcja
% [class, type]=dbscan(x,u,k,Eps)
% Parametry wejścia:
% x – macierz danych (m, n); m - obiekty, n - parametry
% k – liczba obiektów w sąsiedztwie  (minimalna liczba obiektów tworzących grupę)
% Eps – promień sąsiedztwa, jeżeli jest nieznany można wpisać []
% u – tensor, gdzie każda tablica to macierz kowariancji opisująca błąd dla każdego
% obiektu
% Parametry wyjścia:
% class – wektor opisujący przynależność i-tego obiektu do grupy
% type – wektor charakteryzujący i-ty obiekt (obiekt rdzeniowy: 1, obiekt brzegowy: 0,
    szum: -1)

[m,n]=size(x);
[m,n,p]=size(u);

if nargin<3 | isempty(Eps)

  [Eps]=epsilon(x,k);

end

x=[[1:m]' x];
[m,n]=size(x);
type=zeros(1,m);
no=1;
touched=zeros(m,1);

for i=1:m

   if touched(i)==0;

     ob=x(i,:);

    for j=1:p

     d=edist(ob, x, u(i,:,j), u);
     ind=find(D<=Eps);

     if length(ind)>1 & length(ind)<k+1

       type(i)=0;
       class(i)=0;

     end
```

```
end
    if length(ind)==1

      type(i)=-1;
      class(i)=-1;
      touched(i)=1;

    end

    if length(ind)>=k+1;

      type(i)=1;
      class(ind)=ones(length(ind),1)*max(no);

      while ~isempty(ind)

          ob=x(ind(1),:);
          touched(ind(1))=1;
          ind(1)=[];
               d=edist(ob, x(:,:), u(i,:,j), u);

          i1=find(D<=Eps);

          if length(i1)>1

            class(i1)=no;

            if length(i1)>=k+1;

              type(ob(1))=1;

            else
              type(ob(1))=0;

            end

            for i=1:length(i1)

               if touched(i1(i))==0

                 touched(i1(i))=1;
                 ind=[ind i1(i)];
                 class(i1(i))=no;

               end

            end

          end
```

```
        end
        no=no+1;

      end

  end

end

i1=find(class==0);
class(i1)=-1;
type(i1)=-1;


function [Eps]=epsilon(x,k)

% Obliczanie promienia sąsiedztwa Eps
% Funkcja:
% [Eps]=epsilon(x,k)
% Parametry wejścia:
% x – macierz danych (m, n); m - obiekty, n - zmienne
% k – liczba obiektów w sąsiedztwie


[m,n]=size(x);

Eps=((prod(max(x)-min(x))*k*gamma(.5*n+1))/(m*sqrt(pi.^n))).^(1/n);
```

kError

```
function [class,e,M] = ekmeans(x,u,k,iter)

% Algorytm k-średnich MacQueena [79] [87], uwzględniający niepewności
% pomiarowe, u, dla każdego obiektu danych [122].
% Wiersze u są elementami diagonalnymi macierzy kowariancji dla każdego obiektu x.
% Funkcja:
% [class,e,M] = ekmeans(x,k,iter)
% Parametry wejścia:
% x - macierz (m, n); m-obiekty, n-zamienne
% u - macierz (m n), zawierająca w wierszach elementy diagonalne macierzy
% kowariancji
% dla każdego obiektu
% k - skalar, definujący liczbę grup
% iter - skalar, definujący liczbę iteracji
% Parametry wyjścia:
% class, - wektor (1 ,m), definiujący przynależność obiektów do odpowiednich grup
% e - skalar, zawierający sumę odległści obiektów od środków grup do których należą
% M - macierz (k, n), najbardziej prawdopodobne średnie dla każdej grupy

m = size(x,1);

% Losowy wybór środków grup (wybór k obiektów danych stanowiących środki grup)
p = randperm(m);
p = p(1:k);
M = x(p,:);

% Obliczenie odległości Mahalanobisa od każdego obiektu do najbardziej
% prawdopodobnych środków grup
D = mdist(x,u,M);

% Przypisanie każdego obiektu do najbliższego środka
[i1 class] = min(D);

for j = 1:iter

   % Obliczanie najbardziej prawdopodobnych środków grup dla każdej grupy
   for i = 1:k

      M(i,:) = mlmean(x(class==i,:),u(class==i,:));

   end

    % Obliczenie odległości Mahalanobisa od każdego obiektu do najbardziej
    % prawdopodobnych środków grup
   D = mdist(x,u,M);

    % Przypisanie każdego obiektu do najbliższego środka
   [i1 class] = min(D);
```

```
end

% Obliczenie całkowitego błędu w grupowaniu
e = zeros(k,1);

for i = 1:k

    i1 = find(class==i);
    ind = class(i1(1));
    e(ind) = sum(D(ind,i1));

end

e = sum(e);


function D = mdist(x,u,mlm)

% Obliczanie odległości Mahalanobisa pomiędzy środkami grup (najbardziej
% prawdopodobnymi) a obiektami w x biorąc pod uwagę ich niepewności pomiarowe
% wyrażone jako macierz kowariancji dla każdego obiektu [122].
% Wiersze u są elementami diagonalnymi macierzy kowariancji dla każdego obiektu x
% Funkcja:
% d = mdist(x,u,mlm)
% Parametry wejścia:
% x - macierz (m, n); m- obiekty and n- zmienne
% u - macierz (m, n), zawierająca w wierszach elementy diagonalne macierzy
% kowariancji
% dla każdego obiektu
% mlm - wektor (1, n) lub macierz, najbardziej prawdopodobne średnie dla grup(y)
% Parametry wyjścia:
% D - wektor (1, m) lub macierz, odległści Mahalanobisa pomiędzy obiektami,
% a środkami grup

[m,n] = size(x);
d = zeros(1,m);

for j = 1:size(mlm,1)

    C = ones(m,1)*(1./(u(j,:)));
    t = x-ones(m,1)*mlm(j,:);
    d = sum((t.^2).*C,2);
    D(j,:) = d;

end

d = sqrt(d);
```

hError

```
function Z = herror5(x,u)

% Grupowanie hierarchiczne uwzględniające błąd pomiarowy, u, dla każdego obiektu
% [122].
% Funkcja:
% Z = herror5(x,u);
% Parametry wejścia:
% x - macierz (m, n), gdzie m- obiekty and n- zmienne
% u - macierz (m, n), zawierająca błędy dla każdego obiektu (std.^2)
% Parametry wyjścia:
% Z – macierz połączeń obiektów, tak jak w 'linkage' routine w Statistical Toolbox
% (wykorzystanie opcji rysowania dendrogramów )
% Informacja o grupach zawarta jest w macierzy Z m-1 do 3,
% gdzie m stanowi liczbę obserwacji w oryginalnych danych. Kolumna 1i 2 z macierzy
% Z zawierają indeksy (spis) grup łączonych w pary w celu utworzenia binarnego
% drzewa.
% Poszczególne liście (odgałęzienia) drzewa są numerowane od 1 do m. Każdy liść
% stanowi jedną grupę na której budowane są kolejne grupy.
% Każda nowo wprowadzona grupa odpowiada Z(i,:), jest zapisana z indeksem m+i,
% gdzie m jest początkową liczbą liści w drzewie.
% Z(i,1:2) zawiera indeksy dwóch komponentów grup tworzących grupę m+i. Oprócz
% tego
% występuje, n-1 kolejnych grup odpowiadających węzłom wyjściowego drzewa.
% Z(i,3) odpowiada połączeniom odległości pomiędzy dwoma grupami, które łączy się
% w Z(i,:), jeśli całkowita liczba początkowych węzłów wynosi 30 i w kroku 12, grupa
% 5i 7 są łączone za sobą i odległość pomiędzy nimi w tym czasie wynosi 1.5, to 12
% wiersz macierzy Z będzie opisany jako (5,7,1.5). A nowo utworzona grupa będzie
% miała indeks 12+30=42. Jeśli grupa 42 pojawia się w kolejnych wierszach to
% zostanie ona ponownie połączona w większą grupę.

[m] = size(x,1);
mlm = x;
psi = u;

% Inicjalizacja grup w formie komórek:
for i = 1:m

   temp{i} = i;

end

% Obliczenie indeksów odpowiadających najwyższym sekcjom w macierzy
% podobieństw
[I,J] = genindex(m);

d = [];

for i = 1:m
```

```
    i1 = (I==i);
    d = [d; paireddist(x(I(i1),:),x(J(i1),:),u(I(i1),:),u(J(i1),:))];

end

step = 1;

% Identyfikacja 'najlepszych' par obiektów znajdujących się w dwóch komórkach.
[dop i2] = min(d);
linked = sort([I(i2) J(i2)]);
d(I==I(i2) | I==J(i2) | J==I(i2) | J==J(i2)) = [];

[i1,i2] = mlmean(x(linked,:),u(linked,:));

% Aktualizacja informacji o mlm i psi, zastępująca poprzednią
[mlm,psi] = doupdate(mlm,psi,linked,i1,i2);

% Dodanie obiektów do macierzy połączeń (macierz Z)
Z(step,:) = [linked dop];

% Połączenie dwóch komórek zawierających indeksy 'najlepszych' obiektów w jedną,
% jako c{1}
c{1} = [temp{linked}];

% Usunięcie tych dwóch komórek ze zbioru reprezentującego grupy
temp(linked) = [];

% Dodanie nowych komórek do zbioru grup w pierwszej pozycji
cluster = [c temp];

% Budowanie zbioru komórek zawierającego grupy skonstruuowane w i-tym etapie
joined = c;
ww = waitbar(0,'Please wait...');


while step < m-1

    % Wyczyszczenie zmiennych temp
    temp = cluster;

    % Obliczanie indeksów par odległości
    [I,J] = genindex(m-step);

    le1 = m-step-1;

    q = 2*(1./(ones(le1,1)*psi(1,:)+psi(2:end,:)));
    d = [sum(((ones(le1,1)*mlm(1,:)-mlm(2:end,:)).^2).*q,2);d];

    [dop i2] = min(d);
```

```
    linked = sort([I(i2) J(i2)]);

    d(I==I(i2) | I==J(i2) | J==I(i2) | J==J(i2)) = [];

    c{1} = [temp{linked}];
    cl1 = temp{linked(1)};
    cl2 = temp{linked(2)};

    temp(linked) = [];
    cluster = [c temp];
    joined = [joined c];

    % Aktualizacja mlm and psi dla nowej grupy
    [i1,i2] = mlmean(x(c{1},:),u(c{1},:));
    [mlm,psi] = doupdate(mlm,psi,linked,i1,i2);

    le1 = length(cl1);
    le2 = length(cl2);

    % Łączenie pojedynczych obiektów
    if (le1==1 && le2==1)

       Z(step+1,:) = [cl1 cl2 dop];

        % Przyłączenie pojedynczych obiektów do grupy cl1
         elseif (le1>1 && le2==1)

        % Poszukiwanie etapu na którym grupa z elementami cl1 została zbudowana
         w = identifystep(joined,cl1,le1);
        Z(step+1,:) = [cl2 w+m dop];

        % Połączenie dwóch dużych grup zawierających więcej niż jeden element
         elseif (le1>1 && le2>1)

           % Znalezienie etapu na którym została zainicjalizowana budowa dwóch grup
          % z elemntami cl1 I cl2
           w1 = identifystep(joined,cl1,le1);
            w2 = identifystep(joined,cl2,le2);
           Z(step+1,:) = [w1+m w2+m dop];

    end

    step = step+1;
    waitbar(step/(m-1),ww);

end

Z(:,3) = sqrt(Z(:,3));

close(ww)
```

```
function [mlm,psi] = doupdate(mlm,psi,linked,i1,i2)

% Aktualizacja informacji o mlm i psi tak aby nowa aktualizacja zastąpiła poprzednią
mlm(linked,:) = [];
psi(linked,:) = [];
mlm = [i1;mlm];
psi = [i2;psi];


function [I,J] = genindex(m)

% Utworzenie pary indeksów odpowiadających sekcji w symetrycznej macierzy (m, m)
% Parametry wejścia:
% m - skalar, definujący liczbę wierszy I kolumn w symetrycznej macierzy
% Parametry wyjscia:
% I, J - wektory, indeksy definujące pary obiektów, które należy oszacować

pp = (m-1):-1:2;
I = zeros(m*(m-1)/2,1);
I(cumsum([1 pp])) = 1;
I = cumsum(I);
J = ones(m*(m-1)/2,1);
J(cumsum(pp)+1) = 2-pp;
J(1) = 2;
J = cumsum(J);


function w = identifystep(joined,cl,le)

% Identyfikacja początkowego etapu grupowania elementów cl1
% Parametry wejścia:
% joined – komórki tensora, zawierające grupy uporządkowane w kolejności ich
% tworzenia
% cl - wektor, zawierający indeksy obiektów tworzących grupy
% le - długość
% Parametry wyjścia:
% w - skalar, indeks etapów, w których grupa cl1 została utworzona

for k = 1:length(joined)

   if length(joined{k}) == le

      if sum((cl-joined{k})==0)==le

         %sum(ismember(cl1,joined{k}))==le;
         test(k) = 1;

      else
         test(k) = 0;
```

```
        end
    end
end

w = find(test==1);


function d = paireddist(x1,x2,u1,u2)

% Obliczanie odległości pomiędzy obiektami x1 i x2 uwzględniająca niepewności
% pomiarowe (u1 i u2) wyrażonej odpowiednio jako kowariancja obiektu x1 i x2,
% przy założeniu że u1 i u2 są diagonalnymi elementami macierzy kowariancji [122].
% A także, że (U1) = diag([1/u11 1/u22])
% Funkcja:
% d = paireddist(x1,x2,u1,u2)
% Parametry wejścia:
% x1- macierz (m1, n); m- obiekty i n- zmienne
% x2 - macierz (m2, n)
% u1 - macierz (m1, n), błędy pomiarowe dla obiektu x1
% u2 - macierz (m2, n), błąd dla obiektu x2

% q = (1./(u1+u2));
q = 2*(1./(u1+u2));
% połączenie odległości euklidesowej z kowariancją
d = (sum((x1-x2).^2.*q,2));


function [mlm,psi] = mlmean(x,u)

% Obliczanie najbardziej prawdopodobnej średniej z uwzględnieniem błędu
% pomiarowego każdego obiektu wyrażony jako macierz kowariancji, u [122].
% Funkcja:
% mlm = mlmean(x,u)
% Parametry wejścia:
% x - macierz (m, n); m- obiekty i n- zmienne
% u - macierz (m, n), macierz kowariancji dla każdego obiektu
% Parametry wyjścia:
% mlm - wektor (1, n), najbardziej prawdopodobna średnia

p = 1./u;
psi = 1./(sum(p));
mlm = psi.*sum(p.*x);
```

15. Bibliografia

[1] J. T. Liu, R. H. Liu, Enantiomeric composition of abused amine drugs: chromatographic methods of analysis and data interpretation, J. Biochem. Biophys. Methods. 54 (2002), 115–146.

[2] T. R. Sharp, B. L. Marquez, Hyphenated methods, w: S. A. i N. Jespersen (Ed.), Compr. Anal. Chem., Elsevier, (2006), 691–754.

[3] S. Wold, Chemometrics; What do we mean with it, and what do we want from it?, Chemom. Intell. Lab. Syst. 30 (1995), 109–115.

[4] L. Kaufman, P. J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis, John Wiley & Sons, (2009).

[5] D. Yan, J. Wang, Biclustering of gene expression data based on related genes and conditions extraction, Pattern Recognit. 46 (2013), 1170–1182.

[6] M. W. Beck, B. Vondracek, L. K. Hatch, Environmental clustering of lakes to evaluate performance of a macrophyte index of biotic integrity, Aquat. Bot. 108 (2013), 16–25.

[7] A. L. Richards, P. Holmans, M. C. O'Donovan, M. J. Owen, L. Jones, A comparison of four clustering methods for brain expression microarray data, BMC Bioinformatics 9:490 (2008), 1–17.

[8] J. Mazina, M. Vaher, M. Kuhtinskaja, L. Poryvkina, M. Kaljurand, Fluorescence, electrophoretic and chromatographic fingerprints of herbal medicines and their comparative chemometric analysis, Talanta 139 (2015), 233–246.

[9] J. Orzel, M. Daszykowski, M. Kazura, D. de Beer, E. Joubert, A. E. Schulze, et al., Modeling of the total antioxidant capacity of rooibos (Aspalathus linearis) tea infusions from chromatographic fingerprints and identification of potential antioxidant markers, J. Chromatogr. A. 1366 (2014), 101–109.

[10] B. Krakowska, I. Stanimirova, J. Orzel, M. Daszykowski, I. Grabowski, G. Zaleszczyk, et al., Detection of discoloration in diesel fuel based on gas chromatographic fingerprints, Anal. Bioanal. Chem. 407 (2014), 1159–1170.

[11] A. Cygański, Chemiczne metody analizy ilościowej, Wydawnictwo Naukowo-Techniczne, Warszawa (1994).

[12] M. da la Guardia, S. Garrigues, Challenges in Green Analytical Chemistry, Royal Society of Chemistry (2011).

[13] R.A. Nadkarni, Modern Instrumental Methods of Elemental Analysis of Petroleum Products and Lubricants, ASTM International (1991).

[11] J. Solé, L. Bausa, D. Jaque, An Introduction to the Optical Spectroscopy of Inorganic Solids, John Wiley & Sons (2005).

[15] W. Szczepaniak, Metody instrumentalne w analizie chemicznej, Państwowe Wydawnictwo Naukowe, Warszawa/Poznań (1979).

[16] S. Kawakubo, A. Naito, A. Fujihara, M. Iwatsuki, Field Determination of Trace Iron in Fresh Water Samples by Visual and Spectrophotometric Methods, Anal. Sci. 20 (2004), 1159–1163.

[17] F. Perveen, R. Qureshi, F. L. Ansari, S. Kalsoom, S. Ahmed, Investigations of drug–DNA interactions using molecular docking, Cyclic voltammetry and UV–Vis spectroscopy, J. Mol. Struct. 1004 (2011), 67–73.
[18] N. Li, Y. Ma, C. Yang, L. Guo, X. Yang, Interaction of anticancer drug mitoxantrone with DNA analyzed by electrochemical and spectroscopic methods, Biophys. Chem. 116 (2005), 199–205.
[19] K. M. Elkins, Determination of DNA Quality and Quantity Using UV-Vis Spectroscopy, w: K.M. Elkins (Ed.), Forensic DNA Biol., Academic Press, San Diego (2013), 59–62.
[20] Section 4.3: Ultraviolet and visible spectroscopy, Chemwiki.ucdavis.edu. (n.d.). http://chemwiki.ucdavis.edu/Organic_Chemistry/Organic_Chemistry_With_a_Biological_Emphasis/Chapter_04%3A_Structure_Determination_I/Section_4.3%3A_Ultraviolet_and_visible_spectroscopy.
[21] W. R. Melchert, F. R. P. Rocha, A green analytical procedure for flow-injection determination of nitrate in natural waters, Talanta 65 (2005), 461–465.
[22] M. Nagai, M. Sugiyama, T. Hori, Sensitive spectrophotometric determination of phosphate using silica-gel collectors, Anal. Sci. Int. J. Jpn. Soc. Anal. Chem. 20 (2004), 341–344.
[23] G. McMahon, Analytical Instrumentation: A Guide to Laboratory, Portable and Miniaturized Instruments, John Wiley & Sons (2008).
[24] P. Stepnowski, E. Synak, B. Szafranek, Techniki separacyjne, Uniwersytet Gdański, Gdańsk (2010).
[25] A. Braithwaite, J. F. Smith, Chromatographic Methods, 5th Edition, Kluwer Academic Publisher, Dordrecht/Boston/London (1995).
[26] H. M. McNair, J. M. Miller, Basic Gas Chromatography, John Wiley & Sons (2011).
[27] R. P. W. Scott, Liquid Chromatography for the Analyst, CRC Press (1994).
[28] V. R. Meyer, Practical High-Performance Liquid Chromatography, John Wiley & Sons (2004).
[29] F. J. Couper, O. H. Drummer, Gas chromatographic-mass spectrometric determination of ß2-agonists in postmortem blood: application in forensic medicine, J. Chromatogr. B. Biomed. Sci. App. 685 (1996), 265–272.
[30] E. Tanaka, M. Terada, T. Nakamura, S. Misawa, C. Wakasugi, Forensic analysis of eleven cyclic antidepressants in human biological samples using a new reversed-phase chromatographic column of 2 μm porous microspherical silica gel, J. Chromatogr. B. Biomed. Sci. App. 692 (1997), 405–412.
[31] P. Gimeno, A. F. Maggio, C. Bousquet, A. Quoirez, C. Civade, P. A. Bonnet, Analytical method for the identification and assay of 12 phthalates in cosmetic products: Application of the ISO 12787 international standard “Cosmetics–Analytical methods–Validation criteria for analytical results using chromatographic techniques,” J. Chromatogr. A. 1253 (2012), 144–153.
[32] S. M. Karakartal, S. F. Aygün, A. N. Onar, Gas chromatographic separation of PCB and OCP by photocatalytic degradation, Anal. Chim. Acta. 547 (2005), 89–93.

[33] R. Rial-Otero, E. M. Gaspar, I. Moura, J. L. Capelo, Chromatographic-based methods for pesticide determination in honey: An overview, Talanta 71 (2007), 503–514.
[34] Q. Wu, H. Shi, Y. Ma, C. Adams, T. Eichholz, T. Timmons, et al., Determination of secondary and tertiary amines as N-nitrosamine precursors in drinking water system using ultra-fast liquid chromatography–tandem mass spectrometry, Talanta 131 (2015), 736–741.
[35] M. Stanislawska, B. Janasik, W. Wasowicz, Application of high performance liquid chromatography with inductively coupled plasma mass spectrometry (HPLC–ICP-MS) for determination of chromium compounds in the air at the workplace, Talanta 117 (2013), 14–19.
[36] A. Matlack, Introduction to Green Chemistry, CRC Press (2001).
[37] X. Zhang, A. Fang, C. P. Riley, M. Wang, F. E. Regnier, C. Buck, Multi-dimensional liquid chromatography in proteomics-A review, Anal. Chim. Acta. 664 (2010), 101–113.
[38] K. Inoue, H. Tsuchiya, T. Takayama, H. Akatsu, Y. Hashizume, T. Yamamoto, et al., Blood-based diagnosis of Alzheimer's disease using fingerprinting metabolomics based on hydrophilic interaction liquid chromatography with mass spectrometry and multivariate statistical analysis, J. Chromatogr. B. 974 (2015), 24–34.
[39] P. Miralles, A. Chisvert, A. Salvador, Determination of hydroxytyrosol and tyrosol by liquid chromatography for the quality control of cosmetic products based on olive extracts, J. Pharm. Biomed. Anal. 102 (2015), 157–161.
[40] P. Viñas, M. Bravo-Bravo, I. López-García, M. Hernández-Córdoba, Dispersive liquid–liquid microextraction for the determination of vitamins D and K in foods by liquid chromatography with diode-array and atmospheric pressure chemical ionization-mass spectrometry detection, Talanta 115 (2013), 806–813.
[41] E. Heftmann, Chromatography: Fundamentals and applications of chromatography and related differential migration methods-Part B: Applications, Elsevier (2004).
[42] O. D. Sparkman, Z. Penton, F.G. Kitson, Gas Chromatography and Mass Spectrometry: A Practical Guide: A Practical Guide, Academic Press (2011).
[43] M. V. S. Elipe, LC-NMR and Other Hyphenated NMR Techniques: Overview and Applications, John Wiley & Sons (2011).
[44] M. McMaster, LC/MS: A Practical User's Guide, John Wiley & Sons (2005).
[45] D. Orčić, M. Francišković, K. Bekvalac, E. Svirčev, I. Beara, M. Lesjak, et al., Quantitative determination of plant phenolics in Urtica dioica extracts by high-performance liquid chromatography coupled with tandem mass spectrometric detection, Food Chem. 143 (2014), 48–53.
[46] J. L. Cao, J. C. Wei, M. W. Chen, H. X. Su, J. B. Wan, Y. T. Wang, et al., Application of two-dimensional chromatography in the analysis of Chinese herbal medicines, J. Chromatogr. A. 1371 (2014), 1–14.

[47] S. Kinani, S. Layousse, B. Richard, A. Kinani, S. Bouchonnet, A. Thoma, et al., Selective and trace determination of monochloramine in river water by chemical derivatization and liquid chromatography/tandem mass spectrometry analysis, Talanta 140 (2015), 189–197.

[48] E. Shokry, F. Villanelli, S. Malvagia, A. Rosati, G. Forni, S. Funghini, et al., Therapeutic drug monitoring of carbamazepine and its metabolite in children from dried blood spots using liquid chromatography and tandem mass spectrometry, J. Pharm. Biomed. Anal. 109 (2015), 164–170.

[49] R. Jaiswal, E.A. Halabi, M.G.E. Karar, N. Kuhnert, Identification and characterisation of the phenolics of Ilex glabra L. Gray (Aquifoliaceae) leaves by liquid chromatography tandem mass spectrometry, Phytochemistry. 106 (2014), 141–155.

[50] Y. Güzel, E. Aktoklu, V. Roumy, R. Alkhatib, T. Hennebelle, F. Bailleul, et al., Chemotaxonomy and flavonoid profiling of Torilis species by HPLC/ESI/MS2, Biochem. Syst. Ecol. 39 (2011), 781–786.

[51] M. Calderón-Santiago, F. Priego-Capote, B. Jurado-Gámez, M.D. Luque de Castro, Optimization study for metabolomics analysis of human sweat by liquid chromatography–tandem mass spectrometry in high resolution mode, J. Chromatogr. A. 1333 (2014), 70–78.

[52] J. Engel, J. Gerretzen, E. Szymańska, J. J. Jansen, G. Downey, L. Blanchet, et al., Breaking with trends in pre-processing?, TrAC Trends Anal. Chem. 50 (2013), 96–106.

[53] R. F. B. V. J. Barclay, Application of Wavelet Transforms to Experimental Spectra: Smoothing, Denoising, and Data Set Compression, Anal. Chem. 69 (1996).

[54] P.H.C. Eilers, A perfect smoother, Anal. Chem. 75 (2003), 3631–3636.

[55] S. Krishnan, J. T. W. E. Vogels, L. Coulier, R. C. Bas, M. W. B. Hendriks, T. Hankemeier, et al., Instrument and process independent binning and baseline correction methods for liquid chromatography–high resolution-mass spectrometry deconvolution, Anal. Chim. Acta. 740 (2012), 12–19.

[56] B. Walczak, D.L. Massart, Wavelets — something for analytical chemistry?, TrAC Trends Anal. Chem. 16 (1997), 451–463.

[57] Y. Kim, K.A. Schug, S.B. Kim, An ensemble regularization method for feature selection in mass spectral fingerprints, Chemom. Intell. Lab. Syst. 146 (2015), 322–328.

[58] T. G. Bloemberg, J. Gerretzen, A. Lunshof, R. Wehrens, L. M. C. Buydens, Warping methods for spectroscopic and chromatographic signal alignment: A tutorial, Anal. Chim. Acta. 781 (2013), 14–32.

[59] M. Daszykowski, B. Walczak, Analiza czynników głównych i inne metody eksploracji danych, in: Chemom. W Anal. Wybrane Zagadnienia, Wydawnictwo Instytutu Ekspertyz Sądowych (2008).

[60] J. Trygg, J. Gabrielsson, T. Lundstedt, 2.01 - Background Estimation, Denoising, and Preprocessing, in: S.D.B.T. Walczak (Ed.), Compr. Chemom., Elsevier, Oxford, 2009, 1–8.

[61] M. M. Issa, R. M. Nejem, R. I. S. Van Staden, H. Y. Aboul-Enein, New approach application of data transformation in mean centering of ratio spectra method, Spectrochim. Acta. A. Mol. Biomol. Spectrosc. 142 (2015), 204–209.
[62] B. Walczak, ed., Wavelets in Chemistry, 1 edition, Elsevier Science, Amsterdam/ New York (2000).
[63] M. Daszykowski, B. Walczak, Target selection for alignment of chromatographic signals obtained using monochannel detectors, J. Chromatogr. A. 1176 (2007), 1–11.
[64] R. H. Jellema, Variable Shift and Alignment, w: S. D. Brown, R. Tauler, B. Walczak (Ed.), Compr. Chemom., Elsevier (2009), 85–108.
[65] A. M. van Nederkassel, M. Daszykowski, P. H. C. Eilers, Y. V. Heyden, A comparison of three algorithms for chromatograms alignment, J. Chromatogr. A. 1118 (2006), 199–210.
[66] R. Todeschini, D. Ballabio, V. Consonni, Distances and Other Dissimilarity Measures in Chemometrics, w: Encycl. Anal. Chem., John Wiley & Sons (2006).
[67] K. Varmuza, P. Filzmoser, Introduction to Multivariate Statistical Analysis in Chemometrics, 1 edition, CRC Press, Boca Raton (2009).
[68] D. Zuba, A. Parczewski, Chemometria w analityce, Wydawnictwo – Instytut Ekspertyz Sądowych (2015).
[69] R. De Maesschalck, D. Jouan-Rimbaud, D. L. Massart, The Mahalanobis distance, Chemom. Intell. Lab. Syst. 50 (2000), 1–18.
[70] I. Stanimirova, M. Daszykowski, B. Walczak, Metody uczenia z nadzorem - kalibracja, dyskryminacja i klasyfikacja, w: Chemom. w Anal. Wybrane Zagadnienia, Wydawnictwo Instytutu Ekspertyz Sądowych (2008).
[71] M. Daszykowski, B. Walczak, D.L. Massart, Projection methods in chemistry, Chemom. Intell. Lab. Syst. 65 (2003), 97–112.
[72] N. Bratchell, Cluster analysis, Chemom. Intell. Lab. Syst. 6 (1989), 105–125.
[73] S. N. Roy, On a Heuristic Method of Test Construction and its use in Multivariate Analysis, Ann. Math. Stat. 24 (1953), 220–238.
[74] P. J. Huber, Projection Pursuit, Ann. Stat. 13 (1985), 435–475.
[75] M. C. Jones, R. Sibson, What is Projection Pursuit?, J. R. Stat. Soc. Ser. Gen. 150 (1987) 1–37.
[76] M. Daszykowski, From projection pursuit to other unsupervised chemometric techniques, J. Chemom. 21 (2007), 270–279.
[77] D. L. Massart, Y. Vander Heyden, From tables to visuals: principal component analysis, part 1, LC-GC Eur. (2004), 586–591.
[78] S. Wold, K. Esbensen, P. Geladi, Principal component analysis, Chemom. Intell. Lab. Syst. 2 (1987), 37–52.
[79] D. L. Massart, L. Kaufman, The interpretation of analytical chemical data by the use of cluster analysis, John Wiley & Sons, New York (1983).
[80] B. S. Everitt, S. Landau, M. Leese, et al, Cluster Analysis, 5th Edition, John Wiley & Sons, London (2012).

[81] I. Lee, J. Yang, 2.27 - Common Clustering Algorithms, w: S. D. Brown, R. Tauler, B. Walczak (Ed.), Compr. Chemom., Elsevier, Oxford, (2009), 577–618.
[82] J. C. Bezdek, R. Ehrlich, W. Full, FCM: The fuzzy c-means clustering algorithm, Comput. Geosci. 10 (1984), 191–203.
[83] C. H. Hung, H. M. Chiou, W. N. Yang, Candidate groups search for K-harmonic means data clustering, Appl. Math. Model. 37 (2013), 10123–10128.
[84] B. G. M. Vandeginste, D. L. Massart, L. M. C. Buydens, S. de Jong, P.J. Lewi, J. Smeyers-Verbeke, Handbook of Chemometrics and Qualimetrics, Elsevier (1997).
[85] A. Smoliński, B. Walczak, J.W. Einax, Hierarchical clustering extended with visual complements of environmental data set, Chemom. Intell. Lab. Syst. 64 (2002), 45–54.
[86] A. K. Jain, Data clustering: 50 years beyond K-means, Pattern Recognit. Lett. 31 (2010), 651–666.
[87] J. B. MacQueen, Some Methods for classification and Analysis of Multivariate Observations, w: Berkeley, University of California Press (1967), 281–297.
[88] T. M. Martinez, S. G. Berkovich, K. J. Schulten, Neural-gas network for vector quantization and its application to time-series prediction, IEEE Trans. Neural Netw. 4 (1993), 558–569.
[89] B. Fritzke, A growing neural gas network learns topologies, w: Adv. Neural Inf. Process. Syst. 7, Cambridge (1995).
[90] M. Daszykowski, Exploration of multidimensional chemical data; Methods of compresion and visualization, Uniwersytet Śląski (2003).
[91] M. Ester, H. Kriegel, J. Sander, X. Xu, A density-based algorithm for discovering clusters in large spatial databases with noise, Portland (1996).
[92] M. Daszykowski, B. Walczak, D. L. Massart, Looking for Natural Patterns in Analytical Data. 2. Tracing Local Density with OPTICS, J. Chem. Inf. Comput. Sci. 42 (2002), 500–507.
[93] M. Daszykowski, B. Walczak, D. L. Massart, Looking for natural patterns in data: Part 1. Density-based approach, Chemom. Intell. Lab. Syst. 56 (2001), 83–92.
[94] M. Daszykowski, B. Walczak, Density-Based Clustering Methods, w: S. D. Brown, R. Tauler, B. Walczak (Ed.), Compr. Chemom., Elsevier, Oxford (2009), 635–654.
[95] M. Ankerst, M. M. Breunig, H. P. Kriegel, J. Sander, OPTICS: Ordering Points To Identify the Clustering Structure, in: ACM Press (1999), 49–60.
[96] M. Daszykowski, B. Walczak, D. L. Massart, Representative subset selection, Anal. Chim. Acta. 468 (2002), 91–103.
[97] I. Stanimirova, M. Daszykowski, D. L. Massart, F. Questier, V. Simeonov, H. Puxbaum, Chemometrical exploration of the wet precipitation chemistry from the Austrian Monitoring Network (1988–1999), J. Environ. Manage. 74 (2005), 349–363.

[98] K. Yang, Z. Xue, H. Li, T. Jia, Y. Cui, New methodology of hyperspectral information extraction and accuracy assessment based on neural network, Math. Comput. Model. 58 (2013), 644–660.
[99] S. Busygin, O. Prokopyev, P. M. Pardalos, Biclustering in Data Mining, Comput Oper Res. 35 (2008), 2964–2987.
[100] B. Mirkin, Mathematical Classification and Clustering, Springer US, Boston (1996).
[101] M. Wang, X. Shang, X. Li, W. Liu, Z. Li, Efficient mining differential co-expression biclusters in microarray datasets, Gene. 518 (2013) 59–69.
[102] S. C. Madeira, A. L. Oliveira, Biclustering algorithms for biological data analysis:
a survey, IEEE/ACM Trans. Comput. Biol. Bioinform. 1 (2004), 24–45.
[103] Y. Cheng, G. M. Church, Biclustering of expression data, Proc. Int. Conf. Intell. Syst. Mol. Biol. ISMB Int. Conf. Intell. Syst. Mol. Biol. 8 (2000), 93–103.
[104] Y. Kluger, Spectral Biclustering of Microarray Data: Coclustering Genes and Conditions, Genome Res. 13 (2003) 703–716.
[105] R. Bro, E. E. Papalexakis, Coclustering-a useful tool for chemometrics, J. Chemom. 26 (2012) 256–263.
[106] J. Kennedy, R. C. Eberhart, Y. Shi, Particle swarm optimization: developments, applications and resources, San Diego (2001), 81 – 86.
[107] F. Marini, B. Walczak, Finding relevant clustering directions in high-dimensional data using Particle Swarm Optimization, J. Chemom. 25 (2011), 366–374.
[108] T. N. Tran, R. Wehrens, L. M. C. Buydens, Clustering multispectral images: a tutorial, Chemom. Intell. Lab. Syst. 77 (2005), 3–17.
[109] E. A. M. Salah, A. M. Turki, E. M. Al-Othman, Assessment of Water Quality of Euphrates River Using Cluster Analysis, J. Environ. Prot. (2012), 1629–1633.
[110] C. S. Funari, P. J. Eugster, S. Martel, P. Carrupt, J. Wolfender, D. H. S. Silva, High resolution ultra-high pressure liquid chromatography–time-of-flight mass spectrometry dereplication strategy for the metabolite profiling of Brazilian Lippia species, J. Chromatogr. A, (2012),167–178.
[111] T. Li, L. Dai, L. Li, X. Hu, L. Dong, J. Li, et al., Typing of unknown microorganisms based on quantitative analysis of fatty acids by mass spectrometry and hierarchical clustering, Anal. Chim. Acta. 684 (2011), 8–16.
[112] S. Dussert, A. Laffargue, A. de Kochko, T. Joët, Effectiveness of the fatty acid and sterol composition of seeds for the chemotaxonomy of Coffea subgenus Coffea, Phytochemistry 69 (2008), 2950–2960.
[113] A. Brazma, J. Vilo, Gene expression data analysis, FEBS Lett. 480 (2000), 17–24.
[114] I. Granlund, T. Kieselbach, R. Alm, W. P. Schröder, C. Emanuelsson, Clustering of MS spectra for improved protein identification rate and screening for protein variants and modifications by MALDI-MS/MS, J. Proteomics. 74 (2011), 1190–1200.

[115] H. Janečková, K. Hron, P. Wojtowicz, E. Hlídková, A. Barešová, D. Friedecký, et al., Targeted metabolomic analysis of plasma samples for the diagnosis of inherited metabolic disorders, J. Chromatogr. A. 1226 (2012), 11–17.
[116] J. B. Nikas, W. C. Low, Application of clustering analyses to the diagnosis of Huntington disease in mice and other diseases with well-defined group boundaries, Comput. Methods Programs Biomed. 104 (2011), e133–e147.
[117] T. Aste, M. Saadatfar ,T. J. Senden, Geometrical structure of disordered sphere packings, Phys Rev. (2005) 1–15.
[118] T.N. Tran, K. Drab, M. Daszykowski, Revised DBSCAN algorithm to cluster data with dense adjacent clusters, Chemom. Intell. Lab. Syst. 120 (2013), 92–96.
[119] C. Tistaert, B. Dejaegher, Y. V. Heyden, Chromatographic separation techniques and data handling methods for herbal fingerprints: A review, Anal. Chim. Acta. 690 (2011), 148–161.
[120] D. H. Foster, K. Amano, S. M. C. Nascimento, M. J. Foster, Frequency of metamerism in natural scenes., J. Opt. Soc. Am. A. 23 (2006), 2359–2372.
[121] Hyperspectral images of natural scenes (2002); http://personalpages.manchester.ac.uk/staff/david.foster/Hyperspectral_images_of_natural_scenes_02.html.
[122] M. Kumar, N. R. Patel, Clustering data with measurement errors, Comput. Stat. Data Anal. 51 (2007) 6084–6101.