ROCZN. PZH 2008, 59, NR 2, 117-129

WOJCIECH KORCZ, KATARZYNA GÓRALCZYK, KATARZYNA CZAJA, PAWEŁ STRUCIŃSKI, AGNIESZKA HERNIK, TOMASZ SNOPCZYŃSKI, JAN K. LUDWICKI

ZASTOSOWANIE METOD STATYSTYCZNYCH W BADANIACH CHEMICZNYCH

THE APPLICATION OF STATISTICAL METHODS IN CHEMICAL EXPERIMENTS

Zakład Toksykologii Środowiskowej
Narodowy Instytut Zdrowia Publicznego - Państwowy Zakład Higieny
00–791 Warszawa, ul. Chocimska 24
e-mail: wkorcz@pzh.gov.pl
Kierownik: *prof. dr hab. J.K. Ludwicki*

Omówiono kryteria doboru wybranych metod statystycznych w zależności od analizowanych danych pomiarowych. Przedstawiono ich zastosowanie w badaniach chemicznych.

Słowa kluczowe: chemometria, modelowanie zależności, analiza składowych głównych
Key words: chemometry, object modeling, principal compound analysis

WSTĘP

Chemia analityczna obejmuje szerokie spektrum zagadnień związanych m. in. z badaniem różnego rodzaju zależności fizyko-chemicznych, które umożliwiają potwierdzenie tożsamości oznaczanych substancji oraz określenie ich stężenia. W wyniku procesu analitycznego gromadzone są liczne dane, których analiza wymaga zastosowania odpowiednich algorytmów statystycznych. Analiza większości zjawisk, ze względu na ich złożoność wymaga podejścia wielowymiarowego, umożliwiającego ich opis wykorzystujący zależności między zmiennymi. Istnieją metody statystyczne dające możliwość analizy wielowymiarowych zbiorów danych. Wymagają one jednak korzystania ze złożonych obliczeń matematycznych, określonej budowy matrycy danych eksperymentalnych i odpowiedniej jej wielkości. Postęp techniki, który doprowadził do rozpowszechnienia komputerów i rozwoju oprogramowania umożliwiającego zastosowanie algorytmów statystycznych zwiększył dostępność i zastosowanie chemometrii [2, 13, 22]. Chemometria (tak jak pokrewne ekonometria w ekonomii i biometria w biotechnologii) jest działem chemii wykorzystującym matematykę, statystykę, informatykę i teorię podejmowania decyzji do projektowania i optymalizacji warunków doświadczalnych oraz do zdobycia maksimum użytecznej informacji z uzyskanych danych pomiarowych [15, 22].

Jakość przetwarzanych danych pomiarowych ma zasadniczy wpływ na uzyskanie wiarygodnych wyników. Istotne jest więc właściwe zaplanowanie doświadczenia w celu zmini-

malizowania liczby pomiarów koniecznych do uzyskania informacji o mierzonym obiekcie, co w analityce pozwala na ekonomizację czasu i środków. Chemometria nie zajmuje się prowadzeniem pomiarów w przeciwieństwie do metrologii obejmującej rzetelność wykonania zgodnie z odpowiednią metodyką pomiarową i zarejestrowania wyniku pomiaru. Obejmuje więc kolejny etap polegający na sprawdzeniu czy dane otrzymane z pomiarów nie zawierają tzw. błędów grubych lub wyników znacząco odbiegających od pozostałych, mogących zafałszować ostateczny wynik badania [6, 22].

DOBÓR ADEKWATNEGO ALGORYTMU STATYSTYCZNEGO DO ANALIZY DANYCH POMIAROWYCH

Chemometria stosowana jest w celu stworzenia matematycznego modelu zależności między badaną zmienną zależną *y,* lub wieloma zmiennymi zależnymi y_i i licznym zbiorem zmiennych objaśniających x_i (parametry wpływające na pomiar). Wyróżnia się tu dwa przypadki:

(a) liczba zmiennych objaśniających nie przekracza 10 i ustalenie ich wartości jest możliwe. W takim przypadku właściwe zaplanowanie metodyki pomiarowej umożliwia późniejsze wykorzystanie klasycznej analizy regresyjnej.

(b) liczba zmiennych objaśniających jest duża (np. kilkadziesiąt) i wartości niektórych zmiennych nie można określić (zmierzyć). W takim przypadku mamy do czynienia z wielowymiarową matrycą zmiennych objaśniających, co wymaga uwzględnienia wzajemnej korelacji pomiędzy zmiennymi. Ponieważ większość parametrów układu jest w pewnym stopniu skorelowana ze sobą, a liczba zmiennych nieskorelowanych (niezależnych) jest niewielka, należy wykorzystując wybrane algorytmy statystyczne (np. analiza składowych głównych) dokonać transformacji zbioru zmiennych objaśniających w celu otrzymania zbioru nowych zmiennych *p* wzajemnie ortogonalnych, wykorzystywanych w analizie regresyjnej.

Tak uzyskany model zależności może znaleźć również inne zastosowania np. optymalizacja procesu technologicznego, kontrola produkcji, gdy zachodzi potrzeba oceny określonych zmiennych zależnych (np. pomiar stężenia wybranego substratu stosowanego w procesie technologicznym techniką in-line i on-line) [12, 22, 31].

Wnioski wynikające z analizy chemometrycznej mogą być błędne gdy w zbiorze analizowanych danych znajdzie się nieprawdziwy wynik. Aby wyeliminować taką ewentualność stosuje się wstępną kontrolę danych, która pozwala wyeliminować pomyłki powstałe podczas procesu analitycznego, wykryć wyniki odbiegające, określić jednorodność zbioru danych oraz wyeliminować punkty z brakiem danych. Dla większości algorytmów chemometrycznych wymagany jest odpowiedni rozkład populacji zmiennych. Zastosowanie danych o innym rozkładzie, nie spełniającego wymogów zastosowanego algorytmu chemometrycznego, może prowadzić do ich błędnej interpretacji [2, 3, 5, 11].

Ocena rozkładu populacji wyników danych pomiarowych

Właściwości rozkładu danej cechy opisywane są przez wskaźniki położenia i rozproszenia [18, 20, 22, 32].

Wskaźniki położenia to:
- **wartość średnia próby** (średnia arytmetyczna)
- **mediana**, tj. wartość środkowa populacji próbek uporządkowana od wartości najmniejszej do największej
- **wartość modalna** (cecha która w danym rozkładzie występuje najczęściej).

Do wskaźników rozproszenia zaliczane są:
- **rozstęp próby**, tj. różnica między wartością najmniejszą i największą populacji
- **wariancja w próbie**, tj. średnia arytmetyczna kwadratów odchyleń poszczególnych wartości próby od średniej arytmetycznej populacji
- **odchylenie standardowe**, tj. pierwiastek kwadratowy z wariancji, określa zróżnicowanie poszczególnych wartości zmiennych w populacji od średniej arytmetycznej populacji
- **odchylenie przeciętne**, tj. średnia arytmetyczna bezwzględnych odchyleń wartości cechy od średniej arytmetycznej.

Jeżeli rozkład populacji próbek jest niezgodny z rozkładem normalnym konieczne jest zbadanie kierunku zróżnicowania wartości zmiennej. Do liczebnego określenia kierunku i siły symetrii wykorzystywany jest współczynnik skośności q, którego wartość przedstawia asymetrię rozkładu populacji w stosunku do standardowego rozkładu naturalnego (rozkład normalny jest rozkładem symetrycznym) [18, 22, 28, 32].

$$q = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^3}{(n-1)\cdot S^3},$$

gdzie:

q – współczynnik skośności
S – odchylenie standardowe
n – liczebność próby

Niekiedy, w zależności od zastosowanej metody statystycznej, populację próbek należy poddać transformacji, aby osiągnęła rozkład maksymalnie zbliżony do wymaganego w danej metodzie chemometrycznej. Przy zastosowaniu modelu regresyjnego, który wykorzystywany jest przy opracowywaniu większości zastosowań analitycznych wymagany jest rozkład normalny [28].

W przypadku danych doświadczalnych często obserwuje się wyniki leżące z dala od pozostałych rezultatów, które określa się jako punkty odbiegające. Najczęściej nie wiadomo jednak czy jest to wynik błędnego pomiaru, niewłaściwego przygotowania próbki czy też efekt rozkładu danej zmiennej. Ocena tego problemu jest możliwa tylko wtedy, gdy pozostałe wyniki mają rozkład normalny [1, 16, 22].

Kryteria oceny punktów (wyników) odbiegających

Do kontroli punktów odbiegających stosowane są [16, 28, 32]:

- **test *Dixona***, który ma zastosowanie dla małych populacji zmiennych, w którym szereguje się wyniki w kolejności rosnącej. Wynik najmniejszy lub największy kwalifikowany jest jako punkt odbiegający.

$$x_1 \leq x_2 \leq \ldots \leq x_n$$

Korzystając z poniższych wzorów wyznacza się wartość Q.

$$Q = \frac{x_2 - x_1}{x_n - x_1}$$ Punktem odbiegającym jest punkt najmniejszy

$$Q = \frac{x_n - x_{n-1}}{x_n - x_1}$$ Punktem odbiegającym jest punkt największy

Następnie porównuje się wartość Q z Q_{kryt} z tabeli. Punkt odbiegający powinien zostać odrzucony jeżeli $Q \geq Q_{kryt}$.

- **test *t-Studenta***, stosowany dla populacji większej niż 10 wyników.

Zakłada się, że populacja danych, z których wykorzystano *n* wyników, posiada rozkład normalny, charakteryzuje się wartością średnią μ i odchyleniem standardowym σ. Tworzony jest rozkład *Studenta* zawierający 10>n>40 wyników z całej populacji. Wyznacza się wartość średnią *m* i odchylenie standardowe *s* tego zbioru próbek. Przyjmuje się je jako parametry całej populacji wejściowej. Zakłada się możliwość popełnienia błędu, który nie powinien jednak wystąpić z prawdopodobieństwem większym niż poziom istotności α przyjmujący najczęściej wartość 0,05. Korzystając z rozkładu *Studenta* wyznacza się przedział ufności wokół wartości średniej populacji, gdzie z prawdopodobieństwem 1 – α można oczekiwać wszystkich wartości populacji (test t-*Studenta*). Promień przedziału ufności jest wielokrotnością odchylenia standardowego *s*. Krańce przedziału ufności wyznacza się ze wzorów [16, 17, 28, 32]:

$$x_{\min} = m - t_\alpha \cdot s$$

$$x_{\max} = m + t_\alpha \cdot s ,$$

gdzie:

t_α – wartości t dla różnych liczebności populacji i poziomu istotności zawarte są w tabelach statystycznych

m – wartość średnia dla populacji opisanej rozkładem *Studenta*

s – odchylenie standardowe populacji opisanej rozkładem *Studenta*

- **reguła trzech sigm.** Dla populacji o liczebności n>30 rozkład *Studenta* jest zbliżony do rozkładu normalnego. Z teorii rozkładu normalnego przedział ufności tego rozkładu o promieniu równym odchyleniu standardowemu σ zawiera 2/3 populacji rozkładu danej cechy. W przedziale ufności o promieniu 2σ zawarte jest około 90% wartości cechy z danej populacji wyników, a w odległości 3σ około 95% wartości zmiennej [16, 22, 28].

MODELOWANIE BADANEJ ZALEŻNOŚCI

Celem badania powtarzalnego zjawiska (obiektu) zależnego od kilku zmiennych jest stworzenie funkcji opisującej tę zależność, tzn. opracowanie metody oszacowania odpowiedzi badanego zjawiska na podstawie znanych wartości zmiennych objaśniających [13].

Przykładem obiektu może być przyrząd pomiarowy np. chromatograf sprzężony z detektorem. Chromatograf dokonuje rozdziału badanej próbki, a następnie odczytywany jest sygnał (odpowiedź) detektora zależny od stężenia danego analitu. Na stężenie oznaczanego związku chemicznego ma wpływ m. in. matryca, etap przygotowania próbki (np. ekstrakcja), interferencje spowodowane zanieczyszczeniem próbki czy rozkładem analitu. Za czynnik wpływający na pomiar można uznać np. skład fazy stacjonarnej, temperaturę, przepływ, ciśnienie, stabilność przepływu, skład fazy ruchomej oraz czynniki wynikające z budowy i właściwości zastosowanego detektora takie jak czułość, selektywność, poziom szumów oraz dryft [15, 22].

Pierwszym etapem modelowania obiektu jest jego identyfikacja tj. dopasowanie rezultatów modelowania do otrzymanych wyników pomiarowych. Kolejnym etapem jest zbadanie istotności modelu, która polega na porównaniu testem *F- Snedecora* wariancji (wariancji resztowej) odpowiedzi modelu i obiektu. Model jest istotny statystycznie, gdy obliczona wartość F jest większa od wartości krytycznej F_{kryt} odczytanej z tablic statystycznych dla danego poziomu istotności. Następnie, poprzez ocenę dokładności modelu i dokładności pomiaru odpowiedzi obiektu za pomocą testu *F,* ocenia się adekwatność modelu. Zdolność prognostyczną modelu określa się na podstawie dodatkowych pomiarów znanych cech i porównuje się je z rezultatami uzyskanymi w wyniku zastosowania tego modelu [12, 22, 32].

Dla odpowiednio małego przedziału zmiennej objaśniającej każdą funkcję ciągłą i różniczkowalną można przybliżyć (oszacować) wielomianem niskiego stopnia, ponieważ im przedział zmiennej jest niższy, tym niższy jest stopień wielomianu [16, 20, 31]. W takim przypadku najprostszy i najczęściej stosowany w analityce jest model liniowy umożliwiający oszacowanie rzeczywistej odpowiedzi badania [5, 12, 22, 24, 32]:

$$\eta=\beta_0+\beta_1 x_1+\beta_2 x_2+...+\beta_m x_m$$

gdzie:

η – zmienna zależna będącą rzeczywistą odpowiedzią eksperymentalną (badana cecha)
β_m – współczynnik modelu liniowego, parametr regresji liniowej
X_m – zmienna niezależna lub objaśniająca

Zatem wartość mierzona (y) nie jest wartością rzeczywistą, lecz funkcją wartości rzeczywistej pomiaru (η) i błędu pomiarowego (ε):

$y=\eta+\varepsilon$,

wynikającym z niedoskonałości aparatury pomiarowej i wpływu czynników zewnętrznych. Ponieważ nie można zachować idealnie tych samych warunków eksperymentalnych dla wszystkich cech opisujących doświadczenie, powtarzane pomiary dają różne wyniki [1, 4, 5, 6, 16, 24]. Po uwzględnieniu błędu pomiarowego w równaniu liniowym dla i-tego pomiaru równanie przybiera postać[1, 22, 24]:

$y_i=\eta_i+\varepsilon_i=b_o+b_1x_{i1}+b_2x_{i2}...b_mx_{im}+e_i$,

gdzie:

b – oszacowanie współczynnika modelu β wynikające z niemożności wyznaczenia wartości rzeczywistej pomiaru

e – człon równania reprezentujący błąd

x_j – zmienna niezależna lub objaśniająca.

Dla układu złożonego model liniowy może nie wystarczać do opisania zależności badanej cechy. Stosowany jest wtedy model liniowy z członami interakcyjnymi, tj. członami zawierającymi iloczyny zmiennych objaśniających. Kolejnym rozwinięciem modelu liniowego jest model kwadratowy.

Do wyznaczenia b konieczna jest minimalna liczba wykonanych pomiarów, zależnie od liczby zmiennych objaśniających ten układ pomiarowy i charakterystyczna dla wybranego typu modelu (np. model liniowy wymaga $m+1$ pomiarów, gdzie m jest liczbą zmiennych opisujących). W praktyce, do dobrego oszacowania współczynników należy wykonać większą liczbę pomiarów, co pozwoli uśrednić wpływ błędów pomiarowych. Zwiększając liczbę pomiarów ponad niezbędne minimum zwiększa się liczbę stopni swobody i ogranicza wpływ błędów pomiarowych. Konieczne jest, aby zmienne objaśniające x_i (parametry wpływające na pomiar odpowiedzi obiektu) posiadały rozstęp kilkakrotnie większy niż odchylenie standardowe tych zmiennych. Zadowalające oszacowanie współczynników modelu można osiągnąć dla mniejszej liczby pomiarów stosując odpowiedni dobór punktów pomiarowych i korzystając z metod statystycznych [12, 17, 20, 28, 32].

Podstawowym założeniem metod regresyjnych jest jak najlepsze dopasowanie rezultatów modelowania do wyników pomiaru badanej cechy. Miarą tego dopasowania jest suma kwadratów różnic (SKR) [5, 12, 16, 25].

$$SKR = \sum_{i=1}^{n}\left(y_i - \tilde{y}_i \right)^2 = \sum_{i=1}^{n} e_i^2$$

gdzie:

$$\tilde{y}_i = b_0 + b_1x_{i1} + b_2x_{i2} + ... + b_mx_{im}$$

$y_i=\eta_i+\varepsilon_i=b_o+b_1x_{i1}+b_2x_{i2}...b_mx_{im}+e_i$

n – liczebność populacji.

Współczynniki modelu *b* wyznacza się przez założenie minimum SKR i oblicza metodą najmniejszych kwadratów [17, 20, 28, 32]. Suma kwadratów różnic zależna jest od relacji liczby odpowiedzi obiektu i liczby zmiennych opisujących, czyli uwzględnionych parametrów wpływających na pomiar. W celu otrzymania statystycznego miernika jakości dopasowania zostało wprowadzone pojęcie wariancji resztowej [22]:

$$S^2 = \frac{SKR}{n-m-1}$$

gdzie:

n-m-1 – liczba stopni swobody.

Wariancja resztowa wykorzystywana jest do testów statystycznych, takich jak test istotności modelu, który jest statystycznie istotny, jeżeli wyjaśni istotną część zmienności odpowiedzi obiektu. Wariancja odpowiedzi obiektu [17, 22, 32] to:

$$S_y^2 = \frac{\sum\left(y_1 - \tilde{y}\right)^2}{n-1}$$

gdzie:

$$\tilde{y}_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_m x_{im}$$

$$y_i = \eta_i + \varepsilon_i = b_o + b_1 x_{i1} + b_2 x_{i2} \ldots b_m x_{im} + e_i$$

n – liczebność populacji

Na tej podstawie, porównując wariancję odpowiedzi obiektu (S_y^2) i wariancję resztową (S^2) otrzymuje się miarę istotności modelu [16, 25]:

$$F = \frac{S_y^2}{S^2}$$

Miarą jakości dopasowania jest również współczynnik determinacji modelu *D* [17, 20, 28, 32]:

$$D = 1 - \frac{S^2}{S_y^2}$$

Współczynnik determinacji określa jaki ułamek całkowitej zmienności odpowiedzi jest wyjaśniony przez model. Współczynnik determinacji powiązany jest ze współczynnikiem korelacji *R* (r dla modelu liniowego) [22, 27, 32]:

$$R = \sqrt{D} .$$

Analiza danych chemometrycznych polega na ujawnianiu cech najbardziej ze sobą powiązanych. Przy założeniu liniowej zależności stosowana jest analiza korelacji. Miarą współzależności pomiędzy zmienną *x* i *y* jest współczynnik korelacji liniowej Pearsona (*r*).

$$r = \frac{\sum xy - \frac{\sum x \cdot \sum y}{n}}{\sqrt{\left(\sum x^2 - \frac{\left(\sum x\right)^2}{n}\right) \cdot \left(\sum y^2 - \frac{\left(\sum y\right)^2}{n}\right)}}$$

gdzie:

n- liczebność populacji

Może on przyjmować wartość z przedziału od –1 do 1. Dla wartości skrajnych zmienna *x* jest silnie powiązana ze zmienną *y*. Dla wartości r=0 brak jest korelacji liniowej, co nie musi oznaczać niezależności zmiennych tylko fakt, że taka zależność nie jest liniowa. Istotność statystyczną korelacji ocenia się poprzez porównanie współczynnika korelacji (*r*) z wartościami z tablic (wartości krytyczne) dla odpowiedniej liczby stopni swobody i założonego poziomu ufności [4, 17, 20, 28, 32]. Zależność liniowa jest istotna statystycznie, jeżeli obliczona wartość jest większa od wartości krytycznej. W przypadku silnych korelacji pomiędzy zmiennymi objaśniającymi stosowana jest analiza składowych głównych (PCA) i cząstkowa metoda najmniejszych kwadratów (PLS) [7, 8, 11, 14, 17, 22, 23].

Ocena adekwatności modelu polega na sprawdzeniu czy stworzony model w zadowalający sposób odzwierciedla zachowanie obiektu. Model jest adekwatny, kiedy jego dokładność jest tego samego rzędu, co dokładność pomiaru odpowiedzi [12, 32].

Tworzone modele powinny umożliwiać przewidywanie wielkości odpowiedzi obiektu w zakresie zmiennych objaśniających, dla którego została dokonana identyfikacja modelu. Po stworzeniu modelu przeprowadza się serię pomiarów o liczebności *k* pomiarów. Wartości tych pomiarów porównywane są z wartościami przewidzianymi przez model. Jeżeli różnice pomiędzy odpowiedzią modelu i obiektu zbliżone są do błędów pomiarowych, dany model może zostać zastosowany. Miarą prognostycznych zdolności modelu jest współczynnik walidacji wyznaczany poprzez obliczenie sumy kwadratów różnic pomiędzy odpowiedziami obiektu i modelu dla dodatkowej serii pomiarów [18, 22, 32].

$$SKR_{walidacji} = \sum_{i=1}^{k} \left(y_i - \tilde{y}_i \right)^2$$

Następnie wyznacza się wariancję dodatkowej serii pomiarów i współczynnik walidacji Q^2.

$$s_k^2 = \frac{SKR_{walidacji}}{k-1}$$

gdzie:

s_k^2- wariancja pomiarów dodatkowych

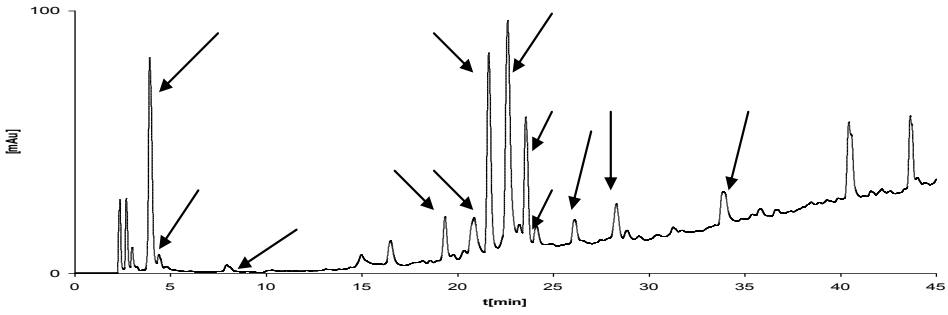
k – liczebność serii pomiarów wykonanych podczas walidacji

$$Q^2 = 1 - \frac{s_k^2}{s_y^2}$$

Model posiada tym większe zdolności prognostyczne im współczynnik walidacji jest bliższy jedności.

PRZYKŁADY ZASTOSOWANIA METOD STATYSTYCZNYCH

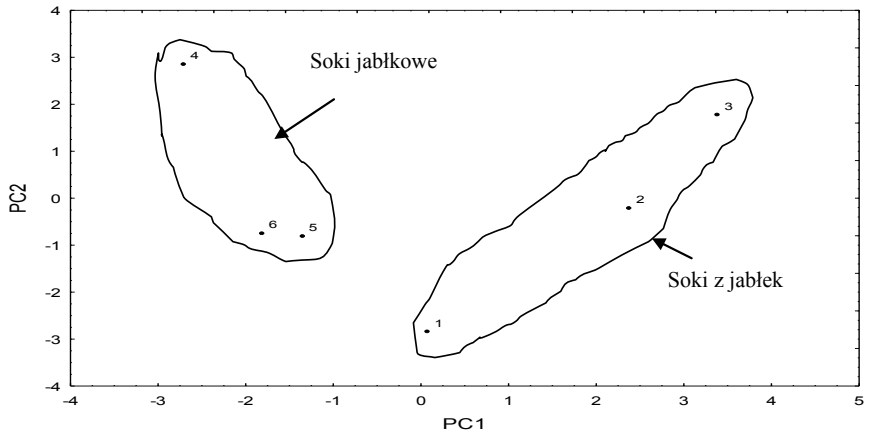
Przykładem zastosowania analizy składowych głównych może być porównanie wyników badań własnych chromatograficznego rozdziału próbek soków wyciśniętych ze świeżych owoców i soków owocowych dostępnych w obrocie [19].



Ryc. 1. Chromatogram soku z jabłka
Fig. 1 Chromatogram of natural apple juice

W tym przypadku stworzono macierz, w której w kolumnach umieszczono powierzchnię pików związków chemicznych „charakterystycznych" dla jabłek wybranych do dalszej analizy chemometrycznej. Tożsamość związku potwierdzano przez porównanie widma absorpcyjnego uzyskanego za pomocą detektora skanującego UV dla wybranego czasu retencji. W wierszach macierzy pogrupowano kolejne próbki handlowych soków jabłkowych i soków wyciśniętych z jabłek. Traktując piki jako zmienne zastosowano analizę składowych głów-

nych. Projekcja dwóch składowych głównych opisujących największy udział wariancji własnych w przestrzeni danych pozwoliła na wizualizację i klasyfikację badanych próbek [19].



Ryc. 2. Klasyfikacja soków wyciśniętych z jabłek i soków jabłkowych. 1 - sok z jabłka odmiany jonagared, 2 – sok z jabłka odmiany cortland, 3 – sok z jabłka odmiany reneta, 4 – sok jabłkowy (producent 1), 5 – sok jabłkowy (producent 2), 6 – sok jabłkowy (producent 3).

Fig. 2 Classification of extracted apple juices and commerce apple juices. 1 – jonagared variety, 2 - cortland variety, 3 – reneta variety, 4 – apple juice (producer 1), 5 – apple juice (producer 2), 6 – apple juice (producer 3)

Ocenę klasyfikacji wykonano wizualnie, chociaż istnieją również algorytmy statystyczne do oceny klasyfikacji wewnątrz grupowej. Przykład ten ilustruje praktyczne zastosowanie analizy składowych głównych.

PODSUMOWANIE

W laboratorium analitycznym rutynowo korzysta się z metod statystycznych w badaniach chemicznych. Znaczna liczba pomiarów umożliwia stworzenie modelu zależności stężenia analitu od odpowiedzi detektora dla pewnego przedziału stężeń, w którym występuje zależność liniowa. Analogicznie jak przy modelowaniu obiektu wykonuje się identyfikację modelu, bada jego istotność, ocenia adekwatność oraz określa zdolność prognostyczną. Jest to wykonywane na etapie walidacji metody analitycznej. Metodyka oceny modelowania w przypadku metody badawczej stosowanej w laboratorium analitycznym opisana jest w odpowiednich normach i zeszytach metodycznych [7, 21, 26]. Stosując test *F-Snedecora* określa się czy wyniki oznaczeń mieszczą się z określonym prawdopodobieństwem w dopuszczalnym dla danej metody przedziale, co stanowi element procesu sterowania jakością badań [16].

Oprogramowanie sprzężone z przyrządem pomiarowym, umożliwia stworzenie modelu zależności, tj. podanie jego równania i obliczenie jego współczynnika korelacji i determinacji.

Wbudowanie metod statystycznych w oprogramowanie sterujące pracą aparatury pomiarowej pozwala na skrócenie pracy analityka poprzez szybszą ocenę wyników badań. Rozwój oprogramowania statystycznego umożliwia stosowanie często bardzo trudnych i złożonych

algorytmów chemometrycznych bez potrzeby znajomości odpowiednich wzorów matematycznych. Istotna jest jednak uwaga, jaką należy przykładać do danych pomiarowych zastosowanych w analizie chemometrycznej, a także znajomość ograniczeń zastosowanych metod statystycznych. Nieumiejętne zastosowanie metod statystycznych prowadzić może do błędnej interpretacji wyników badań [10, 11, 30].

Metody statystyczne znalazły zastosowanie w badaniach chemicznych do gromadzenia i przetwarzania informacji o związkach chemicznych w celu zarządzania przepływem informacji [3, 24, 29, 31]. Pozwalają na przewidywanie fizykochemicznych i biologicznych właściwości związków [29]. Stosowane są w ocenie jakości, pochodzenia i wieku badanych próbek [13, 15]. Metody statystyczne są również wykorzystywane w zarządzaniu jakością w analizie chemicznej. Znalazły także zastosowanie w analizie śladowej przy badaniach produktów spożywczych pod kątem różnych chemicznych zanieczyszczeń, w tym pozostałości pestycydów [10, 30].

W. Korcz, K. Góralczyk, K. Czaja, P. Struciński, A. Hernik,
T. Snopczyński, J.K. Ludwicki

ZASTOSOWANIE METOD STATYSTYCZNYCH W BADANIACH CHEMICZNYCH

Streszczenie

Jakość danych pomiarowych ma zasadniczy wpływ na uzyskanie wiarygodnych wyników. Stosując metody statystyki matematycznej możliwe jest ograniczanie wybranych etapów pracy chemika np. przy klasyfikacji liczebnego zbioru wyników pomiarowych. Metody statystyczne są również wykorzystywane przy wstępnej ocenie jakości uzyskanych danych. W tym przypadku należy sprawdzić czy dane otrzymane z pomiarów nie zawierają tzw. błędów grubych lub wyników znacząco odbiegających od pozostałych, mogących zafałszować wynik doświadczenia. Analiza danych, które podlegają przetwarzaniu technikami chemometrycznymi, polega na ujawnieniu cech najbardziej ze sobą powiązanych. Chemometria stosowana jest w celu stworzenia matematycznego modelu zależności między badaną cechą i licznym zbiorem zmiennych objaśniających. Przy modelowaniu należy dokonać identyfikacji modelu, zbadać jego istotność i adekwatność oraz określić zdolność prognostyczną. Uzyskany model zależności można wykorzystywać do optymalizacji układu, prognozowania wartości zmiennych zależnych na podstawie znanych zmiennych opisujących.

Metody statystyczne znalazły zastosowanie w badaniach chemicznych do gromadzenia i przetwarzanie informacji o związkach chemicznych w celu sprawniejszego zarządzania przepływem informacji. Pozwalają przewidywać fizykochemiczne i biologiczne właściwości związków. Metody statystyczne są również wykorzystywane do zarządzania jakością w analizie chemicznej zanieczyszczeń np. pozostałości pestycydów w żywności.

W. Korcz, K. Góralczyk, K. Czaja, P. Struciński, A. Hernik,
T. Snopczyński, J.K. Ludwicki

THE APPLICATION OF STATISTICAL METHODS IN CHEMICAL EXPERIMENTS

Summary

Quality of the analyzed data has a major impact on reliability of the results. Application of statistical methods allows to reduce some stages of chemist's work, for example classification of the numerous data sets. The statistical methods are applied for preliminary evaluation of the data quality. In this case it is necessary to verify that the raw data base does not include large errors or outliers, which could influence the result of experiment. Data analysis, which is performed by chemometric techniques, rely on finding the most correlated attributes. Chemometry is used towards creation of the mathematical model of relation between analyzed property and numerous sets of described variables (parameters which affect measure). Modeling requires calculations towards model identification, checking its relevance, evaluation of the adequacy and determination of model's prognostic ability. The obtained model of relation could be used for the system optimization in the technological process, forecasting the values subsidiary conditioned upon known values described, also for control of the analytical system. The statistical methods are applied in chemical studies for data collection and analysis of chemical compounds for more efficient management of flow of the information. They allow to foreseen physical and biological properties of chemical compounds. The statistical methods are also applied for quality management in chemical analysis of contaminants including pesticide residues in foodstuff.

PIŚMIENNICTWO

1. *Aleksandrov Y.I., Belyakov V.I.:* Error and Uncertainty in the Results of Chemical Analysis, J. Anal. Chem. 2002, 57, 2, 94 – 103.
2. *Chrétien R.J.:* The state of the art for chemometrics in analytical chemistry, Anal. Bioanal. Chem. 2002, 372, 511-512.
3. *Defernez M., Kemsley E.K.*: The use and misuse of chemometrics for treating classification problems, Trends Anal. Chem. 1997, 16, 4.
4. *Dobecki M.:* Zapewnienie jakości analiz chemicznych, IMP, Łódź 1997.
5. *Dobosz M.*: Wspomagana komputerowo statystyczna analiza wyników badań, EXIT, Warszawa 2001.
6. *Dvorkin V.I.:* Adequacy and Inadequacy in the Metrology of Chemical Analysis, J. Anal. Chem. 2003, 58, 6, 504-508.
7. EA 4/16: EA guidelines on the expression of uncertainty in quantitative testing, December 2003 rev00.
8. *Eilers P.H.C., Marx B.D.*: Multivariate calibration with temperature interaction using two-dimensional penalized signal regression, Chem. Intell. Lab. Sys. 2003, 66, 159–174.
9. *Escandar G.M., Damiani P.C., Goicoecha H.C., Olivieri A.C.:* A review of multivariate calibration methods applied to biomedical analysis, Microchem. J. 2006, 82, 29-42.
10. EURACHEM: Przydatność metod analitycznych do określonych celów, Przewodnik walidacji metod w laboratorium i zagadnienie związane, POLLAB 2 (30).
11. *Frenich A.G., Martinez Vidal J.L., Parrilla P., Martinez Galera M.*: Resolution of folpet, procymidone and triazophos in high performance liquid chromatography diode array detection by using partial least squares calibration to cross sections of spectrochromatograms, J. Chromatogr. A, 1997, 778, 183-192.

12. *Gajek L., Kałuszka M.:* Wnioskowanie statystyczne Modele i metody, Wydawnictwa Naukowo – Techniczne, Warszawa 2001.
13. *Gastaigner J.:* Chemoinformatics: a new field with a long tradition, Anal. Bioanal. Chem. 2006, 384, 57-64.
14. *Gutés A., Ibañez A.B., Céspedes F., Alegret S., del Valle M.*: Simultaneous determination of phenolic coumpounds by means of an automated voltammetric "electronic tongue", Anal. Bioanal. Chem. 2005, 382, 471-476.
15. *Hasegawa T.*: Chemometrics for spectroscopic analysis, Anal. Bioanal. Chem. 2003, 375, 18-19.
16. *Hryniewicz O.:* Nowoczesne metody statystycznego sterowania jakością, Omnitech, Warszawa 1996.
17. *Jóźwiak J., Podgórski J.:* Statystyka od Podstaw, Polskie Wydawnictwa Ekonomiczne, Warszawa 1997.
18. *Karoui R*., De Baerdemaeker J., *Dufour E*.: A comparison and join use of mid infrared and fluorescence spectroscopic methods for differentiating between manufacturing process and sampling zones of ripened soft cheeses, Euro. Food Res. Tech. 2007.
19. *Korcz W.:* Zastosowanie HPLC do potwierdzania autentyczności produktu na przykładzie soków owocowych, Politechnika Warszawska, Wydział Chemiczny, Praca magisterska, Warszawa 2003.
20. *Kuszewski P., Podgórski J*.: Statystyka Wzory i tablice, Szkoła Główna Handlowa, Warszawa 1998.
21. *Ludwicki J.K., Góralczyk K., Hernik A., Czaja K., Struciński P.:* Walidacja metod analitycznych i szacowanie niepewności wyników w badaniach chemicznych zanieczyszczeń żywności, Wydawnictwo Metodyczne Państwowego Zakładu Higieny, Warszawa 2003.
22. *Mazerski J.:* Podstawy Chemometrii, Wydawnictwo Politechniki Gdańskiej, Gdańsk 2000.
23. *Mendieta J.,Diaz-Cruz M.S., Esteban M., Tauler R.*: Multivariete Curve Resolution: A Possible Tool in the Detection of Intermediate Structures in Protein Folding, Biophysical J. 1998, 74, 2876-2888.
24. *Nezhikhovskii G. R.*: Selection of the initial error model in developing analytical chemical measurement procedures, Measure. Tech.1998, 41, 3.
25. *Pappa-Louisi A., Nikita P.*: Statistical tests for the selection of the optimum parameters set in models describing response surfaces in reversed-phase liquid chromatography, Chromatographia 2003, 57, 169-176.
26. PN-EN ISO/IEC 17025:2005: Ogólne wymagania dotyczące kompetencji laboratoriów badawczych i wzorcujących
27. *Rappaport K.D., Kettaneh N., Wold S.:* Perspectives on Implementing Statistical Modeling and design (SMD) in an Industrial/Chemical Environment, American Stat. 1998, 52, 2.
28. *Roeske–Słomka I.:* Podstawy Statystyki, Politechnika Koszalińska, Koszalin 1997.
29. *Rosania G.R., Crippen G., Woolf P., States D., Shedden K.*: A Cheminformatic Toolkit for Mining Biomedical Knowledge, Pharmaceutical Res. 2007, 24, 10.
30. SANCO/10232/2006: Quality control procedures for pesticide residues analysis.
31. *Seasholtz M.B.:* Making money with chemometrics, Chem. Intell. Lab. Sys. 1999, 45, 55-63.
32. *Sobczyk M.:* Statystyka, PWN, Warszawa 1997

Otrzymano: 20.01.2008