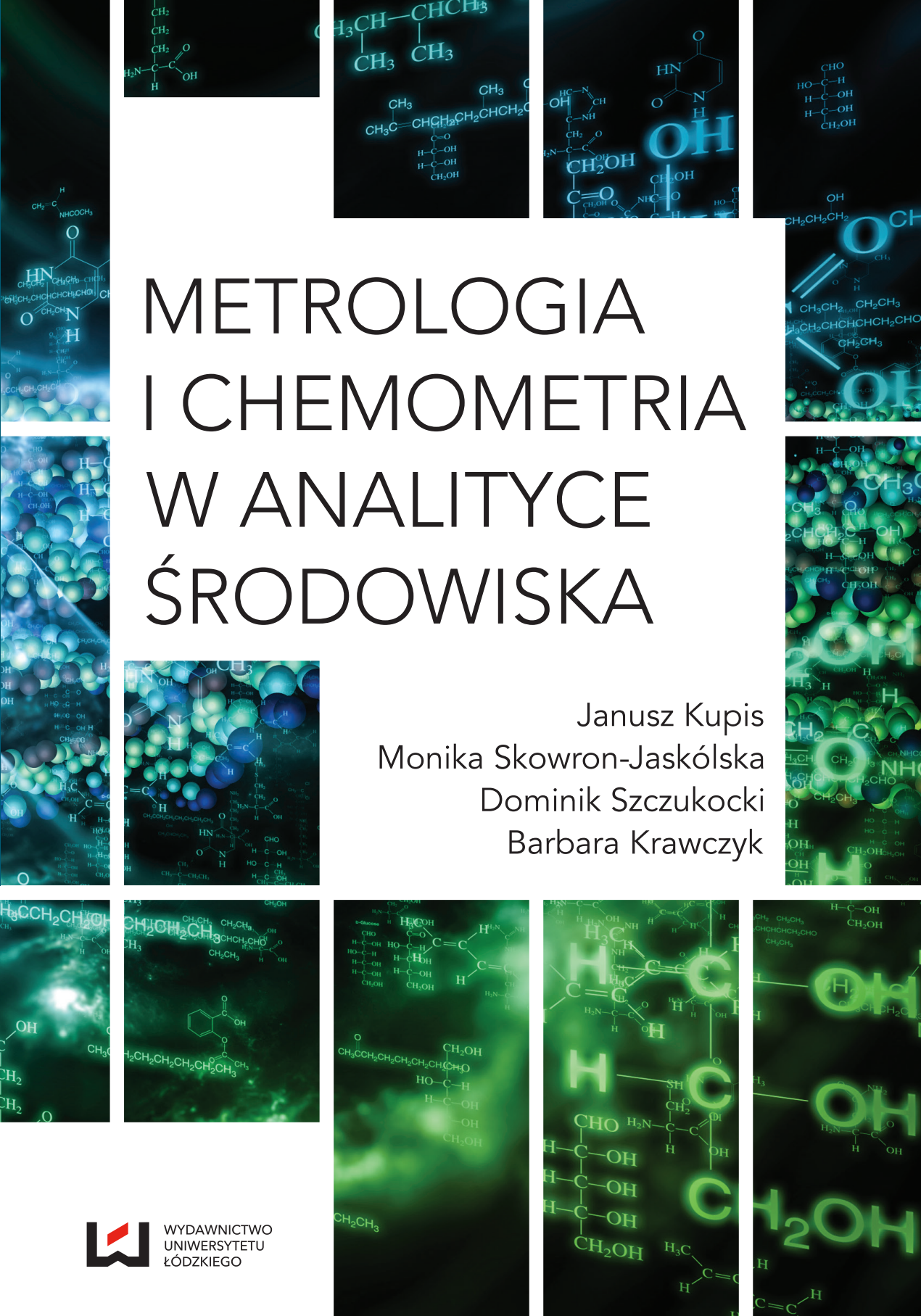


METROLOGIA I CHEMOMETRIA W ANALITYCE ŚRODOWISKA

Janusz Kupis
Monika Skowron-Jaskólska
Dominik Szczukocki
Barbara Krawczyk

WYDAWNICTWO UNIWERSYTETU ŁÓDZKIEGO

http://dx.doi.org/10.18778/8088-176-1.03

CHEMOMETRIA

Janusz Kupis

1 WPROWADZENIE

Dziś każdy, kto rozpoczyna poszukiwania informacji z dowolnej dziedziny nauki, w pierwszym etapie tych poszukiwań wykorzystuje możliwości, jakie dają sieć Internet i silniki wyszukiwarek. To bardzo wygodny, szybki i skuteczny sposób na nieograniczony dostęp do informacji o dokumentach na dany temat, bądź do oryginałów samych dokumentów. Jeśli poszukując w sieci informacji na temat chemometrii w języku polskim, wykorzystamy dowolną frazę określająca tę tematykę, to wynikiem naszych poszukiwań będzie niestety niezbyt pokaźna liczba dokumentów dotyczących tejże dziedziny. Znacznie bogatszym źródłem (z punktu widzenia liczebności publikacji i różnorodności tematyki), są strony angielskojęzyczne.

Publikacją polskojęzyczną, mającą największe znaczenie, na której treściach oparta jest większość pozostałych prac związanych z podstawami chemometrii, jest pozycja autorstwa prof. Jana Mazerskiego z Wydziału Chemii Politechniki Gdańskiej – „*Chemometria praktyczna*". Książka jest doskonałym źródłem pozwalającym na zgłębianie podstaw tego przedmiotu i od początku pojawienia się chemometrii w programie studiów kierunku Analityka chemiczna na Wydziale Chemii Uniwersytetu Łódzkiego, była obok wykładów i instrukcji do ćwiczeń laboratoryjnych podstawową pozycją literaturową, pomocną w nauce chemometrii.

Treść niniejszego rozdziału skryptu oparta jest w części na podstawach przedmiotu zawartych we wspomnianej książce Jana Mazerskiego, ale zawiera też wiele informacji i przykładów wykorzystania danych, których źródłem są pozycje traktujące niektóre obszary chemometrii w bardziej szczegółowy sposób. To bardziej ukierunkowane podejście do zagadnień poruszanych głównie na zajęciach laboratoryjnych wydaje się być uzasadnione, a może nawet konieczne biorąc pod uwagę trudności, z jakimi uczestnicy kursu borykali się w poprzednich dwóch latach prowadzonych zajęć. Nie należy jednak w tym miejscu popadać w euforię myśląc, że rozdział ten będzie lekarstwem na wszelkie dotychczasowe problemy w zrozumieniu wszystkich zagadnień w nim poruszanych. Niezbędnym w tym celu elementem jest więcej niż dobra znajomość podstaw statystyki z pierwszego roku studiów, a także znajomość zagadnień dotyczących macierzy, podstaw rachunku macierzowego oraz podstaw rozwiązywania równań macierzowych.

1.1 Początki i rozwój chemometrii

Zapotrzebowanie na metody pozwalające analizować większe i pod innym kątem, niż ma to miejsce w przypadku statystyki doświadczalnej, zbiory danych pojawiło się w połowie XX wieku. Na ten okres właśnie datuje się powstanie dziedziny nazywanej przez chemików chemometrią. Przez chemików, ponieważ specyfika tej dziedziny, a mówiąc ściślej fakt wykorzystywania praktycznie tych samych metod analitycznych (matematycznych) w innych dyscyplinach naukowych, pozwala ekonomistom nazywać ją ekonometrią, psychologom – psychometrią, archeologom – archeometrią, a biologom – biometrią. Przyrostek 'metria' jest w tym przypadku wskazaniem na charakterystyczne techniki analizy danych, które są wykorzystywane, gdy mamy do czynienia z dużymi, trudnymi do analizy zbiorami danych i które dają inny rodzaj informacji niż uzyskiwany metodami statystycznymi. Gdyby pokusić się o definicję tej dziedziny nauki, to jedną z wielu mogłaby być definicja przytaczana przez Mazerskiego:
Chemometria jest dziedziną zajmującą się wydobywaniem użytecznej informacji z wielowymiarowych danych pomiarowych, wykorzystującą metody statystyki i matematyki.

Ta bardzo ogólna definicja przedmiotu dobrze charakteryzuje wczesną chemometrię, z początków jej zastosowań, kiedy to wykorzystywanymi technikami analizy były te, wywodzące się ze statystyki doświadczalnej. Dziś metody obu dziedzin różnią się zasadniczo. Techniki analizy wykorzystywane w chemometrii zalicza się do technik *eksploracji danych*, określanych mianem *Data Mining* (DM). Ich zastosowanie oraz intensywny rozwój jest ściśle związany z rozwojem technologii informatycznych i możliwościami przechowywania dużych ilości danych cyfrowych. Zatem początki tego okresu to lata 80 poprzedniego wieku. Wtedy to właśnie powstawały relacyjne bazy danych i strukturalne języki zapytań. Następnym elementem, który w latach 90 przyczynił się do rozwoju dziedziny, to hurtownie danych i stanowiące ich nierozłączną część narzędzia do ich analizy, jak poszukiwanie trendów, zależności czy wzorców. Informacje te byłyby trudne do wydobycia, gdyby nie specjalistyczne narzędzia informatyczne – przede wszystkim nowoczesne, zawierające algorytmy technik *Data Mining* oprogramowanie.

Choć sama nazwa chemometria powstała w 1970 roku to, jako odrębna dyscyplina powstawała w latach osiemdziesiątych XX wieku. W tym okresie pojawiły się pierwsze opracowania, czasopisma, książki oraz konferencje dedykowane tylko i wyłącznie temu przedmiotowi. Pierwsze publikacje dotyczyły stosunkowo prostych, analitycznych problemów jak np. rozplot nakładających się pików chromatograficznych. Przegląd literatury z tego okresu pozwoli wyciągnąć wniosek, że to właśnie chromatografia HPLC (High Performance Liquid Chromatography – wysokosprawna chromatografia cieczowa) i NIR (Near Infrared – spektroskopia w bliskiej podczerwieni) miały swój znaczący udział w rozwoju i wyodrębnianiu się chemometrii jako nowej dyscypliny w latach 80.

Siłą napędową rozwoju chemometrii w latach 90 był przemysł farmaceutyczny. Koncerny farmaceutyczne wykorzystywały chemometrię do rozwiązywania nieco bardziej złożonych problemów, jak rozpoznawanie wzorców chromatograficznych (układu pików) preparatów medycznych. Powstawały też algorytmy pozwalające na 'interaktywne' monitorowanie i kontrolę ilościową procesu chemicznego, zachodzącego z udziałem 4–5 możliwych do spektroskopowej identyfikacji reagentów lub produktów. Ten rodzaj rozwiązywanego problemu, z uwagi na dużą złożoność obliczeniową, wymagał rozwoju nowych metod chemometrycznych w zakresie rozpoznawania wzorców i modelowania krzywych. Rosnąca na przestrzeni lat wielkość i stopień skomplikowania głównie biomedycznych baz danych (chromatogramów, spektrogramów) doprowadziła chemometrię do miejsca, w którym można ją uważać za typową dziedzinę wykorzystującą techniki DM. W dzisiejszej chemometrii wykorzystywane są wszystkie możliwe techniki z zakresu eksploracji danych, jak choćby te najbardziej znane: przetwarzanie sygnałów, sieci neuronowe, uczenie maszynowe, drzewa decyzyjne, grupowanie hierarchiczne, itp.

Wielki wpływ uwarunkowań wynikających z imponującego rozwoju technologii IT jest oczywisty i niezaprzeczalny. Dzięki nowym technologiom informatycznym możemy obserwować szybki rozwój metod analitycznych jak np. chromatografia gazowa czy cieczowa z detektorami mas (GC-$(MS)^n$, LC-$(MS)^n$), czy też spektrometria mas ze wzbudzeniem w plazmie indukcyjnie sprzężonej (ICP-MS). Te nowoczesne metody badawcze pozwalają na jednoczesne oznaczanie wielu analitów, przy znacznym obniżeniu granic

wykrywalności oraz zwiększeniu selektywności w stosunku do metod stosowanych uprzednio. Niewątpliwie prowadzi to do zwiększenia możliwości badawczych, ale zwiększa też ilość wytwarzanych informacji. Te nadmiarowe informacje, są często trudne do wykorzystania przez badacza i niekoniecznie zwiększają wiedzę na temat badanych obiektów. Bywa, że wręcz przeciwnie, wprowadzają dodatkowy chaos informacyjny. Powoduje to głównie ich wielowymiarowość, którą trudno zwizualizować wykorzystując tradycyjne wykresy czy tabele.

Nadmiarowość danych jest zwykle niewykorzystywana, a w przypadku niewłaściwie zaplanowanego eksperymentu bywa, że również niepotrzebna. Aby nie dopuścić do zjawiska gromadzenia bezużytecznych danych i ponoszenia kosztów z tym związanych, nowoczesną chemometrię należy postrzegać jako proces. Proces wieloetapowy, będący metodologią planowania badań, ich wykonywania i wreszcie analizy danych, będących wynikiem wcześniej zaplanowanego działania. Proces ciągły, trwający od zrozumienia uwarunkowań ekonomicznych, przez zebranie i zarządzanie danymi, przygotowanie danych, modelowanie i ewaluację modelu chemometrycznego do jego wdrożenia. Zastosowanie chemometrii do analizy zupełnie przypadkowych danych eksperymentalnych (z pominięciem etapu planowania i przygotowania doświadczenia) i potraktowanie jej metod jako jedynie narzędzi typu ‘czarna skrzynka’, prowadzi zwykle do poważnych oraz kosztownych błędów merytorycznych, czyli niewłaściwych wniosków i uogólnień badawczych.

Chemometria jest więc dziś dziedziną, którą trudno jest zdefiniować kilkoma krótkimi zdaniami, stąd mnogość definicji jakie można znaleźć w literaturze dotyczącej przedmiotu. Pełny opis, czym jest dyscyplina zwana chemometrią dałoby Czytelnikowi przeczytanie ich wszystkich. Definicją podawaną dziś przez International Chemometrics Society jest:

Chemometria jest dziedziną chemii, w której stosuje się metody statystyki, matematyki i inne wykorzystujące logikę formalną do

– oceny i interpretacji danych chemicznych (analitycznych)
– optymalizacji i modelowania procesów i eksperymentów chemicznych
– wydobycia możliwie największej ilości informacji z danych eksperymentalnych.

Można odnieść wrażenie, że podana na początku definicja chemometrii, zawarta w książce J. Mazerskiego, w eleganckich i trafnych słowach zawiera wszystko to, o czym czytamy w tej ostatniej…

1.2 Obszary wykorzystania metod Data Mining

Algorytmy wielowymiarowej eksploracji danych cechuje różnorodność. Jednak pomimo różnic, można określić kilka płaszczyzn tematycznych – celów, dla osiągnięcia których zostały stworzone. Główne problemy, jakie można rozwiązywać przy ich pomocy to: wyszukiwanie wzorców i trendów, klasyfikacja obiektów, badanie ich podobieństwa oraz redukcja wymiarowości cech je opisujących. Prowadzi to do powstawania modeli matematycznych reprezentujących obiekty, na podstawie których możemy je grupować, prognozować ich nowe wartości zmiennych, optymalizować i kontrolować warunki jakie na nie wpływają. Wymienione sposoby wykorzystania algorytmów chemometrycznych można ująć w kilka bardziej szczegółowo sformułowanych punktów:

1. **Opis (prezentacja)** – wizualizacja głównych zależności odkrywanych w zbiorach danych (poszukiwanie wzorców i określanie trendów zachowań). Aby była ona możliwa w przypadku obiektów reprezentowanych przez dużą liczbę zmiennych, konieczna jest – redukcja wymiaru przestrzeni cech (inaczej – przestrzeni zmiennych).
2. **Klasyfikacja** – proces przyporządkowania nowych, nieznanych obiektów do pewnych zborów (grup, klas) na podstawie wartości jakościowej zmiennej celu. Zmienną celu w tym przypadku może być cecha bezpośrednio opisująca obiekt, ale także osobna wartość wyznaczona na podstawie składowych wektora cech obiektu. Inaczej – klasyfikacja to przyporządkowanie obiektów do zbiorów na podstawie posiadanej wcześniej informacji o wartości zmiennej celu.
3. **Grupowanie** – poszukiwanie grup lub podobnych struktur danych**.** Różni się od klasyfikacji tym, że w jego przypadku nie ma zmiennej celu. Algorytmy grupujące nie próbują wyznaczać wartości zmiennej celu. Zamiast tego, dzielą cały zbiór danych na stosunkowo zgodne podgrupy lub grupy, gdzie podobieństwo rekordów (wektorów cech) wewnątrz grup jest maksymalizowane, a podobieństwo do rekordów spoza grupy

minimalizowane. Ważną cechą grupowania jest to, że odbywa się ono bez zewnętrznej kontroli, bez nadzoru.

4. **Regresja (estymacja)** – to poszukiwanie funkcji, która będzie zdolna przewidywać rzeczywiste wartości analizowanych zmiennych, minimalizując błąd między wartością rzeczywistą a szacowaną. Szacowanie jest podobne do klasyfikacji z wyjątkiem charakteru zmiennej celu, który jest numeryczny, a nie jakościowy.
5. **Przewidywanie (predykcja)** – przewidywanie jest podobne do klasyfikacji i szacowania, z wyjątkiem faktu, że w przewidywaniu wynik dotyczy przyszłości. Przewidywanie zwykle dotyczy modeli tworzonych dla szeregów czasowych – wartości zmiennej w czasie.
6. **Odkrywanie reguł** – poszukiwanie zależności pomiędzy cechami opisującymi analizowane obiekty. W eksploracji danych polega ono na szukaniu, które atrybuty (zmienne) są 'powiązane ze sobą'. Asocjację taką określa się ilościowo i może ona dotyczyć jednocześnie dwóch lub więcej cech (zmiennych).

Te najważniejsze, wymienione wyżej zastosowania algorytmów *Data Mining* sprawiają, że dzisiaj są one wszechobecne i mają zastosowanie praktycznie w każdej dziedzinie nauki i szeroko rozumianego biznesu. Jako intuicyjnie oczywisty przykład zastosowania DM można podać sektor bankowy. Typowe zastosowania metod eksploracji danych w tym sektorze to ocena ryzyka kredytowego, identyfikacja grup klientów pod kątem sprzedaży produktów finansowych, czy przewidywanie trendów na rynkach finansowych. Nowym przykładem zastosowań eksploracji danych w tym sektorze to systemy detekcji różnego rodzaju przestępstw finansowych, realizowane poprzez analizę wykonywanych operacji bankowych, w celu wykrycia nietypowych wzorców zachowań.

Kolejnym przykładem sektora czerpiącego ogromne korzyści z analizy danych metodami DM jest sektor handlowy. Źródłem wartościowych danych do analizy są oczywiście klienci sieci handlowych. Karty lojalnościowe i choćby sprzedaż online umożliwiają gromadzenie ogromnych ilości danych na temat kupujących: informacji o rodzaju sprzedanych produktów, historii zakupów dotyczących miejsca i czasu, czy śledzeniu tras przesyłek (regionalizacja kupujących określone produkty). Zebrane dane pozwalają określić zachowania konsumentów, wyznaczać grupy docelowe, którym oferowane są

konkretne produkty, tym samym zwiększać sprzedaż i redukować koszty obsługi klientów.

Ciekawym przykładem wykorzystania narzędzi eksploracji danych jest analiza asocjacji, czyli określanie jakie jest prawdopodobieństwo, że klient kupujący produkt X kupi jeszcze produkt Y. Tego typu analizy wykorzystywane są przez systemy, które na podstawie historii zakupów konkretnej osoby są w stanie tworzyć sprofilowane rekomendacje. Trafiają one później do klienta w postaci indywidualnej oferty i różnego rodzaju dedykowanych kuponów czy zniżek, mających na celu przyciągnąć daną osobę do konkretnej sieci handlowej.

Nie można w tym miejscu nie wspomnieć o największej z możliwych baz danych, jaką są dane pochodzące z Internetu. Generują je głównie komunikatory i portale społecznościowe. Zgodnie z najnowszymi szacunkami, w ciągu jednej minuty wykonywanych jest ok. 370 tysięcy rozmów z wykorzystaniem komunikatora Skype. W ciągu jednej minuty wysyłanych jest 198 milionów e-maili i dodawanych jest ponad pół miliona komentarzy na portalu społecznościowym Facebook. Zebrane dane mogą służyć do analizy obciążenia sieci, wykrywania nadużyć i oszustw albo do znajdowania grup klientów, którym można sprzedać konkretny produkt. Mogą też służyć mniej komercyjnym, ważnym z punktu widzenia bezpieczeństwa celom jak wykrywaniu przestępstw oraz przeciwdziałaniu terroryzmowi. Należy w tym miejscu nadmienić, że dane pozyskane z wykorzystaniem Internetu to nie tylko dane niewrażliwe, jak listy połączeń, lokalizacja osób dzwoniących, godziny wykonywania połączeń, czas ich trwania, ale też wiele innych, wkraczających w sferę prywatności autora informacji. Wraz z rozwojem metod DM pojawia się zatem problem prywatności i ochrony danych osobowych, który wymaga prawnego uregulowania, tak aby dostęp do informacji uzyskanych za pomocą metod eksploracji danych miały jedynie uprawione instytucje w uzasadnionych przypadkach oraz aby niemożliwa była sprzedaż takich informacji.

Bez wykorzystania zaawansowanych algorytmów eksploracji danych nie mogłaby dziś istnieć, w postaci jaką znamy, większość dziedzin nauki. Dobrym przykładem może być tutaj biologia molekularna i takie jej działy jak genomika, proteomika czy metabolomika. Eksperymentalne techniki ba-

dawcze jak mikromacierze DNA, metody sekwencjonowania RNA w genomice, czy spektrometrii masowej w proteomice generują ogromne ilości danych, których analiza nie byłaby możliwa bez komputerów i odpowiednio zaprojektowanych algorytmów DM. Ciekawymi przykładami ich wykorzystania w tej gałęzi biologii jest np. możliwość porównywania sekwencji DNA oraz białek w celu znalezienia podobieństwa ich funkcjonowania na podstawie podobieństwa sekwencji, analiza oddziaływań między cząsteczkami, możliwość projektowania leków czy przewidywania struktur białek.

Wielkim zainteresowaniem cieszy się dziś w medycynie wdrażanie różnego rodzaju systemów eksperckich opartych o techniki *Data Mining* wspomagających diagnostykę chorych. Tego typu modele próbują postawić diagnozę na podstawie listy objawów, symptomów oraz wprowadzanych wyników badań. Specjalnie zaprojektowane algorytmy maszynowego uczenia, będącą częścią integralną takiego systemu, analizują przygotowaną wcześniej przez ekspertów, historyczną bazę danych przypadków medycznych. Wynikiem działania takiego systemu jest lista najbardziej prawdopodobnych chorób. Każdy nowy przypadek (objawy – diagnoza) wprowadzony do tak zaprojektowanego systemu jest kolejnym elementem, który powoduje, że system uczy się, stając się coraz bardziej dokładnym i zaawansowanym narzędziem. Jako przykład obecnie wykorzystywanego systemu tego typu można podać komercyjny system Clinical Decision Support and Analytics firmy Alere Analytics. Badania przeprowadzone na grupie 77 tysięcy pacjentów z udziałem 500 lekarzy wykazały dokładność diagnostyczną systemu na poziomie 100%, a jego precyzję na poziomie 75%.

To tylko niewielka część możliwych do przytoczenia przykładów obecności eksploracji danych w życiu codziennym i nauce. Nie dziwi zatem fakt, że jako odrębna gałąź, metody DM pod nazwą *chemometria*, od około 1980 roku znalazły zastosowanie także w chemii.

1.3 Cechy metod chemometrycznych

Co odróżnia chemometrię od typowych metod statystycznych? W jakich sytuacja można wykorzystać metody redukcji wymiaru, klasyfikacji, grupowania, poszukiwania wzorców i trendów? Na takie pytania może dać

odpowiedź analiza różnic zbiorów danych wykorzystywanych w przypadku obu metod.

Wygodnym, powszechnie zaakceptowanym sposobem gromadzenia danych eksperymentalnych jest ich **tabelaryczny układ**. Przy czym każdy wiersz (rekord) takiej tabeli reprezentuje obiekt (w analityce chemicznej zwykle punkt pomiarowy, próbka). Kolumny natomiast zawierają wartości zmiennych opisujące cechy (właściwości) tych obiektów. Dodatkowo, każda kolumna zawsze odpowiada jednej, tej samej zmiennej opisującej obiekt. W przypadku, gdy do analizy zbioru danych wykorzystywane są metody typowe dla statystyki, regułą jest, że liczba ilości pomiarów (obiektów) musi być przynajmniej 4–5 razy większa od liczby mierzonych wielkości. Jak wiadomo z podstaw statystyki związane jest to z akceptowalnym poziomem ufności choćby dla wyliczonej wartości średniej. W przypadku metod chemometrycznych zasada ta zupełnie nie obowiązuje. Dopuszczalna i częsta jest sytuacja odwrotna, w której ilość badanych obiektów jest znacznie mniejsza niż liczba opisujących je zmiennych. Podstawową różnicą między dyscyplinami jest więc stosunek ilości badanych próbek do liczby wykonywanych dla nich pomiarów.

Prostym i obrazującym te różnice przykładem może być analiza chromatogramu zawierającego np. 25 pików odpowiadających składnikom lipidowym jakiegoś tłuszczu zwierzęcego. Wysokość (lub powierzchnię) każdego z pików możemy zmierzyć i wartość zapisać w osobnej kolumnie. Mamy więc jedną próbkę i 25 opisujących ją wielkości. Aby poddać analizie statystycznej tego typu dane (jednoczesna analiza 25 zmiennych), zgodnie z jej regułami powinniśmy wykonać od 100 do 125 chromatogramów badanych próbek. Nie trzeba wspominać, że każdy wykonany pomiar to dodatkowe koszty i potrzebny do tego czas. W takiej sytuacji moglibyśmy również ograniczyć liczbę zmiennych do arbitralnie wybranych. Prowadziłoby to utraty trudnej do określenia ilości informacji zawartej we wszystkich zmiennych. Dzięki specyfice metod z zakresu chemometrii, zestaw 25 zmiennych możemy wiarygodnie analizować korzystając z pomiarów wykonanych zaledwie dla 20–30 próbek. Co ciekawe, wyniki analizy będą oparte na całym zasobie informacji zawartej w próbkach.

Struktura trudnych do analizy z punktu widzenia klasycznej statystyki zbiorów danych, była głównym powodem, dla którego chemometria wypracowała własne metody i algorytmy ich eksploracji. W przeciwieństwie do metod statystycznych, które powstały dla analizy pojedynczych zmiennych lub co najwyżej niewielkiej ich liczby, algorytmy chemometrycznej analizy danych zakładają z góry jednoczesną analizę dużej liczby zmiennych. Zatem specyfika metod to kolejna, zasadnicza różnica między statystyką i chemometrią, a problemy rozwiązywane metodami chemometrycznymi należą do zagadnień obarczonych dużym ryzykiem z punktu widzenia statystyki matematycznej. Gdyby pokusić się o ujęcie najważniejszych cech – zalet algorytmów metod chemometrycznych w kilku punktach, to należałoby pamiętać o:

- możliwości analizy bardzo dużych zbiorów danych
- możliwości jednoczesnej analizy dużej liczby zmiennych, dla niewielkiej liczby badanych obiektów (minimalizacja ilości pomiarów)
- dopuszczalnej dużej zmienności, złożoności i niepewności danych
- możliwości badania istotności stopnia wewnętrznego powiązania zmiennych
- **możliwości uzyskania wysokiej jakości informacji na podstawie danych o dużym poziomie niepewności, co pozwala na rozwiązywanie problemów trudnych, leżących na granicy stosowalności metod statystycznych**.

Szczególną cechą chemometrii jest również to, że gdy mają zostać zastosowane jej algorytmy, powinna ona być obecna na każdym etapie badanego problemu, od jego sformułowania, zaplanowania sposobu jego rozwiązania do wniosków końcowych. Wykorzystanie chemometrii dopiero na etapie opracowania gotowych, niewłaściwie pozyskanych danych często kończy się rozczarowującymi i mało przydatnymi wynikami. Kontrola chemometryczna wszystkich etapów analizy, a zwłaszcza formułowania problemu i planowania eksperymentu pozwala ograniczyć ewentualne niepowodzenia. Typowym oraz częstym błędem praktyków w innych dziedzinach jest właśnie sytuacja, w której wykonana została już seria często kosztownych pomiarów celem potwierdzenia pewnych hipotez, stworzenia modelu zależności, optymalizacji procesu, wykonania prognoz, ale bez wcześniejszego ich zaprojektowania zgodnie z zasadami chemometrii. Badania takie zwykle kończą się niepowodzeniem.

2 KONTROLA DANYCH

2.1 Dokumentacja

Właściwa dokumentacja i czytelny opis danych nie są czynnikami, które mają bezpośredni wpływ na wynik analizy. Niemniej warto poświęcić nieco uwagi tym elementom, ponieważ zaniechania i niestaranność na tym etapie zwykle są źródłem wielu kłopotów i błędnych wyników. Jak już zostało wcześniej wspomniane, najlepszym, powszechnie zaakceptowanym sposobem gromadzenia danych są struktury tabelaryczne. Powinny być one odpowiednio zaplanowane, a ich elementy opisane we właściwy sposób. Każda tabela powinna posiadać:

- nagłówek z informacją o:
  - problemie, którego dotyczy
  - dacie utworzenia i dacie ostatniej modyfikacji
  - identyfikatorze autora danych,
- jednoznacznie opisane kolumny dla każdej zmiennej,
- jednoznaczną, niepowtarzalną nazwę dla każdego obiektu, rekordu (wiersza),
- informacja o pochodzeniu danych, zwłaszcza gdy pochodzą z różnych źródeł:
  - przez kogo wykonane
  - jaką metodą i na jakim przyrządzie
  - kiedy zostały wykonane,
- opis i uzasadnienie ewentualnych modyfikacji danych,
- jeżeli dane są wynikiem obliczeń, należy podać sposób obliczeń.

Jeśli w zbiorze danych istnieje naturalne ich uporządkowanie to powinno ono zostać odwzorowane w przygotowywanej tabeli. Dobrym przykładem są w tym miejscu np. dane pomiarów spektroskopowych, w przypadku których kryterium porządkującym dane jest długość fali lub też dane chromatograficzne, gdzie rolę taką może odgrywać rosnący czas retencji. W tym miejscu warto jeszcze wspomnieć o dobrych zasadach określających nazewnictwo zmiennych i obiektów. Ich podstawą są takie oto elementy:

- nazwy, jeśli to możliwe, muszą się kojarzyć ze zmienną, której dotyczą i to nie tylko autorowi tabeli, ale też innym użytkownikom,
- nazwy powinny być krótkie (najlepiej 2, 3 literowe), aby czytelnie opisywały zmienne czy obiekty na wykresach (najlepszymi nazwami są ogólnie przyjęte skróty),
- nazwy powinny być ciągiem liter i ewentualnie cyfr (bez znaków specjalnych),
- jeśli opracowane według tych zasad skróty nie kojarzą się ze zmienną, powinny zostać koniecznie opisane w dokumencie.

Ponieważ obecnie wszelkie analizy statystyczne, chemometryczne wykonuje się z wykorzystaniem specjalistycznego oprogramowania, dane dobrze jest przygotować w formacie czytelnym dla konkretnej aplikacji. Standardem jest tutaj Excel firmy Microsoft i jego arkusz kalkulacyjny (ewentualnie plik tekstowy w formacie 'CSV'), który jest właściwy dla praktycznie wszystkich aplikacji statystycznych czy *Data Mining.* Należy jedynie pamiętać o kilku podstawowych zasadach pracy z samym arkuszem; jak jednolity format kolumn (liczby albo tekst) i zawsze ten sam typ separatora miejsc dziesiętnych, który można wybrać, jako jedną z opcji ustawienia systemowego lub ustawić bezpośrednio w aplikacji arkusza kalkulacyjnego. Opisy skrótów zmiennych i obiektów (lub jakikolwiek inny tekst) powinny zawsze znajdować się poza pionowym rzutem tabeli z analizowanymi wartościami. Innym rozwiązaniem jest umieszczenie napisów w dowolnym miejscu poza tabelą, ale w taki sposób, aby skopiowana tabela z danymi zawsze nadawała się do analizy statystycznej, chemometrycznej.

Na etapie przygotowywania 'bazy danych' zwykle rozwiązywany jest też problem danych brakujących. Luki w tabeli danych zdarzają się z różnych przyczyn, np. z powodu źle pobranej próbki, jej zanieczyszczenia czy zniszczenia. Efektem tego zawsze jest brak wartości zmiennych opisujących próbkę. W takiej sytuacji, jeśli jest to możliwe, brakujące dane należy uzupełnić. Jeśli nie – miejsce pozostawiamy wolne. Należy zawsze pamiętać, że podstawową zasadą jest, aby **w miejsca brakujących danych nigdy nie wpisywać zer.** Nie należy również wpisywać innych znaków niebędących liczbami – w przypadku zmiennej ilościowej, i nienależących do zbioru dopusz-

czalnych wartości dla zmiennych nominalnych (jakościowych). Wynikiem takiego postępowania są zawsze niepoprawne operacje wykonywane przez program (arkusz), często bez żadnego komunikatu o błędzie.

Luki w danych uzupełniamy wg jednej z następujących zasad:

- zastępowanie braków danych wartością najbardziej dominującą w danym zbiorze, najczęściej występującą – średnią, medianą,
- zastępowanie brakujących danych wartościami najbardziej prawdopodobnymi ale w ramach danego zbioru wartości, np. wartością wygenerowaną losowo z obserwowanego rozkładu zmiennej,
- wykorzystanie metody regresji do oszacowania wartości brakującej danej,
- wykorzystanie metody k–najbliższych sąsiadów do ustalenia wartości najbardziej prawdopodobnej dla brakującej danej.

Istnieje jeszcze przypadek, kiedy zmierzoną wielkość musimy zastąpić inną wartością zmiennej. Ma on miejsce, kiedy wartość mierzonej zmiennej jest mniejsza niż próg oznaczalności metody analitycznej. W takim przypadku **zmierzoną wartość zastępujemy wartością równą połowie progu oznaczalności metody**.

2.2 Kontrola poprawności danych

2.2.1 Rozkład pojedynczej zmiennej, wartości odstające

Kontrola poprawności danych jest najbardziej pracochłonnym i najbardziej żmudnym etapem procesu analizy chemometrycznej. Niemniej etapem koniecznym. Właściwe wykonanie wstępnej obróbki danych jest gwarancją uzyskania poprawnych wyników. Etap ten składa się z kilku czynności, jakie powinny zostać wykonane:

- wykrywanie i usuwanie ewentualnych błędów grubych lub wyników odbiegających (outliers) w istotny sposób od pozostałych (np. z innej populacji),
- badanie rozkładu zmiennej i przeprowadzenie jej transformacji, jeśli zachodzi taka konieczność (normalizacja rozkładu dla testów parametrycznych),

- wykrywanie korelacji pomiędzy zmiennymi opisującymi obiekty,
- skalowanie, autoskalowanie. Wymagają tego niektóre metody DM.

W przypadku pierwszych trzech punktów, bardzo pomocnym narzędziem są różnorodne techniki graficznej prezentacji danych i ich analiza wizualna. Możliwości takie dają nam specjalistyczne programy statystyczne np. 'Statistica', na którą Uniwersytet Łódzki posiada licencję, a oprogramowanie może być wykorzystywane również przez Studentów na zajęciach i w domach.

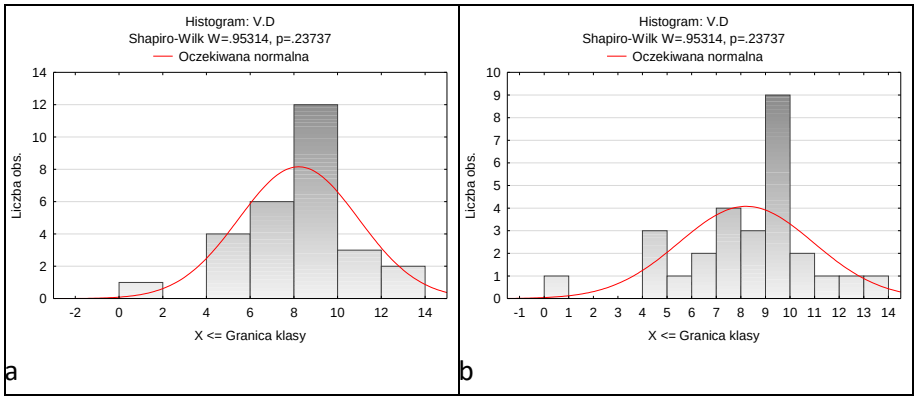
Problem wartości odbiegających (także błędów grubych), leżących z dala od reszty danych jest problemem starym, znanym statystyce doświadczalnej. Aby go rozstrzygnąć, musimy znaleźć odpowiedź na pytanie, czy punkt odbiegający to rzeczywiście niewłaściwy punkt w zbiorze, czy zbiór danych ma taki właśnie rozkład, a punkt ma wartość prawidłową. Znanych jest kilka testów statystycznych, pomocnych w rozwiązywaniu tego problemu. Test Q-Dixona, test G-Grubbsa, test 3-sigma i test przedziału ufności. Wszystkie one zakładają jednak, że pozostałe, nieobarczone błędem wyniki mają rozkład zgodny z normalnym. Najprostszym, graficznym sposobem sprawdzenia tego warunku jest wykonanie histogramu zmiennej. Jeśli wynika z niego, że rozkład zmiennej istotnie odbiega od rozkładu normalnego, zmienną należy poddać odpowiedniej transformacji.

Wykonanie histogramu w programie Statistica praktycznie nie wymaga od nas ingerencji w proces jego tworzenia i przebiega automatycznie. Warto jednak wspomnieć tu o istotnym elemencie, jaki ma wpływ na jego kształt i tym samym na nasze decyzje dotyczące samego rozkładu zmiennej. Jest nim jedynie ilość przedziałów (ilość słupków), z jakich histogram się składa. Prawidłową wartość tej wielkości określają wyrażenia w przybliżeniu równe dla niewielkiej ilości obiektów *n* (dla większych *n* zalecane jest stosowanie zależności **b**):

$$\text{a) } k \leq n/4 \qquad \text{b) } k \approx \sqrt{n} \tag{2.1}$$

gdzie:

n – ilość wartości zmiennej



Rys. 1. Histogramy stężenia witaminy D w surowicy dla 28 przypadków medycznych: a – prawidłowy, b – niewłaściwa ilość przedziałów

Źródło: opr. własne

Przykładowe histogramy pewnej zmiennej wykonane dla różnych wartości parametru *k* przedstawione są na wykresach powyżej. Histogram **b** charakteryzuje zbyt duża liczba przedziałów, przez co w poszczególnych przedziałach rejestrujemy zbyt małą liczbę obserwacji. Histogram taki przyjmuje nieregularny kształt, zależny głównie od czynnika losowego, co utrudnia jego interpretację. Rzadszym błędem jest zbyt mała liczba przedziałów. Otrzymany histogram zawiera wtedy zbyt mało informacji o rozkładzie, a zwłaszcza o wartościach odstających, które są wtedy niewidoczne. Automatyczne wykonanie histogramu w programie Statistica daje nam jednocześnie możliwość wykorzystania testów, których wynik pozwala ocenić, czy rozkład badanej zmiennej jest rozkładem normalnym. Są to znane powszechnie testy Chi-kwadrat, Kołmogorowa-Smirnowa czy Shapiro-Wilka.

Badanie rozkładu zmiennej dające odpowiedź na pytanie, czy jest on rozkładem normalnym możemy przeprowadzić, analizując grupę prostych parametrów, oszacowanych samodzielnie na podstawie zbioru wartości zmiennej. Warto poznać zasady tej analizy. Ich znajomość da nam wyobrażenie, jakimi cechami powinien charakteryzować się rozkład normalny zmiennej.

Wyjściowym parametrem analitycznym jest rozstęp Δ (rozpiętość danych), czyli różnica pomiędzy jego wartością maksymalną (MAX) i minimalną (MIN) w zbiorze. Jeśli wcześniej posiadamy informację, z jakiego zakresu wartości powinny być nasze dane, to już na tym etapie możemy

przeprowadzić wstępną kontrolę pod względem ich poprawności i usunąć dane odbiegające.

Kolejnym wyznaczanym parametrem, jest jeden z parametrów dający informację o typie rozkładu zmiennej. Jest to iloraz wcześniej wyznaczonych wartości MIN i MAX. Bezwzględna wartość MIN/MAX mniejsza od 0.1 jest informacją, że zmienna może posiadać rozkład różniący się od normalnego i w przyszłości może wymagać transformacji. Inną wskazówką, że rozkład zmiennej odbiega od normalnego jest odległe położenie wartości średniej (*m*) zbioru wobec środka przedziału zmienności (*w* = (MAX + MIN)/2). Odległość tę porównuje się z wartością odchylenia standardowego (*s*) pojedynczego pomiaru:

$$\text{zatem jeśli } |m - w| > s \qquad (2.2)$$

to zmienna wymaga dalszego sprawdzenia charakteru rozkładu. Informacji w tym zakresie dostarczyć nam może kolejny parametr, jakim jest iloraz rozstępu i odchylenia standardowego Δ/*s*. Wskaźnik ten nie powinien być spoza przedziału 3–5 (dla małej liczby < 30 wyników od 3 do 4) dla rozkładu normalnego. Inna jego wartość informuje nas o znacznej niejednorodności w rozkładzie.

Odstępstwa od rozkładu normalnego zmiennej, który jest rozkładem symetrycznym, może potwierdzić lub nie, ostatni z parametrów – indeks skośności rozkładu (*q*). Jest to parametr określający asymetrię rozkładu, która wyraża się jego prawo– lub lewoskośnością i wartością indeksu różną od zera. Miara ta jest wykorzystywana w arkuszu kalkulacyjnym Excel. Wartość indeksu skośności mniejsza od –2 sugeruje, że rozkład jest rozkładem lewoskośnym, większa od 2 – prawoskośnym, niewykazującym cech rozkładu normalnego.

Poddanie wszystkich opisanych zmiennych jednoczesnemu testowi, który można dla jasności przedstawić w postaci czterech następujących pytań:

1. Czy wartość MIN/MAX > 0.1 ?

2. Czy |*w* (środek rozkładu) – *m* (średnia)| < *s* (odchylenie std.) ?

3. Czy wartość Δ(rozstęp)/s należy do przedziału (3–5) ?

4. Czy |*q* (skośność)| < 2 ?

pozwala odpowiedzieć na pytanie, czy rozkład zmiennej jest normalny. Jeżeli dla jakiejkolwiek zmiennej, odpowiedź choćby na jedno z powyższych pytań brzmi NIE, wykonujemy jej histogram (i jeśli to możliwe jeden z wymienionych wcześniej testów normalności), ponieważ jest możliwe, że rozkład jest asymetryczny lub wielomodalny. Inną możliwością z jaką możemy mieć do czynienia jest występowanie wartości odbiegających. W zależności od sytuacji stosujemy następujący algorytm postępowania:

1. Jeśli zmienna ma rozkład wielomodalny pozostawiamy ją bez zmian.
2. Jeśli rozkład danej zmiennej jest rozkładem normalnym, ale posiada wartość odstającą potwierdzoną np. testem Grubbsa lub przedziału ufności, usuwamy tę wartość i więcej zmienną się nie zajmujemy.
3. Jeśli rozkład zmiennej nie ma cech rozkładu normalnego (np. potwierdza to nasza analiza lub któryś z testów normalności) i na histogramie widoczna jest wartość mogąca być wartością odbiegającą, należy ją tymczasowo usunąć i wykonać nowy histogram dla zmiennej.
4. Jeżeli po usunięciu wartości odbiegającej rozkład zmiennej nie uległ ‘poprawie’, należy przywrócić usuniętą wartość i dokonać transformacji (normalizacji) zmiennej.
5. Jeżeli po dokonaniu transformacji zmiennej jej rozkład stał się symetryczny, kończymy kontrolę zmiennej. Tak samo postępujemy, gdy po transformacji pojawia się wielomodalność rozkładu.
6. Jeżeli po dokonaniu transformacji zmiennej, na histogramie w dalszym ciągu widoczny jest punkt odbiegający, należy go tymczasowo usunąć i wykonać nowy histogram zmiennej.
7. Jeżeli rozkład transformowanej zmiennej po usunięciu wartości odbiegającej stał się normalny lub przynajmniej symetryczny, jednym z testów należy ocenić, czy odstająca wartość została słusznie usunięta. Więcej zmienną się nie zajmujemy.

Transformacji zmiennej dokonujemy zwykle wtedy, kiedy podejrzewamy ją o rozkład skośny, daleki od normalnego. Częstymi są trudne sytuacje, kiedy rozkład zmiennej wydaje się skośny z powodu pojawiającej się wartości odstającej – i odwrotnie, podejrzewamy istnienie wartości odstającej w zbiorze, ale kształt histogramu jest wynikiem skośności rozkładu. Jedynym wyjściem z takiej sytuacji jest testowanie zmiennej metodą prób

i błędów, ponieważ przedstawiony powyżej optymalny algorytm postępowania, nie zawsze pozwala na rozwiązanie wszystkich napotkanych problemów kontroli zmiennych.

Tab. 1. Tabela najczęstszych funkcji transformujących zmienne

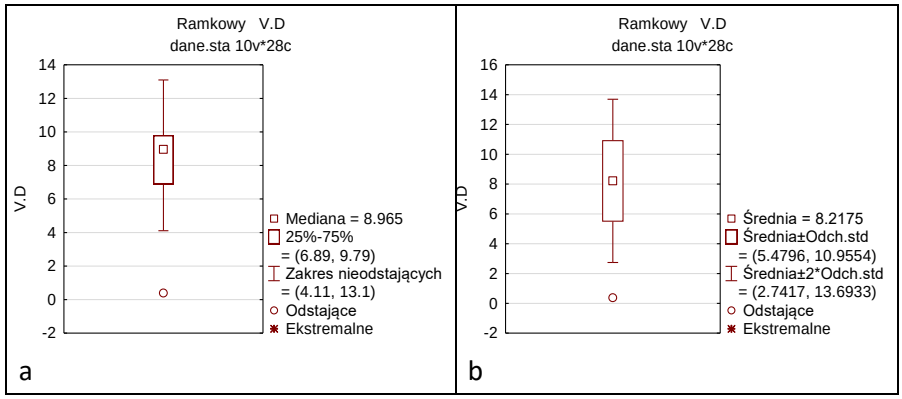
Charakter zmiennej	Przykłady funkcji transformujących
zmienna ma naturalny początek w 0, stosunek MIN/MAX<0.1 i jest silnie prawoskośna	**$x = \log_{10}(x)$,**
zmienna jest silnie prawoskośna	**$x = \log_{10}(x+a)$; $x+a > 0$**
zmienna jest silnie lewoskośna	**$x = \log_{10}(a-x)$; $a > x_{max}$**

Źródło: opr. własne

W tabeli powyżej zamieszczone zostały trzy najczęstsze sytuacje wymagające transformacji zmiennej i funkcje, z jakich należy w danym przypadku skorzystać. Niestety wyznaczenie w sposób analityczny właściwej wartości parametru '*a*' z reguły nie jest możliwe. Dlatego w praktyce stosuje się metodę prób i błędów z zachowaniem koniecznej reguły, aby argument funkcji logarytmicznej był większy od zera. Należy przy tym pamiętać, że wartość poszukiwanego parametru '*a*' może (i z natury musi) być przybliżona. Ma jedynie zapewnić spełnienie warunku braku istotności testów normalności rozkładu transformowanej zmiennej.

Dla rozkładów przynajmniej 'podobnych' do normalnego możemy przeprowadzić wspomniane już wcześniej testy na wartości odbiegające. Warto w tym miejscu wspomnieć o możliwościach graficznych, jakie daje oprogramowanie Statistica w tym zakresie. Przy pomocy prostego i powszechnie stosowanego narzędzia, jakim jest tzw. wykres ramkowy (ang. Box-Whisker) możemy w łatwy sposób wykryć oraz wykluczyć z dalszej analizy wykryte wartości odstające. Na widocznym poniżej przykładzie takich wykresów, dla wcześniej prezentowanych już danych stężenia witaminy D w surowicy, w obu przypadkach analizy, widzimy wartość obarczoną błędem (odstającą od pozostałych). Wartości zakresów dopuszczalnej zmienności badanego parametru liczone są w obu przypadkach inaczej. W pierwszym, na podstawie granic kwartylnych, w drugim, jako przedział ufności (dla współczynnika istotności $\alpha \approx 0.05$) oparty o odchylenie standardowe z próby (pojedynczego pomiaru), co dla ilości stopni swobody = 27 (28 pomiarów) na podstawie wartości parametru krytycznego *t*-Studenta daje

mnożnik dla odchylenia standardowego ≈ 2 (a przedział: 2.7414–13.6933; rysunek 2 b).



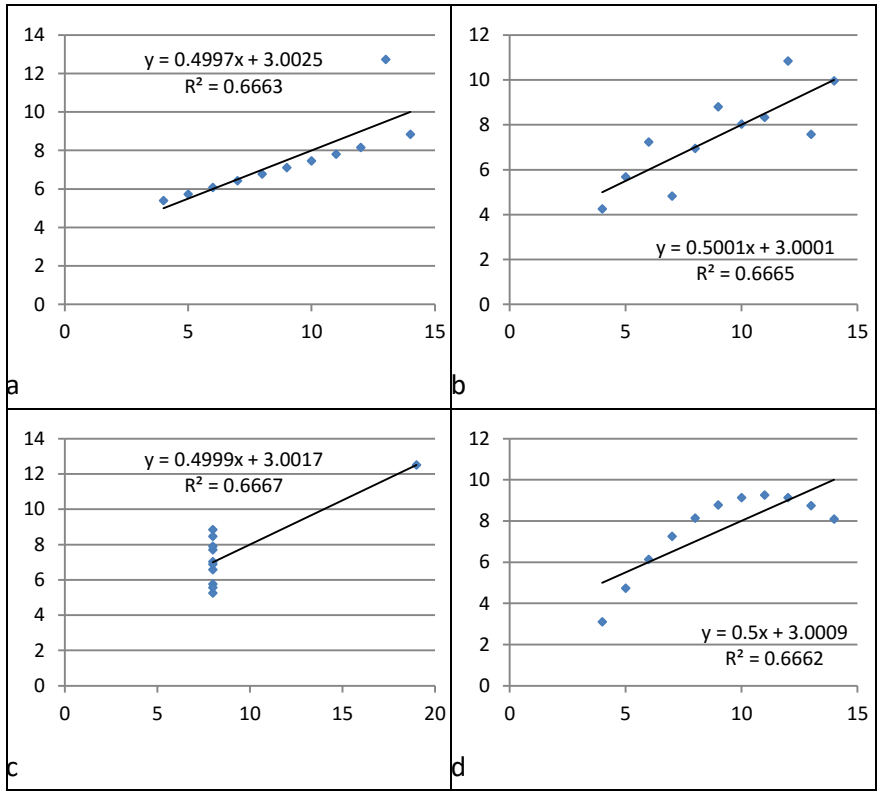
Rys. 2. Wykresy ramkowe (Box-Whisker) wartości odstających dla stężenia witaminy D w surowicy: a – oparty o kwartyle, b – oparty o przedział ufności zm. (2σ, $p_{(27)} \approx 0.95$)

Źródło: opr. własne

2.2.2 Korelacje między zmiennymi

W przypadku danych analizowanych metodami chemometrycznymi mamy zwykle do czynienie z tak zwanymi tablicami szerokimi, posiadającymi ilość kolumn znacznie przewyższającą ilość badanych obiektów (próbek, rekordów). Nigdy w związku z tym nie jest tak, że zmienne niezależne opisujące obiekt, niosą swoistą informację tj. taką, która w najmniejszym stopniu nie pokrywa się z informacją niesioną przez inną zmienną dotyczącą tego obiektu. Rzeczywistość informacji jest wielowymiarowa, ale nie jest ortogonalna. Ta nadmiarowość ma miejsce zwłaszcza, gdy nie mamy wpływu na ortogonalność planu doświadczeń, czyli zwykle wtedy, kiedy dokonujemy eksperymentu na obiektach, których zmiennych zależnych nie kontrolujemy. Dobrym przykładem może tu być pomiar zawartości minerałów, czystości wody w zbiornikach naturalnych. Zdefiniowana w ten sposób współliniowość zmiennych objaśniających (zwykle współczynnikiem korelacji, lub jego kwadratem – współczynnikiem determinacji) jest zjawiskiem wysoce niepożądanym w tworzeniu modeli chemometrycznych, bez względu na to jakiemu celowi mają one służyć. Wymóg reprezentatywności

zmiennych w algorytmach DM oznacza, że zmienne wybrane do tworzonego modelu powinny być ze sobą jak najsłabiej skorelowane. Nadmierna korelacja zmiennych zwykle prowadzi do niestabilności modelu w przestrzeni rozwiązań. Najlepszym tego przykładem może być model regresji wielokrotnej, który w przypadku nadmiarowości informacji niesionej przez zmienne objaśniające prowadzi zwykle do równania prostej, które będzie dobrze szacować wartość prognozowaną nawet wtedy, gdy żadna ze zmiennych nie jest istotna dla zmiennej zależnej (szacowanej). Redukcji niepotrzebnej, nadmiarowej informacji i tym samym redukcji wymiaru przestrzeni zmiennych objaśniających, można dokonać przeprowadzając cały układ zmiennych w ich reprezentację, zwaną czynnikami głównymi (kombinację liniową zmiennych oryginalnych; ang. Principal Component Analysis PCA), ale też badając wstępnie siłę korelacji pomiędzy poszczególnymi zmiennymi metodami statystycznymi.



Rys. 3. Relacje zmiennych – kwartet Anscombe'a

Źródło: opr. własne

Badając zależności między zmiennymi należy zawsze pamiętać, że ich wizualizacja przy dzisiejszych możliwościach obróbki danych nie stanowi żadnego problemu. Same wartości parametrów określających współzależność (współczynnik korelacji) zmiennych nie są dla nas tak wiele mówiące jak rzut oka na wykres. Najlepszym przykładem, przytaczanym w większości pozycji literaturowych zajmujących się tą tematyką jest tzw. kwartet Anscombe'a (Rys. 3).

Jak łatwo zauważyć na zaprezentowanych wykresach, wszystkie zależności cechuje dokładnie taka sama wartość współczynnika korelacji (determinacji), a także identyczne wartości parametrów prostych będących modelem regresji dla zbiorów danych. Na wykresie **'a'** widoczny jest degradujący wpływ błędu grubego lub wartości nietypowej na doskonałą, praktycznie funkcyjną zależność liniową dwóch cech pozostałych obiektów. Z pozornie podobną sytuacją mamy do czynienia w przypadku wykresu **'c'**. Mamy tu wartość odstającą, która wpływa na wartości numeryczne zależności zmiennych, sugerując dość dużą wartość korelacji zmiennych, podczas gdy zależność taka zupełnie nie występuje. Nieco inaczej przedstawia się zależność **'b'.** Może ona być typowym przykładem zależności liniowej punktów doświadczalnych i w jej przypadku wartości numeryczne prawidłowo odzwierciedlają sytuację, z jaką mamy do czynienia. I ostatni przykład – doskonała funkcyjna zależność zmiennych, ale aby ją potwierdzić parametrami regresji liniowej, należy dokonać linearyzacji funkcji, która ją reprezentuje, lub wykorzystać regresję wykładniczą np. arkusza Excela. Ale nie da się tego zauważyć bez wcześniejszej, wizualnej analizy wykresów.

W chemometrii Istnieją pewne zasady dotyczące linearyzacji badanych zależności. Zalecana jest przede wszystkim transformacja zmiennej objaśniającej bez modyfikacji zmiennej zależnej. Pozwala to w łatwy sposób porównywać różne rodzaje transformacji. Transformację zmiennej zależnej stosuje się jedynie w ostateczności. Przykładowe sytuacje i funkcje jakie należy w danych przypadkach stosować przedstawione zostały w tabeli 2.

Tab. 2. Funkcje transformujące dla typowych przypadków linearyzacji zmiennej

Charakter zależności *y = f(x)*	**Funkcja transformująca**
zależność ma asymptotę pionową dla $x_0 = 0$	$x' = \log(x)$ lub $x' = 1/x$
zależność ma asymptotę pionową dla $x_0 = a;\ x_i > a$	$x' = \log(x - a)$ lub $x' = 1/(x - a)$
zależność ma asymptotę pionową dla $x_0 = a;\ x_i < a$	$x' = \log(a - x)$ lub $x' = 1/(a - x)$
zależność ma przebieg sigmoidalny, $a < y_{min}$; $as.\ dolna$ $a > y_{max}$; $as.\ górna$	$y' = \log\left(\frac{y - a}{b - y}\right)$
zależność wzrasta do maximum; y_{max} ; $as.$ $górna$ $b > y_{max}$	$y' = \log(b - y)$
zależność maleje do minimum; y_{min} ; $as.$ $dolna$ a $< y_{max}$	$y' = \log(y - a)$

Źródło: opr. własne

2.2.3 Skalowanie, autoskalowanie (standaryzacja) zmiennej

Jeszcze innego rodzaju transformacjami zmiennych są takie ich przekształcenia, aby ich wartością średnią była wartość bliska zeru, a co ważniejsze, aby niosły one ze sobą porównywalne ładunki informacji. Jest to bardzo częstym i mającym zasadnicze znaczenie wymogiem algorytmów wielu metod chemometrycznych. Pierwszym i najprostszym tego typu zabiegiem, choć chyba najrzadziej stosowanym, jest centrowanie zmiennej, czyli taka jej transformacja liniowa, która sprawia, że jej wartość średnia znajduje się w początku układu współrzędnych (jest równa zero). Jest to warunek konieczny w przypadku takiej metody jak analiza podobieństwa, czy analiza

czynników głównych. Samo centrowanie wykonuje się odejmując od poszczególnych wartości zmiennej jej wartość średnią:

$$X'_{ij} = X_{ij} - \overline{X}_j \tag{2.3}$$

gdzie:
j – jest symbolem zmiennej,
i – jest kolejną jej wartością.

Zasadniczym założeniem typowych przekształceń tego rodzaju jest wspomniana już współmierność zmiennych. Można ją realizować na dwa sposoby: jako skalowanie przedziałowe oraz najczęściej wykorzystywane – autoskalowanie, inaczej nazywane standaryzacją zmiennej. W przypadku skalowania przedziałowego wszystkie wartości danej zmiennej sprowadzane są w sposób proporcjonalny do pewnego, zwykle zawężonego (0-1) przedziału w następujący sposób:

$$X'_{ij} = \frac{X_{ij} - X_j(min)}{X_j(max) - X_j(min)} \tag{2.4}$$

gdzie jak wcześniej:
j – jest symbolem zmiennej,
i – jest kolejną jej wartością.

Ten typ skalowania ma jednak dwie podstawowe wady: wykorzystuje jedynie informację o dwóch wartościach zmiennej – minimum i maksimum, oraz daje złe wyniki w przypadku istnienia wartości odbiegających. Dlatego spośród trzech wymienionych metod skalowania najlepszą jakościowo jest metoda skalowania wariancyjnego zwana autoskalowaniem lub standaryzacją. Tylko ona bowiem, daje gwarancję spełnienia warunku centrowania, współmierności i co bardzo ważne, spełnienia warunku jednakowego zasobu zmienności każdej zmiennej jednocześnie. Zasób zmienności każdej zmiennej zależy od dwóch czynników: jednostek, w jakich wyrażane są zmienne oraz rozkładu wartości zmiennej. Skalowanie przedziałowe eliminuje jedynie wpływ pierwszego z tych czynników, drugi – można wyeliminować wykorzystując do transformacji odchylenie standardowe pojedynczego pomiaru (będące miarą zasobu zmienności) w następujący sposób:

$$X_{ij} = \frac{x_{ij} - \overline{x}_j}{s_j} \quad (2.5)$$

gdzie:

X_{ij} – standaryzowana wartość zmiennej,

s_j – odchylenie standardowe poj. pom.,

j – jest symbolem zmiennej,

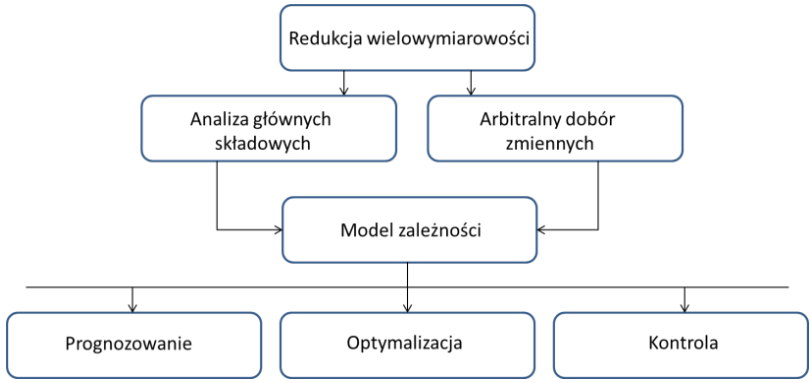
i – jest kolejną jej wartością.

Standaryzacja zmiennych jest zatem transformacją uniwersalną, gdy potrzebujemy, aby zmienne były współmierne i posiadały jednakowy zasób zmienności. Jak wiadomo jej efektem jest wartość średnia zmiennej równa zero a odchylenie standardowe pojedynczego pomiaru (i co oczywiste – wariancja) równa jedności. Porządkując słownictwo dotyczące zmiennych należy w tym miejscu wspomnieć, że oryginalne zmienne poddane skalowaniu lub autoskalowaniu nazywamy zwykle cechami (deskryptorami), aby odróżnić je od zmiennych 'surowych', przed transformacją tego typu.

3 MODELOWANIE ZALEŻNOŚCI – KALIBRACJA

Poznawanie metod chemometrycznych, zasad ich stosowania i sytuacji, w których mogą one przynieść wymierne korzyści, dobrze jest rozpocząć od modelowania zależności, które znane jest ze statystyki klasycznej i odpowiada analizie regresji. Modelowanie zależności jest jednym z podstawowych zastosowań chemometrii i polega na budowie modelu matematycznego, zdolnego przedstawić funkcyjny związek pomiędzy zmienną zależną (szacowaną), a licznym zbiorem zmiennych niezależnych (objaśniających). Stworzony model ma być w tym przypadku narzędziem, które pozwoli na prognozowanie wartości zmiennej zależnej dla zadanych, dowolnych wartości zmiennych objaśniających. Ponadto, może być również wykorzystany do:

- optymalizacji układu, czyli do znalezienia takich wartości zmiennych objaśniających lub ich zakresów, aby zmienna zależna spełniała określone kryteria, przy czym zwykle poszukujemy minimum lub maksimum tej zmiennej,
- do kontroli układu, gdy bezpośrednie wyznaczenie zmiennej zależnej jest pracochłonne lub kosztowne, a zależy nam na szybkiej i taniej ocenie jej wartości, na przykład w trakcie procesu produkcyjnego.



Rys. 4. Tworzenie i główne zastosowania modeli zależności

Źródło: opr. własne

Chemometryczne modele zależności należą do grupy tak zwanych modeli empirycznych i inaczej niż ma to miejsce w statystyce, przy ich tworzeniu nie jest nam potrzebna znajomość teorii opisującej modelowany proces. W zależności od rodzaju (i zwykle ilości) zmiennych poddawanych analizie możemy rozróżnić dwa przypadki modelowania:

- dla niewielkiej liczby kontrolowanych zmiennych objaśniających, dla których mamy możliwość ustalenia z góry ich wartości (np. krzywa wzorcowa),
- dla dużej liczby niekontrolowanych zmiennych objaśniających (kilkadziesiąt lub więcej), na których wartości nie mamy żadnego wpływu.

W przypadku modeli tworzonych dla danych kontrolowanych, otrzymujemy bardzo wiarygodne narzędzie w oparciu o pomiary dla niewielkiej liczby obiektów. Dodatkowo, pomiaru wymaga jedynie odpowiedź badanego obiektu. Całkowicie odmienna sytuacja ma miejsce w przypadku danych niekontrolowanych. Aby uzyskać wiarygodne wyniki i rzetelny model takiego obiektu, należy dysponować znacznie większą liczbą obiektów pomiarowych, a także wiedzą na temat wewnętrznej struktury tego zbioru. Dodatkowo pomiary muszą zostać wykonane zarówno dla zmiennej zależnej jak i zestawu zmiennych niezależnych dla obiektu. Modeli tworzonych dla ($m > 1$) wielowymiarowego zestawu zmiennych niezależnych dotyczy jeszcze problem wzajemnych korelacji pomiędzy nimi. Da przypadku zmiennych kontrolowanych jest on bardzo łatwy do rozwiązania. Zwykle wykorzystuje się możliwość ortogonalizacji wektorów (kolumn macierzy) zmiennych dla obiektów stosując mniej lub bardziej złożone plany czynnikowe (o czym będzie mowa w dalszej części rozdziału). Takiej możliwości nie mamy w przypadku zmiennych niekontrolowanych. Aby choć częściowo wyeliminować problem możliwej silnej korelacji dla takich zmiennych (niepożądanym efektem dla modelu jest wtedy jego niska zdolność prognozowania), można wykorzystać dwie podstawowe procedury. Jedną z nich, najczęściej stosowaną jest wybór odpowiedniego zestawu zmiennych diagnostycznych na podstawie analizy wzajemnych korelacji zmiennych, drugą – analiza PCA (głównych składowych), czyli zamiana pierwotnej ilości m skorelowanych zmiennych, na p z założenia wzajemnie ortogonalnych czynników (nowych, sztucznych zmiennych). Podsumowując, tworząc model zależności, poszukujemy do jego realizacji takiego zestawu zmiennych, który będzie niósł ze

sobą jak najwięcej informacji o zmienności obiektów, a przy okazji będzie zestawem jak najmniej licznym. Spełnienie takich warunków zapewnia ortogonalność zmiennych niezależnych (plany czynnikowe, PCA) lub ich niewielka korelacja (przy arbitralnym wyborze właściwego zestawu dla zmiennych niekontrolowanych). Okazuje się przy tym zwykle, że liczba naprawdę istotnych składowych jest dużo mniejsza niż początkowa liczba zmiennych ($p << m$).

3.1 Modele numeryczne

Jak już wiemy, aby móc przewidzieć zachowanie się obiektu, czyli móc przewidzieć jego odpowiedź na zadane warunki, musimy stworzyć matematyczny model badanego zjawiska:

$$y = f(x_1, x_2, \ldots, x_m) \tag{3.1}$$

W zależności od posiadanej na jego temat wiedzy (funkcji, parametrach) możemy mówić o trzech typach modelowania – trzech typach modeli:

- model w pełni określony

  znamy postać matematyczną funkcji i wartości wszystkich występujących w niej parametrów. Modelami tymi zajmują się nauki podstawowe (prawa fizyczne: grawitacja, elektromagnetyzm itp.);

- model półempiryczny

  znamy z nauk podstawowych postać zależności funkcyjnych, lecz dla konkretnego obiektu brakuje nam informacji o jego parametrach (np. stałej dysocjacji kwasu, czy stałej szybkości danej reakcji chemicznej);

- model empiryczny

  nie znamy zależności funkcyjnych lub są one na tyle skomplikowane, że nie nadają się do zbudowania modelu. Oczywiście, nie znając postaci funkcji nie znamy również jej parametrów.

Modelem interesującym z punktu widzenia chemometrii jest model najtrudniejszy w realizacji – model empiryczny. Nie posiadając zatem żadnej informacji na temat modelowanego zjawiska musimy postawić sobie dwa pytania, na które należy spróbować udzielić odpowiedzi:

- jaka jest postać matematyczna funkcji 3.1,
- jakie są wartości pewnych stałych, zwanych parametrami modelu (funkcji).

Wydaje się to być bardzo karkołomnym zadaniem, jednakże chemometria i jej algorytmy dostarczają nam metod, które potrafią sobie z nim poradzić. Niektóre z nich (np. sztuczne sieci neuronowe) nie odpowiadają nam na zadane pytania wprost. Nigdy nie poznajemy ani funkcji ani tym bardziej jej parametrów pozwalających na modelowanie zjawiska. Są one wprawdzie zakodowane w parametrach samego modelu, ale próby ich bezpośredniego wykorzystania przyniosłoby najczęściej zbyt skomplikowane rozwiązanie, którego uproszczenie z kolei mogłoby znacznie pogorszyć zdolności prognostyczne modelu.

Doświadczenie zdobyte przez nauki przyrodnicze dostarcza nam jeszcze innej możliwości rozwiązywania tego typu problemów. Możemy bowiem spodziewać się, że nasz model może opisywać znana już funkcja z modelu półempirycznego lub w pełni określonego. Modele te stosują zależności funkcyjne będące tzw. funkcjami porządnymi, co oznacza, że są to funkcje ciągłe i różniczkowalne. O funkcjach takich wiemy, że w dostatecznie małym przedziale, każdą z nich możemy przybliżyć wielomianem niskiego stopnia, przy czym **im gładsza jest stosowana funkcja i im mniejszy przedział tym stopień wielomianu może być niższy**. Te dwa spostrzeżenia stanowią podstawę modelowania chemometrycznego i są uzasadnieniem jego uproszczeń. Modelowanie rzeczywistych zjawisk wielomianami niskich stopni ma dodatkową zaletę: pozwala budować modele empiryczne dla niewielkiej liczby pomiarów (doświadczeń).

Tworzenie najprostszego modelu, w którym mamy do czynienia z jedną zmienną niezależną zaczynamy od równania liniowego najprostszej postaci:

$$y = ax + b \tag{3.2}$$

W przypadkach, gdy nie opisuje ono dostatecznie dobrze naszego obiektu możemy zastosować funkcję kwadratową,

$$y = ax^2 + bx + c \tag{3.3}$$

lub podzielić przedział zmiennej objaśniającej na kilka mniejszych, budując różne modele liniowe dla każdego z nich. Jednak w większości przypadków, chemometryczne modele empiryczne to odpowiedniki zjawisk bardziej złożonych, a ich odpowiedź jest wynikiem wpływu znacznie większej liczby niezależnych parametrów. Modele takie budowane są oczywiście z wykorzystaniem funkcji wielu zmiennych. Najprostszym analitycznym opisem zjawiska wieloparametrycznego jest model wielomianowy stopnia pierwszego – model liniowy postaci:

$$\mu = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_m x_m \tag{3.4}$$

gdzie:

μ – zmienna zależna reprezentująca charakterystyczną cechę badanego zjawiska,

β_j – współczynniki, które należy wyznaczyć (j = 0, 1, ..., m),

x_j – zmienne niezależne (objaśniające) (j = 1, 2, ..., m).

Model taki w wielu przypadkach (bardziej złożonych) może okazać się niewystarczający. Pierwszym stopniem jego rozbudowy w praktyce jest tak zwany model liniowy z **interakcjami stopnia pierwszego**. Rozszerzony model liniowy stopnia pierwszego dla np. trzech zmiennych niezależnych zapisujemy w postaci:

$$\mu = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{23} x_2 x_3 \tag{3.5}$$

lub ogólnie (dla dowolnej liczby zmiennych) jako:

$$\mu = \beta_o + \sum_{i=1}^{m} \beta_i x_i + \sum_{i=1}^{m} \sum_{j>1}^{m} \beta_{ij} x_i x_j \tag{3.6}$$

Kolejnym sposobem rozwinięcia modelu liniowego jest uproszczony **model kwadratowy** zawierający oprócz członów liniowych również człony kwadratowe zmiennych niezależnych:

$$\mu = \beta_o + \sum_{i=1}^{m} \beta_i x_i + \sum_{i=1}^{m} \beta_{ii} {x_i}^2 \tag{3.7}$$

Model kwadratowy można dalej rozszerzać, dodając człony interakcyjne i otrzymując model kwadratowy z interakcjami, zwany także **rozszerzonym modelem wielomianowym stopnia drugiego**:

$$\mu = \beta_o + \sum_{i=1}^{m} \beta_i x_i + \sum_{i=1}^{m} \sum_{j>1}^{m} \beta_{ij} x_i x_j + \sum_{i=1}^{m} \beta_{ii} {x_i}^2 \tag{3.8}$$

Model 3.8 to w zasadzie najbardziej skomplikowany model regresji wykorzystywany w chemometrii. Jak łatwo zauważyć wraz ze wzrostem stopnia skomplikowania modelu rośnie ilość współczynników beta, jakie musimy wyznaczyć, aby model był pełny. Zmusza nas to do przeprowadzenia przynajmniej takiej ilości pomiarów, jaka jest ilość wyznaczanych parametrów. Należy też zdawać sobie sprawę, że **regresyjna metoda wyznaczania parametrów modelu może być zastosowana tylko w przypadku, gdy modele są liniowe ze względu na wyznaczane współczynniki (współczynniki β). Dodatkowo, żadna zmienna objaśniająca modelu, nie może być liniową kombinacją jakiejkolwiek innej zmiennej niezależnej modelu**.

Minimalne ilości (*m*) pomiarów koniecznych do zidentyfikowania prezentowanych wcześniej modeli numerycznych zawiera tabela 3.

Liczba koniecznych pomiarów bardzo szybko wzrasta wraz z ilością zmiennych objaśniających tworzących model. W praktyce, zgodnie z regułami statystyki doświadczalnej, dla wyznaczenia wszystkich parametrów modelu regresji na zadowalającym poziomie ufności konieczna jest ilość pomiarów przynajmniej cztero–, a lepiej pięciokrotnie większa niż przedstawiana w tabeli. Przykładowo, dla modelu liniowego z czterema zmiennymi objaśniającymi, należy wykonać co najmniej 20–25 pomiarów. Dzięki sta-

rannemu, zgodnemu z regułami chemometrii doborowi punktów pomiarowych, dla dowolnego modelu można uzyskać wystarczająco dobre oszacowanie współczynników dla znacznie mniejszej liczby pomiarów. Wymaga to jedynie ich właściwego zaplanowania, to znaczy właściwego ich rozmieszczenia w przestrzeni zmiennych objaśniających. Takich możliwości dostarczają nam chemometryczne plany doświadczeń.

Tab. 3. Minimalna, konieczna liczba pomiarów dla danego modelu.

Model	Liniowy	Liniowy z interakcjami	Kwadratowy	Interakcyjny kwadratowy
	m + 1	m+m(m–1)/2+1	2m + 1	2m+m (m–1)/2+1
m = 1	2	2	3	3
2	3	4	5	6
3	4	7	7	10
4	5	11	9	15
5	6	16	11	21
6	7	22	13	28
7	8	29	15	36

Źródło: opr. własne

3.2 Planowanie doświadczeń

Największym błędem prowadzenia badań jest sytuacja, gdy bez dobrze przemyślanego planu eksperymentu wykonana zostanie seria często kosztownych pomiarów i na jej podstawie oczekiwane jest potwierdzenie pewnych hipotez. W takim przypadku nawet najlepsze metody analizy danych nie pozwalają na właściwą ocenę wyników i na uzyskanie na ich podstawie poszukiwanej informacji. Istotne jest zatem odpowiednie zaplanowanie pomiarów, ich wykonanie zgodnie z prawidłami sztuki i zachowaniem zasad metrologii chemicznej. Jedynie w takim przypadku przeprowadzona dalsza analiza chemometryczna umożliwi uzyskanie prawidłowych wyników oraz wyciągnięcie poprawnych wniosków.

Planowanie doświadczeń na gruncie chemometrii łączy się zawsze z dwoma najważniejszymi, na jakie koniecznie musimy zwrócić uwagę, aspektami tego procesu: optymalną liczbą pomiarów i właściwym (możliwie najlepszym) rozmieszczeniem punktów pomiarowych. Optymalizacja liczby pomiarów podyktowana jest naturalną tendencją obniżania kosztów oraz czasochłonności badań. W przeważającej większości przypadków zalecane jest zaplanowanie nieco większej liczby pomiarów niż minimalna. Konieczność taka spowodowana jest nieuniknioną niepewnością pomiarów. Liczba nadmiarowych pomiarów, powyżej koniecznego minimum, nazywana jest w chemometrii i statystyce liczbą **stopni swobody**. Z natury, im większa jest ta liczba, tym skuteczniej możemy ograniczyć wpływ niepewności pomiaru na jakość wyniku. Dlatego zawsze szukać należy kompromisu pomiędzy dokładnością uzyskiwanych wyników a ich kosztami. Praktyka chemometryczna wskazuje, że optymalna liczba stopni swobody powinna kształtować się w granicach od 4 do 10.

Istnieje jeszcze jeden powód zmuszający nas do dalszego zwiększenia liczby pomiarów. Jest nim konieczność sprawdzenia poprawności stworzonego modelu, jego walidacji. Jeśli minimalną ilość pomiarów zwiększymy o ilość stopni swobody, to zbiór taki nazywamy zbiorem uczącym. Aby poddać model walidacji potrzebujemy zbioru testowego pomiarów i zwykle wynosi on około 10% wielkości zbioru uczącego. Jednak nie mniej niż 5 pomiarów. Taką nieskomplikowaną sytuację mamy, gdy tworzymy nasz model na podstawie zmiennych kontrolowanych. Zdarza się jednak, że w pewnych szczególnych przypadkach dla uzyskania wiarygodnych wyników potrzebna jest liczba obiektów kilkukrotnie większa niż wynikałoby to z samej natury problemu. Zdarza się to, gdy wartości zmiennych pochodzą z pomiarów, dla których niepewność jest wielkością tego samego rzędu co ich zmienność. Z taką sytuacją mamy często do czynienia w przypadku zmiennych niekontrolowanych, dlatego zasady budowania planów optymalnych **mogą być omawiane dla modeli opartych o zmienne kontrolowane**. W praktyce chemicznej, w naszych laboratoriach może to odpowiadać np. sytuacji wzorcowania elektrody szklanej.

Drugim ważnym aspektem modelowania wielowymiarowego, o czym wspomniano, jest rozmieszczenie punktów pomiarowych w przestrzeni

zmiennych. Zoptymalizowanie pomiarów pod tym kątem pozwala na uzyskanie dobrej jakości modelu z jednoczesną minimalizacji ilości punków pomiarowych. Nie istnieje jeden uniwersalny sposób rozwiązania tego problemu. Każdy z modeli chemometrycznych najczęściej wymaga odmiennego sposobu testowania przestrzeni zmiennych (planu optymalnego). Zależności uzyskiwane dzięki zastosowaniu modeli numerycznych mają charakter interpolacyjny, dlatego też punkty pomiarowe w przestrzeni zmiennych powinny obejmować cały interesujący nas zakres zmienności każdej cechy. Wynika z tego, że pomimo różnorodności modeli matematycznych, można przyjąć jedną, ogólną zasadę: wybór punktów pomiarowych zwykle powinien dotyczyć krańców przedziałów zmienności. Nie musimy również tworzyć nowego planu doświadczalnego od początku dla każdego testowanego przez nas empirycznego modelu zależności. Istnieją bowiem gotowe rozwiązania, opracowane wcześniej, z których możemy skorzystać. Skorzystać, to znaczy wyznaczyć położenie naszych rzeczywistych punktów pomiarowych na podstawie tzw. zmiennych planu. Zmienne planu, zwykle z przedziału –1 do 1 tworzą **plan optymalny,** czyli plan, który przy danej liczbie punktów doświadczalnych zapewni największą wiarygodność uzyskiwanego rozwiązania – najbardziej wiarygodną przewidywaną wartość zmiennej zależnej.

Przed zbudowaniem od podstaw takiego planu dla konkretnej zależności matematycznej najpierw ustalamy zadowalający nas stopień wiarygodności oczekiwanego rozwiązania. Determinuje to konieczną liczbę stopni swobody a więc liczbę dodatkowych (ponad minimum) pomiarów. Dopiero teraz możemy podjąć próbę rozmieszczenia wszystkich punktów w przestrzeni zmiennych.

3.3 Plany optymalne

Jak tworzy się plan optymalny i jakie parametry pozwalają nam ocenić czy jest on rzeczywiście najlepszy, tj. taki, aby model opracowany na jego podstawie dawał możliwie najbardziej wiarygodne wyniki – prognozy, najłatwiej prześledzić jest na przykładzie prostego modelu liniowego jednej zmiennej.

$$\mu = \beta_0 + \beta_1 x_1 \tag{3.9}$$

Rozwiązanie problemu polega na wyznaczeniu wartości dwóch współczynników: β_0 i β_1. Ponieważ model empiryczny z natury jest jedynie przybliżonym opisem obiektu, a doświadczalne (zmierzone) wartości zmiennej zależnej są przybliżeniem jego rzeczywistej odpowiedzi, współczynniki te możemy jedynie oszacować, co zapisujemy:

$$y = b_0 + b_1 x_1 + e \tag{3.10}$$

gdzie:

e – niepewność pomiaru

Do rozwiązania tak postawionego problemu wykorzystywana jest zwykle metoda regresji, która co warto powtórzyć, wymaga spełnienie warunku liniowości modelu ze względu na szacowane współczynniki β_i (i = 0, 1, 2 …, k). Nie jest przy tym ważne, czy zachowana jest liniowość ze względu na zmienne niezależne. Mogą one występować jako argumenty funkcji nieliniowych jak logarytm, pierwiastek czy funkcja wykładnicza. Szacowane współczynniki β_i argumentami takich funkcji być nie mogą. Metoda regresji wymaga od nas również, aby dowolna zmienna niezależna nie była liniową kombinacją jakichkolwiek innych zmiennych objaśniających ze zbioru.

W praktyce doświadczalnej, dla dowolnego liniowego modelu wielu zmiennych otrzymujemy zwykle serię pomiarów, którą możemy przedstawić w postaci tablicy, jako zestaw wartości zmiennej zależnej y_n oraz wartości zmiennych niezależnych x_{nm}.

$$\begin{array}{cccccc} y_1 & 1 & x_{11} & x_{12} & \dots & x_{1m} \\ y_i & 1 & x_{i1} & x_{i2} & \dots & x_{im} \\ y_n & 1 & x_{n1} & x_{n2} & \dots & x_{nm} \end{array} \tag{3.11}$$

z każdego wiersza takiej tablicy możemy ułożyć równanie dla zmiennej zależnej obciążonej niepewnością pomiaru, opisujące model:

$$y_n = b_0 + b_1 x_{n1} + b_2 x_{n2} + b_3 x_{n3} + \cdots + b_m x_{nm} + e \tag{3.12}$$

otrzymując układ *n* równań, których wygodnym w dalszych rozważaniach zapisem jest zapis macierzowy postaci:

$$\begin{bmatrix} y_1 \\ y_i \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & x_{1m} \\ 1 & x_{i1} & x_{i2} & x_{im} \\ 1 & x_{n1} & x_{n2} & x_{nm} \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ b_m \end{bmatrix} + e \tag{3.13}$$

lub w skrócie:

$$\mathbf{y} = \mathbf{Xb} + \mathrm{e} \tag{3.14}$$

gdzie:

y – wektor kolumnowy wartości zmiennej zależnej

X – macierz danych (z kolumną jedynek dla wyznaczenia wyrazu wolnego)

b – wektor kolumnowy parametrów modelu

e – niepewność pomiaru

Przy czym aby istniała możliwość oszacowania współczynników równania 3.12 musi zachodzić warunek $n \geq m + 1$. Warunek ten jest niczym innym jak koniecznością przeprowadzenia minimalnej ilości pomiarów, zależnej od ilości szacowanych parametrów modelu. Macierz **X** jest z kolei zapisem rozmieszczenia punktów pomiarowych, testujących przestrzeń zmienności danych objaśniających, czyli rzeczywistym planem doświadczenia. **Jest ona zawsze uzupełniana z lewej strony kolumnowym wektorem jedynek pozwalającym na wyznaczenie wyrazu wolnego (b_0) w modelu**.

Przystępując do przykładowych poszukiwań planu optymalnego dla naszego wcześniej zdefiniowanego modelu opisanego zależnością (3.9), musimy pamiętać, że ma on służyć poprawnemu wyznaczeniu parametrów β_i modelu. Kryterium jego poprawności musi zatem dotyczyć wiarygodności wyznaczanych parametrów. Z teorii analizy regresji wiemy, że taką wielkością może być wariancja każdej z wartości β_i. Można ją opisać zależnością:

$$s_{b_i}{}^2 = c_{ii}s^2 \tag{3.15}$$

gdzie:

s^2 – wariancja resztowa modelu (wariancja wartości wektora kolumnowego **y**)

c_{ii} – i–ty element głównej przekątnej macierzy **dyspersji**, którą zapisujemy: $(\mathbf{X}^T\mathbf{X})^{-1}$

i definiujemy jako **macierz odwrotną macierzy informacji** ($\mathbf{X}^T\mathbf{X}$).

Okazuje się, że na wartość wariancji liczonej dla każdego z parametrów β_i znacznie większy wpływ ma wartość wielkości c_{ii} niż wariancja resztowa modelu. Wartości elementów głównej przekątnej macierzy dyspersji zależą tylko i wyłącznie od liczby punktów pomiarowych i ich położenia w przestrzeni zmiennych. Nasze zadanie sprowadza się zatem do znalezienia takiego rozmieszczenia punktów doświadczalnych, to jest takiej macierzy planu **U**, dla której wartości elementów z głównej przekątnej macierzy dyspersji $(\mathbf{U}^T\mathbf{U})^{-1}$ będą jak najmniejsze. Definiując ten warunek, na podstawie teorii rachunku macierzowego będziemy mogli określić dalsze wymogi, jakie będzie musiała spełnić macierz planu optymalnego. Pierwszym z nich jest jak największa wartość wyznacznika macierzy informacji det($\mathbf{U}^T\mathbf{U}$). Oznacza to, że musimy zadbać o **ortogonalność** kolumn macierzy planu (**U**) (wektory są ortogonalne, jeśli ich współczynnik korelacji wynosi zero) oraz o to, aby wariancja każdego wektora kolumnowego zmiennej niezależnej była jak największa. Spełnienie tych dwóch warunków gwarantuje niskie wartości głównej przekątnej macierzy dyspersji c_{ii}, czyli jednocześnie niskie wartości wariancji parametrów modelu β_i.

Ortogonalność zmiennych planu charakteryzują jeszcze dwie cechy bardzo pożądane w procesie tworzenia planów doświadczeń. Jedną z nich jest ortogonalność iloczynu zmiennych, z których powstał. Zatem jeśli wszystkie zmienne w planie są ortogonalne to wszystkie ich iloczyny (odpowiadające członom interakcyjnym) dają kolejne wektory ortogonalnych zmiennych. Drugą jest to, że wartości oszacowanych parametrów modelu są od siebie niezależne. Oznacza to, że nieistotne **zmienne można usuwać z modelu bez konieczności ponownych obliczeń pozostałych jego parametrów**.

Mając na uwadze przedstawione zasady tworzenia planu optymalnego załóżmy, że tworzymy go dla pięciu punktów pomiarowych i zależności:

$$\mu = \beta_0 + \beta_1 x_1 \tag{3.16}$$

Przyjmując zasadę, że plany optymalne buduje się dla zakresu zmiennych od –1 do 1 (łatwo potem przekształcić je do planu opartego o rzeczywiste zmienne), dbając o ortogonalność wektorów kolumnowych macierzy planu rozmieśćmy punkty planu symetrycznie względem jego początku – zera.

$$U = \begin{bmatrix} 1 & -1.0 \\ 1 & -x \\ 1 & 0 \\ 1 & +x \\ 1 & 1.0 \end{bmatrix} \tag{3.17}$$

Dlaczego w ten sposób? Ponieważ wiemy, że w przypadku nieparzystej liczby obiektów, jeden z pomiarów należy wykonać w środku przedziału zmienności. Wiemy też, że dobre cechy prognostyczne modelu zapewnia mu rozmieszczenie punktów pomiarowych na krańcach przedziałów. Symbol *x* natomiast, to wartości zmiennej niezależnej reprezentującej współrzędną punktu, co do którego nie mamy pewności gdzie powinien leżeć. Traktując *x* jako zmienną, można spróbować wyznaczyć jej wartość maksymalizując wyznacznik macierzy informacji ($\mathbf{U}^{\mathrm{T}}\mathbf{U}$):

$$(\boldsymbol{U}^T\boldsymbol{U}) = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ -1.0 & -x & 0 & +x & 1.0 \end{bmatrix} \begin{bmatrix} 1 & -1.0 \\ 1 & -x \\ 1 & 0 \\ 1 & +x \\ 1 & 1.0 \end{bmatrix} = \begin{bmatrix} 5 & 0 \\ 0 & 2 + 2x^2 \end{bmatrix} \tag{3.18}$$

$$det(\boldsymbol{U}^T\boldsymbol{U}) = \begin{vmatrix} 5 & 0 \\ 0 & 2 + 2x^2 \end{vmatrix} = 10 + 10x^2 \tag{3.19}$$

Dla wartości *x* z zakresu –1 do 1, tylko wartość |1| maksymalizuje wartość wyznacznika. Zatem nasz plan optymalny powinien mieć postać:

$$U = \begin{bmatrix} 1 & -1.0 \\ 1 & -1.0 \\ 1 & 0 \\ 1 & 1.0 \\ 1 & 1.0 \end{bmatrix} \qquad (\boldsymbol{U}^T\boldsymbol{U}) = \begin{bmatrix} 5 & 0 \\ 0 & 4 \end{bmatrix} \qquad det(\boldsymbol{U}^T\boldsymbol{U}) = 20 \quad (3.20)$$

Podstawiając za *x* jakąkolwiek inną niż wyznaczona wartość ze 'znormalizowanego' przedziału –1 do 1 zawsze uzyskamy gorszy plan doświadczalny. Należy w tym miejscu zwrócić uwagę, że wariancja zmiennej dla takich wartości jest również maksymalna i równa:

$$s^2 = \frac{\sum(x-\overline{x})^2}{n} = 0.8 \qquad (3.21)$$

Natomiast macierz dyspersji na głównej przekątnej posiada najmniejsze wartości z możliwych:

$$(\boldsymbol{U}^T\boldsymbol{U})^{-1} = \begin{bmatrix} 0.2 & 0 \\ 0 & 0.25 \end{bmatrix} \qquad (3.22)$$

Nie uzyskamy też lepszego planu tworząc go z punktów rozmieszczonych niesymetrycznie względem jego środka, nawet przy zachowaniu ortogonalności zmiennych. Dobrym podsumowaniem tego przykładu będzie wniosek, który można uogólnić na inne modele liniowe: **przy nieparzystej liczbie punktów pomiarowych, jeden należy umieścić w środku planu a pozostałe równomiernie rozłożyć na jego krańcach; w przypadku liczby parzystej – punkty pomiarowe rozmieszczamy po połowie na krańcach planu**.

3.4 Plany czynnikowe 2^m

Plan, którego przykład zaprezentowany został w poprzednim rozdziale jest typem planu 2^m. Jego symbol wskazuje na liczbę punktów pomiarowych planu, równą właśnie 2^m, gdzie m jest liczbą zmiennych niezależnych reprezentujących obiekt. Dodatkowo zapis ten symbolizuje, że wartości

zmiennych planu występują tylko na dwóch poziomach –1 i 1. Tego typu plany optymalne stworzone zostały tylko dla modeli liniowych i modeli z interakcjami.

Tab. 4. Porównanie liczby współczynników kierunkowych w różnych modelach liniowych z liczbą punktów planu 2^m

Model	Liniowy	Liniowy z interakcjami	Liczba punktów planu 2^m
	m + 1	m+m(m–1)/2+1	
m = 1	2	2	2
2	3	4	4
3	4	7	8
4	5	11	16
5	6	16	32
6	7	22	64
7	8	29	128

Źródło: opr. własne

W planie czynnikowym typu 2^m, punkty pomiarowe rozmieszczone są we wszystkich narożach m–wymiarowej przestrzeni planu. Sposób tworzenia planu opartego o liniowy model zależności (bez interakcji) dla dowolnej liczby zmiennych jest prosty i polega na powtarzaniu takiego oto schematu czynności:

– zaczynamy od skrajnej, prawej kolumny macierzy planu wstawiając w nią wartości 1 i -1 (część zapisu 3.23 **a**) co odpowiada planowi 2^1

$$\begin{matrix} 1 \\ -1 \end{matrix}\ \mathbf{a} \qquad \begin{matrix} 1 & 1 \\ 1 & -1 \\ -1 & 1 \\ -1 & -1 \end{matrix}\ \mathbf{b} \qquad \begin{matrix} 1 & 1 & 1 \\ 1 & 1 & -1 \\ 1 & -1 & 1 \\ 1 & -1 & -1 \\ -1 & 1 & 1 \\ -1 & 1 & -1 \\ -1 & -1 & 1 \\ -1 & -1 & -1 \end{matrix}\ \mathbf{c} \qquad (3.23)$$

– dalej (część zapisu 3.23 **b**) kopiujemy wpisane już wartości w tej samej kolumnie poniżej, a w kolumnie po lewej stronie wpisujemy do czterech sąsiednich komórek 1, 1, -1, -1. Otrzymujemy plan czynnikowy dla dwóch zmiennych niezależnych 2^2.
– aby powiększyć plan o kolejną zmienną (do planu 2^3) wykonujemy jeszcze raz czynności opisane w punkcie poprzednim, tj. kopiujemy wszystkie osiem wpisanych już w macierz wartości w komórki poniżej wypełnionych (część zapisu 3.23 **c**) a kolumnę po lewej stronie wypełniamy po połowie wartościami 1 i -1 (1, 1, 1, 1, -1, -1, -1, -1).

Jeśli przyjrzymy się wartościom liczbowym w tabeli powyżej zauważymy, że ilości punktów pomiarowych liczonych według reguły 2^m nie stanowią optimum (zgodnego z regułami opisanymi wcześniej) dla małych $m \leq 2$ i większych $m > 4$ wymiarów przestrzeni zmiennych. W przypadku niskich wartości, liczba punktów pomiarowych jest absolutnym minimum i zapewnia jedynie wyznaczenie wartości parametrów modelu. Wyznaczone w taki sposób wartości β_i zawierają całą niepewności pomiaru, i dla stopnia swobody równego 0 nie da się jej wyliczyć. Należy zatem zwiększyć liczbę punktów pomiarowych. W przypadku, gdy plan na to nie pozwala (m = 1, m = 2), pomiarów dokonuje się kilkukrotnie na krańcach przedziału, gdy chcemy zwiększyć ich liczbę o wartość parzystą i w jego środku, gdy chcemy dodać tylko jeden punkt pomiarowy. **Tylko taki sposób rozbudowy planu nie zaburza jego ortogonalności**.

Z odmienną sytuacją mamy do czynienia, gdy wartość $m > 4$. Wtedy, obliczona na podstawie formuły liczba pomiarów jest większa niż uzasadniona regułami chemometrii. Jeśli wykonamy jedynie część z nich, zmienne przestaną być względem siebie ortogonalne, a plan doświadczenia będzie gorszy niż optymalny. Rozwiązaniem tego problemu są **ułamkowe plany czynnikowe**, oznaczane symbolicznie jako 2^{m-k}. W planach czynnikowych tego typu wykorzystywana jest jedna z cech macierzy o ortogonalnych wektorach kolumnowych, mianowicie, że iloczyn dwóch wybranych wektorów jest kolejnym, ortogonalnym wektorem w zestawie zmiennych. Pozwala to tworzyć plany optymalne dla większej liczby zmiennych z ilością punktów pomiarowych odpowiadającą planom dla mniejszej wartości m. Tę wydawać by się mogło skomplikowaną sytuację najlepiej jest zobrazować przykładem:

Załóżmy, że zamierzamy opisać nasz obiekt (próbkę) z pięciowymiarowej przestrzeni zmiennych za pomocą liniowej zależności bez interakcji, postaci:

$$\mu = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 \qquad (3.24)$$

Ponieważ musimy oszacować wartości sześciu współczynników modelu, optymalną z punktu widzenia chemometrii ilością pomiarów mogłaby być liczba z zakresu np. 10–12. Model typu 2^m dla pięciu zmiennych objaśniających określa 32 punkty pomiarowe, we wszystkich narożnikach (krańcach przedziałów zmienności) przestrzeni zmiennych. Jest to duży nadmiar w porównaniu z naszymi założeniami optymalnej ich ilości. Pierwszym etapem postępowania w rozwiązaniu tego problemu jest znalezienie takiego planu 2^p, który zakłada liczbę pomiarów jak najbliższą liczbie parametrów, które musimy wyznaczyć (6), lecz nie mniejszą. Najbliższym planem spełniającym ten warunek jest plan dla trzech zmiennych niezależnych 2^3 dający 8 punktów pomiarowych. Plan ten dostarcza nam 4 ortogonalnych kolumn, których liczbę można zwiększyć o kolejne 4 tworząc nowe, będące ich iloczynami:

Tab. 5. Macierz planu czynnikowego 2^3 z kolumnami dla członów interakcyjnych

U_0	U_1	U_2	U_3	U_1U_2	U_1U_3	U_2U_3	$U_1U_2U_3$
1	1	1	1	1	1	1	1
1	1	1	-1	1	-1	-1	-1
1	1	-1	1	-1	1	-1	-1
1	1	-1	-1	-1	-1	1	1
1	-1	1	1	-1	-1	1	-1
1	-1	1	-1	-1	1	-1	1
1	-1	-1	1	1	-1	-1	1
1	-1	-1	-1	1	1	1	-1

Źródło: opr. własne

Na podstawie tak przygotowanej macierzy planu możemy wybrać dwie dodatkowe, brakujące ortogonalne kolumny dla naszego planu ułamkowego pięciu zmiennych. Mogą to być dowolne kolumny iloczynów zmiennych wyjściowych planu 2^3 (np. zaznaczone prostokątem). Należy w tym miejscu podkreślić, że w przypadku planów ułamkowych dalsza ich rozbudowa (o kolejne kolumny zmiennych – iloczyny już istniejących) nie jest już możliwa. Można natomiast dokonać dodatkowych pomiarów. Najbezpieczniejszym sposobem ich wykonania jest w naszym, analizowanym przypadku środek przedziału zmienności, czyli punkt [1,0,0,0,0,0].

3.5 Ocena modelu

Miarą dopasowania wyjść stworzonego modelu do odpowiedzi badanego obiektu, zgodnie z założeniami regresji jest suma kwadratów różnic (SKR) pomiędzy tymi wielkościami:

$$SKR = \sum_{i=1}^{n}(y - \tilde{y})^2 \tag{3.25}$$

gdzie:

$\tilde{y}$ – odpowiedź modelu;

y – wartości zmierzone doświadczalnie.

Warunkiem koniecznym, jaki musi spełniać ta wielkość przy wyznaczaniu współczynników β_i modelu jest jej minimum, stąd inna nawa metody, 'metoda najmniejszych kwadratów'.

Jeśli zależność pomiędzy odpowiedzią modelu i jego wyjściami zapiszemy w postaci macierzowej jako:

$$\boldsymbol{y} = \tilde{\boldsymbol{y}} + \boldsymbol{e} = \boldsymbol{Xb} + \boldsymbol{e} \tag{3.26}$$

to, pomijając niepewność pomiaru (którą wyznaczymy na podstawie nadmiarowych ponad minimum pomiarów), aby wyznaczyć wartości wektora kolumnowego współczynników **b** musimy rozwiązać równanie macierzowe postaci:

$$\boldsymbol{y} = \boldsymbol{Xb} \tag{3.27}$$

wiedząc, że macierz **X** z reguły nie jest kwadratowa nie możemy pomnożyć jej przez macierz odwrotną. Możemy ją jednak w taką przekształcić mnożąc obie strony równania lewostronnie przez macierz transponowaną $\mathbf{X^T}$:

$$\boldsymbol{X^T y} = \boldsymbol{X^T Xb} \tag{3.27}$$

i dalej, wiedząc, że macierz kwadratowa pomnożona przez macierz do niej odwrotną daje macierz jednostkową **I**, mnożąc obustronnie ostatnie równanie przez $(\mathbf{X^T X})^{-1}$ możemy je doprowadzić do postaci:

$$\left(\boldsymbol{X^T X}\right)^{-1}\boldsymbol{X^T y} = \left(\boldsymbol{X^T X}\right)^{-1}(\boldsymbol{X^T X})\boldsymbol{b} = \boldsymbol{I b} \tag{3.28}$$

co ostatecznie daje rozwiązanie:

$$\boldsymbol{b} = \left(\boldsymbol{X^T X}\right)^{-1}\boldsymbol{X^T y} \tag{3.29}$$

Warunkiem koniecznym dla wyznaczenia wektora kolumnowego współczynników modelu **b**, jest możliwość obliczenia macierzy odwrotnej do macierzy informacji (X^TX). Dla przypomnienia, jest to możliwe tylko wtedy, gdy macierz **X** zawiera co najmniej tyle wierszy co kolumn oraz gdy żadna kolumna macierzy danych (**X**), nie jest kombinacją liniową pozostałych kolumn (zmiennych). Spełnienie obu tych warunków jest możliwe tylko wtedy, gdy zastosujemy właściwy – **optymalny plan doświadczeń**.

Musimy też pamiętać, że minimalna ilość pomiarów to zawsze liczba stopni swobody równa zero i dodatkowo przypadek, dla którego miara dopasowania modelu do badanego obiektu (SKR) również jest równa zero. Model taki jest wtedy dopasowany do wartości odpowiedzi zawierających niepewności pomiarowe, a nie do rzeczywistych, najbardziej prawdopodobnych odpowiedzi obiektu. Aby stworzyć właściwy model i oszacować niepewności odpowiedzi obiektu musimy dokonać tylu dodatkowych pomiarów ile będzie koniecznych, aby osiągnąć założoną wcześniej dokładność modelu, tożsamą z dokładnością oszacowań jego współczynników.

Dokładność oszacowań współczynników modelu β_i określają szerokości przedziałów, w których z przyjętym wcześniej prawdopodobieństwem (zwykle 95%) znajduje się 'prawdziwa' wartość wyznaczanych parametrów. Im węższy jest ten przedział, tym lepsze oszacowania i tym lepszy model. Natomiast wyznaczane wartości parametrów to środki tych przedziałów, dla których szerokość można policzyć z zależności:

$$r_b = t_\alpha \sqrt{c_{ii}} * s = t_\alpha s_b \tag{3.30}$$

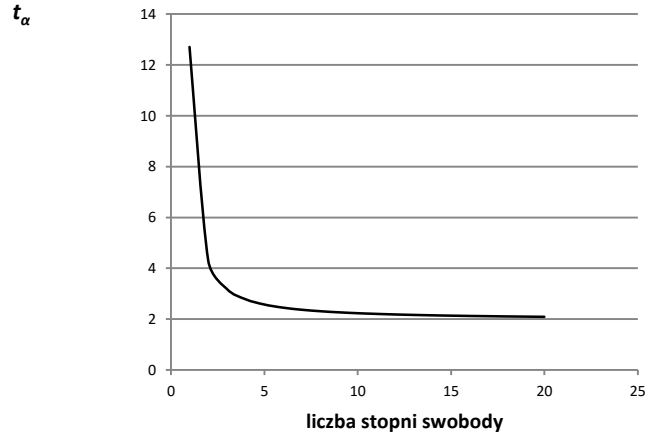
gdzie:

t_α – wartość statystyki *t*-Studenta dla poziomu istotności α i n–(m+1) stopni swobody;

s – odchylenie standardowe modelu (pierwiastek z wariancji resztowej modelu);

c_{ii} – i-ty element głównej przekątnej macierzy dyspersji.

Jak już była mowa, wartości elementów głównej przekątnej macierzy dyspersji zależą jedynie od planu doświadczenia. Możemy je minimalizować poprzez odpowiednie (optymalne) rozmieszczenie punktów pomiarowych w przestrzeni zmiennych. Parametr t_α rozkładu Studenta przy danym poziomie ufności zależy już tylko od ilości stopni swobody. Zależność ta, przedstawiona poniżej uwidacznia, że liczbą stopni swobody, dla których wartość t_α przestaje się silnie zmieniać jest wartość równa 5. Jest to zatem lewy kraniec przedziału liczby dodatkowych pomiarów zwiększających liczbę stopni swobody układu.



Rys. 5. Zależność statystyki t–Studenta od liczby stopni swobody dla α=0.05

Źródło: opr. własne

Wariancja resztowa modelu obliczana jest na podstawie SKR i opisuje ją prosta zależność:

$$s^2 = \frac{SKR}{n-(m+1)} \tag{3.31}$$

gdzie:

n – ilość pomiarów;

m+1 – ilość wyznaczanych parametrów modelu, dlatego *n*-(*m*+1) to ilość stopni swobody układu

Posiadając już wystarczającą, teoretyczną wiedzę chemometryczną dotyczącą modelowania zależności, możemy teraz przystąpić do tworzenia planu optymalnego i dalej, na jego podstawie, do ilościowego określenia modelu badanego zjawiska. Aby tak stworzony model porównać potem z modelem opartym o przypadkowy plan doświadczeń, potrzebny jest nam jeszcze opis matematyczny metod oceny jego wiarygodności i zdolności prognostycznych. Rozpocznijmy zatem od planu optymalnego i dwóch zupełnie dowolnych planów doświadczeń, których celem jest zbadanie tego samego obiektu. Wykorzystamy do tego celu arkusz kalkulacyjny Excel i typowo statystyczne oprogramowanie 'Statistica' firmy StatSoft.

3.6 Przykład oceny modelu opartego o plan doświadczeń 2^3

Plan doświadczeń 2^3 dotyczy, o czym była mowa, zależności liniowej, w której wartość zmiennej zależnej opisywana jest przez trzy zmienne objaśniające:

$$y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 \tag{3.32}$$

Załóżmy teraz, że wykonujemy dwie serie pomiarowe A i B po 10 pomiarów dla tego samego obiektu (opisanego zależnością 3.32), w przypadku których punkty pomiarowe w przestrzeni zmiennych rozmieszczone są zupełnie losowo. Dodatkowo, aby ocenić plany A i B wykonamy też 10 pomiarów wykorzystując optymalny plan czynnikowy – C, 2^3. Wiemy, że optymalną ilością punktów pomiarowych dla tego planu jest wartość 8. Zatem przy 10 pomiarach musimy wykonać dwa pomiary dodatkowe, które najwygodniej jest umieścić w środku planu (punkt 1, 0, 0, 0). Taki plan nazywamy podwójnie centrowanym 2^3. Zmienne dla planu optymalnego, wyrażane w jednostkach planu mogą jak wiadomo przyjmować wartości z krańców przedziałów równe -1 oraz 1. Wiedząc, że wartość 1 odpowiada wartości maksymalnej zmiennej niezależnej a wartość -1 wartości minimum, przy znajomości zakresów wszystkich zmiennych niezależnych (x_1-(4;10), x_2-(3;9), x_3-(2;8)) możemy stworzyć dla nich plan optymalny. Wszystkie porównywane plany A, B, C przedstawione zostały w tabelach poniżej:

Tab. 6. **A** i **B** punkty pomiarowe planów losowych i **C** – punkty planu optymalnego

A

Y	X_1	X_2	X_3
96.63	8.4	8.1	2.6
68.96	5.4	4.8	3.6
64.24	6.1	7.6	5.3
38.82	4.4	7.5	8
54.23	4.8	5.6	7.7
57.11	5.5	3.1	5.5
41.88	6.9	3.3	7.6
97.43	8.5	8.1	3.3
71.96	4.4	7.6	2.8
68.35	4.5	7.9	4.8

B

Y	X_1	X_2	X_3
52.39	5.6	5	7.4
90.40	9.6	4.7	2.9
49.04	5.4	5.5	6.4
52.33	6.4	5.3	5.8
94.00	8.2	5	2.2
76.10	4.7	7.2	3.6
58.54	4.8	3.2	2.9
59.35	6.9	3.6	6.7
85.50	9.3	5.8	2.7
81.12	6.4	7.3	3.6

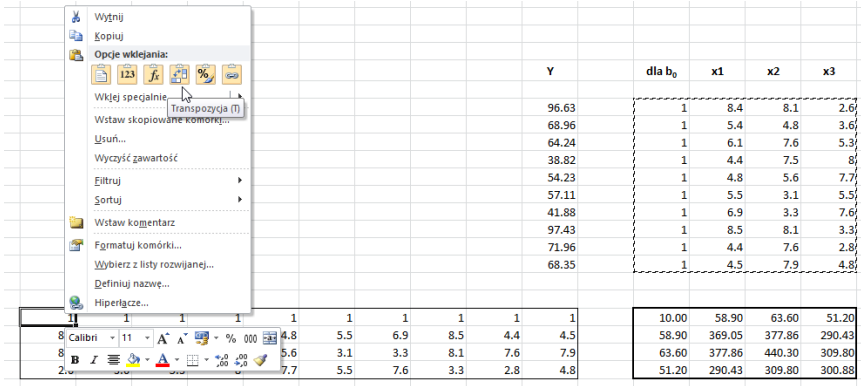
C

Y	X_1	X_2	X_3
77.60	10	9	8
110.05	10	9	2
51.75	10	3	8
80.56	10	3	2
60.03	4	9	8
86.90	4	9	2
27.27	4	3	8
69.59	4	3	2
64.97	7	6	5
72.07	7	6	5

Źródło: opr. własne

Aby wyznaczyć wektory parametrów b_i trzech modeli (na podstawie każdego z planów) musimy wykonać analizę regresji wielokrotnej na danych macierzowych dla każdego przypadku A, B i C osobno. Można to wykonać w prosty sposób wykorzystując do tego dedykowane oprogramowanie statystyczne, lub też w dowolnym arkuszu kalkulacyjnym np. Excel, co ma znacznie większą wartość poznawczą dotyczącą samej metody oraz działań macierzowych. Przykładowe obliczenia wykonane zostaną na danych z serii pomiarowej **A** i zostaną oparte na zależności 3.29. Pierwszym etapem wykonywanych obliczeń jest wyznaczenie macierzy informacji z macierzy danych. Jak wiemy jest to iloczyn macierzy transponowanej $\mathbf{X}^T$ i macierzy wyjściowej **X**. Musimy tu pamiętać o dodatkowej kolumnie jedynek dla wyrazu wolnego b_0 i koniecznie o kolejności mnożenia macierzy. Macierz transponowaną (ikonę transpozycji wskazuje kursor) umieszczamy w arkuszu kopiując do schowka macierz wyjściową i potem tak, aby można było w wygodny sposób zdefiniować obszar macierzy wynikowej (pogrubiona linia ramki), gdyż jest to konieczne przy prawidłowym wypełnieniu komórek z wykorzystaniem funkcji =MACIERZ.ILOCZYN(zakresI;zakresII). Sposób użycia funkcji =MACIERZ.ILOCZYN() jest następujący. 1. Zaznaczamy obszar macierzy będącej wynikiem mnożenia. 2. Wpisujemy odpowiednią formułę w linii poleceń arkusza dla pierwszej komórki wybranego zakresu, zaznaczając jako

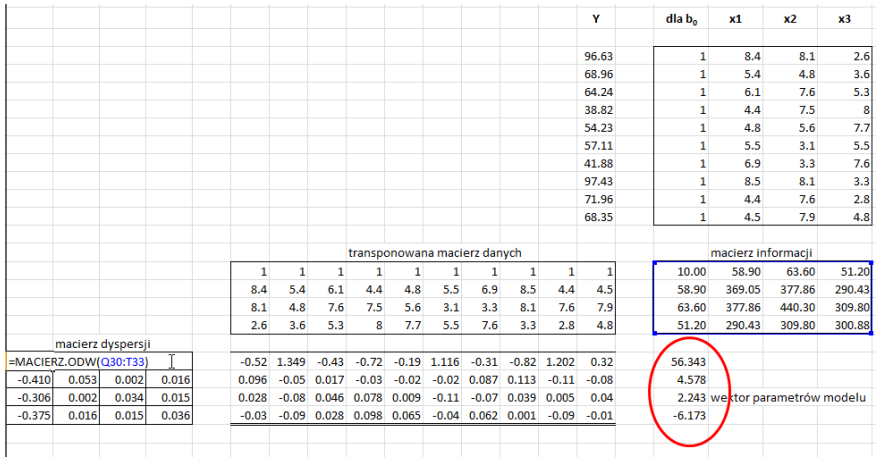
argumenty funkcji obszary macierzy – najpierw transponowanej, później danych (linia przerywana). 3. Kiedy formuła jest gotowa wypełnianie całej macierzy wartościami kończymy jednoczesnym wciśnięciem klawiszy Shift + Ctrl + Enter. Jest to jeden ze sposobów wykorzystywania funkcji macierzowych Excel'a.



Rys. 6. Sposób obliczania elementów macierzy informacji

Źródło: opr. własne

Zgodnie z zależnością 3.29 etapem dalszych obliczeń będzie odwrócenie macierzy informacji (otrzymujemy wtedy macierz dyspersji) i pomnożenie jej prawostronnie przez macierz transponowaną $\mathbf{X}^T$.



Rys. 7. Sposób wyznaczania wartości wektora parametrów b

Źródło: opr. własne

Odwrócenie macierzy informacji wykonuje się w takich samych trzech krokach jak już zostały opisane. Różnica polega jedynie na wykorzystaniu innej funkcji arkusza kalkulacyjnego, mianowicie =MACIERZ.ODW(zakres). Należy tu wspomnieć, że macierz odwrotną możemy wyznaczać jedynie dla macierzy kwadratowej. Wynikiem mnożenia $(\mathbf{X}^T\mathbf{B})^{-1}\mathbf{X}^T$ jest macierz o 4 wierszach i 10 kolumnach (podkreślona podwójną linią), która pomnożona przez wektor odpowiedzi obiektu **Y**, da nam rozwiązanie w postaci kolumnowego wektora parametrów modelu **b** (zaznaczonego elipsą). Przeprowadzona w ten sposób kalkulacja parametrów modelu może zostać wstępnie oceniona na podstawie SKR i wielkości o nią opartej – wariancji resztowej modelu. Można też wyznaczyć promienie ufności dla wartości b_i wykorzystując wartości na głównej przekątnej macierzy dyspersji i zależność 3.30.

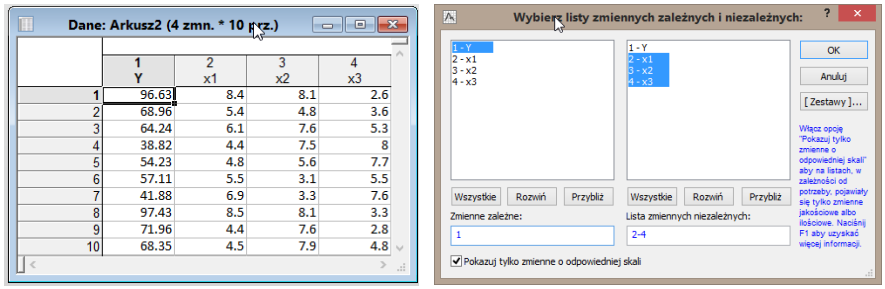
Aby obliczyć wielkość SKR musimy najpierw wyznaczyć odpowiedzi modelu zdefiniowanego na podstawie planu doświadczeń **A**. W tym celu trzeba oczywiście pomnożyć lewostronnie wektor **b** (wyznaczonych parametrów modelu) przez macierz danych **X**. Wektor odpowiedzi modelu posłuży nam dalej do wyznaczenia wartości sumy kwadratów różnic z prezentowanej wcześniej zależności 3.25. Wynik tego mnożenia przedstawia kolejny rysunek (Rys. 8).

						56.343		
						4.578		
						2.243		
						-6.173		
	Y		dla b_0	x1	x2	x3	y	(Y-y)^2
	96.63		1	8.4	8.1	2.6	96.911	0.08015
	68.96		1	5.4	4.8	3.6	69.604	0.41350
	64.24		1	6.1	7.6	5.3	68.593	18.92275
	38.82		1	4.4	7.5	8	43.919	25.96942
	54.23		1	4.8	5.6	7.7	43.341	118.59155
	57.11		1	5.5	3.1	5.5	54.520	6.72629
	41.88		1	6.9	3.3	7.6	48.414	42.69066
	97.43		1	8.5	8.1	3.3	93.047	19.19752
	71.96		1	4.4	7.6	2.8	76.244	18.31131
	68.35		1	4.5	7.9	4.8	65.028	11.02363
							SKR =	261.92677
			macierz informacji					
1	1		10.00	58.90	63.60	51.20		
4	4.5		58.90	369.05	377.86	290.43	S^2 =	43.65446
6	7.9		63.60	377.86	440.30	309.80		
8	4.8		51.20	290.43	309.80	300.88	s =	6.60715

Rys. 8. SKR i wartość odchylenia standardowego reszt dla modelu opartego o plan A

Źródło: opr. własne

Działania przedstawione na Rys. 8 wykonane zostały jedynie po to, aby wyznaczyć wariancję resztową modelu. Wiemy już, że na jej podstawie wyliczyć możemy promień ufności (przedziały ufności) oraz wariancję dla każdego z szacowanych parametrów b_i. Wszystkie te (i jeszcze inne) charakteryzujące model wielkości dostajemy w sposób mniej pracochłonny wykorzystując możliwości programu 'Statistica'. Należy jedynie wprowadzić wszystkie zmienne do arkusza (zależną i niezależne – tym razem bez kolumny jedynek, która dodawana jest automatycznie) i po odpowiednim ich wyborze w panelu analizy regresji wielorakiej (wielokrotnej) wykonać obliczenia, co jest dobrym sposobem sprawdzenia poprawności tych przeprowadzonych w Excel'u.



Rys. 9. STATISTICA – dane i wybór zmiennych w analizie regresji wielorakiej

Źródło: opr. własne

Prowadząc wyżej opisane kalkulacje dla wszystkich planów doświadczeń możemy usystematyzować wiedzę na temat wiarygodności i rzetelności otrzymanych na ich podstawie modeli zależności. Rozpocznijmy zatem nasze porównanie od analizy wielkości SKR i s^2 dla każdego modelu:

Tab. 7. Wariancja resztowa s^2 i SKR modeli dla testowanych planów pomiarowych

Model	SKR	s^2	s
A	261.93	43.65	6.61
B	294.26	49.04	7.00
C – optymalny	105.65	17.61	4.20

Źródło: opr. własne

Wnioski są oczywiste pod warunkiem, że niepewność oszacowania parametrów modelu zależy zawsze od tych samych czynników – warunki pomiaru nie ulegają zmianie. Tylko w przypadku powtarzalnych pomiarów wielkości te mogą być porównywane dla różnych planów doświadczeń. Widzimy teraz, że najgorszym planem doświadczeń jest plan **B**, najlepszym (znacznie odbiegającym od pozostałych) – **C** – optymalny. Skoro wariancje współczynników modelu w dużo większym stopniu zależą od wielkości leżących na głównej przekątnej macierzy dyspersji (zależność 3.15) niż od wariancji resztowej modelu, porównajmy je celem przeprowadzenia bardziej wnikliwej oceny jakości parametrów b_i:

Tab. 8. Wartości głównych przekątnych macierzy dyspersji modeli

Model	$\mathbf{b_o}$	$\mathbf{b_1}$	$\mathbf{b_2}$	$\mathbf{b_3}$
A	6.38	0.053	0.034	0.036
B	6.94	0.045	0.067	0.038
C – optymalny	1.63	0.014	0.014	0.014

Źródło: opr. własne

Różnice między współczynnikami w zależności od zastosowanego planu są widoczne natychmiast, gdy porównamy plany **A** i **B** z planem optymalnym **C**. Wartości elementów głównych przekątnych macierzy dyspersji pozwalają sądzić, że wyliczone na ich podstawie przedziały ufności parametrów modelu będą kilkukrotnie węższe dla modelu opartego o plan doświadczenia **C**. Mniejsze wartości elementów c_{ii} obserwujemy także dla modelu **A** w porównaniu z modelem **B**, z wyjątkiem jednego parametru – b_1, co potwierdza pierwszą, ogólną ocenę rzetelności modeli na podstawie wielkości sumy kwadratów różnic (tym samym wariancji resztowej modelu).

Jednym z kryteriów poprawności modelu chemometrycznego jest istotność wszystkich jego elementów składowych tzn. takie wartości współczynników b_i, dla których przedziały ufności **nie zawierają zera**. Na pytanie czy wszystkie zmienne niezależne są w modelu istotne, natychmiastową odpowiedz daje nam analiza wykonana za pomocą programu ‘Statistica’. Choć można to zrobić w arkuszu Excel wykorzystując zależność 3.30. Wyniki analizy regresji wielokrotnej dla wszystkich modeli przedstawia tabela 9.

Tabela ta jest wynikiem analizy regresji wykonanej z wykorzystaniem modelu zaawansowanego tj. ogólnego modelu regresji – GRM (General Regression Model), tylko dlatego, że dla tej opcji menu programu 'Statistica' wyświetlane są granice przedziałów ufności dla współczynnika istotności $\alpha = 0.05$; ostatnie dwie kolumny. Program posiada tę cechę, że elementy istotne modelu wyświetlane są w kolorze czerwonym (w tekście skryptu – kursywa). Dzięki temu od razu widzimy, że w przypadku modelu **A** istotnymi są wszystkie zmienne z wyjątkiem X_2 oraz wyraz wolny. Ich przedziały ufności nie zawierają w swoich granicach zera co oznacza, że

Tab. 9. Wyniki analizy regresji uzyskane z użyciem programu 'Statistica'

Model A	Podsumowanie regresji zmiennej zależnej: Y R= .96203803 R^2= .92551717 F(3,6)=24.852 *p<.00088* Błąd std. Estymacji, *s* = 6.6072				GRM Ogólne modele regresji przedziały ufności	
	b	Bł. std. (s_i)	t(6)	p	-95.00%	-95.00%
W. wolny	*56.34304*	*16.69379*	*3.37509*	*0.014948*	*15.49481*	*97.19126*
x_1	*4.57784*	*1.52103*	*3.00969*	*0.023710*	*0.85601*	*8.29967*
x_2	2.24251	1.22178	1.83544	0.116109	-0.74708	5.23209
x_3	*-6.17321*	*1.26160*	*-4.89315*	*0.002730*	*-9.26024*	*-3.08618*
Model B	Podsumowanie regresji zmiennej zależnej: Y R= .94392874 R^2= .89100146 F(3,6)=16.349 *p<.00271* Błąd std. estymacji, *s* = 7.0031					
	b	Bł. std. (s_i)	t(6)	p	-95.00%	-95.00%
W. wolny	45.13246	18.44634	2.44669	0.050015	-0.00412	90.26904
x_1	*4.50722*	*1.48107*	*3.04322*	*0.022709*	*0.88317*	*8.13127*
x_2	3.37446	1.81208	1.86220	0.111882	-1.05955	7.80846
x_3	*-5.28014*	*1.36594*	*-3.86557*	*0.008308*	*-8.62248*	*-1.93780*
Model C	Podsumowanie regresji zmiennej zależnej: Y R= .98777221 R^2= .97569395 F(3,6)=80.284 *p<.00003* Błąd std. estymacji, *s* = 4.1962					
	b	Bł. std. (s_i)	t(6)	p	-95.00%	-95.00%
W. wolny	*48.69251*	*5.353676*	*9.0952*	*0.000099*	*35.59254*	*61.79249*
x_1	*3.17322*	*0.494525*	*6.4167*	*0.000676*	*1.96316*	*4.38328*
x_2	*4.39177*	*0.494525*	*8.8808*	*0.000113*	*3.18171*	*5.60183*
x_3	*-5.43550*	*0.494525*	*-10.9914*	*0.000034*	*-6.64556*	*-4.22544*

Źródło: opr. własne

liczba 0 nie należy do populacji wartości współczynników b_0, b_1 i b_3. Model **B** jest jak widać i jak to już zostało powiedziane modelem gorszym niż **A**.

Tylko dwie zmienne X_1 i X_3 są w nim istotne. Najlepszym jest oczywiście model **C** oparty o plan optymalny. Analizując wartości błędów standardowych (odchyleń standardowych) każdego z parametrów w zależności od modelu widzimy, że w przypadku modelu opartego o plan optymalny są one znacznie niższe niż ma to miejsce dla modelu **A** czy **B**. Zawęża to oczywiście istotnie przedziały ufności dla parametrów. Jest to bardzo pożądana cecha z punktu widzenia jakości modelu. Należy też zwrócić uwagę na to, iż wszystkie odchylenia standardowe liczone dla parametrów b_i modelu **C** posiadają tę samą wartość 0.495. Ma to miejsce tylko wtedy, kiedy wektory zmiennych w planie doświadczeń są ortogonalne.

Poza wariancją resztową, ważnymi parametrami ilościowymi, które pozwalają na porównanie i ocenę modeli pod kątem dopasowania ich odpowiedzi do odpowiedzi obiektu są jeszcze takie wielkości jak statystyka F(ilość zmiennych objaśniających (wejść), ilość stopni swobody) oraz parametr *p*, który jest niczym innym jak prawdopodobieństwem prawdziwości hipotezy zerowej. W przypadku badanych modeli hipotezą zerową jaką należy postawić jest 'brak zależności pomiędzy zmienną Y (zależną) i zmiennymi objaśniającymi X_i'. Oczywiście na standardowym poziomie ufności (95%, $\alpha = 0.05$) dla wartości $p < 0.05$ należy ją odrzucić na rzecz hipotezy alternatywnej. Musimy to zrobić dla każdego z prezentowanych modeli (**A:** $p < 0.00088$; **B:** $p < 0.00271$ i **C:** $p < 0.00003$) a w każdym modelu z osobna dla wszystkich istotnych zmiennych X_i (kursywa w Tab. 9). Wartość prawdopodobieństwa słuszności hipotezy zerowej jest także miarą jakości modelu – im wyższa tym model można traktować jako mniej dopasowany do rzeczywistego obiektu.

3.6.1 Istotność modelu i rozkład reszt

Wspomniana wcześniej statystyka F(m,f) jest wielkością określającą ilościowo istotność zidentyfikowanego modelu zjawiska. Jest ona liczona jako iloraz wariancji odpowiedzi obiektu i wariancji resztowej modelu a model uznajemy za statystycznie istotny, jeśli jego wartość jest większa niż stablicowana wartość krytyczna $F_{kr}(m,f)$ dla założonego poziomu ufności $1 - \alpha$:

$$F = \frac{s_y^2}{s^2} \tag{3.33}$$

przy czym wartość $s_{\hat{y}}^2$ liczona jest dla wektora odpowiedzi obiektu z zależności:

$$s_{\hat{y}}^2 = \frac{\sum(y_i - \overline{y})^2}{n-1} \tag{3.34}$$

gdzie:

$\overline{y}$ – wartość średnia odpowiedzi obiektu

Parametr ten został wyliczony dla każdego z przykładowych modeli (Tab. 9) i wiadomo, że jest on w każdym przypadku większy (a model jest istotny) niż odpowiednia wartość krytyczna, ponieważ na jego podstawie liczone są analizowane już wartości prawdopodobieństw słuszności hipotezy zerowej *p*. Wielkość statystyki F(m,f) jest pierwszym kryterium istotności modelu. Jej brak całkowicie dyskwalifikuje zidentyfikowany model i nie może on być brany pod uwagę w jakichkolwiek dalszych analizach chemometrycznych.

Na podstawie statystyki F(m,f) obliczana jest wielkość określana mianem współczynnika determinacji D, który przedstawia zależność:

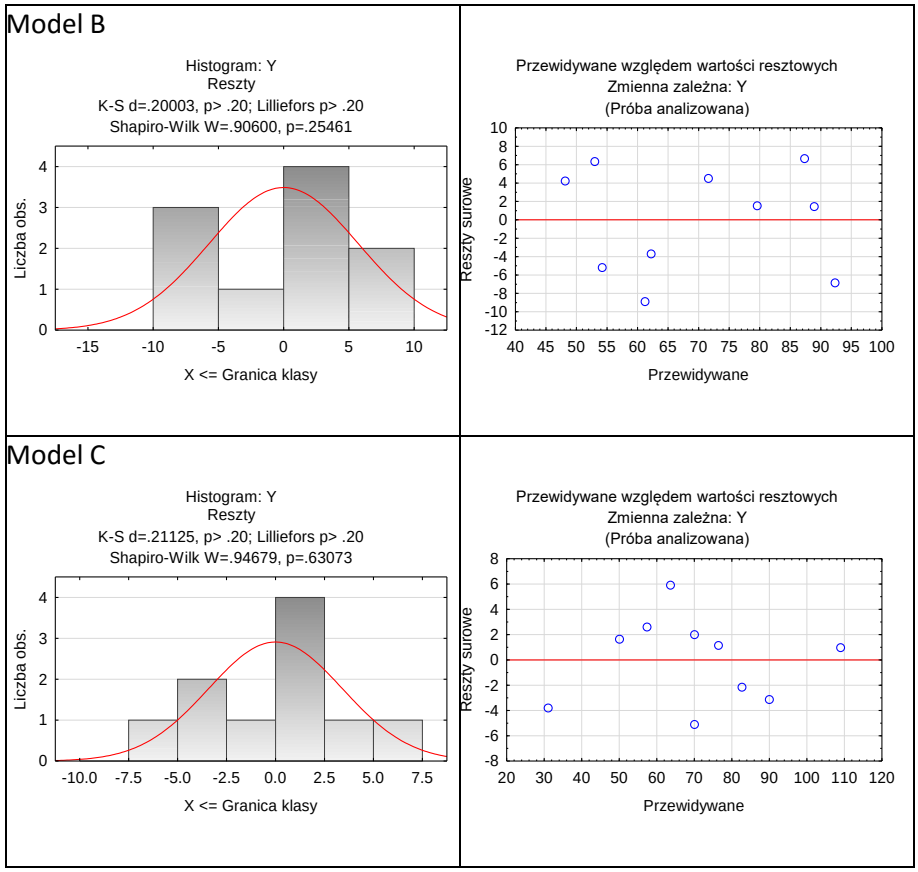
$$D = 1 - \frac{s^2}{s_{\hat{y}}^2} = 1 - \frac{1}{F} \tag{3.35}$$

Ten ilościowy miernik dopasowania modelu do badanego obiektu należy interpretować jako ułamek ogólnej zmienności odpowiedzi obiektu wyjaśniany przez model. Współczynnik determinacji jest niczym innym jak kwadratem dobrze znanego współczynnika korelacji modelu:

$$R = \sqrt{D} \tag{3.36}$$

Wartości współczynnika korelacji i współczynnika determinacji liczone są automatycznie przez program ‘Statistica’ (Tab. 9) i oczywiście im są one większe z tym lepszym modelem mamy do czynienia. Statystyka F(m,f) oraz wielkości liczone na jej podstawie dają nam ogólne informacje dotyczące jakości zidentyfikowanego modelu. Musimy jednak pamiętać o tym, że pojedyncza liczba często nie dostarcza nam wiarygodnego opisu

poprawności modelu badanego obiektu. Często złożoność obiektów i zjawisk wymaga bardziej szczegółowej analizy jego poprawności. Przede wszystkim, analizy pod kątem poprawnego wyrażenia matematycznego opisującego badane zjawisko. Wielu informacji na ten temat daje nam analiza rozkładu różnic pomiędzy odpowiedziami obiektu i wartościami prognozowanymi przez model.



Rys. 10. Rozkład reszt dla modelu opartego o: plan doświadczeń **B** i plan optymalny **C**

Źródło: opr. własne

Z teorii analizy regresji wiemy, że algorytm metody najmniejszych kwadratów wymaga, aby rozkład zmiennej, jaką jest różnica odpowiedzi modelu i obiektu był rozkładem normalnym. Oznacza to, że reszty powinny

rozkładać się symetrycznie wokół wartości zero (ich średnia wartość powinna wynosić 0). Oznacza to również niezależność wartości reszt od badanego zakresu wartości zmiennej zależnej Y. Przykładowe testy reszt badanych modeli **B** oraz **C** przedstawione zostały na rysunku 10. Wynika z nich, że zarówno dla najgorszego modelu **B** jak i najlepszego – **C**, normalność rozkładu reszt jest zachowana. Potwierdzają to wartości testów Kołmogorowa–Smirnowa z poprawką Lillieforsa (przypadek, gdy nie znamy wartości średniej i wartości odchylenia standardowego dla populacji, z której pochodzi próba) oraz bardziej restrykcyjnego (posiadającego większą moc) testu normalności – Shapiro-Wilka. We wszystkich przypadkach wartość *p* wskazuje na nieistotność testów, co nie pozwala na odrzucenie hipotezy zerowej mówiącej o nieistotności różnic pomiędzy rozkładami: normalnym i reszt. Graficznie, brak jakiejkolwiek zależności funkcyjnej pomiędzy wielkościami reszt i wartością przewidywaną (prognozowaną przez model) obrazują wykresy zamieszczone obok histogramów reszt. Podsumowując, **nieprawidłowy rozkład reszt sugeruje konieczność poszukiwania innej zależności matematycznej opisującej modelowany obiekt, zjawisko**.

3.6.2 Adekwatność modelu

Model empiryczny z założenia jest jedynie matematycznym przybliżeniem relacji rządzących zachowaniem się badanego obiektu. Jeżeli zdefiniowaliśmy już model, którego istotność da się potwierdzić statystycznie, ale rozkład reszt nie jest prawidłowy, możemy przetestować jego adekwatność. Kryterium adekwatności modelu określa się, jako istotność różnicy wariancji modelu (reszt) i wariancji metody pomiarowej. Model spełniający to kryterium (wariancja metody musi być istotnie większa od wariancji modelu) nazywamy właśnie modelem adekwatnym. Adekwatność modelu daje nam informację, że niepewność odpowiedzi modelu jest tego samego rzędu, co niepewność pomiaru odpowiedzi obiektu i może nam potwierdzić czy model, który stworzyliśmy jest dostatecznie dobry, czy może gdyby był oparty o bardziej skomplikowaną zależność matematyczną, liniową z interakcjami czy też kwadratową, mógłby być lepszy.

W zależności od tego skąd pochodzi oszacowana wariancja metody (w jaki sposób była liczona), adekwatność modelu testujemy na różne sposoby.

1. Jeśli mamy do czynienia ze standardową metodą, dla której wariancja została oszacowana wcześniej, do określenia adekwatności modelu wykorzystujemy statystykę χ^2 dla wariancji:

$$\chi^2 = \frac{\sum(\tilde{y}_i - y_i)^2}{s_m^2} = \frac{ns^2}{s_m^2} \tag{3.37}$$

gdzie:
s^2 – wariancja resztowa modelu,
s_m^2 – wariancja metody,
n – ilość pomiarów

jeśli wartość statystyki χ^2 przekracza stablicowaną wartość krytyczną $\chi^2_{\alpha,f}$ dla poziomu ufności 1-α i ilości stopni swobody f, model nie jest adekwatny.

2. Jeśli wariancja metody wyznaczana była z dodatkowych pomiarów wykonywanych w planie doświadczeń (w zaplanowanych punktach pomiarowych) to adekwatność modelu wyznaczyć możemy jako statystykę Fishera-Snedecora na istotność różnic wariancji F:

$$F = \frac{s^2}{s_p^2} \tag{3.38}$$

gdzie:
s_p^2 – wariancja dla zbioru powtórzonych pomiarów (wariancja metody).

test jest oczywiście istotny (model nieadekwatny), gdy wyliczona wartość statystyki F jest większa niż stablicowana wartość krytyczna F_{kr}.

3. Jeśli powtórzenia pomiarów realizowane były poza planem doświadczenia w jednym wybranym punkcie x_0, to obliczamy odpowiedzi modelu w tym punkcie $\tilde{y}_0$, ich wartość średnią $\overline{y}_0$ i wyznaczając ich promień ufności r_0 tworzymy wyrażenie:

$$t = \frac{|\tilde{y}_0 - \overline{y}_0|}{r_0} \tag{3.39}$$

którego wartość porównujemy z wartością krytyczną (dla poziomu istotności 0.05 i dla ilości stopni swobody równej ilości pomiarów) testu *t*-Studenta. Jeśli wartość *t* nie jest mniejsza od t_{kr} mamy podstawy do odrzucenia hipotezy o adekwatności modelu.

Musimy pamiętać, że adekwatność modelu nie jest ostatecznym kryterium i nie jest sposobem na wybór jedynego, najlepszego modelu. Istnieje zbiór modeli adekwatnych, z punktu widzenia adekwatności jednakowo dobrych, który charakteryzuje się tym, że zastosowanych w nich przybliżeń matematycznych nie da się wykryć techniką pomiarową wykorzystaną do uzyskania danych do zdefiniowania modelu. Wybór jednego z nich jest możliwy, gdy zastosujemy jeszcze inne, dodatkowe kryteria opisujące zdolność prognostyczną modelu.

3.6.3 Zdolność prognostyczna modelu

Zdolność prognostyczna modelu jest parametrem, który ze zbioru modeli istotnych i adekwatnych pozwala na wybór modelu najlepszego. Jednym ze sposobów oceny zdolności prognostycznych modelu jest wyznaczenie przedziału ufności wyjścia modelu dla określonego punktu w przestrzeni zmiennych $\mathbf{x}_0$. Jego promień r_0 (szerokość) można obliczyć z zależności:

$$r_0 = t_\alpha \sqrt{1 + \boldsymbol{x}_0^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{x}_0} * s \tag{3.40}$$

Jak widzimy promień ten zależy od planu doświadczenia $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$, położenia punktu w przestrzeni zmiennych oraz wariancji resztowej modelu, która jest z kolei zależna od ilości stopni swobody układu, czyli ilości pomiarów ponad konieczne minimum. Tak więc, oszczędne plany doświadczeń dają modele, dla których szerokość przedziału ufności jest zawsze duża. Można je zatem porównywać (promienie) tylko wtedy, gdy dotyczą tych samych punktów w przestrzeni zmiennych i modeli stworzonych w oparciu o plany doświadczeń o tej samej liczbie stopni swobody. Nie jest to zatem

wygodne narzędzie do porównywania jakości modeli nie spełniających tych warunków.

Innym, częściej wykorzystywanym i nieposiadającym tej wady parametrem, pozwalającym na ocenę wartości prognostycznej modelu jest współczynnik walidacji Q^2, który posiada sens matematyczny zbliżony do współczynnika determinacji D. Jednak, aby można było wyznaczyć tę wielkość, rzeczą konieczną jest przeprowadzenie kilku dodatkowych pomiarów, których zbiór nazywany jest zbiorem testowym (walidacyjnym). Należy przy tym pamiętać, że punkty pomiarowe muszą mieścić się w zakresie zmienności zmiennych niezależnych (w przestrzeni planu doświadczeń), ponieważ dla takiego zakresu określany był model zależności. Wyniki tych pomiarów (przeprowadzonych na badanym obiekcie) porównuje się z wyliczonymi odpowiedziami modelu dla tych samych punktów przestrzeni zmiennych niezależnych i dalej wyznacza SKR_t i wariancję reszt s_t^2 dla tego zbioru:

$$s_t^2 = \frac{SKR_t}{n-1} = \frac{\sum(y_i - \tilde{y}_i)^2}{n-1} \tag{3.41}$$

gdzie:
n – ilość pomiarów testowych

Współczynnik walidacji obliczamy wykorzystując zależność taką jak w przypadku współczynnika determinacji, zamieniając jedynie wariancję resztową modelu na wariancję reszt dla zbioru walidacyjnego s_t^2:

$$Q^2 = 1 - \frac{s_t^2}{s_{\tilde{y}}^2} \tag{3.42}$$

gdzie:
$s_{\tilde{y}}^2$ – wariancja wektora odpowiedzi obiektu w punktach testowych (3.34).

Dobry model chemometryczny powinien cechować współczynnik walidacyjny bliski jedności. Dlatego podobnie jak miało to miejsce w przypadku współczynników korelacji i determinacji, im jego wartość jest bliższa jedności, tym model jest lepszy. Spróbujmy teraz potwierdzić najlepszą jakość modelu opartego o plan optymalny dla naszego przypadku trzech planów doświadczeń **A**, **B** i **C** obliczając dla nich parametr Q^2. W tym celu przeprowadzone zostały dwie serie po pięć dodatkowych pomiarów testowych

(1 seria – bliżej i 2 seria – dalej środka planu), w tych samych, losowo wybranych punktach przestrzeni zmiennych dla każdego modelu:

Tab. 10. Wartości współczynników walidacji i determinacji testowanych modeli

1 seria – bliżej środka planu					2 seria – bliżej krańców planu				
Y_{ob}	w. wolny	x_1	x_2	x_3	Y_{ob}	w. wolny	x_1	x_2	x_3
76.088	1	8.5	5.0	4.5	73.294	1	6.5	5.8	4.5
47.353	1	4.2	3.0	4.9	85.477	1	6.3	8.8	4.6
64.021	1	8.5	5.8	7.1	61.190	1	7.7	7.5	7.9
58.559	1	4.6	7.9	7.6	60.672	1	9.3	3.0	5.2
97.620	1	5.7	8.9	2.0	64.012	1	6.9	8.0	7.9

Model	**A**	**B**	**C – optymalny**
R^2 (D)	0.8883	0.8365	0.9635
1 seria - Q^2	0.8452	0.9207	0.9824
2 seria - Q^2	0.0079	0.2255	0.8950

Źródło: opr. własne

Z analizy wartości współczynników walidacji Q^2 możemy wnioskować, że bez względu na położenie punktów pomiarowych w przestrzeni planu zdolność prognostyczna modelu opartego o plan optymalny jest bardzo dobra, a współczynnik walidacji w każdym przypadku wysoki. Pozostałe modele dobrze zachowują się w przypadku pierwszej serii pomiarowej, ale całkowicie zawodzą w przypadku serii drugiej – punktów bardziej odległych od centrum planu. Niewielkie wartości parametru Q^2, zwłaszcza w przypadku modelu **A** (wartość bliska zeru) dla tego zakresu dowodzą, że wartość prognostyczna w tym przypadku nie jest wyższa niż średnia arytmetyczna z dowolnej serii pomiarów w tym zakresie. Dodatkowo należy nadmienić, że usunięcie z zależności zmiennych nieistotnych daje ujemne wartości Q^2, całkowicie dyskwalifikujące modele **A** i **B** bez względu na położenie punktów pomiarowych w przestrzeni planu doświadczenia. Takiego efektu nie obserwowalibyśmy, gdyby nieistotne zmienne były usuwane z modelu stworzonego na podstawie ortogonalnych wektorów zmiennych niezależnych. Tylko w takim przypadku możemy mieć pewność, że są one rzeczywiście nieistotne, a ich usunięcie nie spowoduje konieczności ponownego szacowania pozostałych parametrów modelu.

Przedstawiona metoda walidacji modelu pociąga za sobą konieczność przeprowadzenia przynajmniej kilku dodatkowych pomiarów testowych.

Można je wykonać podczas pomiarów realizowanych w celu identyfikacji modelu lub później, co wydaje się być gorszym rozwiązaniem ze względu na ich powtarzalność. Istnieje też metoda pozwalająca na ocenę prognostyczną modelu, która praktycznie nie wymaga dodatkowych pomiarów odpowiedzi obiektu lub tylko jednego, dla zachowania pełnego zestawu wstępnie wybranych 'uczących' punktów pomiarowych. Jest to krzyżowa metoda oceny modelu (cross–validation). Jest ona bardziej złożona obliczeniowo, ale przy dzisiejszych możliwościach wykorzystania komputerów nie stanowi to wielkiej wady. Najlepszym na to dowodem jest możliwość jej wykorzystania praktycznie w każdym oprogramowaniu statystycznym.

Algorytm krzyżowej oceny modelu jest prosty a jego koncepcja polega na potraktowaniu zbioru danych pomiarowych jednocześnie, jako zbioru danych tworzących model i danych testowych. Realizacja obliczeń polega na wykonaniu *n* razy (gdzie *n* to ilość wszystkich zmierzonych odpowiedzi obiektu) identyfikacji modelu na podstawie zbioru *n* - 1 danych pomiarowych, przy czym wykluczony, każdorazowo inny, jeden pomiar stanowi punkt testowy modelu. Dla tego chwilowo wyłączonego punktu pomiarowego obliczamy odpowiedź modelu w tym punkcie i porównujemy ją z odpowiedzią obiektu. Otrzymany w ten sposób *n* elementowy zbiór różnic stanowi podstawę do obliczenia testowej wariancji resztowej i dalej współczynnika walidacji modelu.

Przy takim sposobie walidacji modelu, interpretacja współczynnika Q^2 i parametrów modelu jest następująca: uzyskany współczynnik walidacji jest oszacowaniem jakości modelu na podstawie pełnego zestawu *n* punktów pomiarowych planu, a parametry modelu są w tym przypadku wartością średnią parametrów modeli cząstkowych, uzyskanych dla każdego *n*–1 zestawu zmiennych. Odchylenia standardowe wektorów parametrów uzyskanych z modeli cząstkowych, służą do wyznaczenia przedziałów ufności każdego z parametrów modelu ostatecznego i co ciekawe, zwłaszcza dla małej liczby stopni swobody, są one węższe niż wyznaczone na podstawie tego samego zbioru danych w sposób standardowy.

4 ANALIZA GŁÓWNYCH SKŁADOWYCH

Pojęcie głównych składowych wprowadzone zostało w roku 1900 przez Karla Pearsona, jednak do rozwoju metody znanej pod nazwą analizy głównych składowych przyczynił się przede wszystkim amerykański statystyk Hotelling. Sama metoda znalazła po raz pierwszy praktyczne zastosowanie w roku 1933 (Hotelling – analiza testów osiągnięć szkolnych) i potem w 1964 (Rao). Analiza głównych składowych (PCA – Principal Components Analysis), podobnie jak dobrze znana cząstkowa metoda najmniejszych kwadratów (PLS – Partial Least Square), czy analiza czynników dyskryminacyjnych (DF – Discriminant Factors), jest jedną z metod analizy czynnikowej. Metody te stanowią zespół procedur matematycznych pozwalających na zredukowanie dużej liczby zmiennych do kilku, z założenia wzajemnie nieskorelowanych czynników – nowych zmiennych. Przy czym kilka pierwszych z nich zachowuje stosunkowo dużą część informacji tkwiących we wszystkich zmiennych pierwotnych i jednocześnie każda z nich jest nośnikiem innych treści merytorycznych. Zapewnia to brak powtarzalności niesionej informacji przez czynniki.

Fakt, że kilka pierwszych głównych składowych zawiera zwykle większość informacji zawartej w danych wyjściowych, można wykorzystać do uzyskania optymalnego rzutowania wielowymiarowej przestrzeni danych na płaszczyznę (2D) lub przestrzeń trójwymiarową (3D). Takie rzuty stanowią wygodną formę graficznej prezentacji danych i są jednym z podstawowych powodów prowadzenia analizy czynników głównych. Drugim z nich, tak samo ważnym jak wizualizacja zależności między zmiennymi, jest redukcja wymiarowości analizowanego problemu (eliminowanie nadmiernej korelacji zmiennych), wykorzystywana często w dalszej analizie chemometrycznej. PCA jest w stanie spełnić oba zadania **tylko wtedy**, gdy w macierzy korelacji zmiennych niezależnych **elementy pozadiagonalne są znacząco różne od zera**. Oznacza to bowiem, że zmienne objaśniające nie są od siebie niezależne i tylko część wnoszonych przez nie informacji o obiekcie jest unikalna.

Informacje wnoszone przez zmienne można w związku z tym podzielić na informacje wspólne i swoiste. Informacją wspólną dwóch zmiennych jest procent zmienności jednej z nich, wyjaśniany zmiennością drugiej.

Miarą tej współzależności zmiennych jest współczynnik determinacji, będący kwadratem współczynnika korelacji Pearsona. Natomiast najlepszą miarą ilości wnoszonej informacji przez pojedynczą zmienną, jak dowodzi się na gruncie statystyki, jest wartość wariancji dla tej zmiennej. Dotyczy to również danych wielowymiarowych, w przypadku których całkowity zakres zmienności, odpowiadający całej informacji jaką niesie kompletny zestaw zmiennych, jest równy sumie wariancji wszystkich zmiennych pod warunkiem ich ortogonalności. Dlatego m.in. ortogonalność (brak korelacji) zmiennych jest podstawowym założeniem metody analizy czynników głównych.

Nadmiarowość informacji jest zjawiskiem występującym powszechnie w przypadku dużej ilości zmiennych. Ma to swoje dobre i złe strony. Podstawową zaletą tego zjawiska jest możliwość minimalizacji wpływu niepewności oznaczenia na wynik pomiarów/badań. Bowiem, inaczej niż ma to miejsce w przypadku informacji wspólnej, niepewność pomiarowa zawarta w każdej zmiennej nie jest skorelowana z niepewnością pomiarową żadnej innej. Ten brak korelacji, pozwala na wyodrębnienie z danych informacji wspólnej, zawierającej jedynie niewielki błąd pomiarowy, a dodatkowo, jest on tym mniejszy im silniejsza jest współzależność zmiennych (silniejsza nadmiarowość). Niekorzystną cechą nadmiarowości informacji są nakłady finansowe na pomiary przeprowadzone celem jej uzyskania oraz czasochłonna analiza własności obiektów opisanych dużą ilością zmiennych.

Jak zostało to już wcześniej powiedziane, w metodzie głównych składowych dąży się do stworzenia nowego układu, nieskorelowanych (ortogonalnych) zmiennych – czynników, dla których podstawą są pierwotne zmienne objaśniające. Nowe zmienne (*pc* –zależność 4.1) są liniową kombinacją zmiennych wyjściowych i charakteryzują się maksymalną z możliwych wariancją pierwszego, a następnie każdego kolejno tworzonego czynnika – pc_1, pc_2 itd.

$$\begin{bmatrix} a_{11} & a_{12} \cdots & a_{1n} \\ a_{21} & a_{22} \cdots & a_{2n} \\ & \vdots & \\ a_{n1} & a_{n2} \cdots & a_{nn} \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} pc_1 \\ pc_1 \\ \vdots \\ pc_n \end{bmatrix} \tag{4.1}$$

Liczba tworzonych czynników jest taka sama jak liczba zmiennych pierwotnych, a macierz **A** jest macierzą współczynników korelacji cząstkowych pomiędzy głównymi składowymi (*pc*) i zmiennymi objaśniającymi, co oznacza, że są to współczynniki kombinacji liniowej (wektory wierszowe) zmiennych objaśniających, definiujących poszczególne główne składowe. Ich wartości wyznaczane są metodą dekompozycji macierzy korelacji dla zmiennych wyjściowych **C**, określonej następującą zależnością:

$$\boldsymbol{C} = \frac{1}{n}\boldsymbol{Z^T Z} \tag{4.2}$$

gdzie:

Z – macierz standaryzowanych danych wyjściowych; n – ilość obiektów

4.1 Wektory i wartości własne

Dekompozycja macierzy kwadratowej (a taką jest macierz korelacji**)**, zwana zagadnieniem własnym, obliczeniowo sprowadza się do rozwiązania układu równań postaci:

$$\begin{cases} (c_{11} - \lambda) \cdot v_1 + c_{12} \cdot v_2 + \cdots + c_{1n} \cdot v_n = 0 \\ c_{21} \cdot v_1 + (c_{22} - \lambda) \cdot v_2 + \cdots + c_{2n} \cdot v_n = 0 \\ \vdots \\ c_{n1} \cdot v_1 + c_{n2} \cdot v_2 + \cdots + (c_{nn} - \lambda) \cdot v_n = 0 \end{cases} \tag{4.3}$$

co możemy zapisać w postaci równania macierzowego jako:

$$(\boldsymbol{C} - \lambda \boldsymbol{I})\boldsymbol{v} = (\boldsymbol{Cv} - \lambda \boldsymbol{v}) = 0 \tag{4.4}$$

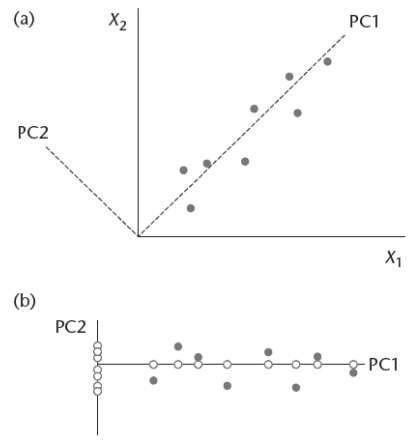
i dalej jako:

$$\boldsymbol{Cv} = \lambda \boldsymbol{v} \tag{4.5}$$

Kolumnowy wektor **v**, spełniający powyższą zależność nazywany jest wektorem własnym macierzy, a związana z nim wartość λ wartością własną macierzy. Cechą wektora własnego macierzy jest to, że nie zmienia on swojego kierunku po przekształceniu opisanym tą macierzą (mnożeniu lewo-

stronnym), zmienia jedynie λ–krotnie swą długość. Ilość wektorów własnych i wartości własnych odpowiada rozmiarowi macierzy poddawanej dekompozycji. Ponadto wektory własne macierzy są wektorami ortogonalnymi, a ich składowe pomnożone przez pierwiastek wartości własnej są równe współczynnikom kombinacji liniowych zmiennych objaśniających tworzących poszczególne czynniki. Odpowiadają one zatem wektorom wierszowym macierzy **A** zależności 4.1. Iloczyn składowej wektora własnego i pierwiastka odpowiadającej mu wartości własnej, jest równy ładunkowi czynnikowemu dla odpowiedniej zmiennej objaśniającej. Taki jest wnoszony przez tę zmienna ładunek informacji w przypadku danego czynnika. Sama wartość własna związana z wektorem własnym jest miarą informacji (zmienności) jaką niesie ze sobą nowa zmienna – składowa główna. Jest kwadratem stosunku długości głównej składowej i długości odpowiadającego jej wektora własnego – zawsze równej jedności.

Przykładem możliwie najprostszej analizy głównych składowych jest przekształcenie dwóch zmiennych objaśniających X_1 i X_2 w zmienne PC_1 i PC_2, zobrazowane graficznie na rysunku 11. Jeśli dla celów poznawczych założymy pewien stopień korelacji między tymi zmiennymi, który jest zresztą widoczny (niech to będzie wartość r = 0.95), to na podstawie macierzy korelacji dla tego prostego przypadku możemy dokonać takiego przekształcenia prowadząc obliczenia samodzielnie.



Rys. 11. Przykład prostego przekształcenia PCA dla dwóch zmiennych objaśniających X1 i X2

Źródło: opr. własne

Pełna macierz korelacji ma zatem postać:

$$\boldsymbol{C} = \begin{bmatrix} 1 & 0.95 \\ 0.95 & 1 \end{bmatrix} \tag{4.6}$$

Aby rozwiązać zagadnienie własne dla tej macierzy (tzn. obliczyć wartości i wektory własne) musimy najpierw rozwiązać równanie charakterystyczne macierzy osobliwej $(\mathbf{C} - \lambda\mathbf{I})$**.** Równanie charakterystyczne to nic innego jak wyznacznik macierzy, a osobliwość oznacza, że wyznacznik ten równy jest zero. Pozwoli nam ono w pierwszym etapie wyznaczyć wartości własne macierzy.

$$det(\boldsymbol{C} - \lambda\boldsymbol{I}) = det\begin{bmatrix} 1-\lambda & 0.95 \\ 0.95 & 1-\lambda \end{bmatrix} = (1-\lambda)^2 - 0.95^2 = 0 \tag{4.7}$$

Rozwiązaniem tego równania są dwie wartości własne λ macierzy **C** (co oczywiste ze względu na rozmiar macierzy) – $\lambda_1 = 1.95$; $\lambda_2 = 0.05$. Oznacza to, że procent całkowitej zmienności (całkowitego ładunku informacji) dla czynnika pierwszego, jest równy $\lambda_1/2 = 97.5\%$.

Obliczone wartości własne należy wykorzystać w kolejnym etapie, pozwalającym na znalezienie wartości składowych wektorów własnych:

$$(\mathbf{C} - \lambda\mathbf{I})\mathbf{v} = \mathbf{0} \qquad \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \tag{4.8}$$

$$\begin{bmatrix} 1-1.95 & 0.95 \\ 0.95 & 1-0.05 \end{bmatrix} = \begin{bmatrix} -0.95 & 0.95 \\ 0.95 & 0.95 \end{bmatrix} \qquad \boxed{\begin{bmatrix} -0.95v_1 + 0.95v_2 \\ 0.95v_1 + 0.95v_2 \end{bmatrix} = 0}$$

Wynikiem mnożenia macierzy opisanych zależnością $(\mathbf{C} - \lambda\mathbf{I})\mathbf{v}$**,** jest wektor kolumnowy (w ramce), którego składowe opisane są równaniami jak wyżej. Po przyrównaniu każdego z nich do zera istnieje możliwość wyznaczenia składowych v_1 i v_2 obydwu wektorów własnych $\mathbf{v}_1$ i $\mathbf{v}_2$. Ponieważ w pierwszym przypadku $\mathrm{v}_1 = \mathrm{v}_2$, w drugim natomiast $\mathrm{v}_1 = -\mathrm{v}_2$, to wektory własne są postaci:

$$\mathbf{v}_1 = [v \;\; v]^T \text{ oraz } \;\; \mathbf{v}_2 = [v \;\; -v]^T \tag{4.9}$$

Z obliczonych wartości v_1 i v_2 widać, że wybór wektorów własnych nie jest jednoznaczny. Składowe tych wektorów mogą być dowolnymi wartościami spełniającymi równania 4.8. Dlatego też, podaje się je, jako wyliczone dla wektorów unormowanych $\|\mathbf{X}\|$=1:

$$\text{dla } \mathbf{v}_1 \text{ i } \mathbf{v}_1\text{: } \sqrt{\mathbf{v}_1^T\mathbf{v}_1} = \sqrt{2v^2} = v\sqrt{2} = 1 \quad \text{to: } v = 0.7071 \qquad (4.10)$$

teraz obydwa wektory własne można przedstawić jako:

$$\mathbf{v}_1 = [0.7071 \quad 0.7071]^T \text{ oraz } \quad \mathbf{v}_2 = [0.7071 \quad -0.7071]^T \qquad (4.11)$$

Obliczona wartość 0.7071 odpowiada kosinusowi kąta 45°. Mogłoby to sugerować, że takie wartości składowych wektorów własnych uzyskujemy tylko dla zależności zmiennych wyjściowych opisanych równaniem $X_1 = X_2$. Tak jednak nie jest. Wartości składowych wektorów $\mathbf{v}_1$ i $\mathbf{v}_2$ zależą tylko i wyłącznie od postaci macierzy korelacji i dla macierzy o rozmiarze 2×2 są zawsze równe 0.7071 i -0.7071. Różnica widoczna jest jedynie dla wartości własnych. Na przykład dla współczynnika korelacji r = 0.9 wartości własne wynoszą $\lambda_1 = 1.9$; $\lambda_2 = 0.1$. Zatem, bez względu na postać równania prostej regresyjnej, dla dwóch zmiennych objaśniających skorelowanych dowolnie silnie, otrzymamy zawsze wektory własne przedstawione równaniami 4.11.

4.2 Przykład PCA z wykorzystaniem oprogramowania Statistica

W oswojeniu się z pewnymi pojęciami związanymi z analizą głównych składowych z pewnością pomoże przykład bardziej skomplikowany, niż ten dla dwóch zmiennych objaśniających, zamieszczony powyżej. Jego analiza, prócz wyznaczenia czynników głównych, pozwoli na zapoznanie się z regułami pozwalającymi na dobór odpowiedniej ich ilości, na właściwe przygotowanie macierzy danych, na zapoznanie się z możliwościami, jakie daje PCA w przestrzeni próbek i wreszcie na zapoznanie się z zasadami interpretacji wyników uzyskiwanych za pomocą tej metody.

Przykładowe dane zaczerpnięte zostały ze statystycznych opracowań Głównego Urzędu Statystycznego i dotyczą procentowego udziału energii

elektrycznej ze źródeł odnawialnych w całkowitym jej zużyciu w kilku wybranych krajach UE, a także dla porównania w całej Unii Europejskiej.

Tab. 11. Struktura pozyskania energii z wybranych źródeł energii odnawialnej w wybranych krajach UE w 2012 roku w [%]

	Kraj 2012	biom	prom	wod	wiatr	biog	biop	geo	okom
UE	UE–28	47.2	5.1	16.2	10	6.8	6.5	3.2	4.9
AT	Austria	50.1	2.1	39.1	2.2	2.2	2.5	0.4	1.5
CZ	Czechy	66.3	6.1	5.6	1.1	11.5	6.7	0	2.6
EE	Estonia	95.9	0	0.3	3.5	0.3	0	0	0
FI	Finlandia	79.7	0	14.6	0.4	0.6	2.8	0	1.9
LT	Litwa	82.8	0	3	3.9	1	9	0.3	0
LV	Łotwa	80.2	0	13.7	0.4	2.2	3.5	0	0
DE	Niemcy	35.9	8.6	5.5	13.2	19.5	9	0.3	7.9
PL	Polska	82.4	0.2	2.1	4.8	2	8	0.2	0.4
SK	Słowacja	55.9	2.9	24.6	0	4.3	10.5	0.4	1.3
SE	Szwecja	51.7	0.1	36.7	3.3	0.7	3.4	0	4.2

Oznaczenia: biom – biomasa, prom – promieniowanie słoneczne, wod – elektrownie wodne, wiatr – wiatrowe, biog – biogaz, biop – biopaliwa, geo – energia geotermalna, okom – odpady komunalne

Źródło: GUS

Analizę czynnikową można prowadzić w oparciu o macierz kowariancji, a także macierz korelacji. Jeżeli analizowane zmienne są porównywalne w tym sensie, że wyrażane są w tych samych jednostkach, a wartości są tego samego rzędu, to do dalszej analizy możemy wykorzystać zarówno macierz korelacji jak i macierz kowariancji. W przeciwnym przypadku analizę składowych głównych przeprowadza się wykorzystując macierz korelacji. Składowe główne otrzymane dla macierzy kowariancji i korelacji nie muszą być takie same w przypadku, gdy zmienne nie spełniają opisanych wyżej warunków.

W programie Statistica możliwy jest wybór pomiędzy jednym i drugim rozwiązaniem. Standardowo, jako macierz poddawana dekompozycji wybierana jest macierz korelacji, dlatego też przy niej pozostaniemy. Jeśli korzystamy z automatycznych rozwiązań (oprogramowania statystycznego), w przypadku metody PCA nie musimy pamiętać o standaryzacji zmiennych objaśniających, która jest konieczna przy wyznaczaniu macierzy korelacji (zależność 4.2). Aplikacja zrobi to za nas. Musimy jedynie pamiętać, że algorytm PCA wbudowany w oprogramowanie do analizy statystycznej zwykle usuwa zmienne i obiekty, które posiadają braki danych. Należy je zatem, jeśli to możliwe, uzupełnić według opisanych wcześniej zasad. W sposób automatyczny otrzymamy również wszystkie wielkości pozwalające na interpretację wyników dla analizowanego zbioru danych wejściowych (objaśniających).

Przystępując do analizy naszych danych warto jest na początku przyjrzeć się ich macierzy korelacji. Pozwoli nam to na wstępną ocenę, czy analiza PCA może być dla nich efektywna:

Tab. 12. Macierz korelacji zmiennych na podstawie danych z tab. 11

Zmienna	Korelacje (energia NOWA.sta) Oznaczone wsp. korelacji są istotne z p < .05000 N=11									
	Średnia	Odch.std	biom	prom	wod	wiatr	biog	biop	geo	okom
biom	66.1909	19.1128	1.0000							
prom	2.2818	3.0459	**-0.7189**	1.0000						
wod	14.6727	13.6054	-0.5406	-0.1126	1.0000					
wiatr	3.8909	4.1889	-0.5187	**0.6229**	-0.2561	1.0000				
biog	4.6455	5.9621	**-0.6244**	**0.9553**	-0.2518	**0.6488**	1.0000			
biop	5.6273	3.3610	-0.3532	0.4671	-0.2241	0.3059	0.4821	1.0000		
geo	0.4364	0.9320	-0.4053	0.3518	0.0778	0.5198	0.1550	0.1912	1.0000	
okom	2.2455	2.5113	**-0.8298**	**0.7824**	0.1414	**0.7593**	**0.7767**	0.2556	0.3620	1.0000

Źródło: opr. własne

W tabeli widocznych jest kilka wartości, dla których współczynniki korelacji mają istotnie dużą wartość (czcionka bold), a także wartość $r > |0.5|$. Duże wartości współczynników są jak wiadomo cechą pożądaną w przypadku analizy PCA, ponieważ tylko wtedy analiza głównych składowych może skutecznie zmniejszyć ilość parametrów opisujących obiekty. Najlepszym wskaźnikiem skuteczności redukcji wymiaru przestrzeni zmiennych pierwotnych są wartości własne każdego z czynników (wektorów własnych) Tab. 12. Im większa wartość własna kilku pierwszych czynników, tym większa ilość informacji nadmiarowej zawarta była w zmiennych wyjściowych.

Należy w tym miejscu przypomnieć, że dla zmiennych standaryzowanych suma wartości własnych wszystkich czynników modelu nie może przekroczyć całkowitej wariancji zmiennych równej liczbowo ilości zmiennych objaśniających. W przypadku naszych przykładowych danych liczba ta jest równa 8.

Tab. 13. Wartości własne głównych składowych modelu

Nr czynnika	Wartości własne (energia 2012.sta)			
	Wartość wł	% ogółu warianc.	Skumul. wartość wł	Skumul. %
1	4.331783	54.14728	4.331783	54.1473
2	1.545248	19.31560	5.877031	73.4629
3	0.945383	11.81729	6.822414	85.2802
4	0.766269	9.57836	7.588683	94.8585
5	0.324382	4.05478	7.913065	98.9133
6	0.078836	0.98546	7.991901	99.8988
7	0.008097	0.10121	7.999998	100.0000
8	0.000002	0.00003	8.000000	100.0000

Źródło: opr. własne

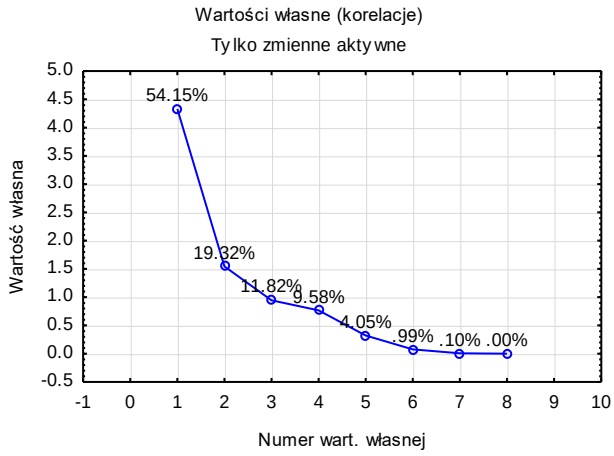
Skuteczność i jednocześnie sens analizy PCA na przytoczonych danych jest zauważalny. Potwierdzają to wartości w pierwszych trzech wierszach tabeli 13. Skumulowane wartości własne dla trzech pierwszych czynników sięgają 85.3% całkowitej wariancji zmiennych. Pierwsze cztery składowe główne wyjaśniają już 94.9% całkowitej zmienności w przestrzeni pierwotnych zmiennych. W tym miejscu pojawia się zawsze pytanie, jaki procent zmienności jest zadowalający – ile czynników należy uwzględnić w modelu powstałym dzięki analizie PCA. Istnieje kilka kryteriów pozwalających podjąć taką decyzję, ale jak można się spodziewać nie są one nigdy wiążące, a stopień wymiarowości przestrzeni głównych składowych jest rzeczą subiektywną i bardzo często zależną od tego, co było celem podstawowym analizy – wizualizacja zależności pomiędzy zmiennymi, czy może wykorzystanie czynników do budowy modelu zależności badanych zmiennych.

Dokonując wyboru ilości istotnych składowych musimy pamiętać istnieniu dwóch kategorii zmienności: właściwej, wynikającej tylko z charakteru badanych obiektów oraz niepożądanej zmienności losowej, jaką obarczone są pomiary wartości cech obiektów. Wybierając składowe dokonujemy zawsze wyboru pomiędzy tymi zawierającymi zmienność pożądaną i wynikającą ze zmienności losowej. Wybór ten, z powodu własności czynników nigdy w zasadzie nie prowadzi do poważnych błędów. Zawsze bowiem wybieramy te o największych wartościach własnych, czyli te, w których zawartość zmienności niepożądanej jest najmniejsza.

W najbardziej typowych sytuacjach liczba uznanych jako istotne nowych zmiennych jest z reguły dużo mniejsza niż całkowita liczba składowych. Podstawowymi, najczęściej uwzględnianymi kryteriami ich wyboru są między innymi: tzw. kryterium poglądowości, zasobu zmienności, spadku wartości własnej, zredukowanych wartości własnych czy wreszcie kilka opartych o wskaźniki liczbowe jak IND i różne kryteria kompozytowe. Kryterium poglądowości jest bezpośrednio związane z celem analizy polegającym tylko i wyłącznie na wizualizacji zależności pomiędzy zmiennymi. Dlatego najwłaściwszym wyborem są tutaj w zależności od rodzaju wykresu dwie lub trzy główne składowe. Kryterium zasobu zmienności można z kolei uznać za najprostsze i w związku z tym często wątpliwe i mało elastyczne. Istnieją przy tym dwa podejścia oceny istotności czynników: w pierwszym z nich, podstawą decyzji jest skumulowana wartość własna wybieranych czynników większa niż 90 – 95 % całkowitej wariancji; w drugim podejściu, jako istotne uznaje się jedynie te czynniki, których wartość własna jest większa od jedności. Odrzucamy w tym przypadku czynniki, dla których zasób zmienności jest mniejszy niż zasób zmienności pojedynczej zmiennej objaśniającej. W przypadku tego kryterium bywa również, że przyjmowane jest inne kryterium wartości progowej – np. 0.75 lub 0.5.

Jednym z ‘graficznych’ kryteriów wyboru właściwej ilości nowych zmiennych (czytaj czynników) jest tzw. kryterium spadku wartości własnej. W tym przypadku decyzję podejmujemy na podstawie zmiany kształtu zależności przedstawiającej wartości własne czynników w funkcji ich numeru. Punktem podziału (odcięcia) czynników na istotne i nie, jest zmiana szybkości spadku wartości własnej na wykresie. Zwykle przebieg takiego wykresu

przedstawia szybki spadek wartości własnych dla kilku pierwszych zmiennych, by w dalszej części przyjąć charakter poziomej linii, obrazującej brak zmian λ.



Rys. 12. Wykres osypiska dla wartości własnych modelu PCA

Źródło: opr. własne

W przypadku prezentowanego przykładu, na podstawie opisanego wyżej kryterium i wykresu 12, wybór czynników istotnych należałoby ograniczyć do pierwszych trzech/czterech głównych składowych.

Większą możliwość wyboru i jednocześnie możliwość podjęcia bardziej obiektywnej decyzji daje sytuacja, w której mamy do czynienia z dużą liczbą zmiennych modelu PCA – czynników. Te bardziej zaawansowane kryteria oparte są o analizę statystyczną rozkładu wartości własnych. Pierwszym z nich jest wskaźnik IND, którego podstawą jest estymacja tak zwanego błędu rzeczywistego RE_k, który jest związany z wielkością sumarycznej wariancji odrzucanych czynników. Dla tzw. szerokich tablic chemometrycznych ($m > n$, gdzie m jest ilością zmiennych objaśniających natomiast n ilością obiektów) błąd ten przedstawia zależność:

$$RE_k = \sqrt{\frac{n - \sum_{i=1}^{k} \lambda_i}{(n-k)m}} \tag{4.12}$$

gdzie:

k – ilość czynników istotnych;

n – ilość obiektów;

m – ilość zmiennych

natomiast wartość ostateczną wskaźnika kryterium IND oblicza się jako:

$$IND_k = \frac{RE_k}{(g-k)^2} \tag{4.13}$$

gdzie: $(\mathrm{g} - \mathrm{k})$ – to ilość nieuwzględnionych głównych składowych

Optymalna liczba składowych głównych odpowiada sytuacji, przy której wartość wskaźnika osiąga minimum lub zaczyna szybko rosnąć.

Kryterium zredukowanych wartości własnych opiera się na założeniu, że wartości te, związane ze zmiennością losową (niepożądaną), są porównywalne i statystycznie dużo mniejsze od zredukowanych wartości własnych składowych istotnych. Zredukowana wartość własna *k*-tej składowej dla szerokiej macierzy danych jest wyrażona zależnością:

$$REV_k = \frac{\lambda_k}{(n-k-1)m} \tag{4.14}$$

Do porównania różnicy pomiędzy wartościami zredukowanymi, wyliczonymi dla ostatniej istotnej składowej i sumy wartości zredukowanych dla pozostałych składowych REV_{g-k}, stosuje się test F Snedecora:

$$F = \frac{REV_k}{REV_{g-k}} \tag{4.15}$$

Obliczoną wartość, zgodnie z zasadami testowania hipotez, porównuje się z odpowiednią, stablicowaną wartością krytyczną, a pierwsze wystąpienie zależności $F < F_{kr}$ jest sygnałem wskazującym na ostatnią istotną składową.

Opisane kryteria bardzo często nie dają wyraźnych sygnałów pozwalających na podjęcie decyzji o istotności czynników. W praktyce chemometrycznej bardzo często konstruuje się wskaźniki będące złożeniem innych. Przykładem takiego kompozytowego, zobiektywizowanego, o większej skuteczności jest np. wskaźnik oparty na dwóch poprzednich. Bez względu na konstrukcję takich wskaźników, nie należy ich traktować jako kryteriów absolutnych. W przypadkach trudnych i wątpliwych najlepszym wskaźnikiem zawsze będzie intuicja, doświadczenie, a przede wszystkim optymalizacja decyzji w zależności od celu prowadzonej analizy.

4.3 Interpretacja wyników analizy głównych składowych

Przekształcenie zmiennych wyjściowych w nieskorelowane główne składowe, pomimo niezaprzeczalnych zalet posiada także pewną wadę. Jest nią brak możliwości fizycznej interpretacji nowego układu zmiennych. Odwzorowanie zmiennych objaśniających obiektów w nowej przestrzeni czynników, zwykle o mniejszym wymiarze, powoduje częściową utratę informacji na temat fizycznego charakteru zmiennych pierwotnych. W nowym układzie zmiennych, wielkościami, które mogą określać stopień powiązania między osiami jednego i drugiego układu są korelacje pomiędzy nimi. Korelacje reprezentowane poprzez tzw. ładunki czynnikowe określane dla nowych zmiennych. W przypadku przykładowych danych z tabeli 11, ich wartości dla każdego czynnika przedstawia tabela poniżej. Cechą każdego wiersza pełnej tabeli ładunków czynnikowych, jest jego jednostkowa wartość, liczona jako pierwiastek sumy kwadratów poszczególnych ładunków dla zmiennej objaśniającej. Przy ograniczonej ilości głównych składowych, wielkość ta, mniejsza niż 1, jest informacją o stopniu wykorzystania zmienności zmiennej wyjściowej przez nowe zmienne – czynniki.

Tab. 14. Ładunki czynnikowe dla zestawu zmiennych wyjściowych na przykładzie tab. 11

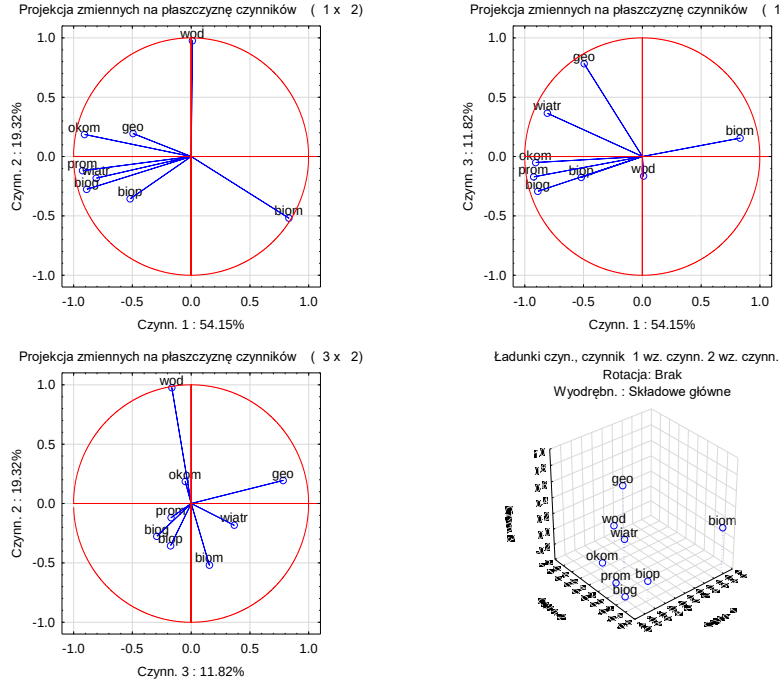
Zmienna	Ładunki czynnik (energia 2012.sta) Wyodrębn. : Składowe główne							
	Czynnik 1	Czynnik 2	Czynnik 3	Czynnik 4	Czynnik 5	Czynnik 6	Czynnik 7	Czynnik 8
biom	0.8319	-0.5193	0.1547	-0.1006	0.0391	-0.0516	0.0085	0.0011
prom	-0.9245	-0.1195	-0.1714	-0.0475	0.3058	0.0501	0.0586	0.0002
wod	0.0104	0.9755	-0.1641	0.1304	-0.0360	0.0543	-0.0015	0.0008
wiatr	-0.8083	-0.1824	0.3657	-0.2236	-0.3323	0.1381	0.0098	0.0002
biog	-0.8893	-0.2757	-0.2929	-0.1414	0.1482	0.0367	-0.0642	0.0003
biop	-0.5195	-0.3568	-0.1760	0.7379	-0.1627	-0.0282	0.0042	0.0002
geo	-0.4946	0.1949	0.7824	0.2381	0.2163	-0.0385	-0.0163	0.0001
okom	-0.9090	0.1872	-0.0498	-0.2563	-0.1498	-0.2192	0.0088	0.0001

Źródło: opr. własne

Analizując wartości ładunków czynnikowych dla poszczególnych składowych głównych, można zauważyć, że aż pięć zmiennych objaśniających

jest ze sobą na tyle silnie skorelowanych, że ich ładunki w pierwszym czynniku są wyższe niż 0.8, a często bliskie wartości 0.9. Pozostałe trzy zmienne wyjściowe dotyczące ‘elektrowni wodnych’, ‘energii geotermalnej’ i ‘biopaliwa’ są silniej skorelowane odpowiednio z czynnikami 2, 3 i 4. Przy czym zmienne objaśniające odpowiadające ‘energii geotermalnej’ i ‘biopaliwu’ leżą wyraźnie pomiędzy dwoma czynnikami, odpowiednio – 1 i 3 oraz 1 i 4. Duży udział w czynniku 2 ma także zmienna odpowiadająca procentowemu udziałowi energii pozyskiwanej z ‘biomasy’.

Wszystkie te zależności można zobrazować graficznie, jako wektory zmiennych wyjściowych w przestrzeni czynników. Dodatkową informacją, jaką dostajemy tworząc wykresy, jest długość tych wektorów. Ich znacznie mniejsza niż promień okręgu długość, jest dowodem na to, iż wychodzą one poza płaszczyznę projekcji określoną wybranymi czynnikami. Sytuację taka obserwujemy dla zmiennych ‘geo’ i ‘biop’ w płaszczyźnie czynników 1 i 2, co jest naturalne z uwagi na duży udział tych zmiennych w czynnikach 3 i 4.



Rys. 13. Projekcje wektorów zmiennych objaśniających w przestrzeni czynników

Źródło: opr. własne

Gdyby pokusić się o krótkie, jakościowe podsumowanie prezentowanych na diagramach zależności, można by było powiedzieć, że procentowy udział energii pochodzącej z takich źródeł odnawialnych jak elektrownie wodne, źródła energii geotermalnej i biopaliwa, w wybranych krajach Unii Europejskiej nie jest skorelowany z udziałem energii odnawialnej pochodzącej z pozostałych źródeł. Struktury procentowego udziału energii 'zielonej' w całkowitej energii wytworzonej w tych krajach, różnią się przede wszystkim pod względem ilości energii wyprodukowanej w elektrowniach wodnych.

Opis interpretacji wyników analizy głównych składowych jest dobrym miejscem, aby jeszcze raz zastanowić się nad fizycznym znaczeniem osi nowego, ortogonalnego układu zmiennych. Jak już wcześniej zostało powiedziane, nadanie czynnikom cech podobnych do tych, jakie posiadały osie pierwotnego układu zmiennych jest w zasadzie niemożliwe. Dlatego też opracowane zostały metody pozwalające na otrzymanie zbioru czynników dających się interpretować w łatwiejszy sposób, a jednocześnie posiadających wszystkie zalety czynników wyjściowych. Jedną z takich metod jest rotacja układu głównych składowych. Najbardziej uznanym i najczęściej stosowanym algorytmem pozwalającym na takie przekształcenie jest metoda VARIMAX. Algorytm ten zakłada możliwość takiej rotacji układu czynników, która prowadzi do jednoczesnej maksymalizacji wariancji (w praktyce uśrednienia) możliwie dużej liczby wybranych wektorów głównych składowych. Stąd nazwa metody.

Wariancja każdego z wektorów jest największa, gdy tylko jedna z jego składowych ma dużą wartość bezwzględną, podczas gdy pozostałe są jak najbliższe zeru. Należy wspomnieć, że inaczej niż jest to opisywane w różnych źródłach, przekształcenie VARIMAX, tak jak inne polegające na zmianie tylko kierunku wektorów głównych składowych, niesie ze sobą brak zmian sumy wariancji wektorów nowego układu. Po takiej transformacji układu czynników będących wynikiem analizy PCA, położenie jego nowych osi pokrywa się z położeniem grup wektorów kierunkowych układu zmiennych objaśniających. Zatem nie jest przypadkowe i nadaje im taką samą interpretację, jaką posiadały odpowiednie wektory kierunkowe układu wyjściowych zmiennych. Nowe kierunki ortogonalnych osi układu odzwierciadlają teraz

w dużym stopniu rzeczywiste relacje, jakie występowały pomiędzy zmiennymi objaśniającymi.

Rotacja układu głównych składowych powoduje, że jego nowe osie nie mogą być dalej nazywane czynnikami/głównymi składowymi. Nie spełniają one bowiem kryterium możliwie maksymalnej zmienności każdej kolejnej osi układu głównych składowych. Nowe osie, po obrocie maksymalizującym ich wariancję, pozostają wciąż nieskorelowane, wzajemnie ortogonalne. Przyjęto dla nich nową nazwę variwektorów. Należy też wspomnieć, że rotacji przestrzeni można dokonać biorąc pod uwagę jedynie pewną liczbę, k istotnych czynników. Łączny ładunek zmienności po obrocie pozostaje zachowany. Dochodzi tylko do bardziej równomiernego jego rozkładu w przestrzenie variwektorów.

Tab. 15. Rozkład ładunków czynnikowych przed i po obrocie VARIMAX dla k czynników

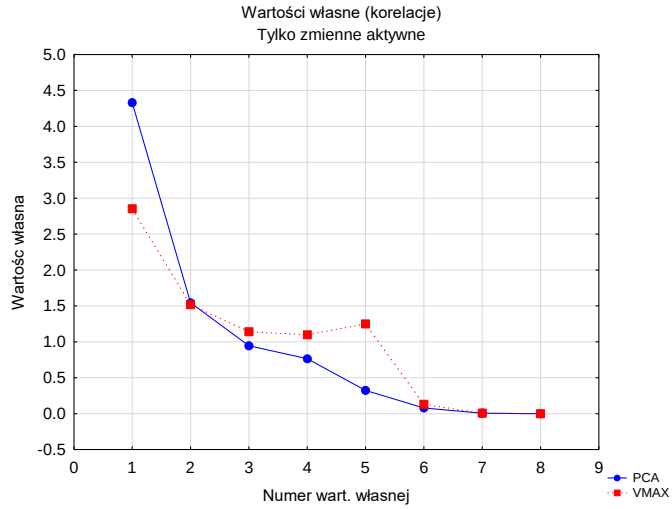
Zmienna	Ładunki czynnik (energia 2012.sta) Wyodrębn. : Składowe główne / variwektory							
	Czynnik 1	Czynnik 2	Czynnik 3	Czynnik 4	Vwektor 1	Vwektor 2	Vwektor 3	Vwektor 4
biom	0.8319	-0.5193	0.1547	-0.1006	-0.7123	0.6181	-0.2396	-0.2213
prom	-0.9245	-0.1195	-0.1714	-0.0475	0.8863	0.0064	0.1368	0.3104
wod	0.0104	0.9755	-0.1641	0.1304	-0.0651	-0.9865	0.0122	-0.1350
wiatr	-0.8083	-0.1824	0.3657	-0.2236	0.7198	0.2372	0.5440	0.0100
biog	-0.8893	-0.2757	-0.2929	-0.1414	0.9293	0.1503	-0.0256	0.2928
biop	-0.5195	-0.3568	-0.1760	0.7379	0.2360	0.0875	0.1020	0.9481
geo	-0.4946	0.1949	0.7824	0.2381	0.1408	-0.1013	0.9536	0.1094
okom	-0.9090	0.1872	-0.0498	-0.2563	0.9082	-0.2073	0.2483	0.0037

Źródło: opr. własne

W przypadku analizowanych w tym rozdziale, przykładowych danych dotyczących źródeł energii odnawialnej w krajach UE widzimy, że metoda obrotu VARIMAX (Tab. 15) powoduje zmiany ilościowe w ładunkach wszystkich czynników. Najbardziej widoczne są one dla czynnika 3 i 4. Prócz tego,

dla czynników 1 i 2 widoczna jest zmiana kierunku wektorów o 180°. Efekt rotacji widoczny jest najbardziej dla wektorów zmiennych objaśniających, które są najmniej skorelowane z pozostałymi. W przykładzie dotyczy to zwłaszcza zmiennych opisujących procentową zawartość wyprodukowanej energii w elektrowniach wodnych, z biopaliw i geotermalnej. Rotacja pozwala na bardziej oczywiste przypisanie 'byłym czynnikom' 2, 3 i 4 charakteru tych zmiennych. Inaczej, kierunki variwektorów lepiej odzwierciedlają kierunki zmiennych wyjściowych (objaśniających).

Po obrocie układu wektorów głównych składowych zmieniają się również ich wartości własne. Jest to oczywiste, ponieważ jak wiemy, są one równe sumie kwadratów ładunków czynnikowych każdego z nowych wektorów. Zmiany tych wartości po przekształceniu VARIMAX są widoczne, gdy porównamy wykresy osypisk dla każdego układu.



Rys. 14. Wykresy spadku wartości własnych dla analizowanych danych przykładowych PCA – wartości własne czynników; VMAX – wartości własne variwektorów

Źródło: opr. własne

Porównanie krzywych spadku wartości własnych w kolejnych czynnikach i variwektorach odpowiednio dla PCA – linia ciągła i PCA–VARIMAX – linia przerywana, pozwala wyciągnąć wniosek, że w przypadku układu po rotacji wartości własne są rozłożone bardziej równomiernie pomiędzy pięć początkowych zmiennych, niż ma to miejsce dla głównych składowych.

W badaniach chemometrycznych wszelkiego rodzaju rotacje wykonuje się przeważnie uwzględniając wszystkie czynniki, tj. dla pełnowymiarowego układu składowych głównych. Takie podejście sprzyja łatwiejszej decyzji dotyczącej wyboru ilości istotnych zmiennych. Wartości własne variwektorów układają się zwykle w bardziej jednoznaczny sposób. Istnieje wyraźniejszy podział na wektory o niskich i wysokich wartościach tej wielkości. Dzięki temu wystarczającym kryterium decyzyjnym może być szybkość spadku wartości własnych w połączeniu z ustaloną wielkością minimalnego zasobu zmienności.

4.4 Analiza przestrzeni obiektów

Macierze ładunków wektorów dla nowych układów zmiennych, bez względu na to czy są nimi variwektory, czy główne składowe zawierają w sobie informację o położeniu badanych obiektów w nowej przestrzeni zmiennych. Współrzędne wskazujące na położenie każdego z obiektów można wyliczyć mnożąc macierz standaryzowanych zmiennych wyjściowych **Z** (objaśniających) przez macierz ładunków czynnikowych wektorów będących nowymi zmiennymi $\mathbf{C_k}$ (czynniki) i $\mathbf{V_k}$ (variwektory).

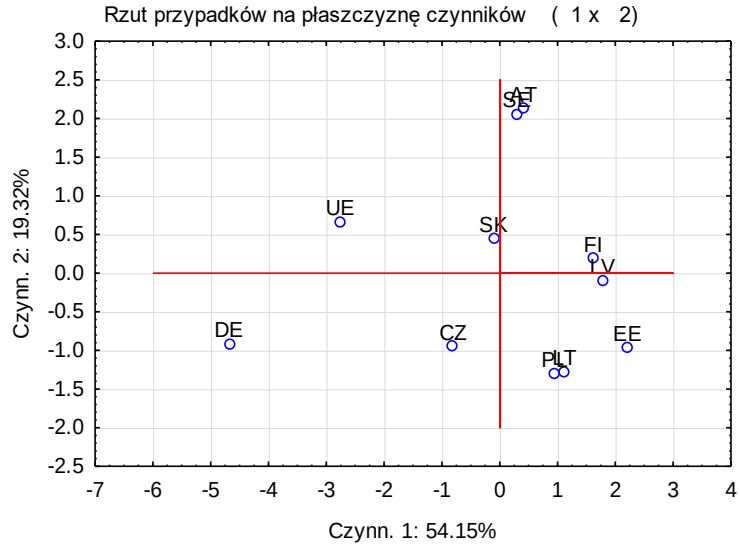
$$\begin{aligned} \boldsymbol{S_k} &= \boldsymbol{ZC_k} \\ \boldsymbol{R_k} &= \boldsymbol{ZV_k} \end{aligned} \tag{4.16}$$

gdzie:

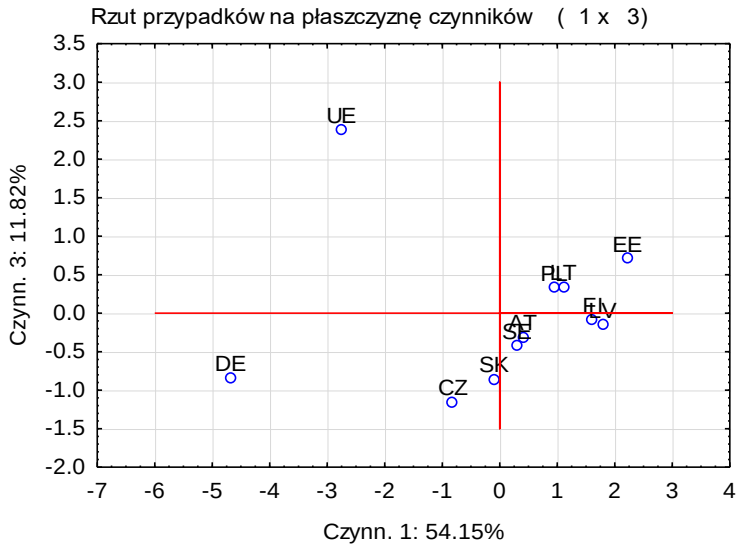
$\mathbf{S_k i\ R_k}$ – to odpowiednio macierz współrzędnych obiektów w przestrzeni głównych składowych i przestrzeni variwektorów.

Obliczone macierze w swoich wierszach zawierają współrzędne obiektów w nowej, k–wymiarowej przestrzeni. Przekształcenie takie, związane zwykle z redukcją ilości zmiennych opisujących obiekty, pozbawia nas niewielką części informacji o nich, ale jego wielką zaletą jest możliwość wykorzystania zdolności naszego mózgu do analizy przestrzeni dwu–, trójwymiarowej. W sytuacji, kiedy dwa, trzy czynniki (variwektory) zawierają w sobie większość zmienności cech badanych obiektów, możemy pokusić się o interpretację zależności między nimi (obiektami) na podstawie obrazów

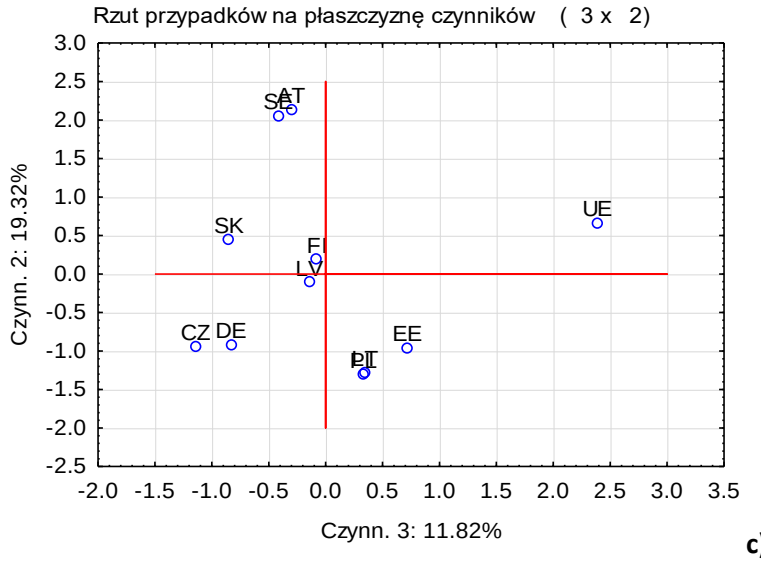
rzutowania ich położenia na dwuwymiarową płaszczyznę wybranych zmiennych, lub analizę ich rozmieszczenia w przestrzeni trójwymiarowej. Rzutowanie takie zapewnia nam maksymalną ilość informacji przy wskazanej ilości istotnych zmiennych w nowej przestrzeni.



a)



b)



Rys. 15. Rzuty położenia obiektów na płaszczyzny czynników: a) 1-2, b) 1-3 i c) 2-3

Źródło: opr. własne

Wykresy zamieszczone na rysunku 15 to właśnie wynik rzutowania położenia obiektów (państw) na płaszczyzny czynników głównych, dla analizowanych przykładowych danych dotyczących energii odnawialnej. Rzutowania takie, nazywane mapami liniowymi, dostarczają przede wszystkim informacji o położeniu obiektów względem siebie. Ich relacje zwykle można podzielić na trzy przypadki: **równomierne rozmieszczenie w przestrzeni**, co pozwala na wniosek, że pomiędzy zmiennymi istnieje co najwyżej zależność liniowa, a obiekty są z tej samej populacji; **widoczna zależność nieliniowa rozmieszczenia** – obiekty rozmieszczone są wzdłuż pewnej krzywej, której kształt dostarcza nam informacji o stopniu nieliniowości zależności i jej typie; **widoczne skupienia obiektów** – na podstawie takiego obrazu mapy liniowej możemy mówić o tendencjach obiektów do tworzenia grup o podobnych wartościach składowych wektorów zmiennych objaśniających, a także o obiektach odosobnionych. Jest to przypadek interesujący, gdy prowadzimy badania pod kątem podobieństwa obiektów.

Analiza map liniowych dla naszych danych pozwala na znalezienie dwóch z wyżej opisanych zależności. Przypadek równomiernego rozmieszczenia obiektów mógłby przedstawiać jedynie wykres 15a (płaszczyzna:

czynnik 1 – czynnik 2). Ostateczne wnioski należy jednak wyciągać na podstawie oglądu większej ilości rzutów, przynajmniej dwóch, co pozwala na 'upewnienie się', czy rzeczywiście rzut na płaszczyznę czynników 1-2 odzwierciedla jednoznacznie równomierne rozmieszczenie punktów/krajów w przestrzeni. Przyglądając się pozostałym rzutom widzimy, że tak jednak nie jest. W perspektywie płaszczyzny czynników 1-3 widzimy dwa punkty odbiegające od pozostałych. Odpowiadają one wektorowi wartości średnich dla Unii i wektorowi wartości zmiennych wyjściowych dla Niemiec. Dlaczego tak jest, odpowiedź daje nam analiza wartości tablicy z danymi wyjściowymi. Dodatkowo, na płaszczyźnie czynników 1-3 widzimy najwyraźniej liniową zależność/zgrupowanie punktów/krajów. Oznacza to ich przynajmniej częściowe podobieństwo z punktu widzenia struktury procentowego udziału energii 'zielonej' w jej całkowitym zużyciu. Częściowe, ponieważ widzimy, że w płaszczyznach prostopadłych do 1-3, skupienie nie są już tak zwarte. Ciekawą cechą jest to, że na wszystkich płaszczyznach rzutowania możemy obserwować bliskie sobie punkty odpowiadające zawsze tym samym krajom. Oznacza to całkowitą zgodność struktury procentowego udziału energii odnawialnej pochodzącej z różnych źródeł w tych krajach. Przyczyna takiego podobieństwa może leżeć na przykład w bliskości ich geograficznego położenia i tym samym na podobnej dostępności określonych źródeł energii. Innym powodem może być podobny stopień rozwoju gospodarczego takich państw i w związku z tym podobny stopień uprzemysłowienia pozwalający na stosowanie tych samych technologii w dziedzinie energetyki. Takie pary to Szwecja i Austria, Litwa i Polska oraz Łotwa i Finlandia.

Podsumowując najważniejsze, zebrane w rozdziale informacje dotyczące metody zwanej analizą głównych składowych, bądź też analizą czynników głównych, musimy podkreślić, że jest to jedna z najczęściej stosowanych metod analizy czynnikowej, która daje możliwość skutecznej analizy wstępnej nieznanego zbioru danych. Analizę taką można ukierunkować na badanie układu zmiennych objaśniających, a także zależności obiektów w nowej przestrzeni zmiennych – czynników/głównych składowych. Najważniejszym celem analizy PCA jest możliwość redukcji wymiarowości przestrzeni zmiennych, co ułatwia tworzenie innych modeli chemometrycznych, a także wizualizacja 2D i 3D problemów opisywanych ich wielowymiarowym układem.

5 ANALIZA SKUPIEŃ

Metody analizy chemometrycznej, znane pod ogólna nazwą analizy skupień (rozpoznawania wzorców, ang. pattern recognition) to metody, których algorytmy dostarczają sposobów pozwalających na grupowanie obiektów opisywanych za pomocą wektora cech, a także na tworzenie skupień (grup, klastrów) zmiennych, jak może to mieć miejsce np. w przypadku analizy podobieństwa. We wszystkich metodach analizy skupień obowiązuje ogólna zasada, że obiekty należące do pojedynczego klastra powinny być do siebie jak najbardziej podobne, natomiast należące do różnych skupień – przeciwnie. Przy czym ich niepodobieństwo powinno być tym większe im większa jest odległość skupień w przestrzeni zmiennych. Nie jest to problem, który łatwo rozwiązać. Istnieje wielka liczba możliwych algorytmów prowadzących do tego celu. W zależności od ilości informacji jakie posiadamy na temat analizowanego zbioru obiektów, rozróżniane są dwie klasy metod rozpoznawania wzorców. Rozpoznawanie nienadzorowane (bez nauczyciela, ang. unsupervised patter recognition) oraz nadzorowane (z nauczycielem, ang. supervised patter recognition). Z każdą z klas związany jest główny algorytm grupowania, określający sposób tworzenia klastrów. Hierarchiczny aglomeracyjny lub podziałowy, generujący sekwencję różnej ilości skupień i tak zwany płaski (uczenie nadzorowane), w przypadku którego liczba klastrów jest z góry założona i nie zmienia się. Rozpoznawanie nienadzorowane prowadzone jest bez wstępnych założeń o ilości i rodzaju możliwych klastrów. Stosujemy je, gdy chcemy rozpoznać wewnętrzną strukturę zbioru elementów, a w szczególności, gdy chcemy wykryć obecność ewentualnych podobieństw obiektów i/lub cech. Jest rzeczą intuicyjnie oczywistą, że w przypadku grupowania nadzorowanego konieczna jest wstępna znajomość struktury zbioru obiektów. Podstawową informacją jest liczba grup, na jakie dzieli się zbiór oraz znajomość cech przedstawiciela każdej z nich. Na tej podstawie algorytmy rozpoznawania nadzorowanego umożliwiają klasyfikację (przyporządkowanie do odpowiednich grup) obiektów nowych, niezdefiniowanych.

Najczęściej wykorzystywanymi przykładami obu rodzajów metod są odpowiednio **analiza podobieństwa** (jako analiza rozpoznawcza) oraz jedna z metod nadzorowanego rozpoznawania obrazów, jaką jest **grupowanie**

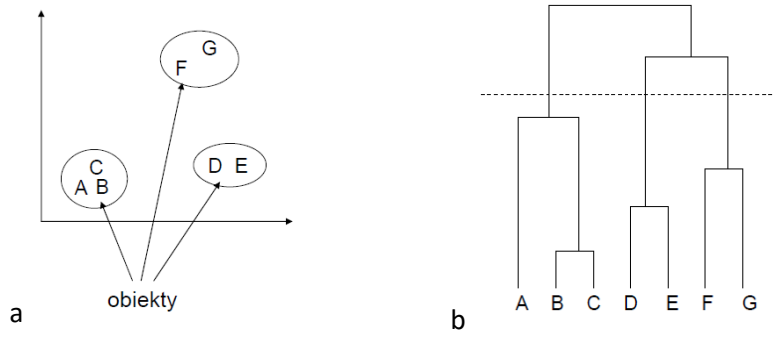
metodą *k*-średnich. Metoda znana jako analiza podobieństwa, zostanie krótko zaprezentowana w dalszej części tekstu wraz z przykładem jej wykonania.

5.1 Analiza podobieństwa (Cluster Analysis)

Analiza podobieństwa jest jedną z najstarszych, najczęściej stosowanych metod chemometrycznych, które noszą wspólna nazwę metod grupowania lub z angielskiego – rozpoznawania wzorców (pattern recognition). Jest metodą analizy danych pozwalającą na tworzenie naturalnych skupień (klastrów) obiektów, które są do siebie podobne a ich podobieństwo można interpretować za pomocą wektora cech. Te naturalne skupienia tworzone są z wykorzystaniem algorytmów, w przypadku których nie musimy z góry dysponować informacją na temat możliwych niejednorodności w analizowanym zbiorze obiektów. Algorytmy analizy podobieństwa wykorzystują pojęcie odległości obiektów (lub cech – co jest istotne w poszukiwaniach zmiennych niosących podobne informacje) w przestrzeni zmiennych przy założeniu, że obiekty podobne będą w niej położone blisko siebie. Przed przystąpieniem do analizy nie jest znana ani liczba klastrów, ani przynależność obiektów do ewentualnych grup. Wynik analizy, stanowi zatem materiał wyjściowy do określenia wewnętrznej struktury wielowymiarowego zbioru obiektów i/lub ich cech. Analiza podobieństwa może być celem sama w sobie, ale najczęściej jest wstępem do kolejnych etapów analizy chemometrycznej. Jej wynik wykorzystywany jest głównie jako źródło danych do redukcji liczby zmiennych niezależnych i tym samym redukcję wymiarowości przestrzeni zmiennych, a także jako sposób na graficzną interpretację wewnętrznej struktury zbioru badanych obiektów.

Algorytm grupowania hierarchicznego aglomeracyjnego, bo to na nim głównie oparta jest analiza podobieństwa, bazuje na pojęciu odległości **obiektów lub zmiennych** w przestrzeni wielowymiarowej i polega na sekwencyjnym grupowaniu kolejnych elementów w klastry. Początkowo każdy z elementów zbioru tworzy osobną grupę. W kolejnym kroku poszukiwane są dwa z nich, najbardziej do siebie podobne (najmniej odległe) i tworzona jest z nich nowa grupa, w tym przypadku dwuelementowa.

Powtarzany w ten sposób krok porównywania odległości pojedynczych elementów lub później ich grup i ich łączenia, prowadzi w końcowym etapie do powstania jednej grupy, składającej się ze wszystkich elementów analizowanego zbioru. Końcowy wynik nie jest oczywiście celem analizy. Na którymś z etapów grupowania (jego wybór jest rzeczą subiektywną) należy bowiem je zakończyć, wybierając najbardziej sensowny w interpretacji podział całego zbioru. Najpopularniejszym, graficznym wynikiem opisanej analizy jest drzewo klastrów zwane dendrogramem.



Rys. 16. Grupowanie hierarchiczne; a – obiekty w przestrzeni, b – dendrogram

Źródło: opr. własne

Oś pozioma dendrogramu (rysunek **b**) nie ma charakteru osi liczbowej i przedstawia zawsze nazwy (etykiety) elementów, których grupowanie dotyczy. Pionowa natomiast jest osią, na której odkładana jest wartość podobieństwa lub odległości pomiędzy obiektami. Przy konstrukcji dendrogramu wykorzystywana jest macierz odległości oparta na dowolnej funkcji matematycznej spełniającej przedstawione dalej warunki.

Przedmiotem analizy jest jak zwykle macierz danych **X**, zawierająca w każdym wierszu wektor zmiennych określających położenie pojedynczego obiektu w przestrzeni cech. Macierz tę, podobnie jak ma to miejsce w przypadku większości metod chemometrycznych, należy do analizy przygotować. Przygotowanie, jak już wcześniej była mowa obejmuje dwa etapy: kompletowanie niepełnych danych (macierz nie powinna zawierać pustych miejsc) oraz transformację zmiennych. Prócz specyficznych transformacji

kolumn macierzy **X** mających na celu ujednolicenie rozkładu błędów w całym zestawie zmiennych, przeprowadza się takie opisane wcześniej transformacje jak skalowanie przedziałowe, skalowanie wariancyjne czy autoskalowanie (standaryzacja). W praktyce stosowane jest zwykle autoskalowanie. Powodem tego jest jego uniwersalność, polegająca na takim przekształceniu wartości zmiennych, po którym posiadają one wszystkie cechy, jakie nadają im skalowanie przedziałowe i wariancyjne jednocześnie. Standaryzacja zmiennych (zależność 2.5) prowadzi bowiem do ujednolicenia ich wariancji i zakresu zmienności, co zapewnia współmierność niesionej przez nie ładunków informacji. W przypadku analizy podobieństwa równość wariancji zmiennych jest cechą pożądaną, podyktowaną specyfiką matematycznych sposobów określania podobieństwa (odległości) obiektów w przestrzeni zmiennych.

Należy w tym miejscu wyraźnie zaznaczyć, że w rozpoznawczej analizie chemometrycznej, a taką jest analiza podobieństwa, możemy rozwiązywać problem podobieństwa **zmiennych (cech)** i/lub podobieństwa **obiektów** w przestrzeni zmiennych. Z matematycznego punktu widzenia, zasadnicza różnica pomiędzy jednym i drugim podejściem ogranicza się w pierwszym przypadku: do analizy podobieństwa wektorów kolumnowych macierzy danych **X**, a w drugim: do analizy podobieństwa jej wektorów wierszowych.

Pojawia się więc konieczność zdefiniowania prawidłowej i miarodajnej wielkości podobieństwa wektorów $\mathbf{z}_R$ i $\mathbf{z}_S$ w wielowymiarowej przestrzeni danych. Pomocne w rozwiązaniu tego problemu są zasady geometrii analitycznej, które określają, że wyrażeniem definiującym odległość dwóch punktów $d(\mathbf{z}_R,\mathbf{z}_S)$ w dowolnej przestrzeni wielowymiarowej jest każda funkcja spełniająca trzy następujące warunki:

1. Odległość punktu od samego siebie jest równa zero

$$d(\boldsymbol{z}_R, \boldsymbol{z}_R) = d(\boldsymbol{z}_S, \boldsymbol{z}_S) = 0 \tag{5.1}$$

2. Odległość punktu R od punktu S jest równa odległości punktu S od punktu R, czyli nie zależy od kierunku pomiaru (kolejności punktów)

$$d(\boldsymbol{z}_R, \boldsymbol{z}_S) = d(\boldsymbol{z}_S, \boldsymbol{z}_R) \tag{5.2}$$

3. Suma odległości punktów R i S od dowolnego punktu T nie może być mniejsza niż odległość bezpośrednia punktów R i S

$$d(\mathbf{z}_R, \mathbf{z}_T) + d(\mathbf{z}_S, \mathbf{z}_T) \leq d(\mathbf{z}_R, \mathbf{z}_S) \tag{5.3}$$

Istnieje więc nieskończenie wiele funkcji, które mogą spełniać te warunki. Mogą one być podstawą do zdefiniowania wykorzystywanej w chemometrii wielkości określającej podobieństwo punktów (obiektów lub cech) w zbiorze. Podobieństwo to w ogólnej postaci możemy zdefiniować jako:

$$S_{RS} = 1 - \frac{d(\mathbf{z}_R, \mathbf{z}_S)}{d_{max}} \tag{5.4}$$

gdzie:
S_{RS} – podobieństwo (wartości od 0 do 1);
d_{max} – największa możliwa odległość punktów.

Algorytm analizy podobieństwa już na początkowych etapach wymaga od nas określenia odległości pomiędzy (początkowo) punktami, potem grupami punktów w przestrzeni wielowymiarowej. O ile nie ma z tym problemu, gdy chodzi o pojedynczy punkt, to w przypadku zbioru punktów pojawiają się już różne sposoby określenia tej wielkości. W praktyce, najczęściej wykorzystywanymi są: **odległość najbliższego sąsiada** w grupach (ang. single link), czyli takich dwóch elementów po jednym z każdej grupy, dla których odległość jest najmniejsza; **odległość najdalszego sąsiada** (ang. complete link), dla której interpretacja jest oczywista, oraz **odległość średnia** (ang. average link), w przypadku której wartość obliczana jest na podstawie zależności:

$$d_{R,S} = \frac{\sum_R \sum_S d_{r,s}}{n_R n_S} \tag{5.5}$$

Istnieje jeszcze wiele metod określania odległości między zbiorami punktów. Jedną, o której warto wspomnieć z uwagi na jej ostatnio rosnącą popularność i obecność w każdej aplikacji z rodzaju statystycznych, jest metoda Ward'a (ang. incremental sum of squares). Jej cechą odróżniająca ją od

wymienionych wcześniej metod jest to, że do oszacowania odległości między skupieniami wykorzystuje analizę wariancji. Metoda ta zmierza do minimalizacji wariancji w grupach i jej maksymalizacji pomiędzy skupieniami, które mogą zostać uformowane na każdym etapie łączenia zbiorów.

Każde z rozwiązań daje w efekcie nieco odmienny diagram. Gdy elementy mają tendencję do tworzenia wyraźnych skupień kulistych, metoda najbliższego sąsiada daje dobre wyniki. W przypadku innych kształtów klastrów, lub gdy zależy nam na podkreśleniu nawet niewielkich różnic pomiędzy elementami, lepsze wyniki wydaje się przynosić zastosowanie odległości najdalszego sąsiada. Metoda Ward'a natomiast, traktowana jest jako bardzo efektywna, dająca najbardziej naturalne skupiska elementów analizowanych zbiorów.

5.1.1 Podobieństwo zmiennych

Analiza rozpoznawcza podobieństwa cech, wykonywana jest zwykle celem określenia możliwości redukcji wymiarowości przestrzeni zmiennych. W przypadku kolumnowych wektorów cech, pomimo różnorodności definicji odległości spotykanych w literaturze, podstawową wielkością, na której oparte są miary odległości cech, wydaje się być współczynnik korelacji zmiennych – ***r***. Jest on uzasadniony pojęciowo, gdyż ma prostą interpretację geometryczną: jest równy kosinusowi kąta α pomiędzy wektorami zmiennych, których podobieństwo jest analizowane.

$$\mathrm{r} = \cos(\propto) \tag{5.6}$$

Wykorzystanie współczynnika korelacji jako miary odległości wektorów nie jest możliwe, ponieważ funkcja *r* nie spełnia warunków przedstawionych zależnościami 5.1 – 5.3. Możliwe jest natomiast wykorzystanie w tym celu funkcji sinus kąta α między wektorami $\mathbf{z}_R$ i $\mathbf{z}_S$ (jej wartości bezwzględnej):

$$d^S = |sin(\alpha)| = \sqrt{1 - r^2} \tag{5.7}$$

Podobną w swych własnościach miarą, mającą najszersze zastosowanie jest funkcja wartości bezwzględnej tangensa kąta α:

$$d^T = |tg(\alpha)| = \sqrt{\frac{1-r^2}{r^2}} \tag{5.8}$$

Obie miary wykazują największą odległość dla zmiennych ortogonalnych ($r = 0$; $\alpha = 90^o$) a najmniejszą, równą 0 (podobieństwo równe 1) dla zmiennych opisywanych wektorami równoległymi lub antyrównoległymi, które zawierają jedynie informację wspólną, w 100% tę samą. Inną jeszcze miarą, wykorzystywaną w programach statystycznych jest dopełnienie do jedności wartości absolutnej współczynnika korelacji *r*:

$$d^D = 1 - |r| \tag{5.9}$$

Jest to prosta obliczeniowo miara, której wartości zawsze zawarte są w przedziale $0 \div 1$. Podobnie jak w przypadku miary tangensowej i sinus wektory równoległe i antyrównoległe mają odległość równą 0, a ortogonalne największą, ale skończoną równą 1.

Warto wspomnieć, że miary odległości oparte na współczynniku korelacji nie wymagają autoskalowania zmiennych. Bez względu na to czy jakakolwiek transformacja z grupy skalowania przedziałowego, czy też skalowania wariancyjnego zostanie przeprowadzona lub nie, wartość współczynnika *r* pozostanie dla każdej z pary wektorów taka sama. Prześledźmy teraz na przykładzie sposób prowadzenia analizy podobieństwa zmiennych i jej efekty w zależności od wyboru metody określania odległości między skupiskami i rodzaju zastosowanej miary odległości. Jako dane do analizy posłużą wyniki badań biegłości laboratoriów, których celem było określenie procentowej zawartości pewnych pierwiastków w tej samej próbce żużlu. W każdym z pięciu laboratoriów pomiary dla tej samej próbki wykonywane były po 3 razy, a wyniki pomiarów (% zawartości pierwiastków) przedstawione zostały w tabeli 16.

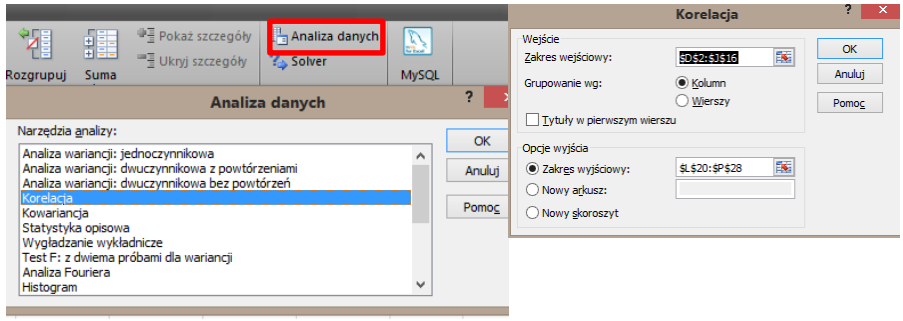
Jeśli celem przeprowadzenia analizy chcielibyśmy wykorzystać jedną z aplikacji statystycznych o darmowym dostępie (wybór jest naprawdę duży) nie zawsze znajdziemy w nich wybór miar odległości opartych na współczynniku korelacji. Podobnie jest w przypadku programu ‘Statistica’. Nie ma możliwości wykorzystania tangensowej miary odległości czy funkcji

Tab. 16. Testy międzylaboratoryjne, badania biegłości

Kod	Lab	Próbka	Si	Al	Fe	Ti	Na	Mg	Ca
A1	A	1	53.30	12.40	10.30	1.20	0.30	2.80	13.90
A2	A	2	52.80	12.30	10.20	1.20	0.20	2.70	13.80
A3	A	3	52.90	12.30	10.20	1.20	0.20	2.70	13.90
B1	B	1	69.00	11.34	9.01	1.00	0.20	2.80	14.10
B2	B	2	57.00	10.35	8.37	1.00	0.20	2.50	13.70
B3	B	3	61.00	10.39	8.44	1.00	0.20	2.60	14.00
C1	C	1	53.30	12.25	10.63	1.20	0.20	2.50	13.60
C2	C	2	53.40	12.47	10.69	1.30	0.40	2.50	13.70
C3	C	3	53.20	12.18	9.85	1.20	0.20	2.30	13.50
D1	D	1	55.30	12.80	10.00	1.17	0.13	3.00	14.22
D2	D	2	54.70	12.40	9.90	1.17	0.13	2.70	13.92
D3	D	3	54.80	12.50	10.00	1.17	0.13	2.81	13.95
E1	E	1	53.90	12.60	9.60	1.40	0.18	3.40	13.00
E2	E	2	54.10	12.80	9.70	1.40	0.19	3.50	13.30
E3	E	3	53.80	12.30	9.50	1.30	0.17	3.30	12.90

Źródło: opr. własne, dane: Doerffel & Zwanziger

sinus. Można natomiast dokonać automatycznej analizy podobieństwa, opartej o wspomniane miary odległości, na podstawie macierzy odległości. Należy ją tylko wcześniej przygotować na przykład w arkuszu kalkulacyjnym. Nie jest to pracochłonne, czy skomplikowane. Kopiujemy zawartość tabeli 16 do Excela, następnie korzystając z funkcji ‘Korelacja’ w module ‘Analiza danych’ (Rys. 17) tworzymy tabelę współczynników korelacji pomiędzy zmiennymi i dalej na jej podstawie macierz odległości tangensowych korzystając z zależności 5.7.



Rys. 17. Wykorzystanie Excel'a do obliczenia macierzy współczynników korelacji

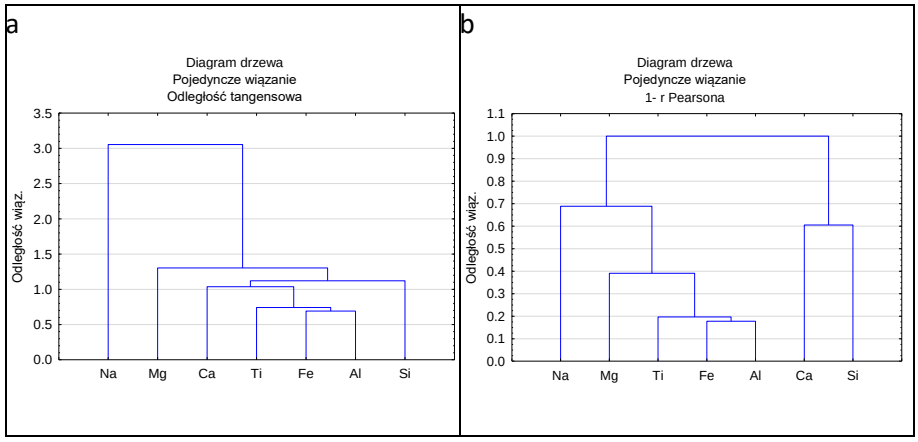
Źródło: opr. własne

Tab. 17. Macierz odległości tangensowych cech

	Si	Al	Fe	Ti	Na	Mg	Ca
Si	0.0000	1.3196	1.2073	1.1205	9.6602	15.7286	2.3264
Al	1.3196	0.0000	0.6920	0.7420	44.4786	2.1072	4.0958
Fe	1.2073	0.6920	0.0000	1.4509	3.0534	15.2829	2657.0934
Ti	1.1205	0.7420	1.4509	0.0000	6.8569	1.3030	1.0369
Na	9.6602	44.4786	3.0534	6.8569	0.0000	3.1613	50.6708
Mg	15.7286	2.1072	15.2829	1.3030	3.1613	0.0000	1.6572
Ca	2.3264	4.0958	2657.0934	1.0369	50.6708	1.6572	0.0000

Źródło: opr. własne

Wykorzystując plik macierzy odległości w programie 'Statistica' należy pamiętać, że musi on posiadać odpowiedni format. Jednym z wymogów jest pełna, symetryczna macierz odległości, tak jak w tabeli 17. Diagram wiązkowy podobieństwa, uzyskany poprzez tworzenie skupień metodą najbliższego sąsiada oraz na podstawie odległości tangensowych zmiennych, przedstawiony jest na wykresie poniżej – wykres a. Obok, celem porównania diagram wykonany tą samą metodą grupowania, ale tworzony na podstawie odległości będącej dopełnieniem do jedności współczynnika korelacji Pearsona.



Rys. 18. Diagramy wiązkowe podobieństwa zmiennych utworzone metodą najbliższego sąsiada: a – tangensowa miara odległości, b – miara odległości 1-r Pearsona

Źródło: opr. własne

Jak należało się spodziewać, wybór funkcji określającej odległość musi mieć wpływ na efekt końcowy analizy. Jego wielkość jest zależna od rodzajów skupień i zwykle jest tym mniejsza im są one wyraźniejsze i bardziej sferyczne. W naszym przykładzie obserwujemy w zasadzie brak różnic grupowania dla czteroelementowej grupy Al, Fe, Ti, Mg. Elementy te są widoczne zarówno na rysunku 18 a jak i 18 b jako praktycznie jedno skupienie. Porównując oba diagramy widzimy też, że wspólny wniosek można wyciągnąć na temat pierwiastka sodu – jest on nieco oddalony od wspomnianej wcześniej grupy. Natomiast różnice między diagramami są znaczące, gdy porównamy je pod kątem zawartości wapnia i krzemu. Dla odległości tangensowej, Ca i Si są położone w tym samym skupieniu co Al, Fe, Ti, Mg, ale dla miary odległości 1 – *r* Pearsona tworzą one dalekie, odrębne skupienie. Takie porównania diagramów i wyciągane na ich podstawie wnioski, często z powodu różnic nie są oczywiste. Ich niejednoznaczność upoważnia do uogólnienia, że w zróżnicowanych wynikach analizy należy raczej szukać podobieństw. Można je wtedy traktować, jako uzasadnienie wniosków najbardziej prawidłowych.

Jeszcze większy wpływ na wynik analizy podobieństwa ma wybór metody obliczania odległości skupień od siebie (najbliższego sąsiada, najdalszego, Ward'a itp.). I znów należy podkreślić, że im podział na klastry jest mniej wyraźny tym jest on większy. Dobrym przykładem różnic może być analiza tego samego zbioru danych wykonana z wykorzystaniem metod najbliższego sąsiada i Ward'a, gdy miarą odległości jest wyrażenie 1 – *r* Pearsona i odległość euklidesowa (5.10), **która zawsze wymaga standaryzacji zmiennych**:

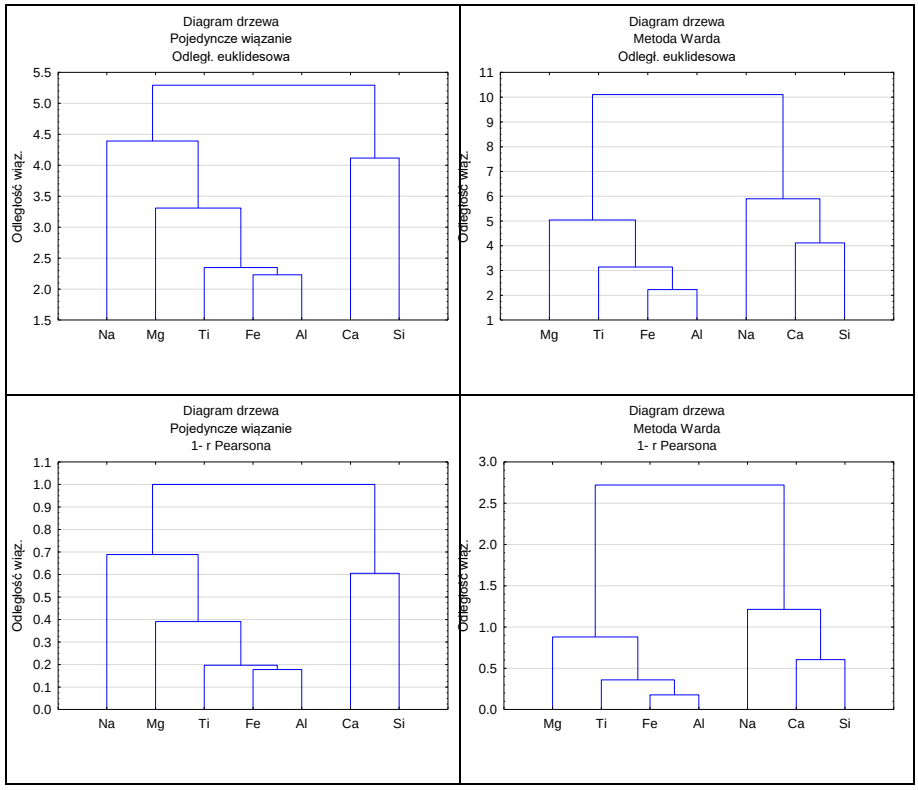
$$d_{ij}^{E} = \sqrt{\sum_{k=1}^{m} \left(\boldsymbol{z}_{ik} - \boldsymbol{z}_{jk}\right)^2} \tag{5.10}$$

gdzie:
m – ilość elementów wektora;
i,j – symbole wektorów

Zależność 5.10 jest niczym innym jak tylko sumą kwadratów różnic odpowiadających sobie elementów wektorów, których odległość jest mierzona.

Porównując diagramy (rysunek 19) można wyciągnąć ogólny wniosek, że większy wpływ na ich kształt ma metoda szacowania odległości niż wyrażenie określające jej miarę. Różnica jest wyraźna dla sodu, który w przypadku metody Ward'a wyraźnie dołącza do grupy Ca, Si, a w przypadku metody pojedynczego wiązania (najbliższego sąsiada) do grupy pierwiastków Al, Fe, Ti, Mg, ale odległość obu grup nie jest tak duża jak w przypadku metody Ward'a.

Z powodu niedostatecznej ilości informacji dotyczących samego pomiaru, wyjaśnienie powodu, dla którego obserwujemy takie właśnie skupienia cech dla przykładowych danych nie jest możliwe. Można natomiast stwierdzić, że metoda Ward'a uwypukla różnice pomiędzy zbiorami elementów i jest najchętniej i najczęściej stosowana w analizie podobieństwa, zwłaszcza w analizie podobieństwa obiektów.



Rys. 19. Różnice w diagramach spowodowane metodą szacowania odległości zbiorów

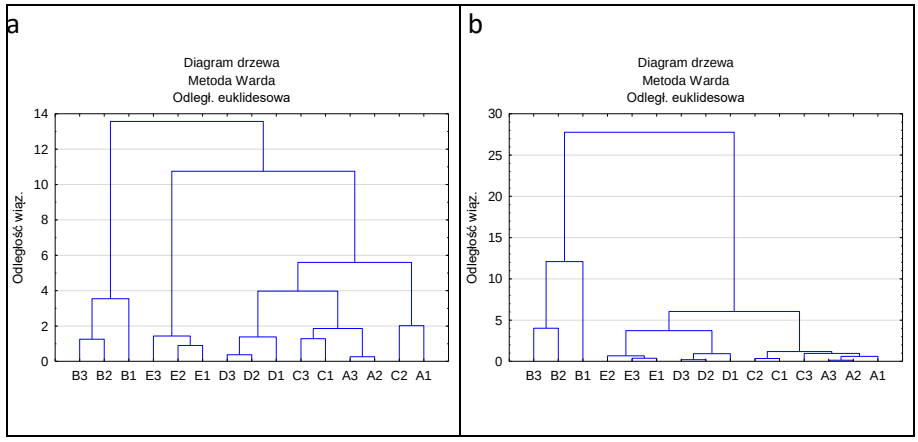
Źródło: opr. własne

5.1.2 Podobieństwo obiektów

Analiza podobieństwa przestrzeni zmiennych objaśniających jest tylko częścią pełnej analizy podobieństwa. Dopełnieniem jest analiza podobieństwa próbek (obiektów), którą można określić, jako poszukiwanie podobieństwa elementów ze względu na wartości opisujących je cech. Ten typ analizy jest próbą znalezienia wspólnej przestrzeni dla różniących się charakterem wielkości fizycznych, chemicznych czy biologicznych jednocześnie charakteryzujących obiekty. W przypadku analizy podobieństwa obiektów, najpowszechniej wykorzystywaną miarą odległości jest miara euklidesowa (5.10). Inne rodzaje miar, które są dostępne na przykład w aplikacji 'Statistica' to odległość miejska 'Manhattan' czy też odległość Czebyszewa.

Wszystkie one wymagają standaryzacji (autoskalowania) zmiennych, która eliminuje nadmierną wagę zmiennych o wysokich wartościach nominalnych.

Wpływ autoskalowania na diagram wiązkowy podobieństwa obiektów (laboratorium – pomiar) dla danych z analizowanego wcześniej przykładu badań kompetencji laboratoriów przedstawia rysunek 20a i 20b.



Rys. 20. Analiza podobieństwa obiektów: a – zmienne standaryzowane, b – zmienne niestandaryzowane

Źródło: opr. własne

Jakościowe różnice wydają się być niewielkie, ale jest to przypadek, gdy rzędy wartości mierzonych wielkości dla pojedynczego obiektu nie są bardzo zróżnicowane. Na obu diagramach możemy zaobserwować najbardziej zgodne ilościowo wyniki dla laboratoriów A i C. Skalowanie uwypukla pewne odchylenia pomiarów C2 i A1 od tej reguły. Bliskie wartości, podobne do tych jakie uzyskują laboratoria A, C, otrzymuje laboratorium D. Na diagramie 20a widzimy, że są one porównywalne do tego stopnia, że tworzą skupienie bliżej niż obciążone większym błędem pomiary C2 i A1. Nie da się tego zaobserwować na mniej szczegółowym diagramie 20b. Największe różnice pomiędzy diagramami obserwujemy dla skupień laboratoriów E i B. Dla danych nieskalowanych laboratorium E nie odstaje ze swymi pomiarami od A, C i D. Inaczej jest w przypadku danych standaryzowanych. Tu wyniki uzyskane przez E tworzą odległe od nich skupienie. Grupa wyników uzyskanych

przez laboratorium B jest najbardziej niezgodna z pozostałymi. Różnicę uwypukla dodatkowo brak skalowania, co oznacza, że mogą one być obciążone większym błędem dla zmiennej o wysokich wartościach nominalnych.

Skupienia utworzone z wykorzystaniem danych przed i po transformacji różnią się jak widać niewiele. Natomiast widoczną na pierwszy rzut oka cechą odróżniającą oba diagramy są wartości odległości między skupieniami i wynikający z tych różnic inny kształt każdego z nich. Widzimy zatem, że ujednolicenie zakresów wartości zmiennych i ich wariancji pozwala na bardziej szczegółową ocenę skupień i tym samym na ocenę podobieństwa między obiektami z 'bliższej perspektywy'.

5.2 Klasyfikacja

Klasyfikacja, nazywana często nadzorowanym rozpoznawaniem wzorców, (obrazów, nadzorowanym uczeniem) jest nieodłączną cechą chemii. Tablica Mendelejewa pierwiastków, grupowanie związków organicznych według ich właściwości, rodzaje reakcji chemicznych są niczym innym jak przykładami klasyfikacji w chemii. Klasyfikacja nie straciła na znaczeniu, wręcz przeciwnie, także w przypadku nowoczesnej chemii. Nowoczesne zastosowania klasyfikacji pozwalają na odpowiedź na wiele pytań analitycznych, jak choćby czy spektrogram uzyskany metodą NMR odpowiada konkretnemu związkowi, czy na podstawie chromatogramu materiału biologicznego pacjenta należy zakwalifikować do grupy chorych, lub obciążonych ryzykiem zachorowania, i w końcu czy obserwowany w podczerwieni proces technologiczny produkcji elementu pozwala zaklasyfikować go do zbioru elementów bez wad technologicznych. Podobnych przykładów można podawać wiele…

Dlatego klasyfikacja obiektów jest jednym z typowych zastosowań technik chemometrycznych. Polega ona na przypisaniu obiektu opisanego zestawem mierzalnych cech do jednej z rozłącznych klas. Klasyfikacja ma ponadto charakter jakościowy tj. obiekt należy lub nie należy do danej klasy. Nie może należeć do niej częściowo, czyli należeć jednocześnie do kilku klas. Warunkiem koniecznym klasyfikacji jest znajomość liczby klas i znajomość zestawu cech przedstawiciela każdej z nich. Dowolność dotyczy jedynie wy-

boru reguł (metody) klasyfikacji. Ich cechą charakterystyczną z punktu widzenia chemometrii jest to, że zmienne objaśniające nie definiują bezpośrednio klas odpowiedzi obiektu. Definiuje je model matematyczny oparty o zmienne.

5.2.1 Nadzorowane rozpoznawanie wzorców

Spełnienie warunków znajomości liczby klas i znajomości składowych wektora cech przedstawicieli każdej z nich, nazywane jest uczeniem nadzorowanym (nadzorowanym rozpoznawaniem wzorców). Metoda ta składa się z trzech zależnych od siebie kroków, które z racji realizowanych funkcji przyjmują ich nazwy. Są to: przetwornik, preprocesor i klasyfikator. Bez względu na rodzaj metody klasyfikacji każdy z bloków, pomimo częstych różnic samych operacji, wykonuje zawsze podobne działania.

Do zadań **przetwornika** należy przygotowanie danych do dalszej analizy. Jak wiemy składają się one z prostokątnej macierzy obserwacji **X** o *n* wierszach i *m* kolumnach oraz (dotyczy uczenia nadzorowanego) dodatkowego wektora kolumnowego **q** przynależności klasowej o *n* wierszach. Dla tak przygotowanych danych, każdemu obiektowi odpowiada *m* elementowy **wektor obrazu** w *m* wymiarowej, euklidesowej przestrzeni obrazu. Każdemu obiektowi (punktowi) odpowiada również, opisana jakościowo elementem wektora **q,** przynależność do określonej klasy. Dodatkowo, przetwornik realizuje zadanie uzupełniania brakujących danych. Obie macierze, **X** oraz **q** nie mogą zawierać pustych miejsc, a gdy nie jest możliwym podanie pełnej informacji dla któregoś z wierszy (obiektów), wiersz ten należy wykluczyć ze zbioru danych.

Rolą **preprocesora** jest opracowanie wektorów obrazu pod kątem wyodrębnienia jedynie przydatnych cech obiektów. Zwykle polega ona na selekcji najważniejszych zmiennych objaśniających, a także ewentualnych ich transformacjach, mających na celu zwiększenie skuteczności klasyfikacji. Warto w tym miejscu zwrócić uwagę na jeszcze jeden (prócz wcześniej omówionych) sposób transformacji, jakim jest skalowanie ważone. Pozwala ono bowiem na uwypuklenie cech (wcześniej poddanych standaryzacji), które maja większy wpływ na jakość klasyfikacji niż inne. Ważenie cech polega na

przemnożeniu każdej z wartości cechy przez wcześniej odpowiednio zdefiniowaną dla niej wagę. W efekcie cechy silnie różnicujące obiekty zostają 'rozciągnięte', co zwykle ułatwia pracę klasyfikatora. Ponadto wagi pozwalają na selekcję cech, co umożliwia ograniczenie liczby zmiennych do minimum przy jak najmniejszej utracie informacji przez nie niesionej. Należy w tym miejscu zadać istotne pytanie, jakimi sposobami można wyznaczyć właściwe wagi cech? Istnieją dwie podstawowe metody ich wyznaczania. Pierwsza oparta jest na analizie rozkładów wartości cechy w poszczególnych klasach, druga na analizie wariancji cechy i wykorzystaniu w tym celu statystyki F Fishera-Snedecora.

Analiza rozkładów wartości cech w klasach jest najmniej skomplikowana w przypadku dwóch skupień. Najczęściej stosowanym sposobem oceny cechy do różnicowania klas jest w tym przypadku tzw. **stosunek Fishera**, **będący jednocześnie jej wagą**. Można go wyznaczyć dla każdej z cech (po standaryzacji zmiennych – ważne: dla wszystkich elementów zbioru) z zależności:

$$w = F_{(A,B)} = \frac{|\overline{z}_A - \overline{z}_B|}{\sqrt{s_A^2 + s_B^2}} \tag{5.11}$$

gdzie:
$\overline{z}_A$ i $\overline{z}_B$ średnie wartości standaryzowanej cechy odpowiednio dla klasy A i B;
s_A^2, s_B^2 – wariancja cechy odpowiednio w klasie A i B.

W przypadku, gdy mamy do czynienia z ilością klas większą niż dwie, wyrażenie na wagę klasy przybiera postać:

$$w = \frac{\sum_{j=1}^{P} F_{j(A,B)}}{P} \tag{5.12}$$

gdzie:
$P = q(q-1)/2$ – to ilość wszystkich par klas;
q – ilość klas.

Jak widać, jest to wartość średnia wielkości *w* wyznaczonych z zależności 5.11 dla wszystkich par klas. Im jest ona większa od zera tym większą moc różnicująca obiekty posiada dana cecha.

Wagi cech oparte o analizę wariancji obliczane są, jako wartości statystyki F Fishera-Snedecora (w statystyce wykorzystywana do oceny istotności różnic wariancji). Jest to iloraz wariancji pomiędzy klasami do sumy wariancji wewnętrznych we wszystkich klasach. Równania opisujące te wielkości to odpowiednio 5.13 i 5.14:

$$s_k^2 = \frac{\sum_{i=1}^{k} n_i(\overline{z}_i - \overline{\overline{z}})^2}{k-1} \tag{5.13}$$

$$s_\Sigma^2 = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(z_{ji} - \overline{z}_i\right)^2}{\sum_{i=1}^{k} n_i - k} \tag{5.14}$$

gdzie:
k – to ilość klas,
i – numer klasy,
n_i – liczba próbek w klasie i,
j – numer próbki w klasie i,
z_{ji} – wartość cechy dla j–tej próbki i–tej klasy.

Ważenie cech to bardzo skuteczne narzędzie analityczne. Daje ono informacje, czy dany zestaw zmiennych objaśniających zawiera wystarczającą ilość informacji dla przeprowadzenia klasyfikacji. Daje również, o czym była mowa, możliwość selekcji najważniejszych z punktu widzenia klasyfikacji cech. Selekcja cech to jeszcze jeden proces realizowany przez preprocesor. Oparty jest on na powszechnie przyjętych założeniach, że dostateczną zdolność do klasyfikacji posiadają zmienne z wagami powyżej wartości 0.5, gdy szacowane są na podstawie stosunku Fishera. W przypadku analizy wariancji wagi zmiennych 'istotnych' powinny być większe od 2.

Ostatnim blokiem procesu klasyfikacji, który na podstawie zbioru uczącego realizuje zadanie wypracowania reguł decyzyjnych jest **klasyfikator**. Reguły te, to matematyczne zależności określające przynależność do klas. Wartości obliczane na ich podstawie dla określonego obiektu decydują o jego przynależności do jednego ze zbiorów.

Sam proces poszukiwania właściwych funkcji rozstrzygających nazywany jest procesem uczenia klasyfikatora. Zwykle jest to proces iteracyjny, który zatrzymywany jest dla warunku z góry określonej liczby przypadków

prawidłowych klasyfikacji obiektów zbioru uczącego. W każdym kroku warunek ten jest sprawdzany i w zależności od wyniku podejmowana jest dalsza decyzja, co do sposobu działania. Może to być polecenie dalszej nauki klasyfikatora, ponowne rozpoczęcie nauki z innego punktu początkowego lub też ponowne przygotowanie danych zlecane preprocesorowi i zmiana typu klasyfikatora.

Po zakończeniu procesu uczenia, klasyfikator poddawany jest ocenie. Za pomocą testowego zbioru obiektów, zbioru niebiorącego udziału w uczeniu, sprawdzana jest skuteczność klasyfikatora. Obiekty w takim zbiorze muszą być znane użytkownikowi kontrolującemu proces tworzenia klasyfikatora. Tylko wtedy możliwa jest ocena jego pracy. Dopiero zadowalająca ocena klasyfikatora pozwala na jego zastosowanie do przewidywania przynależności nieznanych obiektów do odpowiednich klas.

Uczenie i jego przebieg jest cechą zależną od rodzaju klasyfikatora. Nadzorowana analiza skupień oparta jest dziś na wielu różnych algorytmach realizujących proces klasyfikacji i proces poszukiwania funkcji rozstrzygających. Każdy z nich posiada określone wady i zalety. Zostaną one omówione na przykładzie kilku najpopularniejszych, najczęściej stosowanych metod klasyfikacji.

5.2.2 Metody klasyfikacji

Klasyfikator Bayesa

Pomimo braku możliwości praktycznego zastosowania, prezentację najbardziej popularnych metod klasyfikacji warto jest rozpocząć od klasyfikatora Bayesa. Jest to klasyfikator ‘teoretyczny’, który wniósł znaczący wkład w rozwój teorii klasyfikacji. Największą jego zaletą w badaniach teoretycznych jest brak procesu iteracyjnego. Funkcja decyzyjna otrzymywana jest w tym przypadku w sposób analityczny, przy założeniu, że obiekty w każdej klasie mają określony rozkład prawdopodobieństwa znalezienia się w niej. Przy tym najczęściej zakłada się, że jest to wielowymiarowy, zgodny z wymiarem przestrzeni cech, rozkład normalny. Największa zaleta klasyfikatora Bayesa jest jednocześnie jego podstawową wadą, która nie pozwala na jego uczenie się.

Sposób działania klasyfikatora ogranicza się do dwóch etapów. Pierwszy z nich polega na wyznaczeniu parametrów rozkładów cech obiektów w każdej klasie, czego rezultatem jest znalezienie elementu reprezentującego środek każdej klasy. Dalej, na podstawie macierzy kowariancji ($\mathbf{Z}^{\mathrm{T}}\mathbf{Z}$, – gdzie **Z** to macierz standaryzowanych cech obiektów w klasie**)** klasyfikator szacuje prawdopodobieństwo przynależności obiektu do każdej z klas. Ostatecznie, na podstawie tej wartości obiekt jest przypisywany do jednej z klas – oczywiście tej o najwyższym prawdopodobieństwie przynależności. Prócz tego znając położenie środków klas i parametry rozkładów cech w klasach klasyfikator może określić granice hiperprzestrzeni (hiperobszarów), w których prawdopodobieństwa przynależności do klas są sobie równe. Kształty powierzchni hiperobszarów zależą tylko i wyłącznie od wartości elementów w macierzy kowariancji. Jeśli macierze są identyczne, istnieje tylko jedna powierzchnia rozdzielająca obszary. Dla dwuwymiarowej przestrzeni cech jest nią linia prosta, dla przestrzeni trójwymiarowej – płaszczyzna, dla przestrzeni wielowymiarowej – oczywiście hiperpłaszczyzna. Natomiast kształt powierzchni granicznej jest hiperpłaszczyzną stopnia drugiego, gdy macierze kowariancji nie są sobie równe. Podsumowując, warto dodać, że wymóg znajomości rozkładów zmiennych w klasach jest rzeczą kłopotliwą, dlatego klasyfikator Bayesa rzadko znajduje zastosowanie w praktyce.

Liniowa maszyna ucząca się (LLM – Linear Learning Machine)

Najprostszym klasyfikatorem, opartym na liniowej funkcji decyzyjnej, tworzącej w przestrzeni wielowymiarowej hiperpłaszczyznę dzielącą ją na dwa obszary jest liniowa maszyna ucząca się. Funkcję decyzyjną można opisać w tym przypadku zależnością:

$$a_0 + a_1 z_1 + a_2 z_2 + \cdots + a_k z_k = 0 \qquad (5.15)$$

Klasyfikator taki jest liniowym ze względu na wszystkie cechy, jakie opisują obiekt. Klasyfikowane obiekty mogą się znaleźć w klasie A lub B zależnie od wartości funkcji 5.15. Zwykle kryterium podziału jest znak wartości funkcji wynikającej z oszacowania. Odpowiednie wartości współczynników równania 5.15 nie są możliwe do wyznaczenia na drodze analitycznej. W tym celu wykorzystywany jest iteracyjny algorytm ‘uczenia’ klasyfikatora.

Jest on realizowany poprzez sprzężenie zwrotne i obejmuje trzy etapy. **Pierwszym** z nich jest wybór początkowych wartości parametrów funkcji. Istnieje tu zwykle całkowita dowolność, chociaż dowodzi się, że położenie punktu początkowego ma wpływ na szybkość uczenia się klasyfikatora i jakość uzyskiwanych potem wyników jego pracy. Dlatego najczęstszym sposobem jego wyboru jest wykorzystanie składowych wektora łączącego środki obu klas. W **drugim** etapie, dla kolejnych punktów zbioru uczącego obliczamy wyrażenie definiujące jego przynależność do zbioru (przynależność ta jest oczywiście znana):

$$S = \sum_{i=0}^{n} a_i z_i \tag{5.16}$$

gdzie: $z_0 = 1$

W **trzecim** kroku, na podstawie obliczonej wartości *S* sprawdzamy zgodność przynależności elementu uczącego do właściwego zbioru. Jeśli nie jest ona właściwa, modyfikowane są parametry funkcji decyzyjnej 5.15:

$$\boldsymbol{a}' = \boldsymbol{a} + c\boldsymbol{z} \tag{5.17}$$
$$c = -\frac{\alpha S}{\boldsymbol{z}^T \boldsymbol{z}}$$

gdzie:

$\mathbf{a}$ – stary wektor współczynników,

$\mathbf{a}'$ – nowy wektor współczynników,

$\mathbf{z}$ – wektor współrzędnych źle sklasyfikowanego obiektu,

α – parametr zależny od wersji algorytmu.

Powyższe etapy powtarzane są dopóki zachodzi konieczność korekty parametrów funkcji decyzyjnej. Brak konieczności korekty jest warunkiem zakończenia procedury uczenia klasyfikatora. Jeśli nie jest on spełniony przy założonej z góry ilości iteracji wynik uczenia jest negatywny. Może on wtedy sugerować, że nie istnieje hiperpłaszczyzna poprawnie dzieląca obiekty ze zbioru uczącego na klasy. Uzyskanie pozytywnego wyniku uczenia klasyfikatora LLM jest możliwe tylko wtedy, gdy zbiór elementów jest liniowo separowalny, tzn. elementy są tak rozłożone w przestrzeni, że nie zachodzi nakładanie się klas, a obie klasy tworzą zwarte, sferyczne skupienia. Można wykazać, że dla normalnych rozkładów punktów w populacjach obu klas,

warunkiem liniowej separowalności jest równość macierzy kowariancji obu klas (w innych przypadkach jest to hiperpowierzchnia stopnia drugiego). Klasyfikator LLM możemy też wykorzystywać do klasyfikacji obiektów należących do więcej niż dwóch klas. Uzyskujemy wtedy bardziej skomplikowany podział przestrzeni zmiennych, a ich interpretacja matematyczna jest utrudniona.

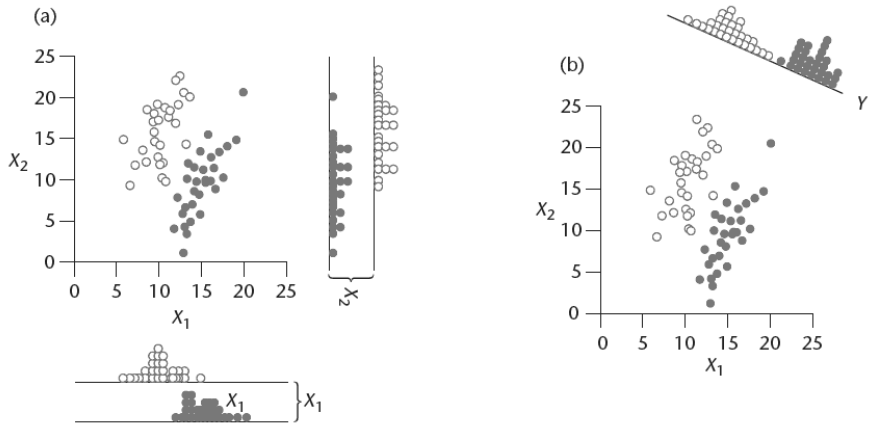
Klasyfikator LLM posiada także cechę pewnej niejednoznaczności podziału zbioru elementów na klasy. Jest to jego mankament i wynika z faktu możliwości wyboru różnych punktów początkowych w procesie uczenia. Otrzymujemy wtedy wiele rozwiązań spełniających kryteria poprawnej klasyfikacji. Dodatkowo dla tego typu klasyfikatora nie istnieje narzędzie teoretyczne pozwalające na oszacowanie jakości prognostycznej tych równorzędnych hiperpłaszczyzn decyzyjnych. Można to zrobić jedynie metodą walidacji z wykorzystaniem zbiorów testowych obiektów, których przynależność do zdefiniowanych klas jest wcześniej znana.

Analiza dyskryminacyjna

W przestrzeni wielu zmiennych objaśniających, jako klasyfikator stosowano początkowo opisaną wyżej liniową maszynę ucząca się – LLM. Często brak zadowalających wyników i wymienione wady klasyfikatora były powodem poszukiwań nowego algorytmu dla klasyfikatora liniowego. Ich efektem był algorytm operujący w przestrzeni jedynie kilku najważniejszych cech opisujących obiekty. Ponieważ wybór cech dokonywany jest w tym algorytmie poprzez odrzucenie zmiennych mniej różnicujących obiekty, sam algorytm nazwany został **liniową analizą dyskryminacyjną** (LDA – Linear Discriminant Analysis).

W literaturze dotyczącej statystyki panuje obecnie pogląd, że analiza dyskryminacyjna to m. in. metoda klasyfikacji będąca niezwykle efektywnym narzędziem eksploracji danych. Zadaniem klasyfikatora opartego o algorytmy dyskryminacyjne jest rozstrzyganie, które cechy obiektu w najskuteczniejszy sposób różnicują elementy zbioru, tj. dzielą konkretny zbiór przypadków na występujące w naturalny sposób klasy. Innymi słowy, analiza dyskryminacyjna umożliwia podział elementów zbioru ze względu na wartości wybranych cech (w procesie eliminacji innych) i dzięki temu, na predykcję przynależności dowolnego obiektu do danej grupy. Zasadniczą

częścią algorytmu jest zatem znalezienie takich zmiennych, zazwyczaj różniących się znacznie wartościami średnich, które będą podstawą do określenia funkcji dyskryminacji, będącej liniową kombinacją wybranych cech. Prostym przykładem większych możliwości funkcji dyskryminacji (zdolności klasyfikacyjnych) w porównaniu z pojedynczymi zmiennymi, jest podział obiektów, które reprezentowane są przez dwie, wybrane w tym celu cechy. Punkty odpowiadające klasyfikowanym obiektom można przedstawić w tym przypadku w układzie współrzędnych XY (2D).



Rys. 21. a) Dwie zmienne i ich rozkłady analizowane osobno, b) rozkłady zmiennej Y dla funkcji dyskryminacyjnej (zawsze liniowej) dla każdego ze zbiorów

Źródło: [6]

Liniowa kombinacja wartości zmiennych X_1 i X_2, pozwala na określenie funkcji dyskryminacji postaci:

$$Y = a_1 X_1 + a_2 X_2 \tag{5.19}$$

Jak łatwo zauważyć (rysunek 21) zmienna Y jest zdecydowanie skuteczniejszym klasyfikatorem, niż każda ze zmiennych X_1 i X_2 osobno. Jest to oczywisty wniosek, kiedy przyjrzymy się rozkładom wszystkich wspomnianych zmiennych (Y, X_1 i X_2). Elementem, który różni je zasadniczo jest poło-

żenie ich środków (wartości oczekiwanych). Największą ich różnicę obserwujemy dla obliczanej na podstawie wartości funkcji dyskryminacji zmiennej Y. Funkcja dyskryminacji maksymalizuje bowiem różnice 'uniwersalnej' zmiennej Y między klasami obiektów.

Niewiele bardziej skomplikowanym od przedstawionego przypadkiem jest klasyfikacja obiektów w trójwymiarowej przestrzeni zmiennych. Działanie algorytmu i wykorzystanie go w celu określenia wagi zmiennych wyjściowych oraz eliminacji zmiennych nieistotnych dla funkcji decyzyjnej, przedstawione zostanie krótko na danych przykładowych. Danych dotyczących zawartości cukrów i sorbitolu w soku jabłkowym pochodzącym z różnych źródeł, regionów kraju. Celem takiej analizy będzie oczywiście stworzenie modelu pozwalającego na określenie regionu pochodzenia owoców na podstawie składu ich cukrów.

Tab. 18. Zawartość (g/dm^3) cukrów i sorbitolu w jabłkach z różnych regionów kraju

region	sacharoza	glukoza	fruktoza	sorbitol
A	20	6	40	4.3
A	27	11	49	2.9
A	26	10	47	2.5
A	34	5	47	2.9
A	29	16	40	7.2
B	6	26	49	3.8
B	10	22	47	3.5
B	14	21	51	6.3
B	10	20	49	3.2
B	8	19	49	3.5
C	8	17	55	5.3
C	7	21	59	3.3
C	15	20	68	4.9
C	14	19	74	5.6
C	9	15	57	5.4

Źródło: opr. własne

Aby przeprowadzić klasyfikację obiektów z grup A, B i C, wykorzystany zostanie w tym celu program Statistica i jego moduł analizy wielowymiarowej, w który to znajdziemy analizę dyskryminacyjną. Musimy pamiętać, że w przypadku trzech zbiorów obiektów otrzymamy trzy proste decyzyjne, dla których algorytm każdorazowo określi nam wagę zmiennych wyjściowych w modelu. Na tej podstawie oraz na podstawie pozostałych parametrów

(statystyka Fishera, p, lambda Wilksa) będziemy mogli podjąć decyzję o ich pozostawieniu bądź usunięciu z równania funkcji dyskryminacji.

Najprostszy z możliwych sposobów analizy pozwolił na uzyskanie takich oto wyników dotyczących parametrów dla zmiennych wyjściowych:

Tab. 19. Parametry opisujące istotność oryginalnych zmiennych w modelu

N=15	Podsumowanie analizy funkcji dyskryminacyjnej. (analiza dyskryminacyjna.sta) Zm. w modelu: 4;Grupująca: probka (3 grup) Lambda Wilksa: .01630 przyb. F (8,18)=15.373 p<.0000				
	Cząstkowa Lambda Wilksa	F usun.(2,9)	p	Toler.	1-Toler. (R-kwad)
sach	**0.234112**	**14.72154**	**0.001453**	**0.534504**	**0.465496**
gluk	0.595076	3.06207	0.096733	0.861879	0.138121
fruk	**0.185312**	**19.78335**	**0.000508**	**0.544379**	**0.455621**
sorb	0.729875	1.66544	0.242448	0.722961	0.277039

Źródło: opr. własne

Tab. 20. Wagi poszczególnych zmiennych w równaniach dyskryminacji

Zmienna	Funkcje klasyfikacyjne; (analiza dyskryminacyjna.sta) – wagi zmiennych std.		
	A	B	C
sach	**15.0393**	**-3.69699**	**-11.3423**
gluk	-1.8291	2.93096	-1.1018
fruk	**-9.6115**	**0.36291**	**9.2486**
sorb	-2.1914	-0.22939	2.4207
Stała	-15.6370	-3.53711	-9.8807

Źródło: opr. własne

Analiza danych pozwala ustalić, że istotnymi cechami w modelu są jedynie dwie zmienne – zawartość sacharozy oraz zawartość fruktozy (wartości pogrubione w tabeli 14). Wniosek taki pozwalają wyciągnąć odpowiednio niskie i wysokie wartości parametrów: cząstkowa Lambda Wilksa i statystyki F usunięcia, potwierdzające wysoką moc dyskryminacyjną zmiennych. Na istotność wybranych zmiennych wskazuje również parametr p, który przyjmuje w obu przypadkach wartości poniżej współczynnika istotności dla poziomu ufności 95%. Podobną wartość tego parametru obserwujemy dla całego modelu, p = 0.000.

W tabeli 20 zamieszczone zostały wartości będące wagami dla poszczególnych zmiennych objaśniających. Należy je interpretować jako

współczynniki kierunkowe prostych decyzyjnych w wielowymiarowej przestrzeni zmiennych. Sposób ich wykorzystania dobrze obrazować może przykład klasyfikacji jabłek o takich oto wartościach badanych związków w sokach: 11; 23; 59; 3.9 g/dm^3 odpowiednio sacharozy, glukozy, fruktozy i sorbitolu. Chcąc dokonać klasyfikacji (określenia regionu pochodzenia owoców) można wybrać jedynie dwie z tych wielkości, odpowiadające istotnym zmiennym w modelu, co upraszcza obliczenia. Można również wykorzystać pełen zestaw zmiennych, ponieważ nie jest on w tym przypadku zbyt liczny.

Obliczone wartości liniowej funkcji dyskryminacji dla każdej pary skupień, z wykorzystaniem jedynie dwóch istotnych, standaryzowanych zmiennych przedstawiają się następująco:

A: -15.64 + 15.04 * 11 – 9.61 * 59 = -417.2

B: -3.54 – 3.70 * 11 + 0.36 * 59 = -22.8

C: -9.88 – 11.34 * 11 + 9,25 * 59 = **411.1**

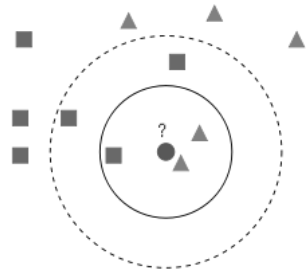
Takie wartości funkcji dyskryminacji z największym prawdopodobieństwem pozwalają sądzić, że owoce o takiej zawartości cukrów i sorbitolu pochodzą z regionu C – najwyższa wartość funkcji decyzyjnej.

Ogólne równanie funkcji dyskryminacyjnej jest równaniem podobnym do równania regresji liniowej wielu zmiennych. Współczynniki kierunkowe tego równania nazywane są dyskryminacyjnymi, często wagami. Określają one ważność oryginalnych zmiennych składowych wchodzących w jego skład. Wagi zmiennych objaśniających szacuje się dopóki funkcja liniowa niewystarczająco dobrze separuje istniejące, naturalne grupy analizowanych obiektów. Inaczej, wartości funkcji dyskryminacji wyliczane dla obiektów z różnych grup powinny się różnić między sobą możliwie jak najbardziej. Należy w tym miejscu zaznaczyć, że separacja większej ilości klas w przestrzeni zmiennych jest w przypadku klasyfikatorów liniowych zadaniem stosunkowo skomplikowanym. Zmuszeni jesteśmy do określenia funkcji dyskryminacji dla każdej pary skupień, co prowadzi do podziału wielowymiarowej przestrzeni zmiennych złożonym układem hiperpłaszczyzn decyzyjnych.

Metoda *k*-najbliższych sąsiadów (kNN – k – Nearest Neighbour)

Algorytm *k* najbliższych sąsiadów jest ogólnie rzecz biorąc algorytmem regresji nieparametrycznej. W statystyce wykorzystywany jest do predykcji wartości jakościowej zmiennej losowej. Jego zastosowanie jako klasyfikatora wynika z jego prostoty i braku konieczności wykonywania złożonych obliczeń statystycznych. Jako klasyfikator, metoda kNN jest znana i wykorzystywana w chemii i innych dziedzinach nauki od około 30 lat. Podstawowym założeniem algorytmu tego klasyfikatora jest to, że obiekty położone blisko siebie w przestrzeni zmiennych należą do tej samej klasy – są obiektami podobnymi. Konsekwencją jaka z tego wynika, jest możliwość klasyfikacji obiektów na podstawie znajomości przynależności klasowej wybranej liczby obiektów sąsiadujących z klasyfikowanym.

Podstawowym problemem, jaki należy rozwiązać w przypadku kNN jest liczba obiektów k, potrzebnych do przypisania kolejnego elementu do właściwego zbioru. Liczba ta w dużej mierze zależy od samej struktury skupień. Jeśli klasy (skupienia) są wyraźnie odseparowane, wtedy do podjęcia decyzji o przynależności nowego obiektu do klasy, wystarczy analiza przynależności jednego, najbliższego sąsiada; k = 1. W przypadkach bardziej skomplikowanych, gdy odległości między obiektami są porównywalne z odległościami skupień, wybór jednego sąsiada może powodować niewłaściwą, przypadkową klasyfikacje nowego elementu. Dlatego też dużo bardziej prawidłowe wyniki klasyfikacji uzyskuje się, gdy ocenimy przynależność do klasy większej ilości obiektów sąsiadujących z klasyfikowanym. Dobrą regułą jest przyjęcie nieparzystej liczby obiektów sąsiadujących. Pozwala ona na uzyskanie rozwiązania dla obiektów sąsiadujących przynależnych do różnych skupień. Z praktyki wynika, że najlepsze wyniki uzyskuje się dla niewielkiej, nieparzystej wartości parametru k = 3, k = 5. Obrazem przykładowego działania klasyfikatora 3 NN jest rysunek poniżej (Rys. 22).



Rys. 22. Schemat działania algorytmu 3-NN (linia ciągła) i 5-NN (linia przerywana)

Źródło: opr. własne

Wykorzystywane miary odległości obiektów klasyfikowanych od ich sąsiadów w przypadku algorytmu kNN są typowe. Podstawą jest odległość euklidesowa (zależność 5.10) a także jej kwadrat. Jeżeli cechy opisujące obiekty wyrażane są w różnych jednostkach, to celem zmniejszenia wpływu ich wartości nominalnych na mierzoną odległość można zastosować miarę zwaną **ważoną odległością euklidesową:**

$$d_{ij}^{E} = \sqrt{\sum_{k=1}^{m} \frac{1}{w_i^2} \left(\boldsymbol{z}_{ik} - \boldsymbol{z}_{jk}\right)^2} \tag{5.18}$$

gdzie:
m – ilość elementów wektora;
i,j – symbole wektorów;
w – waga zmiennej

Ciekawą cechą omawianego klasyfikatora kNN jest to, że nie wymaga on uczenia. W tym sensie jest on podobny do klasyfikatora Bayesa. Nie wymaga natomiast, co jest niewątpliwą zaletą, znajomości typów rozkładów zmiennych dla obiektów w skupieniach. Dlatego jest on uważany za typowy klasyfikator nieparametryczny. Cechą dającą mu przewagę na klasyfikatorem LLM i LDA jest zdolność rozróżniania klas przy braku ich liniowej separowalności. Pozwala to na uzyskanie poprawnych wyników w przypadkach tak skomplikowanych jak ‘wyspowy’ charakter klasy (otoczonej obiektami innych klas), niespójności przestrzennej obszarów klasy czy wreszcie, gdy dochodzi do częściowego pokrywania się przestrzeni zmiennych różnych klas.

Łatwość wykorzystania omawianego klasyfikatora okupiona jest pewnymi jego wadami. Głównym mankamentem jest fakt, że nie definiuje on w sposób matematyczny linii podziału pomiędzy skupieniami. Odpowiedzi kNN mają charakter jedynie jakościowy, co oznacza, że aby otrzymać odpowiedź klasyfikatora dotycząca nowego obiektu, zawsze musimy dysponować pełnym zbiorem sklasyfikowanych już przypadków.

W przypadku kNN istnieje również problem oceny wiarygodności klasyfikacji tą metodą. Nie istnieje bowiem żadna ogólnie przyjęta metoda i miara walidacji klasyfikacji pojedynczego obiektu.

Klasyfikator SIMCA

Klasyfikator SIMCA (Simple Modeling of Class Analogy) z punktu widzenia specyfiki algorytmu jest najbardziej uniwersalnym, łączącym w sobie zalety wszystkich wcześniej wymienionych typów klasyfikatorów. Jego uniwersalność pozwala na rozwiązywanie takich zagadnień chemometrycznych jak określenie przynależności danej próbki do konkretnej populacji. Możemy tu mówić na przykład o problemie wartości odbiegających, błędów grubych. Innym problemem, możliwym do rozwiązania za pomocą tego klasyfikatora, jest określenie stopnia dopasowania obiektu do określonej klasy na podstawie jego cech. Takich możliwości nie dawały nam nieparametryczne klasyfikatory liniowe, kNN i parametryczny klasyfikator Bayesa.

Algorytm klasyfikatora SIMCA oparty jest o metodę głównych składowych. Dla każdej z klas tworzony jest jej indywidualny model w oparciu jedynie o kilka istotnych składowych. Modelem tak stworzonej klasy jest pewna objętość (hiperobjętość w przestrzeni wielu składowych), którą można utożsamiać z przedziałem ufności dla wartości będącej środkiem danej klasy. Przy czym środek ten wyznaczany jest, jako punkt przestrzeni określony wartościami średnimi każdego z czynników. W tak stworzonym modelu klasy, z określonym poziomem prawdopodobieństwa powinny się znaleźć wszystkie obiekty do niej należące. Niezależna analiza PCA prowadzona dla każdej klasy osobno, pozwala na minimalizację rozmiarów przestrzeni każdej klasy i bardziej precyzyjne dopasowanie należących do niej obiektów.

Kryterium przynależności obiektu do danej klasy, jest stosunek wariancji resztowej dla badanego obiektu w klasie np. A $\left(s_i^2\right)_A$ i wariancji resztowej dla tej klasy $(s_0^2)_A$.

$$F = \frac{\left(s_i^2\right)_A}{\left(s_0^2\right)_A} \tag{5.19}$$

Stosunek ten (F) to statystyka Fishera-Snedecora, dająca wiarygodne wyniki, gdy rozkład różnic e_{ik}(pomiędzy rzeczywistą wartością *k*–tej zmiennej *i*–tego obiektu a wartością obliczoną na podstawie modelu) nie odbiega w sposób istotny od rozkładu normalnego. Wariancje wykorzystywane w teście opisują wyrażenia:

wariancja resztowa klasy A – pierwiastek z niej to promień ufności klasy

$$(s_0^2)_A = \frac{\sum_{i=1}^{N}\sum_{k=1}^{M}(e_{ik})_A^2}{(N-S-1)(M-S)} \tag{5.20}$$

gdzie:

N – liczba obiektów w klasie A;

M – liczba cech w klasie A;

S – liczba istotnych składowych w klasie A;

e_{ik} – różnica między rzeczywistą wartością *k*–tej zmiennej *i*–tego obiektu a wartością obliczoną na podstawie modelu.

wariancja resztowa *i*–tego obiektu klasy A

$$\left(s_i^2\right)_A = \frac{\sum_{i=1}^{M}(e_{ik})_A^2}{(M-S)} \tag{5.21}$$

Jeśli test F (5.19) nie wykaże istotności różnic wariancji resztowych modelu klasy i nowo klasyfikowanego obiektu, to obiekt uznajemy za przynależny do danej klasy. W przeciwnym przypadku ($F \geq F_{kr}$) uznajemy go za obiekt odosobniony, który może być przedstawicielem obiektów nowej klasy lub, jak często się zdarza, obiektem obciążonym błędem grubym. Potwierdzeniem przynależności (lub nie) obiektu do klasy może być porównanie jego wartości cech z przeciętnym zakresem cech dla klasy (określonym

promieniem ufności – $(s_0)_A$). Najtrudniejszym, jeśli chodzi o podjęcie decyzji przynależności obiektu do klasy, jest przypadek, kiedy da się stwierdzić przynależność do dwóch/kilku klas jednocześnie. Sytuacja taka wskazuje na nakładanie się przestrzenne klas i zwykle zwiększenie liczby składowych dla poszczególnych modeli skupień rozwiązuje problem niejednoznaczności.

Zastosowanie klasyfikatora SIMCA w przypadku klas tworzących bardziej lub mniej zwarte odseparowane skupienia nie nastręcza większych kłopotów. Ciekawą jego cechą jest możliwość zastosowania, kiedy obiekty jednej z klas nie tworzą widocznego skupienia, ale są równomiernie rozłożone w całej przestrzeni zmiennych – tak zwany **przypadek asymetryczny**. Przypadek taki występuje często w badaniach środowiskowych, np. jakości wód, gleby powietrza. Parametry opisujące jakość środowiska mogą przyjmować w zasadzie dowolne wartości, jednak obiekty 'jakościowo dobre' cechuje zestaw cech tylko z pewnego przedziału, który może być podstawą do stworzenia zwartego modelu klasy. Rezygnacja z budowy modelu dla klasy rozproszonej pozwala na eliminację obiektów odległych od 'jakościowo dobrych', czyli eliminację obiektów rozproszonych z przedziału ufności dla klasy.

Jak w przypadku każdego klasyfikatora możemy mówić, a także dokonać walidacji skuteczności klasyfikacji z wykorzystaniem algorytmu SIMCA. Wielkością oceniającą ilościowo jego zdolność do podziału obiektów '*a* , *b*' na dwie klasy A i B jest parametr opisany równaniem:

$$D_{a,b} = \sqrt{\frac{(s_b^2)_A + (s_a^2)_B}{(s_a^2)_A + (s_b^2)_B}} \tag{5.22}$$

gdzie: $\left(s_b^2\right)_A$ – wariancja resztowa obiektu 'b' w klasie A; itd. …

Im większa wartość parametru $D_{a,b}$, tym większa zdolność klasyfikatora do prawidłowego podziału elementów na klasy, co w dużej mierze zależy od samej struktury danych wejściowych i prawidłowej ilości istotnych składowych klasy.

6 BIBLIOGRAFIA

[1] Mazerski J., *Chemometria praktyczna*, Warszawa, Malamut, 2009.

[2] Larose D.T., *Odkrywanie wiedzy z danych*, Warszawa, PWN, 2006.

[3] Larose D.T., *Metody i modele eksploracji danych*, Warszawa, PWN, 2008.

[4] Rencher A.C., *Methods of Multivariate Analysis,* Wiley-Interscience, 2002.

[5] Brereton R.G., *Applied Chemometrics for Scientists*, Wiley & Sons, 2007.

[6] Miller J.N., Miller J.C., Statistics and Chemometrics for Analytical Chemistry, Pearson, 2010.

[7] Einax J.W., Zwanziger H.W., Geis S., *Chemometrics in Environmental Analysis*, VCH, Wiley Company, 1997.

Jednymi z najistotniejszych elementów wykształcenia chemika są umiejętności prawidłowego prowadzenia pomiarów, jak i interpretacji uzyskanych wyników.
Badania analityczne prowadzone są za pomocą nowoczesnych technik pomiarowych, które skracają czas analiz i dostarczają dużych ilości wyników. Metrologia i walidacja to wciąż rozwijające się dziedziny, łączące wiedzę z zakresu chemii analitycznej i fizycznej, matematyki, statystyki. Pozwalają one na weryfikację uzyskiwanych danych i umożliwiają porównywanie ich między laboratoriami.
Obszerne zbiory danych pomiarowych są trudne do efektywnej interpretacji. Z pomocą przychodzi tu chemometria – nowa dziedzina wiedzy, która zajmuje się wydobywaniem użytecznych informacji z wielowymiarowych danych pomiarowych, bazując na metodach m.in. statystki i matematyki.
Skrypt skierowany jest głównie do studentów kierunków chemicznych i przyrodniczych, ale będzie przydatny również tym wszystkim, którzy w swoim życiu zawodowym wykonują pomiary chemiczne oraz zajmują się interpretacją uzyskanych wyników badań.

Projekt finansowany ze środków funduszy norweskich oraz środków krajowych

Publikacja realizowana w ramach Projektu finansowanego ze środków funduszy norweskich oraz środków krajowych na Rozwój Polskich Uczelni

Książka dostępna również jako e-book

WYDAWNICTWO UNIWERSYTETU ŁÓDZKIEGO

www.wydawnictwo.uni.lodz.pl
e-mail: ksiegarnia@uni.lodz.pl
tel. (42) 665 58 63