



You have downloaded a document from  
**RE-BUŚ**  
repository of the University of Silesia in Katowice

**Title:** Modelowanie nieselektywnych sygnałów analitycznych w kontekście kontroli jakości wybranych produktów

**Author:** Joanna Orzeł

**Citation style:** Orzeł Joanna. (2015). Modelowanie nieselektywnych sygnałów analitycznych w kontekście kontroli jakości wybranych produktów. Praca doktorska. Katowice : Uniwersytet Śląski

© Korzystanie z tego materiału jest możliwe zgodnie z właściwymi przepisami o dozwolonym użytku lub o innych wyjątkach przewidzianych w przepisach prawa, a korzystanie w szerszym zakresie wymaga uzyskania zgody uprawnionego.



UNIWERSYTET ŚLĄSKI  
W KATOWICACH



Biblioteka  
Uniwersytetu Śląskiego



Ministerstwo Nauki  
i Szkolnictwa Wyższego

PRACA DOKTORSKA

---

Modelowanie nieselektywnych sygnałów  
analitycznych w kontekście kontroli jakości  
wybranych produktów

---

mgr Joanna Orzeł

Promotor pracy:

dr hab. prof. UŚ Michał Daszykowski

Instytut Chemii  
Wydział Matematyki, Fizyki i Chemii  
Uniwersytet Śląski  
Katowice, 2015

### ***Składam serdeczne podziękowania***

*Panu dr hab., prof. UŚ MICHAŁOWI DASZYKOWSKIEMU  
za umożliwienie realizacji moich badań, wszelką pomoc, nieustanną mobilizację  
do pracy, a także cenne uwagi, wskazówki i rady udzielone zarówno w trakcie  
studiów jak i podczas pisania niniejszej rozprawy.*

### ***Dziękuję także***

*Pani dr IVANIE STANIMIROVEJ-DASZYKOWSKIEJ  
za życzliwość, zaangażowanie, pomoc w badaniach i poświęcony czas.  
Pani prof. dr hab. BEACIE WALCZAK  
za merytoryczne uwagi i wskazówki udzielone podczas realizacji badań.  
Pani dr KATARZYNIĘ PYTLAKOWSKIEJ  
za życzliwość, pomoc w badaniach i poświęcony czas.  
Panu dr hab. MIECZYŚLAWOWI SAJEWICZOWI oraz Panu dr SŁAWOMIROWI MAŚLANCE  
za pomoc w realizacji projektów badawczych.*

### ***Dziękuję również za owocną współpracę i wszelkie cenne wskazówki***

*Panu dr IRENEUSZOWI GRABOWSKIEMU z Izby Celnej w Białej Podlaskiej  
Pani dr MARIOLI BARTOSZEK oraz Pani dr JUSTYNIE POLAK z Uniwersytetu Śląskiego  
Pani dr DALENEE DE BEER oraz Pani dr ELIZABETH JOUBERT z Uniwersytetu w Sellenbosch  
Pani prof. dr hab. BOGUSŁAWIE CZARNIK-MATUSEWICZ z Uniwersytetu Wrocławskiego*

### ***Swoje podziękowania kieruję również do***

*NARODOWEGO CENTRUM NAUKI oraz UNII EUROPEJSKIEJ  
za wsparcie finansowe w trakcie realizacji badań  
oraz Firmy GLOKOR Sp. z o.o. za współpracę w ramach projektu DoktoRIS*

### ***Szczególnie dziękuję***

*KOCHANYM RODZICOM, BABCIOM I DZIADKOWI, którzy zawsze wspierali mnie  
w moich wyborach, bez których nie miałabym możliwości realizacji swoich marzeń.  
WUJKOWI WIESŁAWOWI mojemu największemu fanowi.  
KARINIE I MARCIE bez których początek, środek i koniec tej pięknej podróży miałby  
zupełnie inny kształt.*

*Asia*

## Streszczenie

Ocena jakości produktów, która zależy od wielu parametrów (w tym składu chemicznego) z definicji wymaga podejścia globalnego. W większości przypadków nie jest to możliwe poprzez kontrolę wyłącznie jednego czy dwóch parametrów. Z tego względu produkty opisuje się przez sygnał analityczny (potencjalnie bogaty w informację chemiczną). Sygnał taki będzie złożony, gdyż zawiera wkłady pochodzące od indywidualnych komponentów badanych produktów. Często do opisu produktów używa się **sygnałów nieselektywnych** (np. chromatogramy czy widma spektroskopowe). Zasób informacji w nich zawartej pozwala rozpatrywać je jako tzw. **chemiczne odciski palca**, które oddają unikalność badanych próbek. Ze względu na złożony skład chemiczny często obserwuje się interferencje sygnałów pochodzących od poszczególnych komponentów. **Techniki chemometryczne** umożliwiają izolację użytecznej informacji, która z tego powodu nie jest bezpośrednio dostępna. Zalety chemicznych odcisków palca powodują, że są one coraz częściej wykorzystywane do oceny jakości produktów, zwłaszcza w kontekście badań próbek spożywczych, analityki procesowej czy monitoringu środowiska.

Celem zrealizowanych badań było **efektywne wykorzystanie informacji** zawartej w nieselektywnych sygnałach analitycznych do rozwiązywania problemów szeroko pojętej chemii analitycznej. Opracowano szereg rozwiązań analitycznych służących ocenie jakości: cukru (ze względu na zabarwienie i zawartość popiołu), oleju napędowego (ze względu na normy prawne dotyczące wysokości podatku akcyzowego) oraz produktów spożywczych takich jak kawa, mięta pieprzowa, bazylia, oregano, pasta pomidorowa i herbata typu rooibos (ze względu na całkowitą zawartość antyoksydantów i polifenoli). Produkty scharakteryzowano poprzez nieselektywne sygnały analityczne takie jak: fluorescencyjne widma synchroniczne – fluorescencyjne obrazy, widma rejestrowane w zakresie podczerwieni oraz chromatogramy zarejestrowane z wykorzystaniem techniki wysokosprawnej chromatografii ciekowej sprzężonej z detektorem DAD (HPLC-DAD). Dla każdego celu badawczego wybrano niezbędne techniki chemometryczne, które umożliwiły opracowanie funkcjonalnych metodyk analitycznych i sformułowanie konstruktywnych wniosków. Metodyki uzupełniono dokonując walidacji podstawowych parametrów charakteryzujących procedury analityczne.

Proponowane rozwiązania eliminują konieczność wykonywania kosztownych i czasochłonnych analiz dzięki zastosowaniu sygnałów nieselektywnych. Wszystkie eksperymenty projektowane były tak, aby ograniczyć produkcję szkodliwych odpadów chemicznych – zgodnie z zasadami zielonej chemii. Sygnały nieselektywne, tam gdzie to możliwe, rejestrowano dla próbek bez wstępnego przygotowania do analizy lub używając wody destylowanej jako jedyne rozpuszczalnika. Wysoka efektywność proponowanych rozwiązań sprawia, że mogą one być rekomendowane do zastosowań praktycznych w laboratoryjnych badaniach rutynowych oraz typu on-line, do kontroli jakości produktów. Taka adaptacja wymaga jednak przeprowadzenia dodatkowych badań, zwłaszcza z uwzględnieniem możliwych zmian składu chemicznego próbek.

## Summary

Quality of products depends on many parameters including their unique chemical composition. By definition, the global approach for the quality evaluation is necessary. It often requires simultaneous control of more than one parameter. Thus, the evaluated product is described by instrumental signal(s) possibly reach in chemical information (containing the inputs of several chemical components included in the product). The nonselective signals (e.g. chromatograms or spectra) are used frequently. They are considered as chemical fingerprints since the information they contain is unique. Nevertheless, the complex chemical composition of products leads to overlaps of particular signals characterizing different components. Isolation of significant information from such complex data facilitates the application of chemometric approaches. Considering their advantages the chemical fingerprints are used to evaluate quality of food products, in process analytical technology, and to monitor the environment.

The aim of the research was to extract the relevant information from the nonselective signals in order to develop novel analytical approaches for the quality evaluation. Different approaches aiming at evaluation of the quality of sugar (in course of its color and ash content), diesel oil (in course of the level of the imposed excise duty) and food products as coffee, pepper mint, oregano, basil, tomato paste, and rooibos tea (in course of their total antioxidant capacity and total polyphenol content) are developed. The nonselective signals like excitation-emission fluorescence matrices, high performance liquid chromatograms detected by diode array detector (HPLC-DAD) and infrared spectra were registered for examined products. Obtained signals were efficiently modelled using selected chemometric methods. Moreover, the developed new chemical approaches have been validated according to their chemometric and analytical parameters.

The application of the nonselective signals in the developed approaches significantly reduces or fully eliminates the use of classic, laborious, time and money consuming analytical methods. All of performed experiments were designed in accordance with the green chemistry principles, i.e. to limit the amount of harmful chemical wastes. Thus, the nonselective signals were collected for raw samples (without any laboratory preparation) or with the use of distilled water as the only solvent.

Regarding to the high efficiency of the developed methodologies they can be recommended for the routine laboratory or on-line quality evaluation. It should be emphasized, the presented results refer to the preliminary studies and the application must be preceded by the additional experiments (taking into account the various changes of chemical composition of samples).

---

## Spis treści

Spis treści .....	5
Wykaz skrótów stosowanych w pracy .....	7
Notacja matematyczna stosowana w pracy .....	8
1. Wstęp .....	9
2. Nieselektywne sygnały analityczne i ich przykłady .....	11
2.1. Chromatogramy .....	11
2.2. Widma spektroskopowe .....	12
2.2.1. Fluorescencyjne obrazy .....	12
2.2.2. Widma rejestrowane w zakresie podczerwieni .....	14
3. Przygotowanie nieselektywnych sygnałów do analizy .....	16
3.1. Fluorescencyjne obrazy .....	18
3.2. Widma rejestrowane w zakresie podczerwieni .....	21
3.3. Chromatogramy .....	22
3.4. Zestawienie sygnałów analitycznych .....	22
4. Modelowanie sygnałów nieselektywnych .....	25
4.1. Metody modelowania danych bez nadzoru .....	25
4.1.1. Analiza czynników głównych .....	25
4.1.2. Dekompozycja sygnałów z użyciem metody współbieżnej analizy czynnikowej .....	28
4.2. Metody modelowania danych z nadzorem .....	29
4.2.1. Kalibracja wybranej własności .....	29
4.2.2. Klasyfikacja/dyskryminacja próbek .....	39
4.3. Algorytmy obliczeniowe użyte w pracy .....	43
5. Nieselektywne sygnały i podejścia chemometryczne, a kontrola jakości .....	44
5.1. Kontrola jakości produktów spożywczych .....	44
5.2. Ocena jakości środowiska naturalnego .....	46
5.3. Analiza jakości produktów energetycznych .....	46
5.4. Dziedziny nauk typu „omika” .....	47
5.5. Ocena jakości w badaniach farmakologicznych .....	47
6. Cel pracy .....	49
7. Opublikowane badania własne .....	50

---

7.1. Publikacja I .....	51
<i>Kontrola jakości cukru na podstawie fluorescencyjnych odcisków palca i stabilnej kalibracji</i>	
7.2. Publikacja II .....	62
<i>Metoda równoczesnego oznaczania Solvent Yellow 124 i Solvent Red 19 w oleju napędowym z wykorzystaniem spektroskopii fluorescencyjnej i chemometrii</i>	
7.3. Publikacja III .....	75
<i>Identyfikacja nielegalnego procederu usuwania komponentów fiskalnych z oleju napędowego o obniżonej aktywności</i>	
7.4. Publikacja IV .....	86
<i>Szybka metoda oceny zdolności antyoksydacyjnej żywności używając fluorescencyjnych obrazów na przykładzie ekstraktów kawy i mięty pieprzowej</i>	
7.5. Publikacja V .....	96
<i>Metoda przewidywania zawartości hydrofilowych antyoksydantów zawartych w paście pomidorowej na podstawie widm z zakresu podczerwieni oraz fluorescencyjnych obrazów zarejestrowanych dla próbek surowych i ekstraktów</i>	
7.6. Publikacja VI.....	107
<i>Modelowanie całkowitej zdolności antyoksydacyjnej naparów herbaty typu rooibos (Aspalathus linearis) używając chromatograficznych odcisków palca i identyfikacja potencjalnych substancji o właściwościach antyoksydacyjnych</i>	
8. Badania własne nieopublikowane .....	126
<i>Efektywność oceny całkowitej zdolności antyoksydacyjnej na podstawie fluorescencyjnych obrazów dla próbek bazylii, oregano oraz suplementów diety</i>	
9. Podsumowanie i wnioski .....	129
10. Finansowanie .....	133
11. Literatura .....	134
12. Curriculum Vitae .....	141

---

## Wykaz skrótów stosowanych w pracy

Skrót	Nazwa polska	Nazwa angielska
<sup>1</sup> HNMR	protonowy rezonans magnetyczny	<i>hydrogen – 1 nuclear magnetic resonance</i>
MS	spektrometria mas	<i>mass spectrometry</i>
IR	podczerwień	<i>infrared</i>
NIR	bliska podczerwień	<i>near infrared</i>
MIR	właściwa podczerwień	<i>middle infrared</i>
ATR	osłabione całkowite odbicie	<i>attenuated total reflectance</i>
S/N	stosunek sygnału do szumu	<i>signal to noise ratio</i>
PCA	analiza czynników głównych	<i>principal component analysis</i>
P-ALS	metoda asymetrycznych najmniejszych kwadratów z funkcją kary	<i>penalized asymmetric least squares</i>
MCR-ALS	wieloparametrowa dekonwolucja z wykorzystaniem iteracyjnej metody najmniejszych kwadratów	<i>multivariate curve resolution alternating least squares</i>
PARAFAC	współbieżna analiza czynnikowa	<i>parallel factor analysis</i>
COW	metoda zoptymalizowanego nakładania sygnałów maksymalizująca ich wzajemną korelację	<i>correlation optimized warping</i>
PLS	regresja częściowych najmniejszych kwadratów	<i>partial least squares regression</i>
N-PLS	N-modalna regresja częściowych najmniejszych kwadratów	<i>N-way partial least squares regression</i>
MLR	metoda regresji wielorakiej	<i>multiple linear regression</i>
PC	czynnik główny	<i>principal component</i>
PCR	regresja czynników głównych	<i>principal component regression</i>
RMSE	średni błąd kwadratowy	<i>root mean square error</i>
RMSEP	średni błąd kwadratowy przewidywania	<i>root mean square error of prediction</i>
RMSCV	średni błąd kwadratowy walidacji krzyżowej	<i>root mean square error of cross-validation</i>
PRM	stabilny wariant regresji częściowych najmniejszych kwadratów	<i>partial robust M-regression</i>
UVE-PLS	regresja częściowych najmniejszych kwadratów z eliminacją zmiennych nieistotnych	<i>uninformative variable elimination partial least squares</i>
LDA	liniowa analiza dyskryminacyjna	<i>linear discriminant analysis</i>
D-PLS	dyskryminacyjna metoda regresji częściowych najmniejszych kwadratów	<i>discriminant partial least squares regression</i>
CART	drzewa klasyfikacji i regresji	<i>classification and regression trees</i>
SIMCA	metoda modelowania indywidualnych grup próbek	<i>soft independent modelling of class analogies</i>
ORAC	zdolność pochłaniania rodników tlenowych	<i>oxygen radical absorbance capacity</i>
TAC	całkowita zdolność antyoksydacyjna	<i>total antioxidant capacity</i>



---

## Notacja matematyczna stosowana w pracy

$x$	skalar
$\mathbf{x}$	wektor
$\mathbf{x}_{[m,1]}$	wektor o wymiarach $m \times 1$
$\mathbf{X}$	macierz
$\mathbf{X}_{[m,n]}$	macierz o wymiarach $m \times n$
$\underline{\mathbf{X}}$	tensor
$\mathbf{X}_{[m,n,o]}$	tensor o wymiarach $m \times n \times o$
$ \otimes $	iloczyn Khatri-Rao
$\ \bullet\ $	norma euklidesowa wektora
$ \bullet $	wartość bezwzględna
$\mathbf{x}^T$	operacja transpozycji wektora $\mathbf{x}$

## 1. Wstęp

Ocena właściwości fizycznych i chemicznych próbek o złożonym składzie chemicznym jest niewątpliwie jednym z aktualnych wyzwań analizy chemicznej. Zagadnienie to obejmuje oznaczenia jakościowe i ilościowe wybranych składników chemicznych, jak również ocenę właściwości wynikających z interakcji różnych komponentów zawartych w próbce. Parametry charakteryzujące próbkę, zależące od jej ogólnego składu chemicznego to np. smak, zapach, całkowita zdolność antyoksydacyjna, konsystencja czy wartość energetyczna. Niektóre z tych parametrów są wyznacznikami pewnej cechy próbki jaką jest jej jakość. Ponieważ jakość często zależy od wielu parametrów, a także od ogólnego składu chemicznego produktu, jej ocena wymaga podejścia globalnego. Założenie, że będzie to możliwe poprzez kontrolę tylko jednego czy dwóch parametrów jest często błędne. Z tego powodu próbki zazwyczaj opisuje się przez sygnał analityczny, który jest potencjalnie bogatym nośnikiem informacji chemicznej. Taki sygnał będzie złożony, ponieważ zawiera wkłady pochodzące od wielu indywidualnych komponentów próbki. Sygnały selektywne, w kontekście badania próbek złożonych, mogą być potencjalnie rejestrowane tylko z użyciem technik rozdzielczych np. chromatografii, umożliwiających izolowanie poszczególnych składników próbki i ich ilościowe oznaczenie. Uzyskanie sygnału selektywnego w dużym stopniu zależy od liczby składników próbki i parametrów pomiaru (np. rozdział chromatograficzny musi być optymalny – to wymaga odpowiedniego doboru kolumny, składu fazy ruchomej, warunków rozdziału itp.). Dlatego też, całkowity rozdział składników próbek naturalnych takich jak ropa naftowa, żywność czy próbki biologiczne jest często bardzo trudny, a nawet niemożliwy do osiągnięcia. Efektywność rozdziału chromatograficznego, a przez to jakość rejestrowanego sygnału, można poprawić m.in. poprzez:

- stosowanie określonych procedur laboratoryjnych (np. poprzez wstępne oczyszczenie, ekstrakcję czy zagęszczanie próbki),
- zwiększenie zdolności rozdzielczej układu chromatograficznego (np. poprzez stosowanie ortogonalnych systemów chromatograficznych, dodatkowego rozdziału, technik sprzężonych czy zaawansowanej detekcji),
- wykorzystanie technik chemometrycznych w celu izolacji użytecznej informacji, która jest bezpośrednio niedostępna z powodu interferencji sygnałów pochodzących od wszystkich komponentów mieszaniny.

Stale rosnąca popularność metod chemometrycznych, wspomagających interpretację, eksplorację, klasyfikację i kalibrację złożonych sygnałów analitycznych przyczynia się także do coraz częstszego wykorzystywania nieselektywnych sygnałów analitycznych do oceny wybranych parametrów (w tym jakości) wybranych produktów. Sygnał nieselektywny równocześnie definiuje wiele składników chemicznych próbki, pośród których znajdują się te, które bezpośrednio wpływają na kontrolowany parametr. Zasięg informacji zawartej w nieselektywnych sygnałach pozwala rozpatrywać je jako tzw. **chemiczne odciski palca** badanych próbek, charakteryzujące ich unikalność. Dobór odpowiednich metod chemometrycznych umożliwia modelowanie parametrów i właściwości objaśnianych przez

różne regiony sygnału analitycznego. Dzięki takiej możliwości sygnały nieselektywne coraz częściej są używane w badaniach próbek spożywczych, kontroli stanu środowiska naturalnego czy na potrzeby analityki procesowej.

W niniejszej pracy doktorskiej omawiam możliwości zastosowania różnych sygnałów nieselektywnych do skutecznej kontroli wybranych parametrów opisujących jakość wybranych produktów.

Charakterystykę nieselektywnych sygnałów instrumentalnych, a w szczególności, tych które użyłam w moich badaniach przedstawiam w rozdziale drugim. Metody chemometryczne, które użyłam w pracy, dostosowane do analizy wybranych chemicznych odcisków palca i wydobycia informacji o jakości badanych próbek opisałam w rozdziałach trzecim i czwartym. Rozdziały te wzbogaciłam przykładami zastosowań poszczególnych sygnałów lub metod w przeprowadzonych badaniach.

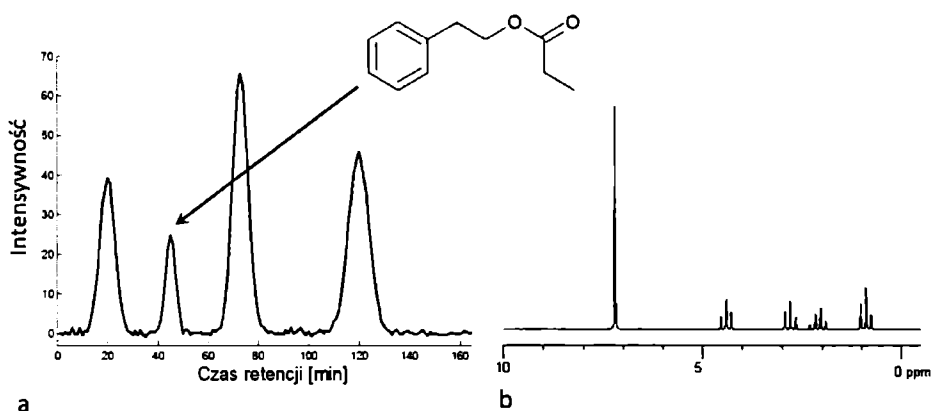
Przykłady zastosowań zaznaczyłam pionową linią znajdującą się po lewej stronie fragmentów pracy, w których je przytaczam.

Następnie, w rozdziale piątym omawiam pięć dziedzin, w których jakość jest definiowana różnorodnie, a do jej oceny wykorzystywane są sygnały nieselektywne i podejścia chemometryczne. Badania własne, będące ilustracją możliwości wykorzystania sygnałów nieselektywnych do oceny jakości różnych produktów przedstawiam w sześciu artykułach naukowych składających się na cykl publikacyjny stanowiący podstawę niniejszej pracy oraz w rozdziale pt. *Badania własne nieopublikowane*. Pracę zamykają rozdziały zawierające podsumowanie i wnioski oraz pozycje literaturowe, na które powołuję się w prowadzonej dyskusji. Dołączam również listę moich osiągnięć naukowych oraz podziękowania za uzyskane wsparcie finansowe badań zrealizowanych w ramach pracy doktorskiej.

## 2. Nieselektywne sygnały analityczne i ich przykłady

Każdy sygnał instrumentalny, który zawiera informacje o chemicznych komponentach próbki i równocześnie nie pozwala na bezpośrednie oraz jednoznaczne oznaczenie jakościowe bądź ilościowe wybranego składnika lub też określonej właściwości analizowanej mieszaniny chemicznej można nazwać nieselektywnym sygnałem analitycznym.

Techniki instrumentalne, przy pomocy których rejestrowane są sygnały nieselektywne dla próbek pochodzenia naturalnego to między innymi spektroskopie w obszarze bliskiej podczerwieni i protonowego rezonansu magnetycznego (ang. *hydrogen-1 nuclear magnetic resonance*,  $^1\text{H NMR}$ ) lub techniki chromatograficzne. Nie podlega wątpliwości, że sygnał analityczny rejestrowany dla tej samej substancji przy użyciu dwóch różnych technik będzie się różnił. Przykładem takiej sytuacji są chromatogram i widmo  $^1\text{H NMR}$  zarejestrowane dla tej samej substancji. Na chromatogramie będzie ona reprezentowana przez jeden pik, a na widmie spektroskopowym będziemy obserwować zbiór pików charakterystyczny dla struktury chemicznej tej substancji (rysunek 1).



**Rysunek 1** a) przykładowy chromatogram mieszaniny pestycydów, na którym drugi pik odpowiada propionianowi 2-fenyletylu oraz b) zbiór pików tworzących widmo  $^1\text{H NMR}$  reprezentujące ten sam związek chemiczny [1].

### 2.1. Chromatogramy

Rozdział chromatograficzny to jedna z operacji analitycznych, wykonywana najczęściej w celu ustalenia składu chemicznego próbek złożonych. Skuteczny rozdział komponentów chemicznych zawartych w badanej próbce jest możliwy dzięki zastosowaniu układów chromatograficznych charakteryzujących się różnymi właściwościami retencyjnymi. Pełniejszą informację na temat składu próbki, wstępnie rozdzielonej chromatograficznie, można uzyskać stosując różne typy zaawansowanych detektorów, np. detektor z matrycą diodową (ang. *diode array detector*, DAD) lub detektor mas (ang. *mass spectrometry detector*, MS). Jednakże, identyfikacja poszczególnych substancji wykrytych podczas

rozdziłu chromatograficznego jest zadaniem często skomplikowanym i wymaga stosowania kosztownych substancji wzorcowych. W przypadku próbek o złożonym składzie chemicznym uzyskanie skutecznego rozdziłu chromatograficznego jest trudne lub wręcz niemożliwe. Pomimo braku optymalnego rozdziłu, chromatograficzne odciski palca próbek są efektywnie wykorzystywane do konstrukcji chemometrycznych modeli, które wspomagają kontrolę jakości. Dzięki odpowiedniemu doborowi metod przygotowania i modelowania chromatogramów jest to możliwe nawet wtedy, gdy rozdział chromatograficzny nie jest optymalny i/lub gdy poszczególne piki chromatograficzne nie zostały zidentyfikowane.

## 2.2. Widma spektroskopowe

Widma spektroskopowe charakteryzują interakcje pomiędzy badaną próbką, a promieniowaniem elektromagnetycznym. Informacje zawarte w widmie spektroskopowym mogą odzwierciedlać właściwości absorpcyjne i/lub emisyjne badanej materii. W zależności od przyjętej metodyki pomiaru badana jest interakcja składników próbki z promieniowaniem elektromagnetycznym o:

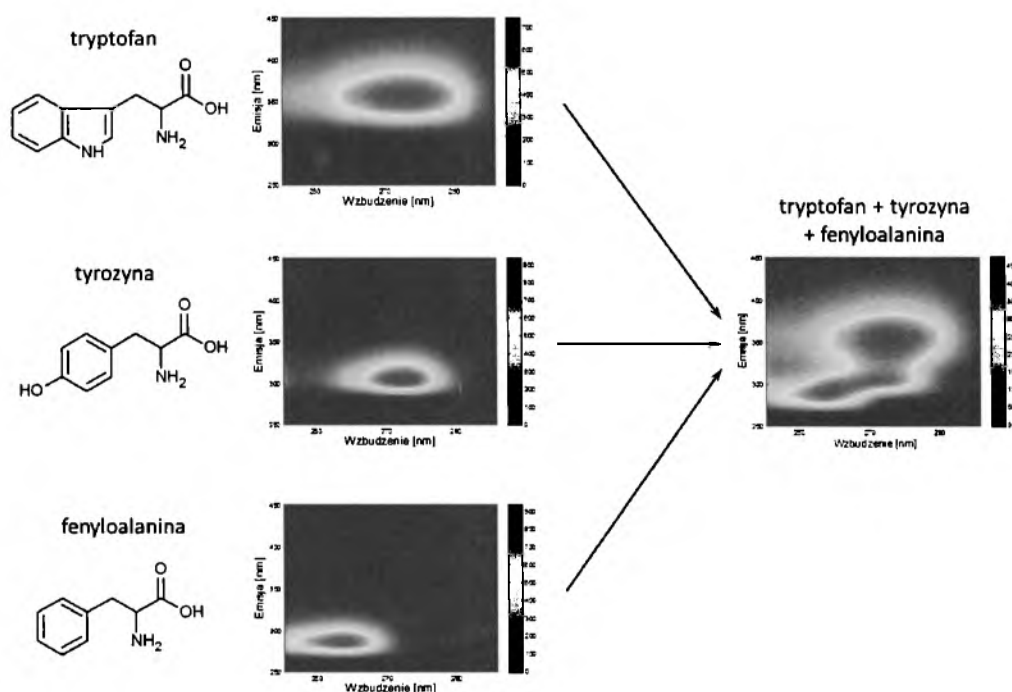
- określonej energii (jedna wybrana długość fali elektromagnetycznej) w zdefiniowanym zakresie pomiarowym, np. widma absorpcyjne UV-Vis lub w zakresie podczerwieni (ang. *infra red*, IR) bliskiej (ang. *near infrared*, NIR) czy właściwej (ang. *middle infrared*, MIR), a także fluorescencyjne widma emisyjne;
- różnej energii (wybrany zakres długości fal elektromagnetycznych) w zdefiniowanym zakresie pomiarowym, np. fluorescencyjne obrazy próbek.

Bez względu na to, czy badana jest interakcja pomiędzy składnikami próbki, z jedną czy wieloma długościami fali elektromagnetycznej, rejestrowane widma mogą mieć charakter nieselektywny. Widma rejestrowane dla szerokiego zakresu pomiarowego z reguły zawierają więcej chemicznej informacji, a zatem sygnał jest bardziej charakterystyczny. Przykładem nieselektywnych sygnałów spektroskopowych o bogatym zasobie informacji są fluorescencyjne obrazy lub widma absorpcyjne rejestrowane w zakresie podczerwieni.

### 2.2.1. Fluorescencyjne obrazy

Fluorescencyjne obrazy nazywane również fluorescencyjnymi widmami synchronicznymi (ang. *excitation-emission fluorescence spectroscopy matrices* lub *fluorescence landscapes*), to rodzaj pomiarów instrumentalnych, których wynikiem są dwuwymiarowe sygnały analityczne. W trakcie pomiaru próbka jest wzbudzana promieniowaniem elektromagnetycznym o różnej długości fali z zakresu UV-Vis. Dla pojedynczej fali wzbudzającej rejestrowane jest widmo emisyjne w wybranym zakresie spektralnym z przedziału UV-Vis. Dzięki takiemu sposobowi rejestracji sygnału analitycznego uzyskuje się informacje zarówno o właściwościach absorpcyjnych jak i emisyjnych komponentów zawartych w badanej próbce. Na rysunku 2 przedstawiono wybrane przykłady

fluorescencyjnych obrazów zarejestrowanych dla układów chemicznych zawierających aminokwasy. Dwuwymiarowe widmo mieszaniny trzech aminokwasów umożliwia rejestrację sygnałów wszystkich składników próbki podczas jednego pomiaru. Stanowi to ważną zaletę fluorescencyjnych obrazów, ze względu na pominięcie etapu zmiany ustawień przyrządu pomiarowego lub położenia próbki względem układu optycznego instrumentu, dokonywanego przez analityka, a tym samym polepszenie jakości sygnałów dzięki jednoczesnej rejestracji widm emisyjnych. Konsekwencją takiego sposobu rejestracji fluorescencyjnych obrazów jest również ich większy zasób informacji w porównaniu z widmami fluorescencyjnymi rejestrowanymi dla pojedynczej fali wzbudzającej.



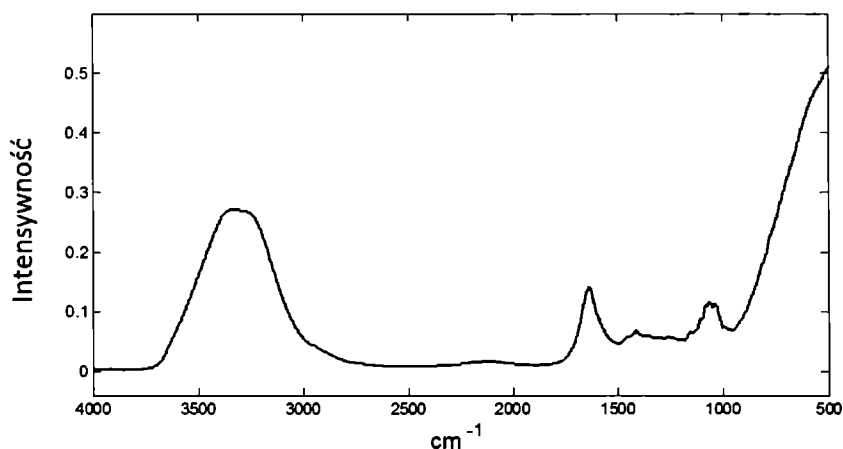
**Rysunek 2** Fluorescencyjne obrazy aminokwasów: fenyloalaniny, tyrozyny i tryptofanu oraz ich równomolowej mieszaniny, zarejestrowane dla tego samego zakresu fal wzbudzenia (240 nm – 300 nm,  $\Delta = 1$  nm) i emisji (250 nm – 450 nm,  $\Delta = 1$  nm) z zastosowaniem układu pomiarowego o kącie prostym pomiędzy źródłem wzbudzenia i detektorem [2].

Kolejne zalety fluorescencyjnych obrazów wynikają z charakterystyki techniki analitycznej jaką jest spektroskopia fluorescencyjna. Umożliwia ona prowadzenie pomiarów uzyskując wysoką czułość bez niszczenia badanego materiału, co przekłada się na ich wysoką użyteczność. Do przykładowych zastosowań fluorescencyjnych obrazów możemy zaliczyć ilościowe oznaczenie zawartość tokoferoli w olejach [3], ustalenie pochodzenia geograficznego mleka [4], czy badania mieszanin oleju napędowego i nafty [5].

### 2.2.2. Widma rejestrowane w zakresie podczerwieni

Widma absorpcyjne rejestrowane dla promieniowania elektromagnetycznego z zakresu podczerwieni  $14000 - 10 \text{ cm}^{-1}$  (około  $740 - 30000 \text{ nm}$ ) zawierają piki reprezentujące charakterystyczne ugrupowania atomów analizowanych substancji. Atomy węgla połączone wiązaniami pojedynczymi z atomami wodoru w zależności od otoczenia chemicznego wykazują zdolność do absorpcji promieniowania o różnej energii. Podobnie ugrupowania zawierające azot czy grupy hydroksylowe, mogą być zidentyfikowane na podstawie widm rejestrowanych w omawianym zakresie promieniowania elektromagnetycznego. W kontekście analizy próbek złożonych, zawierających różne związki organiczne, sygnały IR mają jednak charakter nieselektywny, którego przyczyną jest występowanie tych samych ugrupowań w różnych związkach chemicznych zawartych w analizowanym materiale. Brak możliwości identyfikacji jakościowej czy ilościowej poszczególnych związków chemicznych nie utrudnia oceny właściwości próbki o charakterze globalnym. Dodatkowo, rejestracja widma jest szybka i nie powoduje zniszczenia badanego materiału, co przekłada się na dużą popularność zastosowań spektroskopii w zakresie podczerwieni.

Widma w zakresie średniej podczerwieni wykorzystałam do oceny całkowitej zdolności antyoksydacyjnej oraz zawartości związków fenolowych w pastach pomidorowych. Badania opisałam w **publikacji V**. Na rysunku 3 zaprezentowane jest przykładowe widmo jednej z badanych past pomidorowych, która została wyprodukowana w Polsce.



**Rysunek 3** Widmo IR zarejestrowane dla pasty pomidorowej w zakresie  $500 - 4000 \text{ cm}^{-1}$  ( $\Delta = 4 \text{ cm}^{-1}$ ) przy użyciu spektrometru Nicolet iS50 wyposażonego w przystawkę ATR (ang. *attenuated total reflectance*) – spektroskopia osłabionego całkowitego odbicia z jednoodbiciowym kryształem diamentowym. Widmo zarejestrowano dla próbki bezpośrednio po wyjęciu z opakowania, w którym została zakupiona (bez przygotowania próbki do analizy).

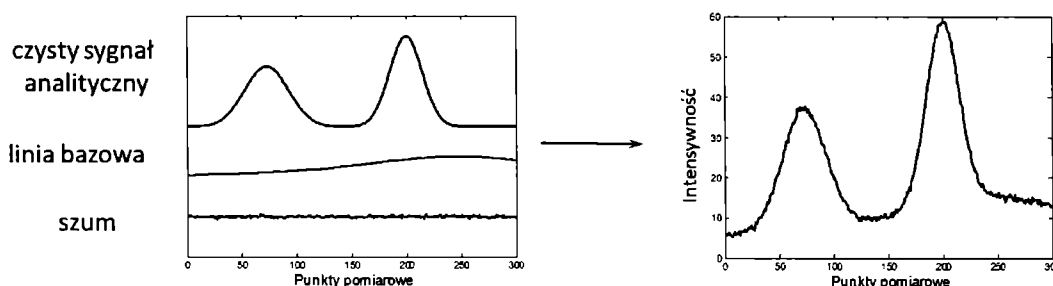
Większość z sygnałów rejestrowanych przez aktualnie wykorzystywane techniki spektroskopowe to sygnały nieselektywne. Jest to efekt m.in. ograniczania pracy laboratoryjnej, a co za tym idzie, wstępnego przygotowania próbki przed analizą do niezbędnego minimum. Nie mniej jednak, nieselektywne widma rejestrowane przy użyciu spektroskopii fluorescencyjnej, podczerwieni (w jej zakresie bliskim lub właściwym) czy spektrometrii mas, widma NMR oraz chromatogramy, są chemicznymi odciskami palca próbek, dzięki czemu często znajdują zastosowanie w szeroko pojętej kontroli jakości.



### 3. Przygotowanie nieselektywnych sygnałów do analizy

Wszystkie sygnały analityczne, oprócz informacji o składnikach badanej próbki, zawierają nieistotne chemiczne elementy – szum i linia bazowa – wynikające z niedoskonałości używanych przyrządów pomiarowych czy samej metody pomiarowej. Dodatkowo, każda technika pomiarowa może być obciążona charakterystycznymi dla niej komponentami sygnału, które nie są chemicznie istotne i mogą negatywnie wpływać na interpretację chemicznej informacji. Dlatego ważnym etapem warunkującym właściwe wykorzystanie sygnałów instrumentalnych jest ich wstępne przygotowanie do dalszej analizy.

Każdy sygnał instrumentalny składa się z trzech komponentów: szumu, linii bazowej i czystego sygnału analitycznego, czyli informacji o chemicznych składnikach próbki [6] (rysunek 4).



**Rysunek 4** Elementy składowe sygnału instrumentalnego tj. szum, linia bazowa i czysty sygnał analityczny na przykładzie chromatogramu dwuskładnikowej mieszaniny.

Szum i linia bazowa mogą istotnie zniekształcać piki poszczególnych składników próbki. Wówczas jakościowa i ilościowa analiza jest znacznie utrudniona, a czasem wręcz niemożliwa do przeprowadzenia. Często usunięcie linii bazowej oraz szumu przed dalszą analizą jest konieczne.

Źródłem szumu, komponentu sygnału o najwyższej częstotliwości, jest ograniczona czułość używanego detektora oraz różne zjawiska i oddziaływania zachodzące w trakcie dokonywanego pomiaru. Definiuje się go jako odchylenie standardowe sygnału rejestrowanego przez przyrząd pomiarowy [7].

Poziom szumu w wyjściowym sygnale określa parametr nazywany stosunkiem sygnału do szumu (ang. *signal to noise ratio*, S/N) wyrażany jako stosunek wartości średniej z sygnału do jego odchylenia standardowego. Wyróżnia się kilka rodzajów szumu. Szum biały, proporcjonalny do sygnału czy skorelowany. Niezależnie od rodzaju szumu należy zawsze sprawdzić wartość parametru S/N, aby ocenić konieczność jego usunięcia z wyjściowego sygnału. Jest to niezbędne, ponieważ szum może zafałszować prawdziwe wyniki analizy, pomimo iż jest komponentem niezależnym od chemicznego składu badanych

próbek. Szum można eliminować stosując, np. filtr skonstruowany z użyciem mediany, lub średniej ważonej albo kompresując dane używając analizy czynników głównych (ang. *principal component analysis*, PCA) [8], lub transformacji Fouriera czy falkowej [9].

Składnikiem sygnałów, który charakteryzuje się najmniejszą częstotliwością, jest linia bazowa. Jest to odpowiedź detektora obserwowana podczas nieobecności analitów w próbce. Linia bazowa może ulegać zmianom w czasie, dlatego jej kształt bywa różny dla sygnałów rejestrowanych przy użyciu tego samego detektora. Przed dokonaniem analizy chemometrycznej linia bazowa powinna być zidentyfikowana i usunięta z sygnału, jeżeli może negatywnie wpłynąć na interpretację informacji zawartych w sygnałach. Korekcję linii bazowej najczęściej przeprowadza się wykorzystując nieliniowe filtry lub metodę asymetrycznych najmniejszych kwadratów z funkcją kary (ang. *penalized asymmetric least squares*, P-ALS) [10].

Usunięcie linii bazowej było niezbędne podczas analizy sygnałów chromatograficznych opisanej w **publikacji VI**. Wykorzystano metodę P-ALS (pierwsza pochodna sygnału została użyta do oceny jego gładkości, wartość parametru kary –  $\lambda$  (ang. *penalty parameter*) wynosiła 10000). W **publikacji V** linię bazową sygnałów IR korygowano dla każdego sygnału oddzielnie – przed dokonaniem pomiaru widma próbki za każdym razem rejestrowano widmo tła. Sygnały fluorescencyjne użyte w **publikacjach I – V** nie wymagały korekcji linii bazowej.

Najważniejszy z punktu widzenia analizy chemicznej komponent sygnału instrumentalnego to informacja o składnikach próbki – czysty sygnał analityczny. W zależności od złożoności składu chemicznego badanej próbki w czystym sygnale pik/piki reprezentujące poszczególne składniki mogą być wyraźnie rozdzielone lub mogą nakładać się na siebie w różnym stopniu, co utrudnia identyfikację poszczególnych komponentów oraz interpretację zarejestrowanej informacji.

Nakładanie się pików to efekt np. zbliżonych właściwości spektralnych lub powinowactwa do fazy chromatograficznej rozdzielanych substancji chemicznych tworzących mieszaninę. Efekt ten można wyeliminować dokonując dekonwolucji sygnału. Jest to operacja matematyczna przy pomocy której sygnał przedstawiany jest jako suma wkładów jego składników. Operacja dekonwolucji może zostać przeprowadzona poprzez estymację funkcji matematycznych opisujących piki (np. funkcji Gaussa), obliczenie pierwszej pochodnej lub pochodnej wyższego rzędu analizowanego sygnału. Można również użyć bardziej zaawansowanych metod chemometrycznych, jak np. wieloparametrowa dekonwolucja z wykorzystaniem iteracyjnej metody najmniejszych kwadratów (ang. *multivariate curve resolution alternating least squares*, MCR-ALS) [11], współbieżna analiza czynnikowa (ang. *parallel factor analysis*, PARAFAC) [12] lub transformacja sygnału do domeny częstotliwości [6], które umożliwiają estymację widm czystych składników mieszaniny. Dodatkową zaletą transformacji sygnału dokonanej metodami Fouriera czy falkową [9] jest możliwość estymacji i usunięcia wcześniej omawianych komponentów o najwyższej i najniższej częstotliwości – szumu i linii bazowej.

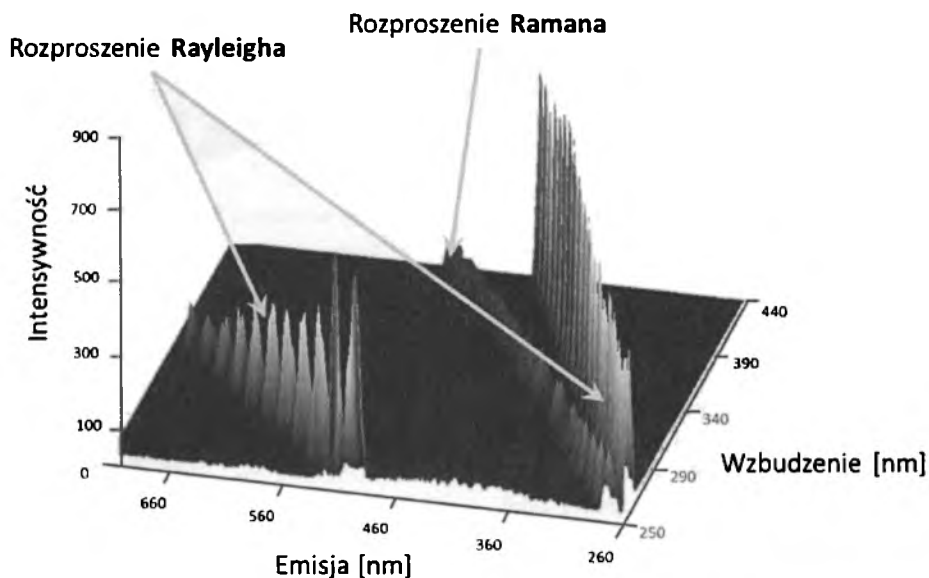
Omówione powyżej komponenty są typowe dla każdego sygnału instrumentalnego. Analiza porównawcza, modelowanie i końcowa interpretacja informacji zawartej w sygnałach, może być również zaburzona przez dodatkowe efekty występujące w sygnałach, które są charakterystyczne dla wybranej metody pomiarowej. Dlatego, przygotowanie danych do analizy powinno obejmować również ocenę wpływu i ewentualne usunięcie ze zbioru analizowanych sygnałów efektów specyficznych.

Badania omówione w niniejszej pracy dotyczą modelowania i analizy nieselektywnych sygnałów instrumentalnych takich jak fluorescencyjne obrazy, widma rejestrowane w zakresie podczerwieni oraz chromatogramy. Z tego względu, w dalszej części scharakteryzuję efekty specyficzne dla tych sygnałów.

### 3.1. Fluorescencyjne obrazy

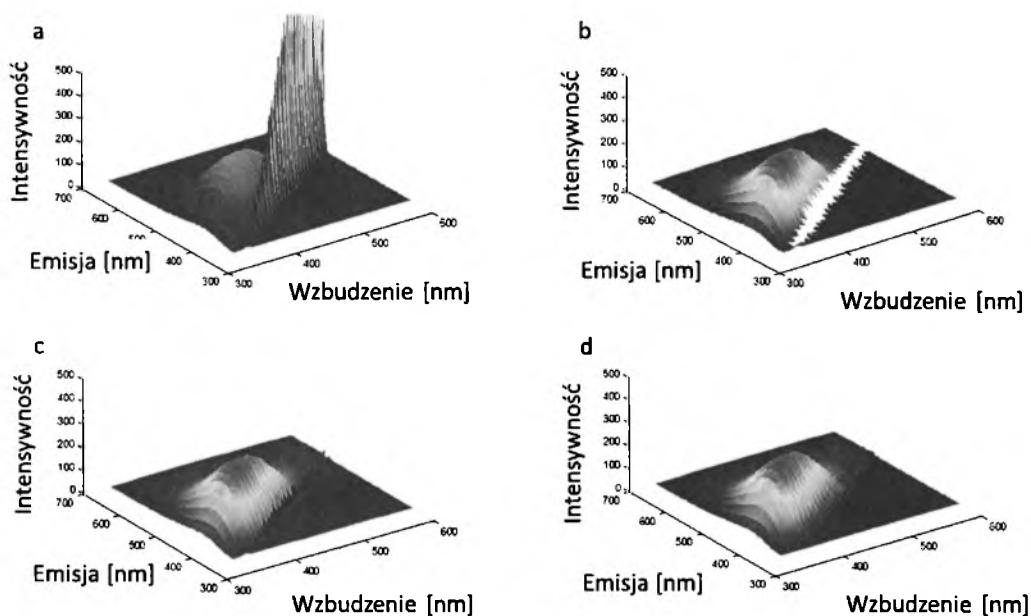
Wiązka promieniowania elektromagnetycznego wzbudzająca próbkę w trakcie rejestracji sygnałów fluorescencyjnych na skutek kontaktu z próbką ulega różnym zjawiskom fizycznym. Może zostać zaabsorbowana przez komponenty próbki (następstwem tego zjawiska jest często emisja promieniowania fluorescencyjnego, które rozchodzi się w każdym kierunku przestrzeni), przejść przez próbkę bez zmiany energii albo ulec rozproszeniu na cząsteczkach w niej zawartych.

W przypadku promieniowania przechodzącego przez próbkę, stosowana geometria układu pomiarowego (źródło promieniowania i detektor są ustawione względem siebie zazwyczaj pod kątem prostym) zapobiega jego rejestracji podczas pomiaru. Niestety, promieniowanie rozproszone w każdym kierunku przestrzeni dociera do detektora podobnie jak promieniowanie emisyjne. Jeżeli zjawisko rozproszenia ma charakter sprężysty na widmie obserwuje się piki rozproszenia Rayleigha. W przypadku rozproszenia niesprężystego są to piki rozproszenia Ramana. Bez względu na charakter zjawiska rozproszenia, na fluorescencyjnych obrazach specyficzne efekty reprezentują diagonalne linie pików [13] (rysunek 5). Są one nieistotne z chemicznego punktu widzenia, mogą interferować z sygnałami pochodzącymi od fluoroforów zawartych w próbce lub zaburzać konstrukcję modeli chemometrycznych. Z tych powodów, przed modelowaniem i interpretacją sygnałów powinny one zostać usunięte.



**Rysunek 5** Linie rozproszenia promieniowania elektromagnetycznego sprężystego (Rayleigha) i niesprężystego (Ramana), obserwowane na fluorescencyjnym obrazie zarejestrowanym dla próbki wody destylowanej w zakresie emisyjnym od 260 nm do 650 nm,  $\Delta = 2$  nm, dla fal wzbudzenia w zakresie od 250 nm do 440 nm,  $\Delta = 10$  nm.

W literaturze opisano różne sposoby eliminacji pików rozproszenia. Najprostszym z nich jest zastąpienie pików rozproszenia przez zera lub brakujące elementy [14,15]. Niemniej jednak takie podejście może powodować utratę istotnej informacji chemicznej, zwłaszcza gdy piki rozproszenia interferują z pikami pochodzącymi od różnych komponentów próbki. Dodatkowo, wprowadzenie do widma brakujących elementów znacznie ogranicza wachlarz metod użytecznych do modelowania tak przygotowanych danych. Aby tego uniknąć, po usunięciu wybranych fragmentów widma należy dokonać interpolacji naruszonego sygnału np. wykorzystując triangulację Delaunaya [16]. W tej metodzie wybrany obszar jest interpolowany poprzez konstrukcję sieci nieregularnych trójkątów, budowaną na bazie pozostałych w sygnale punktów pomiarowych. Sieć budowana jest tak, aby punkty z nienaruszonej części sygnału nie zostały do niej włączone, a wyłącznie stanowiły jej punkty brzegowe. Rysunek 6 przedstawia przykładowy fluorescencyjny obraz próbki piwa zawierający piki rozproszenia Rayleigha i ich eliminację z wykorzystaniem różnych metod.



**Rysunek 6** Fluorescencyjny obraz próbki piwa, zarejestrowany w zakresie emisyjnym od 320 nm do 700 nm,  $\Delta = 2$  nm, dla fal wzbudzenia w zakresie od 300 nm do 600 nm,  $\Delta = 10$  nm a) przed i po usunięciu pików rozproszenia promieniowania elektromagnetycznego różnymi metodami: b) wstawiając brakujące elementy, c) wstawiając zera oraz d) stosując triangulację Delaunaya.

Często piki rozproszenia charakteryzują się wysoką intensywnością w porównaniu z intensywnością pików charakteryzujących fluorofory zawarte w badanej próbce. Usunięcie fragmentu widma reprezentującego zjawisko rozproszenia sprawia, iż informacje o fluoroforach zawarte w widmie stają się wyraźniejsze i łatwiejsze do interpretacji.

Korekcja sygnałów fluorescencyjnych ze względu na piki rozproszenia Rayleigha i/lub Ramana była niezbędna podczas interpretacji chemicznej informacji zawartej w sygnałach wykorzystanych do weryfikacji hipotez badawczych postawionych w pięciu publikacjach z cyklu prezentowanego w niniejszej pracy.

W **publikacji I**, poruszającej zagadnienia oceny jakości cukru ze względu na zabarwienie i zawartość popiołu, w miejsce występowania pików rozproszenia Rayleigha wstawiłam zera, co umożliwiło efektywną kalibrację widm.

W **publikacjach II i III**, w których oceniałam jakość oleju napędowego ze względu na ilościową zawartość komponentów akcyzowych (**publikacja II**) oraz identyfikowałam nielegalny proceder jego odbarwiania (**publikacja III**), zastosowałam metodę korekcji sygnałów zaproponowaną przez Zeppa i współpracowników [16]. Przedstawiony w publikacji algorytm (zapisany w środowisku programowania Matlab) do korekcji pojedynczego dwuwymiarowego sygnału dostosowałam do pracy z sygnałami zestawionymi w trójwymiarowy tensor. Następnie wybrałam parametry szerokości i lokalizacji pików rozproszenia tak, aby zapewniały one optymalny sposób korekcji sygnałów, tzn. jak najmniejsze naruszenie pików fluoroforów zawartych w widmach. Efektywność działania tego algorytmu (optymalna interpolacja naruszonego sygnału oraz szybkość matematycznych

operacji dokonywanych na wielowymiarowych danych) sprawiła, że wykorzystałam go także w **publikacjach IV i V**. Sprawdził się zarówno dla sygnałów zarejestrowanych dla wodnych ekstraktów kawy i mięty pieprzowej, jak i fluorescencyjnych obrazów past pomidorowych.

Kolejnym obserwowanym efektem istotnym podczas pracy z sygnałami fluorescencyjnymi jest zjawisko wygaszania promieniowania emisyjnego (ang. *quenching*). Jest to osłabienie intensywności sygnału fluorescencyjnego na skutek oddziaływań zachodzących w próbce, zwłaszcza jeżeli charakteryzuje się ona wysokimi stężeniami analitów i/lub złożonym składem chemicznym. Powodem wygaszania może być interakcja promieniowania emisyjnego z fluoroforami obecnymi w próbce (jego absorpcja) lub tworzenie kompleksów fluoroforów ze składnikami matrycy, które są niezdolne do fluorescencji. Jest to zjawisko szczególnie niekorzystne w przypadku dokonywania analizy ilościowej składu próbki. Eliminuje się je stosując odpowiednie rozcieńczenie próbki lub wstępny rozdział analitów od matrycy [13].

Optymalizacja rozcieńczenia próbek była niezbędna podczas badań wodnych ekstraktów produktów spożywczych (kawa, mięta pieprzowa, pasta pomidorowa) przedstawionych w **publikacjach IV i V**, jak również podczas badań ekstraktów próbek bazylii i oregano oraz wybranych suplementów diety.

### 3.2. Widma rejestrowane w zakresie podczerwieni

Sygnały spektroskopowe rejestrowane w zakresie podczerwieni są czułe na obecność w otoczeniu przyrządu pomiarowego takich składników jak para wodna czy dwutlenek węgla. Jeżeli pomiar prowadzony jest w pomieszczeniu gdzie nie ma zapewnionych stałych poziomów stężeń tych substancji lub obojętnej atmosfery dla przyrządu pomiarowego należy sygnały przed analizą chemometryczną skorygować w odpowiednich obszarach. Korekcji takiej można dokonać również w trakcie pomiaru, rejestrując sygnał tła każdorazowo przed pomiarem próbki. Szybkość rejestracji widm IR oraz ich niedestrukcyjny charakter umożliwia uzyskanie satysfakcjonującej wartości stosunku sygnału do szumu już w wyjściowych sygnałach. Osiąga się to poprzez wielokrotną rejestrację widma dla pojedynczej próbki (np. 32 lub 64 powtórzenia) w wybranym zakresie spektralnym, a końcowy sygnał jest średnią arytmetyczną ze wszystkich dokonanych pomiarów. Często etapem przygotowania sygnałów IR do analizy jest także wybór zakresu spektralnego charakterystycznego dla modelowanej właściwości. Taka redukcja wymiarowości danych niejednokrotnie prowadzi do obniżenia kompleksowości modelu chemometrycznego oraz poprawienia jego właściwości dopasowania do danych i przewidywania.

W **publikacji V** porównałam efektywność modelowania fluorescencyjnych sygnałów i widm podczerwieni w kontekście oceny całkowitej zdolności antyoksydacyjnej past pomidorowych. Aby uzyskać sygnał rejestrowany w obszarze podczerwieni (zakres 4000 do 400  $\text{cm}^{-1}$ ) o najlepszej wartości parametru S/N, dla każdej próbki wykonano 64 powtórzenia. Przed pomiarem widma każdej próbki rejestrowane było tło w celu korekcji linii bazowej sygnału. Sprawdziłam efektywność modelowania całkowitej zawartości

antyoksydantów w oparciu o całe sygnały IR oraz o wybrany (sugerowany w literaturze [17]) zakres widma tj. od 2000 do 900  $\text{cm}^{-1}$ . Potwierdziłam dokonane przez Lu i współpracowników obserwacje większej efektywności wieloparametrowych modeli kalibracyjnych zbudowanych w oparciu o sygnały IR w zawężonym zakresie spektralnym.

### 3.3. Chromatogramy

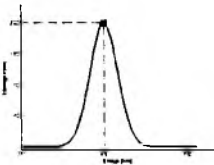
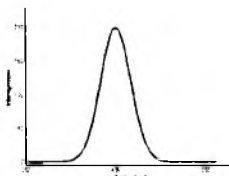
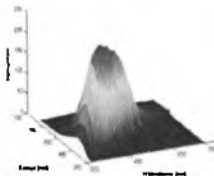
Chromatogramy zarejestrowane dla tej samej próbki (lub zbioru próbek) mogą różnić się liczbą punktów pomiarowych (mieć różną długość). Wówczas zestawienie ich w jedną macierz (operacja jest konieczna do przeprowadzenia jakiegokolwiek analizy chemometrycznej) jest niemożliwe. Dlatego należy wszystkie chromatogramy ujednolicić tak, aby miały tę samą liczbę punktów pomiarowych.

Innym efektem charakterystycznym dla chromatogramów jest przesunięcie pików reprezentujących tę samą substancję. Przyczyną takiego zjawiska jest starzenie złoża kolumny chromatograficznej lub zmienność warunków rozdzielania (np. wahania składu fazy ruchomej). By dokonać analizy porównawczej należy nałożyć na siebie chromatogramy, czyli ustalić pozycje odpowiadających sobie pików na osi czasu elucji. Jedną z metod chemometrycznych używanych w tym celu jest metoda zoptymalizowanego nakładania sygnałów maksymalizująca ich wzajemną korelację (ang. *correlation optimized warping*, COW) [18].

Metoda COW została zastosowana do nałożenia 228 chromatogramów zarejestrowanych dla próbek naparów herbaty typu rooibos. Szczegóły przygotowania tych sygnałów do analizy przedstawione są w **publikacji VI**.

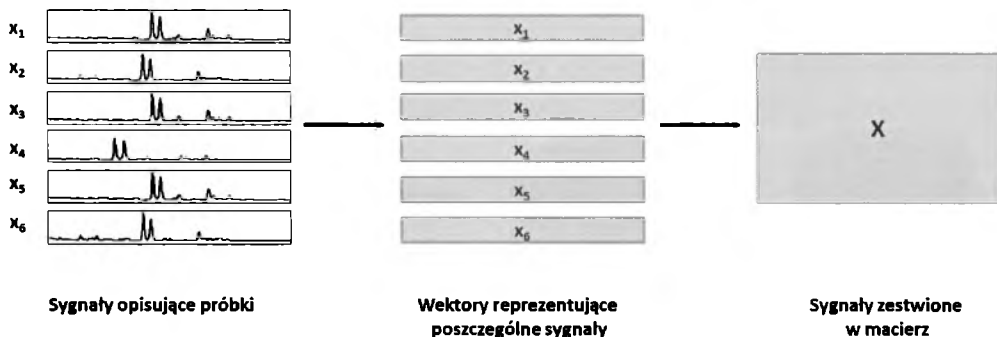
### 3.4. Zestawienie sygnałów analitycznych

Aby dokonać analizy chemometrycznej należy zestawić ze sobą wszystkie sygnały reprezentujące analizowane próbki. To w jaki sposób powinny zostać zestawione sygnały jest uzależnione od postaci w jakiej sygnał jest reprezentowany (wpływającej na rzędowość danych, rysunek 7) oraz od metody jaka ma zostać użyta do ich analizy.

Typ danych	Skalar	Wektor	Macierz
Postać numeryczna	201,5	[0,5 ... 98,0 ... 201,5 ... 39,9 ... 0,3]	$\begin{bmatrix} 0,3 & \dots & 11,2 & \dots & 17,0 & \dots & 10,1 & \dots & 0,5 \\ 0,5 & \dots & 98,0 & \dots & 201,5 & \dots & 39,9 & \dots & 0,3 \\ \dots & & & & & & & & \\ 0,1 & \dots & 15,7 & \dots & 38,0 & \dots & 25,1 & \dots & 1,1 \end{bmatrix}$
Postać graficzna			
Rząd	Zerowy	Pierwszy	Drugi

Rysunek 7 Rzędowość danych przedstawiona na przykładzie danych fluorescencyjnych.

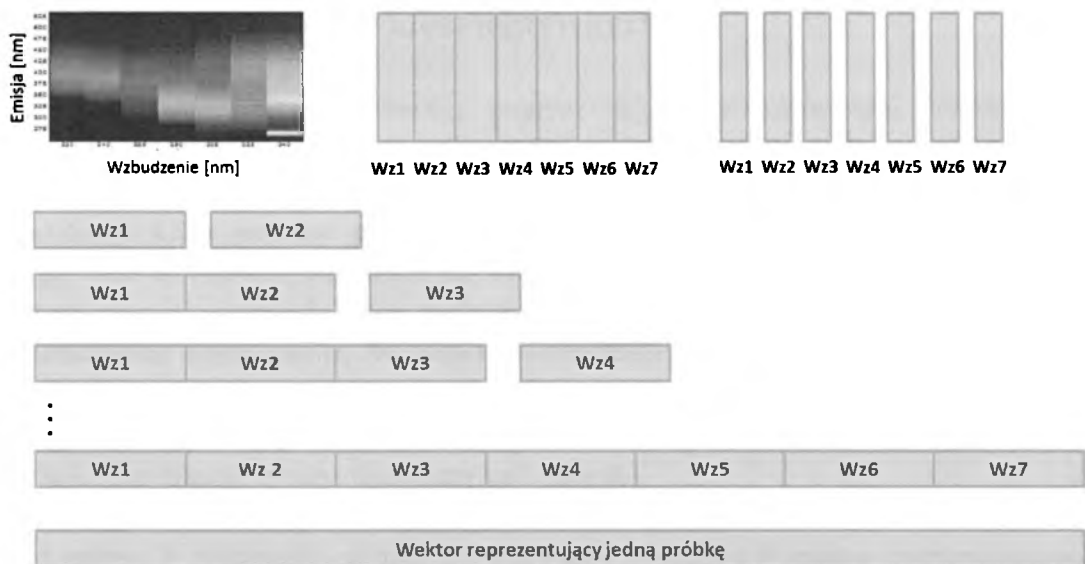
Jeżeli sygnały instrumentalne mają postać wektora (np. widma NIR, UV-Vis, chromatogramy) to organizuje się je w macierz, której kolejne wiersze zawierają sygnały zarejestrowane dla poszczególnych próbek (rysunek 8). Przygotowana w ten sposób macierz sygnałów nadaje się do analizy przeprowadzanej z zastosowaniem klasycznych metod chemometrycznych, jak np. analiza czynników głównych [8] czy regresja częściowych najmniejszych kwadratów (ang. *partial least squares*, PLS) [19].



Rysunek 8 Schemat zestawiania przykładowych sygnałów instrumentalnych w macierz danych.

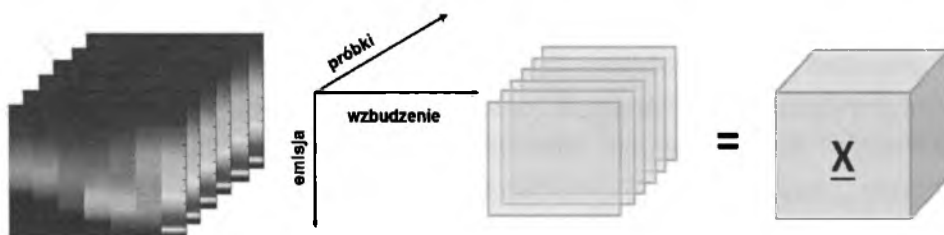
W przypadku sygnałów o dwuwymiarowym charakterze, np. fluorescencyjne obrazy, można je organizować w różny sposób. Sygnały można rozwinąć, tak aby utworzyły sygnały jednowymiarowe jak przedstawia rysunek 9 i zestawić do postaci macierzy użytecznej do modelowania klasycznymi metodami chemometrycznymi.





**Rysunek 9** Proces rozwijania fluorescencyjnego obrazu; Wz oznacza kolejne sygnały emisyjne zarejestrowane dla różnych długości fali wzbudzającej.

Sygnały można również pozostawić w formie dwuwymiarowej i zestawić w trójwymiarową strukturę (tensor), której kolejne wymiary odpowiadają długościom fal wzbudzenia, emisji oraz próbkom (rysunek 10). Zestawione w ten sposób sygnały można modelować wykorzystując metody dedykowane dla danych N-modalnych, np. metodę współbieżnej analizy czynnikowej, PARAFAC [12] lub N-modalną regresję częściowych najmniejszych kwadratów (ang. *N-way partial least squares*, N-PLS) [20].



**Rysunek 10** Konstrukcja trójwymiarowej struktury danych – tensora zawierającego widma fluorescencyjne.

W publikacjach I – V, w których jako dane wykorzystałam fluorescencyjne obrazy, dane zestawiałam w macierze zawierające rozwinięte sygnały, zarejestrowane dla badanych próbek oraz w trójwymiarowe tensory. W każdej z publikacji porównałam efektywność modelowania sygnałów rozwiniętych w wektory i w postaci macierzy. Widma rejestrowane w zakresie podczerwieni i chromatogramy zestawione w macierze były przedmiotem badań opisanych w publikacjach V i VI.

## 4. Modelowanie sygnałów nieselektywnych

Każdy sygnał instrumentalny powstaje poprzez rejestrację odpowiedzi detektora dla wybranej liczby punktów pomiarowych, zdefiniowanej przez przyjęty krok pomiarowy. Taki sposób prowadzenia pomiaru powoduje, że w większości przypadków pik pojedynczego analitu opisuje kilka punktów pomiarowych. W związku z tym, sygnały instrumentalne zawierają wiele skorelowanych zmiennych, które uniemożliwiają użycie ich do modelowania np. metody regresji wielorakiej (ang. *multiple linear regression*, MLR) [19], ze względu na jej matematyczne ograniczenia. Wówczas wykorzystuje się zmienne niezależne wybrane z całego sygnału lub skonstruowane na jego podstawie, które w zależności od tego jakie informacje chcemy z sygnału nieselektywnego uzyskać, wykorzystywane są do celów eksploracji, klasyfikacji/dyskryminacji czy kalibracji [6].

Metody używane do uzyskania z sygnałów istotnej chemicznie informacji można podzielić na dwie grupy. W zależności od tego czy używamy wyłącznie sygnałów instrumentalnych czy też równocześnie uwzględniamy dodatkowe informacje o modelowanym układzie, np. pochodzenie, stężenie wybranej/wybranych substancji czy przynależność do zdefiniowanych klas (np. stopień palenia kawy, wysokość podatku akcyzowego nałożonego na olej napędowy, pochodzenie geograficzne próbki) mówimy odpowiednio o metodach modelowania danych bez nadzoru i metodach modelowania danych z nadzorem.

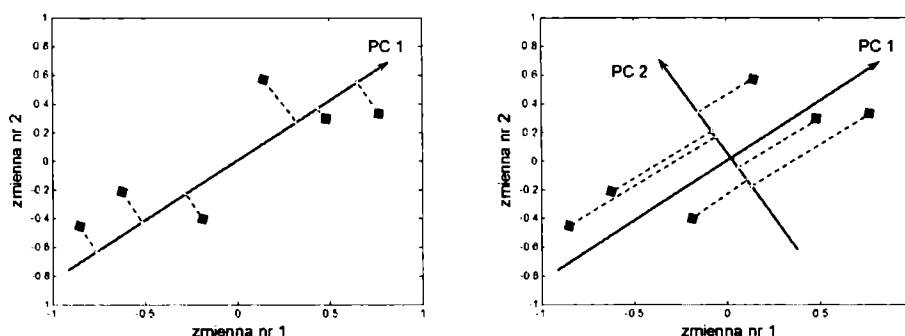
### 4.1. Metody modelowania danych bez nadzoru

Tę grupę metod wykorzystuje się do eksploracji struktury danych. Jest to najczęściej pierwszy etap interpretacji informacji w nich zawartej. Celem eksploracji jest wizualizacja struktury danych. Umożliwia to ocenę podobieństw próbek i kompleksowości badanego układu (zjawiska), jak również ocenę tego, czy w eksplorowanych danych znajdują się obiekty odległe – oddalone od większości danych (ang. *outlying objects*). Analiza eksploracyjna determinuje wybór metod modelowania z nadzorem i dalszą interpretację uzyskanych wyników. Podczas pracy z sygnałami nieselektywnymi, które są wieloparametrowe i zawierają zmienne skorelowane, do wizualizacji informacji w nich zawartych, niezbędna jest redukcja ich kompleksowości. Narzędziami eksploracyjnym wykorzystywanymi najczęściej do tego celu są między innymi metody projekcji, tj. analiza czynników głównych, PCA [8] czy współbieżna analiza czynnikowa, PARAFAC [12], dedykowana analizie danych N-modalnych.

#### 4.1.1. Analiza czynników głównych

Analiza czynników głównych to metoda umożliwiająca modelowanie i wizualizację struktury danych wieloparametrowych, zawierających zmienne skorelowane. Wizualizacja danych jest możliwa dzięki konstrukcji nowych zmiennych będących liniową kombinacją oryginalnych zmiennych, nazywanych czynnikami głównymi (ang. *principal components*, PC). Czynniki główne konstruuje się tak, aby maksymalizowały opis wariancji (zmienności) danych.

Algorytm konstrukcji nowych zmiennych zakłada, że każdy nowy czynnik główny opisuje część wariancji, która nie została opisana przez wcześniej skonstruowane czynniki. W związku z tym, czynniki główne są wzajemnie ortogonalne co przedstawia rysunek 11.



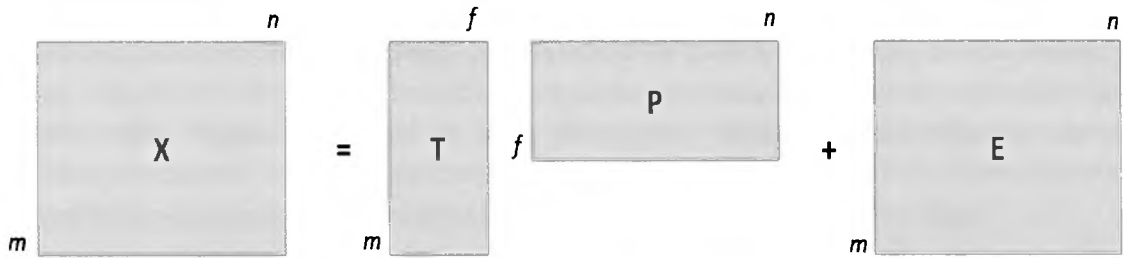
**Rysunek 11** Konstrukcja dwóch pierwszych czynników głównych (PC 1 oraz PC 2) w dwuwymiarowej przestrzeni opisanej przez zmienną nr 1 i zmienną nr 2.

Każdy czynnik główny można przedstawić jako iloczyn dwóch wektorów zawierających odpowiednio współrzędne obiektów i parametrów na nowej zmiennej. Liczba czynników głównych, opisująca satysfakcjonujący poziom wariancji danych,  $f$ , nazywana jest kompleksowością modelu. Dla danych o  $m$  obiektach (próbki) i  $n$  parametrach (punktach pomiarowych),  $\mathbf{X}_{[m,n]}$ , równanie opisujące dekompozycję oryginalnych zmiennych ma następującą postać:

$$\mathbf{X}_{[m,n]} = \mathbf{T}_{[m,f]} \mathbf{P}_{[f,n]}^T + \mathbf{E}_{[m,n]} \quad (1)$$

gdzie,  $\mathbf{T}_{[m,f]}$  i  $\mathbf{P}_{[f,n]}$  to macierze wyników i wag zawierające odpowiednio współrzędne obiektów i parametrów na nowych zmiennych (czynnikach głównych), natomiast  $\mathbf{E}_{[m,n]}$  to macierz zawierająca tę część wariancji, która nie została opisana przez model PCA o  $f$  czynnikach, tzw. reszty od modelu.

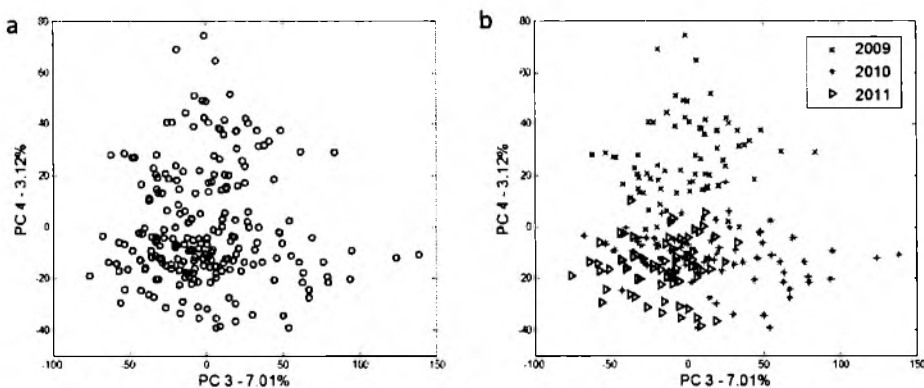
Rysunek 12 przedstawia graficzną postać operacji dekompozycji danych wykorzystując model PCA opisany równaniem (1).



**Rysunek 12** Dekompozycja macierzy danych  $X$  do macierzy wyników  $T$ , wag  $P$  oraz reszt  $E$  wykorzystując model PCA o  $f$  czynnikach głównych.

Dokonując projekcji obiektów na czynniki główne ocenia się to, czy dane wykazują tendencję do grupowania. Wzbogacając projekcję o dodatkowe informacje można zaobserwować, czy występują w nich trendy, np. zależne od czasu lub pochodzenia geograficznego próbek (rysunek 13). Interpretacja informacji zawartej w wagach ułatwia ocenę tego, które parametry wykazują największy wkład do konstrukcji wybranego czynnika głównego, a tym samym odpowiadają za występowanie zaobserwowanego trendu w danych.

We wszystkich problemach badawczych, które podejmuję w ramach niniejszej pracy doktorskiej, wykorzystuję dane wielowymiarowe. Metoda analizy czynników głównych jest jednym z podstawowych narzędzi chemometrycznych użytych przeze mnie do eksploracji struktury każdego zestawu sygnałów nieselektywnych analizowanych w **publikacjach I–VI**. Jednym z przykładów praktycznego wykorzystania metody PCA jest wizualizacja struktury danych opisanych w **publikacji VI** (rysunek 13).



**Rysunek 13** Projekcja próbek herbaty typu rooibos na trzeci i czwarty czynnik główny (PC 3 i PC 4) a) bez zaznaczenia informacji o roku produkcji próbek, b) z uwzględnieniem informacji o roku produkcji próbek.

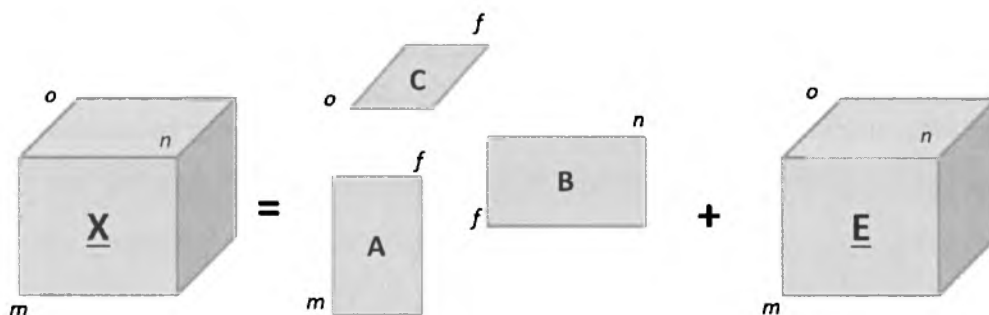
Projekcja obiektów na płaszczyznę zdefiniowaną przez trzeci i czwarty czynnik główny umożliwiła zaobserwowanie tendencji różnicowania się próbek ze względu na rok produkcji. Próbki z lat 2010 i 2011 miały inną charakterystykę chemiczną niż próbki wyprodukowane w 2009 roku. Projekcja wag na tę samą płaszczyznę umożliwia identyfikację czterech charakterystycznych polifenoli (orientyny, izoorientyny, hiperozydu i 3-*O*-robinobiozydu kwercetyny) odpowiedzialnych za wystąpienie obserwowanego w danych trendu.

#### 4.1.2. Dekompozycja sygnałów z użyciem metody współbieżnej analizy czynnikowej

Metoda współbieżnej analizy czynnikowej, PARAFAC może być rozważana jako rozszerzenie metody PCA do dekompozycji danych o  $N$ -modalnej strukturze. Oryginalne dane przedstawiane są w postaci nowych zmiennych opisujących obiekty i wagi parametrów. Liczba zbiorów wag jest zależna od wymiarowości danych. W przypadku struktury danych o trzech wymiarach konstruowane są nowe zmienne (tzw. triady), które reprezentują trzy wymiary danych i przedstawiają informację o obiektach oraz wagi dla dwóch pozostałych kierunków (rysunek 14). Matematycznie dekompozycja rozwiniętego tensora danych jest opisana równaniem:

$$\mathbf{X}_{[m \times n \times o]} = \mathbf{A}_{[m, f]}(\mathbf{C}_{[n, f]} \otimes \mathbf{B}_{[o, f]})^T + \mathbf{E}_{[m \times n \times o]} \quad (2)$$

gdzie,  $\mathbf{X}_{[m \times n \times o]}$  to rozwinięta do postaci macierzy o wymiarach  $m \times n \times o$  trójwymiarowa struktura danych o oryginalnych wymiarach  $m \times n \times o$ ,  $\mathbf{A}_{[m, f]}$ ,  $\mathbf{B}_{[n, f]}$  i  $\mathbf{C}_{[o, f]}$  to macierze zawierające  $f$  triad,  $\otimes$  reprezentuje iloczyn Khatri-Rao, natomiast  $\mathbf{E}_{[m \times n \times o]}$  to macierz zawierająca reszty od modelu PARAFAC o  $f$  triadach.



**Rysunek 14** Graficzna postać modelu PARAFAC.  $\mathbf{X}_{[m, n, o]}$  zawiera oryginalne zmienne zestawiane w trójwymiarową strukturę,  $\mathbf{A}_{[m, f]}$ ,  $\mathbf{B}_{[n, f]}$  i  $\mathbf{C}_{[o, f]}$  to macierze zawierające  $f$  triad,  $\mathbf{E}_{[m, n, o]}$  zawiera reszty od modelu PARAFAC o  $f$  triadach.

Ważną różnicą pomiędzy metodami PARAFAC i PCA jest to, że poszczególne triady nie muszą być wzajemnie ortogonalne (jak w przypadku kolejnych czynników głównych). Tego typu analiza eksploracyjna, zastosowana do fluorescencyjnych obrazów, umożliwia m.in. uzyskanie tzw. czystych profili spektralnych fluoroforów zawartych w analizowanych próbkach. Jednak poprawna interpretacja wyników wymaga gruntownej wiedzy na temat składu chemicznego próbki, co jest trudne zwłaszcza wtedy, gdy próbki są pochodzenia naturalnego. Dlatego podobnie jak PCA, model PARAFAC jest najczęściej wykorzystywany do eksploracji struktury wielomodalnych danych.

## 4.2. Metody modelowania danych z nadzorem

Są to metody, w których sygnały instrumentalne modeluje się wykorzystując dodatkową informację o próbkach (oprócz zestawu wybranych parametrów lub sygnałów instrumentalnych je opisujących). W zależności od charakteru dostępnej informacji dzieli się je na metody kalibracyjne, gdzie dodatkowa informacja ma postać zmiennej ciągłej (np. nasycenie koloru, stężenie komponentów) oraz metody klasyfikacyjne/dyskryminacyjne, w których zmienna opisująca próbki ma charakter dyskretny (klasy próbek, np. pochodzenie geograficzne, rok produkcji, warunki przechowywania). Wybór metody modelowania z nadzorem zależy od przyjętego celu badawczego.

### 4.2.1. Kalibracja wybranej własności

Konstrukcja modelu kalibracyjnego polega na ustaleniu zależności pomiędzy informacją o właściwościach próbki (fizycznych lub chemicznych), a zawartością badanego składnika/składników lub wartościami modelowanej własności, na którą wpływ mają różne komponenty chemiczne próbki i/lub zjawiska w niej zachodzące (np. całkowita zdolność antyoksydacyjna). Z matematycznego punktu widzenia, konstruuje się równanie regresyjne, w którym informacje zawarte w zmiennej niezależnej (objaśniającej),  $x$ , umożliwiają estymację wartości zmiennej zależnej (objaśnianej),  $y$ . Jest to możliwe dzięki wyznaczeniu współczynnika(ków) regresji,  $b$  ( $b=[b, b_0]$ ), czyli rozwiązywane jest klasyczne równanie:

$$y = bx + b_0 \quad (3)$$

Chemometryczne metody kalibracyjne to np. regresja głównych składowych (ang. *principal component regression*, PCR) [21] czy regresja częściowych najmniejszych kwadratów, PLS [21]. Znane są również metody dedykowane modelowaniu danych zestawionych w trójwymiarową strukturę, np. N-modalna regresja częściowych najmniejszych kwadratów, N-PLS [20].

## 4.2.1.1. Regresja częściowych najmniejszych kwadratów

Regresja częściowych najmniejszych kwadratów jest to jedna z najpopularniejszych metod konstrukcji modeli kalibracyjnych, stosowana do danych wieloparametrowych zawierających skorelowane zmienne. Model budowany z użyciem algorytmu PLS oparty jest na tzw. zmiennych ukrytych, czyli nowych zmiennych, skonstruowanych na podstawie zmiennych oryginalnych. Nowe zmienne konstruowane są tak, aby każda następna była ortogonalna do poprzednich (podobnie jak w metodzie PCA), dobrze opisywała wariancję zawartą w  $\mathbf{X}$  oraz w  $\mathbf{y}$ , a także uwzględniała kowariancję pomiędzy  $\mathbf{X}$  a  $\mathbf{y}$ . Model może być zbudowany dla zmiennej zależnej w postaci wektora (wariant PLS-1) lub macierzy (wariant PLS-2). W niniejszej pracy konstruowałam wyłącznie modele PLS-1.

Model PLS-1 można wyrazić następująco:

$$\mathbf{X}_{[m,n]} = \mathbf{T}_{[m,f]} \mathbf{P}_{[n,f]}^T + \mathbf{E}_{[m,n]} \quad (4)$$

$$\mathbf{y}_{[m,1]} = \mathbf{T}_{[m,f]} \mathbf{q}_{[f,1]} + \mathbf{r}_{[m,1]} = \mathbf{X}_{[m,n]} \mathbf{b}_{[n,1]} + \mathbf{r}_{[m,1]} \quad (5)$$

gdzie,  $\mathbf{X}_{[m,n]}$  to macierz danych zawierająca  $m$  obiektów i  $n$  parametrów,  $\mathbf{T}_{[m,f]}$  i  $\mathbf{P}_{[f,n]}$  reprezentują wagi modelu PLS,  $\mathbf{q}_{[f,1]}$  zawiera współczynniki regresji dla modelu PLS skonstruowanego używając  $f$  nowych zmiennych,  $\mathbf{E}_{[m,n]}$  i  $\mathbf{r}_{[m,1]}$  to różnice pomiędzy wyjściowymi wartościami  $\mathbf{X}_{[m,n]}$  i  $\mathbf{y}_{[m,1]}$ , a tymi uzyskanymi z modelu,  $\mathbf{b}_{[n,1]}$  to wektor współczynników regresji dla oryginalnej macierzy  $\mathbf{X}$ , obliczany według poniższego wzoru:

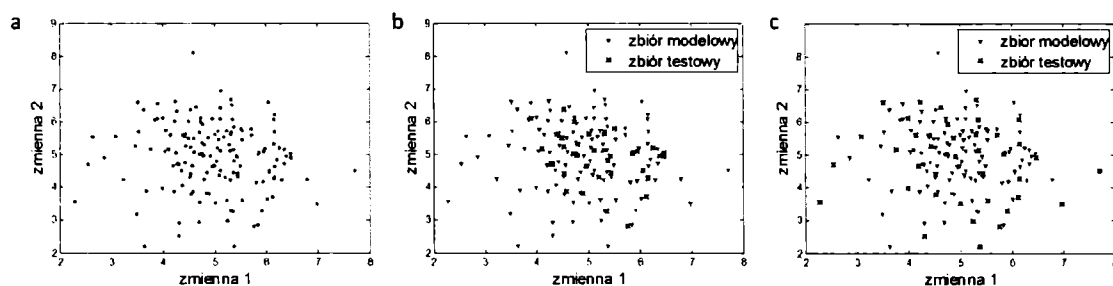
$$\mathbf{b}_{[n,1]} = \mathbf{W}_{[n,f]} (\mathbf{P}_{[f,n]}^T \mathbf{W}_{[n,f]})^{-1} \mathbf{q}_{[f,1]} \quad (6)$$

gdzie,  $\mathbf{W}_{[n,f]}$  reprezentuje wagi maksymalizujące kowariancję pomiędzy  $\mathbf{X}_{[m,n]}$  a  $\mathbf{y}_{[m,1]}$ .

Podobnie jak w przypadku modelu PCA,  $f$  określa optymalną liczbę czynników – nowych zmiennych – niezbędną do uzyskania modelu kalibracyjnego o satysfakcjonujących właściwościach dopasowania do danych i przewidywania. Właściwości te wyznacza się dla dwóch niezależnych podzbiorów danych oryginalnych, z których jeden służy do konstrukcji modelu (nazywany jest zbiorem modelowym), drugi natomiast do oceny jego właściwości przewidywania modelowanej własności (nazywany zbiorem testowym).

Obiektywnego wyboru próbek (obiektów) włączonych do zbiorów modelowego i testowego dokonuje się używając przeznaczonych do tego celu algorytmów. Jednymi z najpopularniejszych są algorytmy Kennarda i Stona [22] oraz Duplex [23]. W obydwu algorytmach, podziału próbek dokonuje się w oparciu o wzajemną odległość próbek.

Pierwszy z nich zapewnia włączenie do zbioru modelowego próbek reprezentujących całą przestrzeń eksperymentalną, dzięki czemu zbiór ten reprezentuje wszystkie możliwe źródła zmienności. Do zbioru testowego włączane są natomiast pozostałe próbki o mniej zróżnicowanej charakterystyce. W przypadku algorytmu Duplex zbiór modelowy i testowy zawierają próbki o podobnej reprezentatywności, gdyż zbiory te konstruowane są równocześnie (na podstawie obliczonych odległości próbki włączane są naprzemiennie do zbioru modelowego i testowego). Rysunek 15 przedstawia podział zbioru próbek opisanych przez dwa parametry na zbiory modelowy i testowy z użyciem dwóch omówionych algorytmów.



**Rysunek 15** a) rozproszenie próbek na płaszczyźnie zdefiniowanej przez dwie zmienne, podział próbek na zbiory modelowy i testowy z użyciem algorytmów b) Kennarda i Stona oraz c) Duplex.

W każdej publikacji cyklu zastosowałam algorytmy wyboru zbiorów modelowego i testowego, dzięki czemu mogłam zweryfikować jakość skonstruowanych modeli zarówno pod względem dopasowania do danych ze zbioru modelowego jak i pod względem właściwości przewidywania ocenionej na podstawie niezależnych zbiorów (próbek testowych).

W przypadku, gdy w zbiorze danych znajdują się replikaty próbek (laboratoryjne lub techniczne) podczas podziału danych na zbiory modelowy i testowy należy zwrócić uwagę na to, aby wszystkie replikaty danej próbki zostały włączone do tego samego podzbioru. Jest to ważne ze względu na ocenę jakości modelu. Jeżeli replikaty tej samej próbki będą zarówno w zbiorze modelowym jak i testowym, właściwości predykcyjne modelu mogą zostać w sposób fałszywy uznane za lepsze niż w rzeczywistości.

Takie podejście do konstrukcji zbiorów modelowego i testowego opisałam w publikacjach II, III, IV oraz V.

Dopasowanie modelu do danych i jego właściwości predykcyjne wyrażone są jako średnie błędy kwadratowe pomiędzy wartościami zmiennej zależnej obserwowanej jako wynik eksperymentu,  $y$ , oraz przewidzianej na podstawie modelu,  $\hat{y}$ , obliczane dla każdego z podzbiorów osobno, zgodnie z poniższymi wzorami:

$$RMSE(f) = \sqrt{\frac{1}{k} \sum_{i=1}^k (y_i - \hat{y}_i)^2} \quad (7)$$



$$RMSEP(f) = \sqrt{\frac{1}{t} \sum_{i=1}^t (y_i - \hat{y}_i)^2} \quad (8)$$

gdzie, RMSE oznacza średni błąd kwadratowy (ang. *root mean square error*), RMSEP to średni błąd kwadratowy przewidywania (ang. *root mean square error of prediction*),  $k$  oraz  $t$  to liczba obiektów włączona odpowiednio do zbiorów modelowego i testowego, a  $f$  określa optymalną kompleksowość modelu.

Optymalna kompleksowość modelu jest wybierana na podstawie walidacji krzyżowej (ang. *cross-validation*) [21]. Procedura polega na iteracyjnym podziale zbioru modelowego na dwa niezależne podzbiory. Pierwszy z nich używany jest do konstrukcji modelu, drugi zaś do walidacji jego właściwości przewidywania. W zależności od liczby obiektów włączonej do podzbioru walidacyjnego,  $p$ , procedura nazywana jest walidacją krzyżową typu wyrzuć jeden obiekt (ang. *leave-one-out*) lub wyrzuć więcej obiektów (ang. *leave-more-out*). Obiekty mogą być wyrzucane kolejno lub w sposób przypadkowy. Wówczas mówi się o walidacji krzyżowej typu Monte Carlo [24]. Pary podzbiorów budowane są do momentu, aż wykorzystane zostaną wszystkie możliwości lub osiągnięty zostanie limit założony przez użytkownika procedury,  $c$ . Dla każdej pary podzbiorów budowane są modele ze wzrastającą liczbą czynników,  $f$ , użytych do budowy modelu. Dla określonej liczby czynników oblicza się średni kwadratowy błąd walidacji krzyżowej (ang. *root mean square error of cross-validation*, RMSECV) zgodnie z następującym równaniem:

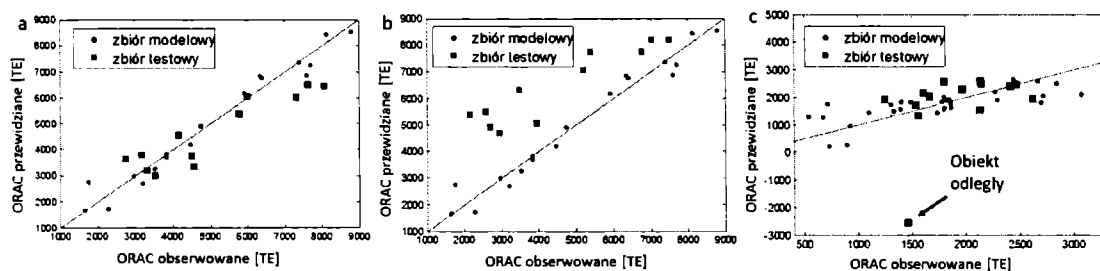
$$RMSECV(f) = \sqrt{\frac{1}{p} \sum_{i=1}^c (y_i - \hat{y}_i)^2} \quad (9)$$

Ostateczny model kalibracyjny budowany jest z użyciem tej liczby czynników, dla której wartość średniego błędu walidacji krzyżowej osiągnęła minimum, uznawanej za liczbę optymalną.

Oceny optymalnej kompleksowości modeli kalibracyjnych i dyskryminacyjnych, przedstawionych w publikacjach I – VI, dokonałam używając tej samej metody – walidacji krzyżowej typu wyrzuć jeden obiekt.

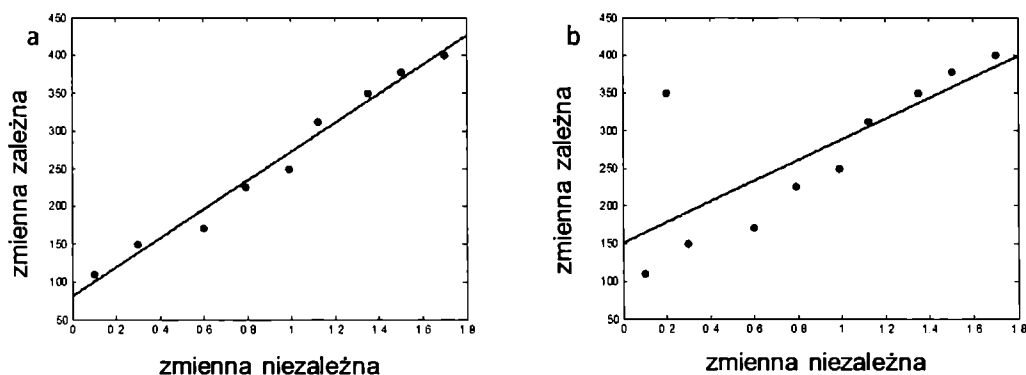
Modele PLS można przedstawić graficznie jako wykres wartości zmiennej zależnej ( $y$  obserwowane), które wyznaczono eksperymentalnie i wartości przewidzianych z użyciem skonstruowanego modelu ( $\hat{y}$  przewidziane). Konstrukcja takiego wykresu dla dwóch niezależnych zbiorów modelowego i testowego ułatwia interpretację jakości modelu, jak również ustalenie, dla których próbek model wykazuje gorsze właściwości przewidywania. Na rysunku 16 przedstawione są wykresy obrazujące właściwości dopasowania do danych i przewidywania trzech modeli regresyjnych. Na każdym z wykresów oprócz punktów reprezentujących poszczególne próbki (włączone do zbiorów modelowego i testowego)

umieszczono linię reprezentującą idealną sytuację, w której wartość obserwowana jest równa wartości przewidzianej. Na podstawie rozproszenia próbek wokół tej linii można dokonać wizualnej oceny parametrów dopasowania do danych i przewidywania, które charakteryzują oceniany model. Równomierne rozłożenie próbek włączonych do zbiorów modelowego i testowego na rysunku 16 a) świadczy o dobrym dopasowaniu modelu do danych i jego dobrych właściwościach przewidywania. Podobnie, model przedstawiony na rysunku 16 b) charakteryzuje się dobrym dopasowaniem do próbek zawartych w zbiorze modelowym, zdecydowane oddalenie wszystkich próbek zawartych w zbiorze testowym od idealnej linii dowodzi jego złych właściwościach przewidywania. Mogą one być przyczyną np. nieodpowiedniej kompleksowości modelu lub braku reprezentatywności zbioru modelowego. Rysunek 16 c) przedstawia model o dobrych właściwościach dopasowania do danych i przewidywania. W odróżnieniu od modelu przedstawionego na rysunku 16 a) w zbiorze testowym zawarty jest obiekt dla którego model wykazuje gorsze właściwości przewidywania. Taki obiekt nazywany jest obiektem odległym.



**Rysunek 16** Wykresy ilustrujące modele N-PLS skonstruowane dla fluorescencyjnych obrazów próbek wodnych ekstraktów mięty pieprzowej i informacji o całkowitej zdolności antyoksydacyjnej a) o dobrych właściwościach dopasowania do danych i przewidywania, b) o dobrych właściwościach dopasowania do danych i niesatysfakcjonujących właściwościach przewidywania oraz c) model PLS skonstruowany dla fluorescencyjnych obrazów wodnych ekstraktów kawy i informacji o całkowitej zdolności antyoksydacyjnej, zawierający obiekt odległy włączony do zbioru testowego. Wykresy zostały skonstruowane dla zestawu danych omawianego w **publikacji IV**.

Model regresyjny konstruowany jest w taki sposób, aby dobrze opisywał informacje o wszystkich obiektach użytych do jego konstrukcji. Takie kryterium jest zadowalające i efektywne wówczas, gdy w zbiorze modelowanych danych wszystkie obiekty są w sposób równomierny rozproszone w przestrzeni eksperymentalnej. Często, w danych eksperymentalnych, zwłaszcza opisujących próbki pochodzenia naturalnego, zdarzają się obiekty oddalone od większości próbek, zwane obiektami odległymi. Próbkę tego typu są wynikiem błędu laboratoryjnego lub unikatowego składu chemicznego próbki. Obecność obiektów odległych negatywnie wpływa na właściwości dopasowania i przewidywania modeli regresyjnych opartych na założeniu minimalizacji sumy kwadratów reszt od modelu. Model ulega „przekrzywieniu” w celu zminimalizowania sumy kwadratów reszt dla wszystkich obiektów, przez co nie opisuje dobrze większości danych (rysunek 17). Błędna estymacja współczynników regresji prowadzi do modeli o złych właściwościach predykcyjnych.



**Rysunek 17** Rysunki przedstawiają przykładowe modele kalibracyjne utworzone metodą najmniejszych kwadratów dla danych: a) bez obiektów odległych oraz b) z obiektem odległym.

W praktyce nie można z góry zakładać, że w danych eksperymentalnych obiekty odległe będą nieobecne. Wstępna analiza struktury danych, dokonana metodami eksploracyjnymi rzadko pozwala na ich skuteczną detekcję. Dodatkowo obiekty, które uznano za odległe mogą dostarczyć cennych informacji podczas konstrukcji modelu, dlatego opracowano metody regresyjne, które są nieczułe na ich obecność np. stabilny wariant regresji częściowych najmniejszych kwadratów (ang. *partial robust M-regression*, PRM) [25].

#### 4.2.1.2. Stabilny wariant regresji częściowych najmniejszych kwadratów

Stabilny model kalibracyjny PRM konstruowany jest tak, aby jak najlepiej opisać kowariancję pomiędzy zmiennymi zależnymi i niezależną reprezentowaną przez większość danych zawartych w zbiorze modelowym i zarazem minimalizuje negatywny wpływ obiektów odległych. W trakcie budowy modelu obiekty odległe nie są eliminowane ze zbioru modelowego. Dzięki użyciu stabilnych estymatorów i przypisaniu wagi każdemu obiektowi użytemu do konstrukcji modelu negatywny wpływ obiektów odległych na model jest eliminowany.

Wagi poszczególnych obiektów obliczane są w sposób iteracyjny jako iloczyn wag definiowanych w przestrzeni modelu,  $w_i^x$ , i wag dla reszt,  $w_i^r$ . Wagi wpływu  $i$ -tego obiektu na model obliczane są zgodnie z następującymi równaniami:

$$w_i^x = \psi\left(\frac{\|t_i - L1med(T)\|}{med_i \|t_i - L1med(T)\|}, c\right) \quad (10)$$

oraz

$$\psi(z, c) = \frac{1}{(1 - \frac{|z|}{c})^2} \quad (11)$$

gdzie,  $\|\bullet\|$  to norma euklidesowa wektora  $t_i$  zawierającego wyniki modelu PLS o wybranej kompleksowości, 'L1med' to mediana L1 będąca stabilnym estymatorem środka danych [26],  $\psi$  to funkcja ważąca (ang. "Fair" function),  $c$  stała (w większości zastosowań  $c = 4$ ) [27].

Waga reszt dla  $i$ -tego obiektu jest obliczana według następującego równania:

$$w_i^r = \psi\left(\frac{r_i}{\sigma_{\text{MAD}}}, c\right) \quad (12)$$

gdzie,  $r_i$  to reszty od modelu (podniesione do kwadratu różnice pomiędzy przewidzianą na podstawie modelu i obserwowaną wartością odpowiedzi dla  $i$ -tego obiektu),  $\sigma_{\text{MAD}}$  to stabilny estymator skali danych wyrażony jako mediana z bezwzględnej wartości odchyżeń od mediany [27]:

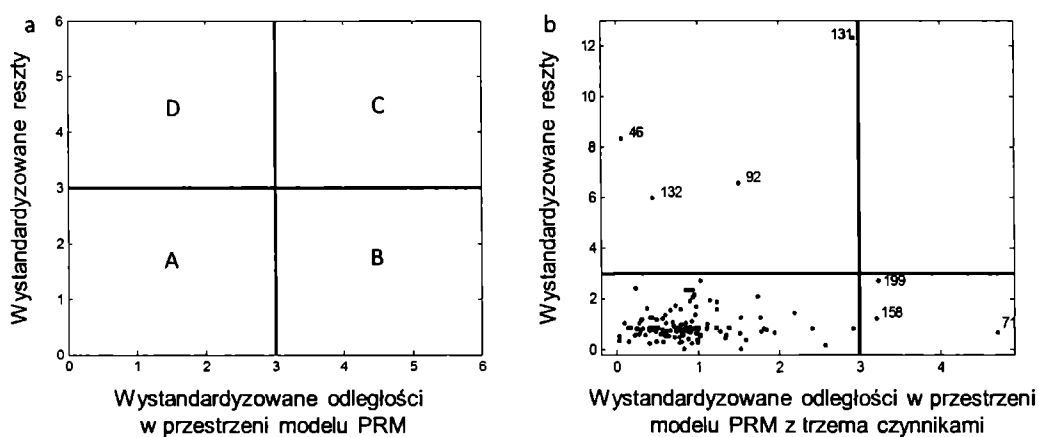
$$\sigma_{\text{MAD}} = \text{med}_i |r_i - \text{med}_j(r_j)| \quad (13)$$

gdzie, 'med' oznacza stabilny estymator położenia środka danych – medianę.

Wagi obliczane są iteracyjnie, do momentu w którym zbieżność algorytmu jest osiągnięta, tj. kiedy różnice pomiędzy normami wektorów współczynników regresji modelu PRM w dwóch następujących po sobie iteracjach są na dostatecznie niskim poziomie (np.  $10^{-2}$ ). Następnie, wyznaczone wagi są używane do konstrukcji wieloparametrowego modelu. Oszacowane parametry modelu PRM można zastosować do identyfikacji obiektów odległych oraz do ustalenia zakresu ich oddziaływania na model. Wykorzystuje się do tego dwie wystandaryzowane wartości oszacowane dla każdego z obiektów – stabilne odległości w przestrzeni modelu oraz reszty. Dla obydwu wartości, wyznacza się wartości progowe zakładając rozkład normalny dla większości obiektów użytych do budowy modelu, które dzielą przestrzeń tego modelu na cztery części. Wizualizacji dokonuje się na wykresie konstruowanym dla wartości odległości względem wartości reszt (rysunek 18 a). Wpływ danego obiektu określa się w oparciu o to, w której części przestrzeni się znajdzie.

Ze względu na wartości obydwu parametrów obiektu dzieli się na:

- A. obiekty o małych wartościach reszt i zlokalizowane w centrum przestrzeni modelu (ang. *regular objects*);
- B. obiekty o małych wartościach reszt, ale oddalone w przestrzeni modelu (ang. *good leverage objects*), obiekty które mogą pozytywnie wpływać na parametry modelu i zwiększać zakres jego stosowalności;
- C. obiekty oddalone w przestrzeni modelu od większości danych oraz posiadające duże wartości reszt (ang. *bad leverage objects*), negatywnie wpływające na parametry modelu;
- D. obiekty o dużych wartościach reszt, ale nie oddalone w przestrzeni modelu (ang. *high residual objects*).



**Rysunek 18** a) Podział przestrzeni stabilnego modelu PRM na cztery obszary ze względu na oddziaływanie obiektów na model, b) wizualizacja rozkładu obiektów zbioru modelowego w przestrzeni modelu PRM zbudowanego z użyciem trzech czynników (szczegółowa dyskusja znajduje się w **publikacji I**).

Stabilny wariant modelu kalibracyjnego wykorzystałam do rozwiązania problemu z dziedziny analityki procesowej, który wyczerpująco omówiłam w **publikacji I**. Ocena jakości próbek cukru (wyrażonej jako kolor lub zawartość popiołu) pobranych bezpośrednio po procesie rafinacji tego produktu dokonana została w oparciu o modele kalibracyjne zbudowane na podstawie ich fluorescencyjnych obrazów. W publikacji przedstawiłam porównanie działania metod PLS, N-PLS i stabilnej metody kalibracyjnej PRM. Prawdopodobieństwo występowania obiektów odległych w przypadku próbek pochodzących bezpośrednio z procesu produkcyjnego jest większe niż wtedy gdy do budowy modelu użyte są próbki przygotowane w laboratorium. Lepsze parametry dopasowania do danych i przewidywania uzyskane dla modeli stabilnych (w porównaniu z metodami PLS i N-PLS) potwierdziły powyższe założenie i umożliwiły kontrolę parametrów jakości w oparciu o nieselektywne sygnały – fluorescencyjne obrazy.

## 4.2.1.3. N-modalna regresja częściowych najmniejszych kwadratów

Innym typem modelu w trakcie konstrukcji którego wykorzystuje się kryterium częściowych najmniejszych kwadratów jest jego N-modalna wersja [20], dedykowana modelowaniu danych, w których zmienne niezależne są tensorem (mają wymiarowość danych większą niż dwa, czyli są drugiego lub wyższego rzędu). Podobnie jak w przypadku klasycznego modelu PLS, regresja zmiennej objaśnianej ( $\mathbf{y}$ ) dokonywana jest w oparciu o zmienne ukryte, utworzone wg. zasady maksymalizacji wariancji zawartej w zmiennych objaśniających i kowariancji pomiędzy zmiennymi objaśniającymi i objaśnianą. W notacji wektorowo macierzowej model N-PLS można przedstawić w następujący sposób:

$$\mathbf{X}_{[m,n \times o]} = \mathbf{T}_{[m,f]}(\mathbf{W}_{[o,f]} |\otimes| \mathbf{W}_{[n,f]})^T + \mathbf{E}_{[m,n \times o]} \quad (14)$$

$$\mathbf{y}_{[m,1]} = \mathbf{T}_{[m,f]} \mathbf{b}_{[f,1]} + \mathbf{e}_{[m,1]} \quad (15)$$

gdzie,  $\mathbf{X}_{[m,n \times o]}$  to rozwinięta postać tensora  $\mathbf{X}_{[m,n,o]}$  zmiennych objaśniających,  $\mathbf{T}_{[m,f]}$ ,  $\mathbf{W}_{[n,f]}$  oraz  $\mathbf{W}_{[o,f]}$  to macierze zawierające  $f$  triad (na które zdekomponowana została macierz  $\mathbf{X}$ ),  $|\otimes|$  reprezentuje iloczyn Khatri-Rao,  $\mathbf{E}_{[m,n \times o]}$  oraz  $\mathbf{e}_{[m,1]}$  zawierają reszty od modelu,  $\mathbf{y}_{[m,1]}$  to zmienna objaśniana,  $\mathbf{b}_{[f,1]}$  to wektor zawierający współczynniki regresji modelu.

W przypadku fluorescencyjnych obrazów, metody N-modalne pozwalają uwzględnić podczas konstrukcji modelu kalibracyjnego trójliniową strukturę danych. Trójliniową strukturę danych determinuje:

- liniowa zależność sygnału analitu od jego stężenia;
- dwuliniowość sygnału (ang. *bilinearity*) dla każdej próbki (tzn. sygnał może być przedstawiony jako iloczyn dwóch macierzy, które pomnożone przez siebie odtworzą ten sygnał);
- stałość profili analitów dla różnych próbek (tzn. kształt profilu dla pojedynczego komponentu jest identyczny we wszystkich rejestrowanych sygnałach, a intensywność sygnału zależy od stężenia tego analitu) [28].

Często, modele chemometryczne konstruowane na podstawie fluorescencyjnych obrazów posiadają zaletę drugorzędowości, tzn. nie są czułe na obecność ewentualnych interferentów w badanych próbkach. Metody PARAFAC i N-PLS wykorzystują dekompozycję trójliniowej struktury danych dzięki czemu zaleta drugorzędowości jest zachowana. Jednak nie każdy model skonstruowany na podstawie fluorescencyjnych obrazów będzie posiadał zaletę drugorzędowości. Modelowanie rozwiniętych widm metodą PLS często prowadzi do modeli

kalibracyjnych o bardzo dobrych właściwościach predykcyjnych. Niestety zaburzenie trójliniowej struktury danych (przez ich rozwinięcie) powoduje utratę zalety drugorzędowości.

N-modalną regresję częściowych najmniejszych kwadratów wykorzystałam do konstrukcji modeli dla fluorescencyjnych obrazów zestawionych w trójwymiarowe tensory. W każdym przypadku parametry modeli N-PLS porównałam z parametrami koncepcyjnie prostszej metody PLS, stosowanej dla rozwiniętych sygnałów dwuwymiarowych. Porównania dotyczące efektywności modeli klasycznych i N-modalnych przedstawiłam w pięciu publikacjach zawartych w cyklu, w których omawiam następujące zagadnienia:

- w **publikacji I** modelowanie jakości cukru, ocenionej jako jego kolor i zawartość popiołu;
- w **publikacjach II i III** ocenę jakości oleju napędowego, którą definiują wymagania stawiane przez Ministerstwo Finansów Rzeczypospolitej Polskiej;
- w **publikacjach IV i V** ocenę jakości produktów spożywczych wyrażoną jako zawartość związków o właściwościach antyoksydacyjnych lub polifenoli.

#### 4.2.1.4. Regresja częściowych najmniejszych kwadratów z eliminacją zmiennych nieistotnych.

Kolejnym z wariantów klasycznej metody PLS jest metoda umożliwiająca eliminację zmiennych nieistotnych – UVE-PLS (ang. *uninformative variable elimination*) [29]. Metoda ta ułatwia eliminację zmiennych niezależnych, które charakteryzują się wysoką wariancją, ale małą kowariancją z modelowaną zmienną zależną. Dzięki temu kompleksowość modelu zbudowanego na podstawie zmiennych istotnych jest zazwyczaj niższa od tego, który został zbudowany dla danych oryginalnych, przez co interpretacja uzyskanych wyników staje się bardziej przejrzysta.

W trakcie tworzenia modelu UVE-PLS do oryginalnych zmiennych objaśniających (zawartych w zbiorze modelowym) wprowadzane są nieistotne zmienne o rozkładzie normalnym i bardzo małej amplitudzie. Ze zmodyfikowanego zbioru modelowego, w sposób iteracyjny, usuwane są kolejne obiekty. Dla pozostałych obiektów obliczane są współczynniki regresji odpowiadające modelowi o optymalnej kompleksowości. Na podstawie współczynników regresji uzyskanych w kolejnych iteracjach obliczana jest ich stabilność (wyrażana jako stosunek wartości średniej współczynników regresji do ich odchylenia standardowego). Porównując stabilności współczynników dla zmiennych oryginalnych z tymi obliczonymi dla dodanych zmiennych nieistotnych wybierane są zmienne istotne (mające bezwzględne wartości stabilności większe niż wartość progowa ustalona na podstawie sztucznie dodanych zmiennych). Następnie, używając wyłącznie zmiennych istotnych, buduje się model PLS, który często charakteryzuje się niższymi (lub porównywalnymi) wartościami RMSE, RMSEP oraz kompleksowością w porównaniu do modelu zbudowanego dla oryginalnych danych.

Metodę UVE-PLS wykorzystałam w analizie problemu opisanego w **publikacji VI**. Przedstawiłam możliwość oceny jakości wyrażonej jako całkowita zdolność antyoksydacyjna naparów herbaty rooibos na podstawie danych chromatograficznych dwóch rodzajów: 1) tabeli pików zawierającej informacje o ilościowej zawartości 12 związków fenolowych uznawanych za posiadające właściwości antyoksydacyjne oraz 2) nieselektywnych chromatogramów. Zastosowanie metody UVE-PLS umożliwiło wskazanie jednego związku polifenolowego (witeksyny), który nie jest istotny podczas modelowania wybranego parametru w związku z czym jego ilościowe oznaczenie może w kolejnych badaniach zostać pominięte. Dodatkowo eliminacja nieistotnych zmiennych z nieselektywnych chromatogramów umożliwiła wskazanie pięciu pików chromatograficznych reprezentujących dodatkowe substancje (nie ujęte w tabeli pików), które mogą być potencjalnymi markerami zdolności antyoksydacyjnych i na zawartość których należy zwrócić uwagę w kolejnych badaniach. Wstępna analiza dodatkowych pików wykazała że są wśród nich np. aspalalinina i pochodne eriodicytolu.

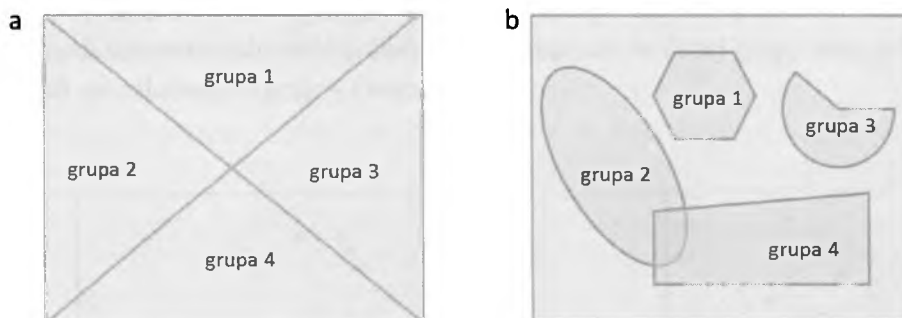
#### 4.2.2. Klasyfikacja/dyskryminacja próbek

Metody klasyfikacyjne/dyskryminacyjne mają na celu ustalenie przynależności próbki do z góry zdefiniowanych grup próbek, na podstawie utworzonych reguł logicznych. Przykładowe problemy badawcze, do rozwiązania których można użyć metod klasyfikacyjnych lub dyskryminacyjnych to np.:

- ocena stanu zdrowia pacjenta, dokonywana na podstawie chromatogramu zarejestrowanego dla wydychanego przez niego powietrza. W oparciu o skład chemiczny tworzy się reguły logiczne i dyskryminuje próbki oddechu pacjentów tzn. przypisuje do jednej ze zdefiniowanych grup: zdrowy lub chory [30];
- ocena pochodzenia geograficznego próbek żywności. Na podstawie badań składu chemicznego ocenia się przynależność produktów do jednej, kilku lub żadnej ze zdefiniowanych grup określających pochodzenie geograficzne [31].

Reguły logiczne, na podstawie których próbki zostaną przypisane do wybranych grup mogą być konstruowane aby zapewnić dyskryminację lub modelować poszczególne grupy próbek. W pierwszym przypadku, cała przestrzeń eksperymentalna (opisana przez próbki zawarte w zbiorze danych) jest dzielona na założoną liczbę wzajemnie wykluczających się podprzestrzeni (definiujących obszary grup próbek), rysunek 19 a). Wtedy próbka zostanie przypisana tylko do jednej z możliwych grup. W drugim przypadku, każdej z grup przypisany jest fragment przestrzeni eksperymentalnej z możliwością pokrywania się fragmentów przynależących do dwóch różnych grup. Pozostawiona jest także przestrzeń eksperymentalna, która nie należy do żadnej z nich, rysunek 19 b). Wówczas próbka może być przypisana do jednej, kilku lub żadnej z rozważanych grup.





**Rysunek 19** Przykłady podziału przestrzeni eksperymentalnej ze względu na sposób tworzenia reguł logicznych: a) dyskryminacyjny, b) modelowanie poszczególnych grup.

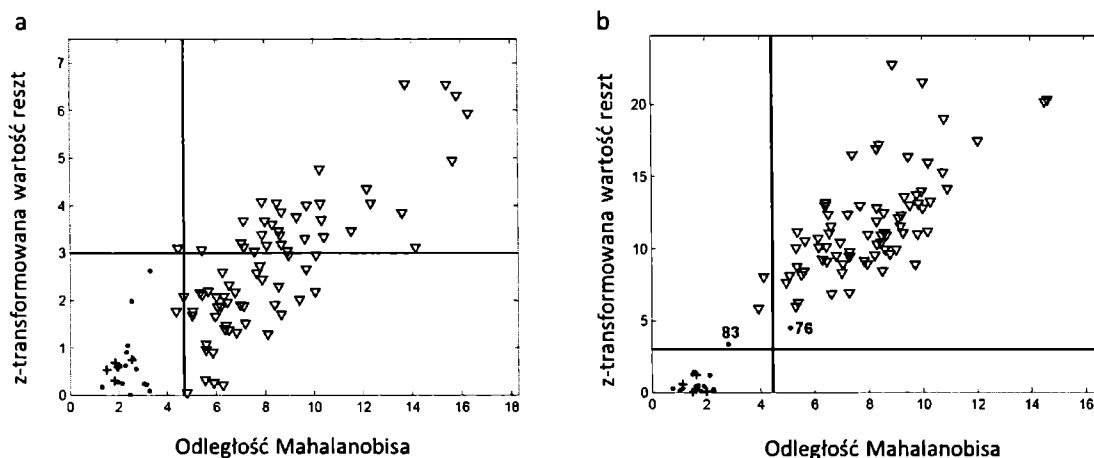
Używając sygnałów nieselektywnych, modele klasyfikacyjne/dyskryminacyjne konstruuje się wykorzystując takie chemometryczne techniki jak np.: liniowa analiza dyskryminacyjna (ang. *linear discriminant analysis*, LDA) [32], dyskryminacyjna metoda regresji częściowych najmniejszych kwadratów (ang. *discriminant partial least squares*, D-PLS) [33], drzewa klasyfikacji i regresji (ang. *classification and regression trees*, CART) [34] czy metoda modelowania indywidualnych grup próbek (ang. *soft independent modelling of class analogies*, SIMCA) [35].

#### 4.2.2.1. Metoda modelowania indywidualnych grup – SIMCA

SIMCA jest przykładem metody klasyfikacyjnej, w której dla każdej z rozpatrywanych grup tworzony jest odrębny model tzn. odrębny zestaw reguł, który umożliwia ocenę przynależności nowej próbki do danej grupy. Oceny dokonuje się na podstawie parametrów ją opisujących (np. zestaw różnych parametrów fizykochemicznych jak pH, zawartość wybranych substancji czy barwa lub parametrów zawartych w sygnale instrumentalnym).

Reguły klasyfikacji w metodzie SIMCA konstruuje się w oparciu o parametry modeli PCA, budowanych osobno dla zbiorów modelowych reprezentujących każdą rozpatrywaną grupę próbek. W przestrzeni (wybranego) modelu PCA o optymalnej liczbie czynników (ustalonej np. na drodze walidacji krzyżowej) definiuje się kryteria klasyfikacyjne, decydujące o przynależności próbek do danej grupy. Jednym kryterium klasyfikacyjnym jest wartość reszt od modelu PCA o wybranej kompleksowości, drugim natomiast odległości Mahalanobisa od środka danych w przestrzeni tego modelu. Wartości progowe wyznaczające granice danej grupy ustala się dla założonego poziomu prawdopodobieństwa poprawnego przypisania do tej grupy, korzystając z wybranych wartości stabilizowanych (np. zakładając normalny rozkład reszt od modelu PCA i odległości w przestrzeni tego modelu). Nowa próbka będzie należała do danej grupy jeżeli wartości wybranych współczynników klasyfikacyjnych, obliczonych w przestrzeni modelu PCA o założonej kompleksowości będą mniejsze niż ustalone wartości progowe. SIMCA należy do metod, w których stosuje się tzw. „miękkie” reguły klasyfikacji. Oznacza to, iż nowa próbka może zostać przypisana do jednej, kilku lub żadnej z rozpatrywanych grup.

Oceny jakości modelu klasyfikacyjnego SIMCA dokonuje się na podstawie niezależnych zbiorów testowych stanowiących zbiór próbek przynależących do danej grupy oraz próbki, które należą do innych modelowanych grup – rysunek 20.

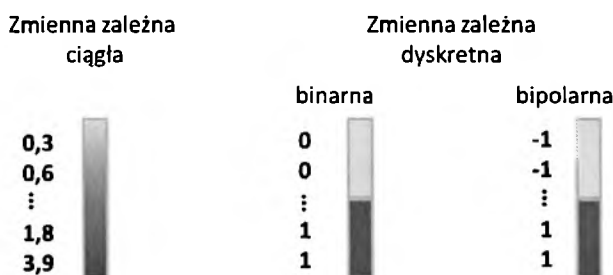


**Rysunek 20** Wyniki klasyfikacji próbek dokonanej metodami a) SIMCA i b) jej stabilnym wariantem [31]. Próbki należące do zbioru modelowego oznaczono jako ( $\bullet$ ), próbki należące do zbioru testowego modelowanej grupy oznaczono jako ( $+$ ), a próbki należące do zbioru testowego zawierającego wszystkie próbki przynależące do innych klas niż modelowana oznaczono jako ( $\nabla$ ). Dla odległości Mahalanobisa wartość progową ustalono na podstawie wartości krytycznej rozkładu  $\chi^2$ , dla liczby stopni swobody równej kompleksowości modeli klasyfikacyjnych oraz poziomu ufności 99,9%.

Rysunek 20 został przedstawiony w publikacji pt. *Improvement of classification using robust soft classification rules for near-infrared reflectance spectral data* [31], której jestem współautorem. Metoda SIMCA i jej stabilny wariant (nieczuły na obecność obiektów odległych) zostały wykorzystane do identyfikacji sposobu karmienia świń na podstawie zmian obserwowanych w wątrobie będących konsekwencją stosowanej podczas hodowli karmy. Klasyfikacji dokonano na podstawie nieselektywnych sygnałów instrumentalnych – widma z zakresu bliskiej podczerwieni – zarejestrowanych dla próbek wątroby. Skonstruowane modele SIMCA umożliwiły odróżnienie trzech sposobów hodowli, a w szczególności karmy podawanej zwierzętom (karma zawierała dodatki: oleju rzepakowego, mieszaniny olejów rzepakowego i sojowego, oraz nielegalny dodatek – tłuszcz pochodzenia zwierzęcego). Zagadnienie będące przedmiotem publikacji jest istotne ze względu na regulacje chowu zwierząt przyjęte przez Unię Europejską.

#### 4.2.2.2. Dyskryminacyjny wariant regresji częściowych najmniejszych kwadratów

W pierwszym zamyśle, metoda PLS została zaproponowana do modelowania zależności liniowych pomiędzy zmiennymi niezależnymi  $X$  i zmienną zależną  $y$  o charakterze ciągłym. Później zaczęto stosować tę metodę do modelowania zależności, w których zmienna  $y$  ma charakter dyskretny – binarny (przyjmuje wartości 0 lub 1) lub bipolarny (przyjmuje wartości 1 lub -1) – rysunek 21.



Rysunek 21 Zmienna zależna o charakterze ciągłym i dyskretnym.

W takim przypadku, zmienna zależna zawiera informacje o przynależności opisywanych próbek do jednej z możliwych grup, a model PLS skonstruowany na jej podstawie umożliwi dyskryminację obiektów opisanych przez zmienne niezależne. Uogólniając, zadaniem dyskryminacyjnego wariantu regresji częściowych najmniejszych kwadratów – D-PLS jest konstrukcja modelu regresji w oparciu o nowe zmienne maksymalizujące różnice pomiędzy próbkami przynależącymi do różnych grup.

W moich badaniach metodę D-PLS wykorzystywałam do konstrukcji modeli dyskryminujących próbki oleju napędowego ze względu na proceder nielegalnego usunięcia dodatków akcyzowych (barwnika – Solvent Red 19 i znacznika – Solvent Yellow 124), opisaną w publikacji III. Do budowy modeli użyłam fluorescencyjne obrazy zarejestrowane dla próbek oleju napędowego, który został poddany symulacji usunięcia komponentów akcyzowych na drodze sorpcji. Ze względu na obowiązujące w Polsce przepisy prawne i przyjęty plan eksperymentu skonstruowałam cztery bipolarne zmienne dyskretny, zawierające informacje o przynależności do jednej z dwóch możliwych grup – paliwo o obniżonej i o regularnej wysokości podatku akcyzowego. Otrzymane modele dyskryminacyjne umożliwiły rozróżnienie próbek zawartych w niezależnych zbiorach testowych na dwie grupy ze skutecznością od 77% do 100%.

Jakość modelu klasyfikacyjnego lub dyskryminacyjnego oceniana jest na podstawie parametrów uwzględniających liczbę poprawnie (lub niepoprawnie) zakwalifikowanych do danej grupy próbek osobno dla zbiorów modelowego i testowego. Najpopularniejszym parametrem jest stopień poprawnej dyskryminacji (ang. *correct discrimination rate*) wyrażający liczbę poprawnie przypisanych do wybranej grupy próbek (najczęściej wyrażaną jako wartość procentowa w odniesieniu do wszystkich ocenianych próbek). Inne parametry

charakteryzujące jakość modelu dyskryminacyjnego (lub klasyfikacyjnego) to jego czułość (nazywana również wskaźnikiem liczby próbek prawdziwie pozytywnych, ang. *true positive rate*) lub specyficzność (nazywana również wskaźnikiem liczby próbek prawdziwie negatywnych, ang. *true negative rate*). Czułość wyrażona jest jako stosunek liczby próbek poprawnie przypisanych do danej grupy (ang. *true positive*) do liczby wszystkich próbek które oceniono jako przynależące do tej grupy. Natomiast specyficzność to stosunek liczby próbek źle przypisanych do grupy (ang. *true negative*) do liczby wszystkich próbek, które rozpoznano jako przynależące do danej grupy z użyciem modelu.

### 4.3. Algorytmy obliczeniowe użyte w pracy

Wszystkie obliczenia wykonane w ramach niniejszej pracy wykonałam używając algorytmów zaimplementowanych w środowisku obliczeniowym Matlab.

Do przygotowania danych do analizy wykorzystałam: algorytm opracowany przez Zeppa i współpracowników [16] (do korekcji sygnałów fluorescencyjnych) oraz algorytm zoptymalizowanego nakładania sygnałów maksymalizujący ich wzajemną korelację, COW [36] (do korekcji sygnałów chromatograficznych).

Do konstrukcji modeli eksploracyjnych, dyskryminacyjnych i kalibracyjnych wykorzystałam pakiet TOMCAT [37] oraz pakiet N-way toolbox [38].

W miarę potrzeb dokonywałam samodzielnych modyfikacji dostępnych algorytmów, przygotowywałam własne skrypty obliczeniowe lub korzystałam z algorytmów udostępnionych przez współpracowników.

## 5. Nieselektywne sygnały i podejścia chemometryczne, a kontrola jakości

Kontrola jakości wybranych produktów to zagadnienie, któremu trudno jest przypisać sztywne ramy. Ze względu na różnorodność produktów i parametrów charakteryzujących ich jakość, jej ocena wymaga użycia rozmaitych rozwiązań analitycznych. Z uwagi na liczne zalety jakie oferują metody opierające się na rejestracji nieselektywnych sygnałów analitycznych oraz odpowiednio zastosowane metody chemometryczne są one często wykorzystywane do kontroli jakości.

### 5.1. Kontrola jakości produktów spożywczych

Jest to jeden z obszarów badawczych, w którym omawiane podejście do analizy jakości jest niezwykle szeroko wykorzystywane. W ramach weryfikacji jakości produktów spożywczych rozpatruje się szereg cech fizyko-chemicznych, które o niej decydują. Najważniejszym zagadnieniem jest **spełnianie wymogów** sanitarno-epidemiologicznych i tych, które odnoszą się do składu chemicznego żywności (zwłaszcza zawartości konserwantów, dodatków „polepszających” smak i zapach, a także różnorodnych wypełniaczy) stosowanych na terenie danego państwa lub wspólnot np. Unii Europejskiej. W literaturze naukowej przedstawiono wiele metod dedykowanych kontroli nielegalnych dodatków wprowadzanych do żywności. Przykładowo, zawartość nielegalnego barwnika – sudan-I, wprowadzanego do różnych produktów spożywczych, oceniana jest na podstawie widm spektroskopii Ramana [39]. Wykorzystując widmo UV-Vis i metody chemometryczne można równocześnie ocenić ilościową zawartość pięciu innych barwników wprowadzanych do żywności [40]. Nie mniej ważny jest **skład chemiczny żywności** informujący konsumenta o jej jakości np. informacja o zawartości wybranych składników odżywczych. Szeroko stosowane są połączenia chromatografii lub spektroskopii w zakresie bliskiej lub średniej podczerwieni i metod konstrukcji wieloparametrowych modeli kalibracyjnych do oceny całkowitej zawartości antyoksydantów, np. [41 – 43]. Te same techniki spektroskopowe używane są do oceny zawartości wody i kwasów organicznych w pomidorach [44], zawartości kwasów tłuszczowych trans w olejach jadalnych [45] czy całkowitej zawartości tłuszczu w majonezach i sosach sałatkowych (bez konieczności przygotowania próbki i wyjmowania jej z opakowania) [46]. Fluorescencyjne obrazy służą z kolei do oceny takich parametrów jak zawartość tokoferoli w olejach [3] czy ryboflawiny i aminokwasów zawierających pierścienie aromatyczne w piwie [47], jak również do oceny jakości różnych produktów zaliczanych do nabiału [48]. Inne substancje, których zawartość jest wyznacznikiem jakości ocenianej subiektywnie przez konsumenta (ze względu na obniżenie kaloryczności jedzenia czy walory smakowe), to tzw. substancje słodzące. Jedną z nich jest acesulfam-K którego ilość w produktach uznawanych za dietetyczne można ocenić w sposób szybki i niedestruktywny używając widm w zakresie podczerwieni modelowanych metodą PLS [49]. Zawartość witamin w suplementach diety i napojach energetycznych jest również subiektywnym (z punktu widzenia konsumenta) wyznacznikiem jakości, który można oceniać używając widm absorpcyjnych UV-Vis [50].

Kolejnym ważnym zagadnieniem w kontekście oceny jakości produktów spożywczych jest **weryfikacja ich pochodzenia geograficznego** [51] i często za nią idące **różnice w jakości produktów** oferowanych przez producentów z różnych krajów lub regionów świata. Możliwość identyfikacji miejsca produkcji czy pochodzenia danego towaru jest ceniona zwłaszcza ze względu na regulacje Unii Europejskiej i innych instytucji, które przyznają między innymi unikatowym regionalnym produktom spożywczym ochronę prawną. Wyróżnia się takie znaki jak **PDO** odnoszący się do unikatowych produktów ściśle związanych z danym regionem (ang. *protected designation of origin*), **PGI** odnoszący się do produktów produkowanych w danym regionie geograficznym (ang. *protected geographical indication*), czy **TSG** odnoszący się do produktów charakteryzujących się tradycyjnym sposobem produkcji, składem lub wykonaniem z tradycyjnych surowców (ang. *traditional specialities guaranteed*) [52,53]. Ze względu na wprowadzaną ochronę prawną, weryfikacja pochodzenia żywności może również przyczynić się do **wykrycia procederu fałszowania** wybranych, unikalnych i chronionych produktów spożywczych. Na przykład, pochodzenie miódów polskich oceniano na podstawie widm jądrowego rezonansu magnetycznego poddanych kalibracji z użyciem metody PLS [54]. W ramach międzynarodowego projektu TRACE badano między innymi pochodzenie miódów wyprodukowanych w różnych krajach na podstawie ich chromatograficznych odcisków palca [55]. Z kolei geograficzne pochodzenie mleka może być ocenione przy użyciu modelu dyskryminacyjnego i fluorescencyjnego obrazu próbki [4]. Nie mniej ważnym zagadnieniem jest celowe **fałszowanie produktów** spożywczych nieobjętych specjalną ochroną prawną np. poprzez dodatek tańszych wypełniaczy [56]. Nieselektywne sygnały analityczne są wykorzystywane podczas oceny autentyczności wołowiny ze względu na dodatek wieprzowiny [57] lub koniny [58]. Widma spektroskopii Ramana znalazły zastosowanie do oceny zafałszowania masła innym niż zawartym w mleku tłuszczem zwierzęcym [59], a spektroskopia w zakresie podczerwieni pozwala wykryć dodatek margaryny w maśle [60] lub mleka sojowego w mleku krowim [61]. Analizując widma fluorescencyjne można kontrolować skład chemiczny olejów roślinnych [62] i oliwy z oliwek [63]. Sygnały instrumentalne rejestrowane w zakresie podczerwieni używane są do wykrycia procederu fałszowania szafranu [64] czy wprowadzania dodatku pszenicy lub kukurydzy do zmielonej kawy [65].

Jakość produktów ocenia się również ze względu na przemiany chemiczne zachodzące w żywności na przestrzeni czasu kontrolując **termin przydatności do spożycia** (ang. *shelf life*). Widma w zakresie podczerwieni wspomagają ocenę jakości owoców [66] i warzyw [67] w zależności od warunków ich przechowywania. Nieselektywne sygnały są podstawą konstrukcji modeli dyskryminacyjnych stosowanych do oceny świeżości mięsa wołowego [68], drobiowego [69] i ryb [70]. Widma fluorescencyjne były również pomocne w ocenie stopnia świeżości jogurtów [71]. Pomiary jakości wykonywane na przestrzeni czasu znalazły swoje zastosowanie także **podczas kontroli procesu produkcyjnego**. Ze względu na szybkość rejestracji widm w zakresie podczerwieni, które są źródłem informacji o jakości produktu i zarazem procesie, stosuje się je np. do oceny procesu dojrzewania piwa [72], wyrobu ciasta chlebowego [73] czy oceny tekstury sera topionego [74].

Podejmowane są również **próby zastąpienia analizy sensorycznej** (dokonywanej przez wykwalifikowanych materiałoznawców tworzących tzw. panele sensoryczne) przez metody oparte na sygnałach nieselektywnych wspieranych modelami chemometrycznymi np. do oceny parametrów jakości związanych z reakcjami Maillarda [75].

## 5.2. Ocena jakości środowiska naturalnego

To kolejny obszar, w którym nieselektywne sygnały analityczne i towarzyszące im rozwiązania chemometryczne znalazły swoje zastosowania. Są to zwłaszcza **badania jakości wód i gleb**. Wśród technik najczęściej używanych do badań próbek wód ważne miejsce znajduje spektroskopia fluorescencyjna w szczególności fluorescencyjne obrazy stosowane do oceny śladowych i ultraśladowych zawartości substancji organicznych [76–78], a zwłaszcza wielopierścieniowych węglowodorów aromatycznych [16,76,79]. Badania gleb pod kątem informacji o zanieczyszczeniach takich jak substancje organiczne [80], czy nieorganiczne (azotany) [81] najczęściej prowadzi się przy użyciu spektroskopii w zakresie bliskiej podczerwieni. Technikę tę wykorzystuje się także do jakościowej charakterystyki podłoża [82]. Zanieczyszczenia spowodowane pozostałościami paliw kontroluje się na podstawie modeli kalibracyjnych konstruowanych dla fluorescencyjnych obrazów [83] lub widm w zakresie bliskiej podczerwieni [84]. Opracowano również czułą technikę umożliwiającą ilościową ocenę zawartości pochodnych naftalenu w glebie na podstawie pomiarów fluorescencyjnych [85]. Dane analityczne uzyskane przy użyciu technik chromatograficznych, modelowane z użyciem wieloparametrowych metod kalibracyjnych są stosowane do kontroli jakości osadów rzecznych i ścieków np. ze względu na zawartość herbicydów [86].

Innym aspektem jakości środowiska naturalnego ocenianym poprzez rejestrację nieselektywnych sygnałów analitycznych są **badania bioindykatorów**. Gatunki zaliczane do tej grupy często występują w bardzo mało licznych populacjach. Metody oparte na pomiarze widm NIR i modelach kalibracyjnych [87] charakteryzują się większą czułością niż standardowe testy biologiczne używane do oceny liczebności bioindykatorów [88].

## 5.3. Analiza jakości produktów energetycznych

To problem analityczny, w którym nieselektywne sygnały rejestrowane głównie w zakresie bliskiej podczerwieni, są również intensywnie wykorzystywane do oceny parametrów różnych produktów zaliczanych do grupy paliw. Sygnały NIR stosuje się do oceny właściwości fizycznych paliwa lotniczego (np. temperatur zapłonu czy zamarzania) [89], w połączeniu z metodami kalibracyjnymi są także pomocne przy kontroli jakości mieszanin olejów napędowych [90] oraz węgla pod względem jego wykorzystania w elektrowniach [91]. Dzięki drugorzędowej strukturze danych, fluorescencyjne obrazy z powodzeniem są używane do badań mieszanin oleju napędowego i nafty [5] oraz paliwa lotniczego i oleju napędowego [92].

#### 5.4. Dziedziny nauk typu „omika”

Stosunkowo młode dziedziny nauk z grupy „omika” takie jak np. genomika, proteomika czy metoabolomika stają się coraz bardziej popularne, stanowiąc nowoczesne platformy dla realizacji m.in. badań medycznych i biotechnologicznych [93]. Analizę próbek w tych dziedzinach przeprowadza się wykorzystując metody instrumentalne generujące takie nieselektywne sygnały jak dwuwymiarowe żele elektroforetyczne [94] czy widma masowe pochodzące z pomiarów typu MS/MS [95]. Ekstrakcja użytecznej informacji zawartej w takich sygnałach nie byłaby możliwa, gdyby nie metody chemometryczne. Uzyskanie istotnej informacji z takich sygnałów jest trudne, zwłaszcza ze względu na ich dużą złożoność pod względem liczby analizowanych substancji, nakładanie się sygnałów analitów o podobnych właściwościach czy rozmiar danych. Dlatego, zaawansowana analiza danych i krytyczne podejście do uzyskanych wyników są w tego typu badaniach niezbędne [96].

Wydawać by się mogło, że wykorzystanie podejść „omika” do oceny jakości jest mało popularne. Jednak, w dobie rozwoju inżynierii genetycznej i wprowadzania na rynek konsumencki coraz większej ilości żywności modyfikowanej genetycznie wiedza na temat **genomu lub proteomu produktów pochodzenia roślinnego i zwierzęcego** jest potrzebna. Nazwana z języka angielskiego **fudomika** jest stosunkowo nową składową „omik”, która zajmuje się między innymi tego typu problemami analitycznymi [97]. W literaturze są opisane metody umożliwiające ocenę czy próbka żywności jest transgeniczna [98], jak również pozwalające kontrolować zawartość takich związków jak pestycydy, alergeny czy antybiotyki [99].

Innym obszarem aplikacyjnym podejść typu „omika” jest **kontrola procesów biofarmaceutycznych**, w których jakość stosowanych leków pochodzenia biologicznego lub uprawianych roślin ma kluczowe znaczenie w zrozumieniu kontrolowanych mechanizmów biologiczno-chemicznych [100].

#### 5.5. Ocena jakości w badaniach farmakologicznych

Nieselektywne sygnały stosuje się przede wszystkim, do oceny **jakości leków** zawierających pojedyncze substancje aktywne [101,102] jak i tych wieloskładnikowych, charakteryzujących się matrycą o skomplikowanym składzie [103]. Spektroskopia w zakresie bliskiej podczerwieni cieszy się największym zastosowaniem w analizie jakości leków [104]. Wykorzystując tę technikę opracowano rozwiązania takich problemów badawczych jak kontrola procesu produkcji leków i ocena ilościowa substancji aktywnej w matrycy w formie tabletek [105] czy mieszanie i ujednorodnianie leków w postaci proszków [106]. Kolejne ważne problemy występujące podczas produkcji leków, do których spektroskopia NIR ma swoje zastosowanie, to ocena zawartości wilgoci w tabletkach [107], a także określenie postaci polimorficznych substancji aktywnych [108,109].



Wysokosprawną chromatografię cieczową z odpowiednio dobranym sposobem detekcji jest z kolei szeroko stosowana do **oceny jakości ziół** o zastosowaniu leczniczym [109], do **charakterystyki leków stosowanych w medycynie chińskiej** [110], a także do oceny czystości leków [111].

Fluorescencyjne obrazy dzięki własności drugorzędowości, która uwidacznia się podczas konstrukcji modeli kalibracyjnych, są z powodzeniem stosowane do **kontroli dynamiki procesu uwalniania i dystrybucji leków** w ludzkim organizmie. Oceny takiej dokonuje się w oparciu o próbki płynów ustrojowych, które mają bardzo bogatą matrycę. W serum krwi bada się uwalnianie i metabolity leków przeciwnowotworowych [112], lub przeciwkaszlowych [113], a także zawartość innych substancji biologicznie aktywnych jak popularny antyoksydant resweratrol [114]. W moczu z kolei oznaczano np. zawartość leków stosowanych w leczeniu chorób psychicznych [115], chorób serca [116] czy popularnego leku przeciwzapalnego – paracetamolu [117]. Fluorescencyjne obrazy znalazły również zastosowanie w ocenie ilościowej zawartości testosteronu w kosmetykach [118].

Prowadzone są także badania nad wykorzystaniem nieselektywnych sygnałów w **diagnostyce chorób i poszukiwaniach potencjalnych biomarkerów** wybranych schorzeń [119 – 121].

Duża liczba publikacji poświęconych użyciu nieselektywnych sygnałów i technik chemometrycznych w kontekście kontroli jakości, świadczy o potrzebie tego typu rozwiązań. Najczęściej, do opracowywania nowych metodyk analityki procesowej (dział nauki opisujący zagadnienia związane z kontrolą procesów produkcyjnych) wykorzystywane są sygnały instrumentalne uzyskiwane dzięki technikom niedestrukcyjnym, które można zastosować bezpośrednio w analizach typu „on-line” (np. spektroskopie fluorescencyjna, NIR). Ze względu na stały wzrost wymagań stawianych różnorodnym produktom, jak również na konieczność dokonywania analiz jakości w odniesieniu do badań z różnorodnych dziedzin np. środowiskowych, biologicznych czy kryminalistycznych, chemometryczne opracowanie sygnałów nieselektywnych jest zagadnieniem niezwykle istotnym.

## 6. Cel pracy

**Celem mojej pracy było efektywne wykorzystanie informacji zawartej w nieselektywnych sygnałach analitycznych poprzez ich przetwarzanie z użyciem narzędzi chemometrycznych w kontekście rozwiązywania problemów szeroko pojętej chemii analitycznej.**

Wśród nich są problemy związane z kontrolą jakości wybranych produktów spożywczych oraz badaniem jakości i autentyczność produktu jakim jest olej napędowy.

W trakcie moich badań, opracowałam szereg rozwiązań analitycznych użytecznych do:

- oceny jakości cukru ze względu na jego zabarwienie i zawartość popiołu,
- oceny jakości oleju napędowego o obniżonej akcyzie ze względu na zgodność poziomu stężenia dodatków akcyzowych (Solvent Red 19 i Solvent Yellow 124) z wymaganiami stawianymi przez rozporządzenia Ministerstwa Finansów oraz możliwość poddania próbki oleju procedurze nielegalnego usuwania dodatków akcyzowych na drodze sorpcji,
- oceny jakości różnych produktów spożywczych (kawa, trzy rodzaje ziół, pasty pomidorowe, suplementy diety oraz herbata typu rooibos) wyrażonej jako całkowita zawartość antyoksydantów.

By osiągnąć ustalony cel zaplanowałam i przeprowadziłam eksperymenty konieczne do pozyskania zbioru danych opisujących poszczególne problemy badawcze. W dwóch przypadkach, skorzystałam z zestawów danych eksperymentalnych otrzymanych od zespołów badawczych z Kopenhagi i Sellenbosch. Następnie, dane przygotowywałam do analizy i modelowałam używając różnych metod chemometrycznych, adekwatnych do rozwiązania wybranego problemu badawczego.

Ze względu na atrakcyjne właściwości spektroskopii fluorescencyjnej i zalety charakteryzujące fluorescencyjne obrazy (w porównaniu z widmem emisyjnym rejestrowanym dla pojedynczej fali wzbudzenia) są to nieselektywne sygnały analityczne, na których skupiam największą uwagę w mojej pracy. Nie mniej jednak, zastosowanie innych nieselektywnych sygnałów (np. chromatogramów czy widm rejestrowanych w zakresie podczerwieni) w kontekście kontroli jakości również było obiektem moich badań.

Nieprzypadkowo, wybrane do badań produkty charakteryzują się skomplikowaną matrycą chemiczną. **Celem przeprowadzonych badań było również udowodnienie, że odpowiednio zastosowane narzędzia chemometryczne umożliwiają uzyskanie istotnej informacji, nawet jeżeli analizowany sygnał jest złożony.** Dlatego, wykorzystywane przeze mnie nieselektywne sygnały analityczne oprócz poszukiwanej informacji o jakości zawierają dodatkowe informacje, nie zawsze istotne z punktu widzenia modelowanego parametru.

## 7. Opublikowane badania własne

W tym rozdziale charakteryzuję zrealizowane przeze mnie badania, stanowiące podstawę niniejszej pracy doktorskiej. Wyniki badań zostały opisane w sześciu artykułach, które opublikowano w czasopismach z Listy Filadelfijskiej. Sumaryczna wartość współczynników *Impact Factor* wszystkich publikacji wynosi 19,345\*.

Poniżej, przedstawiam krótki opis badań w nich zawartych oraz oryginalne publikacje wraz z deklaracjami wkładu wszystkich współautorów, którzy przyczynili się do ich powstania.

\*Wartości współczynników *Impact Factor* dla roku w którym wydano publikację.

## 7.1. Publikacja I

### ***Kontrola jakości cukru na podstawie fluorescencyjnych odcisków palca i stabilnej kalibracji***

---

Tytuł oryginalny: ***Controlling sugar quality on the basis of fluorescence fingerprints using robust calibration***

Autorzy: Joanna Orzeł  
Michał Daszykowski  
Beata Walczak

Czasopismo: Chemometrics and Intelligent Laboratory Systems

Rok wydania: 2012

Tom: 110

Strony: 89 – 93

Wartość  
współczynnika  
*Impact Factor* : 2,291

---

W **publikacji I** wykorzystałam możliwość zastosowania fluorescencyjnych obrazów w kontekście kontroli jakości cukru w trakcie procesu produkcyjnego. Nieselektywne widma fluorescencyjne wykorzystałam do oceny dwóch wyznaczników jakości cukru – jego barwy i zawartości popiołu. Metody chemometryczne, które użyłam do konstrukcji modeli kalibrujących wybrane parametry jakości to podstawowy (klasyczny), stabilny i N-modalny wariant regresji częściowych najmniejszych kwadratów. Użycie stabilnej metody konstrukcji modeli kalibracyjnych (PRM) umożliwiło włączenie do modeli większej liczby źródeł zmienności próbek. Dzięki temu zarówno błędy dopasowania jak i przewidywania, które charakteryzują modele stabilne, skonstruowane do oceny koloru i zawartości popiołów są niższe w stosunku do parametrów charakteryzujących modele PLS i N-PLS. Stosując proponowane w publikacji rozwiązanie można ocenić zabarwienie cukru z błędem 3,24%, natomiast zawartość popiołu z błędem 4,37%. Uzyskanie wysokiej dokładności oceny parametrów jakości cukru sprawia, że proponowane rozwiązanie może znaleźć zastosowanie praktyczne.



Contents lists available at SciVerse ScienceDirect

## Chemometrics and Intelligent Laboratory Systems

journal homepage: [www.elsevier.com/locate/chemolab](http://www.elsevier.com/locate/chemolab)

## Controlling sugar quality on the basis of fluorescence fingerprints using robust calibration

J. Orzel, M. Daszykowski\*, B. Walczak

Department of Analytical Chemistry, Chemometric Research Group, Institute of Chemistry, The University of Silesia, 9 Szkolna Street, 40-006 Katowice, Poland

### ARTICLE INFO

#### Article history:

Received 11 July 2011

Received in revised form 28 September 2011

Accepted 3 October 2011

Available online 8 October 2011

#### Keywords:

Process analytical technology

Partial least squares

Partial robust M-regression

N-PLS

### ABSTRACT

The aim of our study was to highlight the benefits of robust calibration in the context of process control. Two properties were monitored – the color and ash content of sugar samples. It was shown for the data being studied that robust models, constructed using the partial robust M-regression technique, have a better fit to the majority of the data and prediction properties than the classic partial least squares and N-way partial least squares models. In particular, the constructed calibration models were characterized by a root mean square errors improved by 1.60% and 1.82% and a root mean square errors of prediction (for independent test samples) improved by 2.39% and 1.11% compared to classic partial least squares models constructed for color and ash content, respectively.

© 2011 Elsevier B.V. All rights reserved.

### 1. Introduction

Nowadays, the quality of a given food product is a significant criterion that is considered by consumers. A high product quality implies its safety and satisfactory sensory properties. The safety of food products is controlled by authorized laboratories on the basis of regulations for describing the attributes of food products. These regulations are established by local and international organizations e.g. the International Standardization for Organization (ISO), the Quality Control Council of the United States, the Polish Standardization Committee, etc. On the other hand, sensory evaluation is often more definitive than the laboratory one because some product features such as color, texture and smell can be judged by consumers themselves. Therefore, it is very important to control the final product and its production process in order to obtain products that are virtually identical from batch to batch. This philosophy is the core of process analytical technology (PAT). PAT efficiently combines knowledge of analytical chemistry and the use of chemometric tools to design and monitor the quality of manufactured products. The PAT methodology is greatly appreciated in many branches of industry, including the food, pharmaceutical and chemical industries [1–5]. The sugar industry is one of many examples where the PAT concepts are exercised extensively [6,7]. The main properties that are important for sugar quality assessment are its color and purity (defined as the total amount of organic and inorganic impurities). To monitor changes in these two parameters, multivariate calibration models such as

principal component regression (PCR) [8] and partial least squares regression (PLS) [8] can be used.

Recently, there has been a steadily growing interest in the use of chemical fingerprints for describing different processes and their end products for the purpose of process control. Spectroscopic techniques, such as near-infrared spectroscopy (NIR) and fluorescence spectroscopy have gained a great deal of attention in the field of PAT. They are relatively inexpensive, allow for the rapid acquisition of spectroscopic fingerprints (also on-line) and are non-destructive techniques.

In this study, the use of fluorescence fingerprints in PAT is explored. Excitation and emission fluorescence spectra of samples (the so-called fluorescence landscapes) that are collected simultaneously contain richer chemical information than a single excitation or emission fluorescence spectrum. Therefore, they seem to be attractive for the purposes of process control. To date, fluorescence fingerprints have been used to construct multivariate calibration models that are able to predict the riboflavin content in yogurt [9] and to monitor the deterioration of extra virgin olive oil during heating [10]. In the context of sugar quality control, fluorescence fingerprints have been used to construct calibration models to describe the color of raw cane sugar [11] and the physicochemical quality parameters of thick juice and beet sugar [12].

The presence of small amounts of impurities, such as phenols, amino acids, products of their reactions and inorganic residuals, in sugar is responsible for its quality. Therefore, their levels should be monitored. As illustrated in [13], owing to its high sensitivity, fluorescence spectroscopy can be used for this purpose.

In this study, the possibilities offered by the robust calibration technique, called partial robust M-regression (PRM), which enables

\* Corresponding author.

E-mail address: [michal.daszykowski@us.edu.pl](mailto:michal.daszykowski@us.edu.pl) (M. Daszykowski).

the color and purity of sugar to be controlled on the basis of the excitation–emission fluorescence fingerprints, were examined. It is argued that in the field of PAT, robust modeling techniques are highly desirable because process data may contain outlying samples due to new sources of variation, e.g. unstable process conditions, changes in the chemical composition of substrates, etc. The presence of outlying samples in process data affects the construction of all classic calibration models based on the least squares principle [14]. Robust calibration approaches are unaffected by outliers and additionally they allow outliers to be detected [15]. It is the conviction of the authors that robust models are valuable tools for PAT and that their use and role should be popularized and extended.

2. Theory

2.1. Partial least squares regression

The aim of partial least squares regression, PLS, is to describe the relationship between a set of explanatory variables, **X** and a response variable(s) **y**. In this study one continuous response variable, **y**, is modeled (the so-called PLS-1 variant). Using the PLS approach, it is possible to construct a calibration model for data containing strongly correlated variables. This problem is always met when the number of variables exceeds the number of samples, as in the modeling of spectroscopic fingerprints. In order to deal with this issue, during the PLS modeling explanatory variables are replaced by a number of latent factors (new variables), **T**, that maximize covariance with **y**. The PLS model can be described as:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E} \tag{1}$$

$$\mathbf{y} = \mathbf{T}\mathbf{q} + \mathbf{r} = \mathbf{X}\mathbf{b} + \mathbf{r} \tag{2}$$

where **X** is the data matrix containing *m* samples and *n* explanatory variables, elements of vector **q** are regression coefficients for PLS factors, **T**, **P** holds the PLS loadings, **E** are the differences between the observed and reconstructed **X**, **r** are the differences between the observed and predicted **y** and **b** is the vector of regression coefficients for explanatory variables, calculated as:

$$\mathbf{b} = \mathbf{W}(\mathbf{P}^T\mathbf{W})^{-1}\mathbf{q} \tag{3}$$

where **W** holds the loadings that maximize the covariance criterion and **P** is given as:

$$\mathbf{P} = \mathbf{X}^T\mathbf{X}\mathbf{W} \tag{4}$$

2.2. N-way partial least squares

The N-way partial least squares, N-PLS, is an extension of classic PLS regression for modeling higher order data arrays [16]. The three way data array, **X**, of size *I* × *J* × *K*, is first rearranged (unfolded) into a two-way data matrix of size *I* × *JK* and then decomposed into a set of so-called triads. Each triad contains one score vector and two weight vectors. The N-PLS model of data array **X** can be presented as:

$$\hat{x}_{ijk} = t_i w_j^l w_k^k \tag{5}$$

where, *i* = 1, 2, ..., *I*, *j* = 1, 2, ..., *J* and *k* = 1, 2, ..., *K*, **X** is three-way data array arranged as *samples* (*I*) × *emission spectra* (*J*) × *excitation spectra* (*K*), *t<sub>i</sub>* are scores, *w<sub>j</sub>* and *w<sub>k</sub>* are loadings.

During model construction weight vectors are optimized to obtain maximal covariance of score vector with response variable:

$$\max_{\mathbf{w}_a^j, \mathbf{w}_a^k} \{ \text{cov}(\mathbf{t}_a, \mathbf{y}) \} \min \sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K (x_{ijk} - t_{aj} w_{aj}^j w_{ak}^k)^2 \tag{6}$$

where **y** is response variable with dimensions (*I* × 1), *t<sub>a</sub>* are N-PLS scores (*I* × 1), *w<sub>a</sub><sup>j</sup>* (*J* × 1) and *w<sub>a</sub><sup>k</sup>* (*K* × 1) are N-PLS weight vectors for the *a*th model factor.

The number of factors can be selected using for example the leave-one-out cross-validation procedure.

2.3. Robust variant of the partial least squares regression

The goal of partial robust M-regression, PRM, is to describe the relationship represented by the majority of data between the explanatory variables and a response variable [17]. In PRM a weighting scheme is introduced in order to reduce the negative influence of outlying objects on the estimates of regression coefficients. The global weights, which are a product of leverage, *g<sub>i</sub><sup>\*</sup>*, and residual weights, *g<sub>i</sub><sup>r</sup>*, for objects are estimated iteratively. The leverage weight of the *i*th object is defined as:

$$g_i^* = \psi \left( \frac{\|\mathbf{t}_i - \text{L1med}(\mathbf{T})\|}{\text{med}_i \|\mathbf{t}_i - \text{L1med}(\mathbf{T})\|} \right) \tag{7}$$

and

$$\psi(z, c) = \frac{1}{(1 - |z|)^2} \tag{8}$$

where  $\|\bullet\|$  is the Euclidean norm, *t<sub>i</sub>* are the *f* PLS scores of the *i*th object, 'L1med' is the L1-median robust estimator of the data center [18],  $\psi$  is the weighting function called the "Fair" function and *c* is a tuning constant (in most applications *c* = 4) [19].

The residual weight of the *i*th object is given by the following expression:

$$g_i^r = \psi \left( \frac{r_i}{\sigma_{\text{MAD}}}, c \right) \tag{9}$$

where *r<sub>i</sub>* are the squared differences between the observed and predicted response value of the *i*th object,  $\sigma_{\text{MAD}}$  is the robust estimator of the data scale expressed as the median of absolute deviation around the median [19]:

$$\sigma_{\text{MAD}} = \text{med} \left| r_i - \text{med} (r_j) \right| \tag{10}$$

where 'med' is the median estimator of data center.

The PRM model is built iteratively. The weights of objects are recalculated until the algorithm's convergence is reached, i.e. when the difference between the norms of the regression coefficients of the PRM factors in two subsequent steps is relatively small, e.g. 10<sup>-2</sup>. The PRM model is constructed in the following steps:

0. Prior to construction of the PRM model, **X** is centered using either the coordinatewise median or L1-median, whereas response variable **y** is centered using the median. Initialize leverage and residual weights (see Eqs. (7) and (10)) and calculate square root of their product (global weights).
1. Build the classic PLS-1 model for weighted rows of **X** and weighted elements of **y** using the global weights.
2. Compute residuals from the model for each object, *r<sub>i</sub>*, using the actual PLS model and update global weights according to Eqs. (7)–(10).
3. Return to step (1) until convergence is achieved.

Initialization of leverage and residual weights is an important step that has an effect on method convergence. Initial estimates of leverage weights can be found according to Eq. (7) by replacing the PRM scores with explanatory variables and Eq. (8). Initial residual weights can be estimated using Eq. (9), where residuals are substituted by centered elements of the response variable around median and then Eq. (8).

The PRM method is one example of a robust variant of the classic PLS-1 method. More robust calibration techniques are presented and discussed in [20–23]. Compared to other robust variants of PLS, PRM has a high computational speed and withstands a relatively large number of outliers [19].

#### 2.2.1. Identification of outliers/studying outlying character of objects

Once the PRM model is constructed and using its parameters, it is possible to identify outlying objects in a calibration set (and test set, if deemed necessary) and/or to find out reason for their outlyingness. For objects from the calibration set, their robust standardized distances in the model space of  $f$  PRM latent factors,  $d_i^T$  and standardized residuals,  $d_i^R$  are evaluated:

$$d_i^T = \frac{\|t_i - L1\text{med}(T)\|}{\alpha_{Qn}(\|t_i - L1\text{med}(T)\|)} \quad (11)$$

$$d_i^R = \frac{|r_i - \text{med}(r)|}{\alpha_{Qn}(r)} \quad (12)$$

where  $\alpha_{Qn}$  is the robust Qn estimator of the data scale [19].

For both distances,  $d_i^T$  and  $d_i^R$ , the cutoff values are defined by assuming normal distribution of data majority. The cutoff values divide the model space into four regions containing different types of objects. They can be visualized in a so-called distance–distance plot, presented in Fig. 1, where the standardized residual distances of objects are plotted versus their standardized leverage distances. With respect to  $d_i^T$  and  $d_i^R$  one can distinguish among three types of outlying objects: bad leverage objects (with long leverage distances, i.e. located far away in the PRM model space from data center and with long residual distances), high residual objects (with long residual distances and short leverage distances) and good leverage objects (with long leverage distances but short residual distances). By setting the cutoff value to 3, it is expected that 99.9% of objects will have  $d_i^T$  and  $d_i^R$  distances shorter than the cutoff value.

In order to rapidly grasp why objects from the model set are outlying, one can inspect leverage and residual weights of the PRM model simultaneously (using a plot similar to the distance–distance plot). Depending on the weight type, a large magnitude indicates

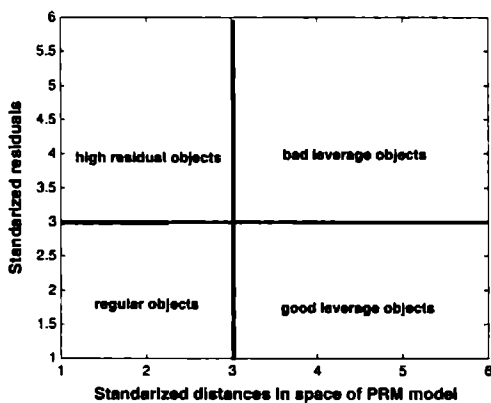


Fig. 1. Example of a distance–distance plot for the robust partial least squares model.

the outlying character of an object in the model space and/or with respect to the model fit (residuals).

In relation to objects from test set, it is also possible to find out the reason for their outlyingness by evaluating the respective residuals and PRM scores. For this purpose, test set objects are projected into the PRM model space in order to obtain their scores and then robust distances (assuming weights for the test samples equal to one). However, it should be emphasized that diagnostics of outlying objects in a test set using the distance–distance plot can provide some insight only at the stage of robust model construction, i.e. response values are available for objects from test set (e.g., after splitting the available data into model and test sets).

The final PRM model is constructed when the algorithm's convergence is achieved. Then, the weights for all objects are established and influence of outlying objects on the model is reduced. Although the final PRM model is optimal, it is also possible to discard outliers using the distance–distance plot to construct the final model using the classic PLS approach. This step can be regarded as fine tuning the final calibration model, since with the classic PLS model more efficient estimates of the regression coefficient are derived as compared to PRM.

#### 2.4. Selection of the model set

Construction of any calibration starts with the selection of model and test sets. Objects in the model set are used for construction of the model and samples from the test set for validation of the model. Construction of a reliable calibration model – with good prediction properties – requires uniform sampling of the experimental domain as well as including all of the sources of data variation into the model set [24]. There are several efficient uniform subset selection algorithms well suited for this purpose [24]. Among the most popular ones are the Kennard and Stone (K&S) and the Duplex algorithms [25,26]. In both algorithms, samples are selected into a model set on the basis of the maximal dissimilarity among samples scored using the Euclidean distance. In the context of calibration, the main advantage of the Duplex algorithm over K&S is that the model and test sets are representative (i.e. samples in the model and test sets are uniformly scattered over the experimental domain).

#### 2.5. Selection of model complexity and scoring its prediction ability

One of the most important steps during the construction of the calibration model is the choice of the optimal number of latent factors,  $f$ , which is frequently supported by using the cross-validation procedure (CV) [27]. During cross-validation, a subset of  $p$  objects is removed from the data and these objects form the test set. A calibration model is created for the remaining objects and the root mean square error of cross-validation, RMSECV, is calculated for the test set objects. The procedure is repeated and the next subset of objects is removed from the data for which a number of models with an increasing number of latent factors are constructed. This is done until all possibilities for the construction of the subsets are exhausted. For each latent factor, the corresponding RMSECV is reported as an average value and calculated as:

$$\text{RMSECV}(f) = \sqrt{\frac{1}{p} \sum_{i=1}^p (y_i - \hat{y}_i)^2} \quad (13)$$

where,  $y_i$  is the experimental value of the response variable and  $\hat{y}_i$  is the predicted value of the response variable.

There are different variants of the cross-validation procedure, e.g. the leave-one-out cross-validation (LOOCV), the leave-more-out cross-validation (LMOCV) and the Monte Carlo cross-validation (MCCV). LOOCV is the most popular approach; however, LMOCV or MCCV are recommended since they provide more accurate estimates

of a model's complexity [28,29]. MCCV and LOOCV were used in this study in order to obtain robust estimates of RMSECV (for PRM) and to evaluate the complexity of N-PLS. To construct robust estimates of RMSECV, the trimming procedure was used [17]. Assuming that some objects are outliers (e.g. 5%), they have large residuals from the robust model and as a result they significantly affect estimates of RMSECV. Therefore, an a priori defined fraction of outliers is omitted during the calculation of RMSECV. Root mean square error (RMSE) and root mean square error of prediction (RMSEP) describing the fit and prediction properties of a calibration model, respectively, are defined as:

$$\text{RMSE}(f) = \sqrt{\frac{1}{m} \sum_{i=1}^m (y_i - \hat{y}_i)^2} \quad (14)$$

$$\text{RMSEP}(f) = \sqrt{\frac{1}{m_t} \sum_{i=1}^{m_t} (y_i' - \hat{y}_i')^2} \quad (15)$$

where,  $m$  is the number of samples in the model set (or the number of samples in the model set after trimming),  $m_t$  is the number of samples in the test set (or the number of samples in the test set after trimming),  $y_i'$  is the experimental response value and  $\hat{y}_i'$  is the predicted response value from the model with  $f$  latent factors for the  $i$ th object from the test set.

### 3. Data set and data pretreatment

The advantages of robust calibration for the quality control are demonstrated on the process data describing production of sugar [30]. They were collected by process sampling (during 8-hour shifts) at a Scandinavian plant. During one 3-month campaign in late autumn 268 sugar samples were collected from a centrifuge (the final step of the process). Each sample was dissolved in un-buffered water (2.25 g/15 ml), placed in a 10 × 10 mm cuvette and fluorescence spectra were measured using a PE LS50B spectrofluorometer. The emission spectra were recorded from 275 to 560 nm in 0.5 nm intervals (571 wavelengths) at seven excitation wavelengths (230, 240, 255, 290, 305, 325 and 340 nm). In addition to the fluorescence spectra, the ash content and color were determined for each sample. Ash content was determined using conductivity by reflecting the total amount of inorganic impurities in the refined sugar expressed as mass percentages of sample. The color of a sample (membrane-filtered solution of sugar adjusted to pH = 7) was described by the absorption observed at 420 nm.

Sugar fluorescence landscapes had a relatively high signal to noise ratio. Therefore, neither noise elimination nor baseline correction of signals was necessary.

To present the benefits of robust calibration, outliers were kept in the data, in contrast to the strategy presented in [14]. Prior to PLS and PRM modeling, fluorescence landscapes were unfolded, as is shown in Fig. 2. For the N-PLS model, fluorescence landscapes were arranged as a three-way data array with dimensions *samples* × *emission spectra* × *excitation spectra*.

As illustrated in Fig. 2, emission spectra obtained using excitation wavelengths 230, 240 255 and 290 nm contain the Rayleigh scattering. This effect does not reinforce models, therefore prior to construction of PLS and PRM models these parts of spectra were removed, whereas for N-PLS they were substituted with missing values (NaN).

### 4. Methods

The calibration models presented in this paper were constructed using freely available toolboxes developed for MATLAB: the TOMCAT toolbox [31] and the N-way toolbox [32].

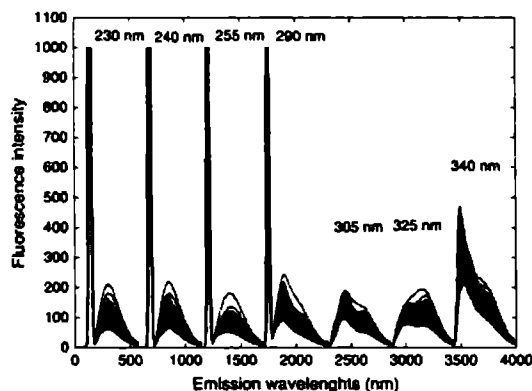


Fig. 2. Unfolded fluorescence excitation–emission spectra of sugar samples collected every eight hours (excitation wavelengths are indicated at the top of each emission spectrum).

### 5. Results and discussion

One of the first steps in data modeling is the selection of the model and test sets from the available data. This is done when designing a calibration domain is impossible. In the literature, different approaches are discussed [24]. For instance, samples to model and test sets are sometimes selected randomly, but with a limited opportunity to guarantee the representativeness of a model set. Another possibility is to design a model set uniformly using the Kennard and Stone algorithm [26]. This allows all samples with unique characteristics (all potential outlying objects) to be included into a model set, whereas the remaining ones serve as test samples. Such a model set is representative, in contrast to the test set. Therefore, the final calibration model usually has a better fit (small RMSE) than its prediction error for new samples. Possibly the best way to design model and test sets, and as a result to obtain reliable estimates of a model's fit and its prediction, is to select samples for the model and test sets uniformly using the Duplex algorithm because both sets will be representative (potential outliers will be present in the model and test sets). In this study, the Duplex algorithm was used to split samples, which were described by unfolded fluorescence landscapes (except the input data to N-PLS). Each set contained a total of 134 samples. Two response variables, namely color and ash content in sugar samples, were modeled, using different multivariate calibration techniques. To compare the fit and prediction properties of models, they were constructed for the same model and test sets.

Firstly, the PLS method was chosen as the standard chemometric modeling approach to model the color and ash content of sugar samples. The optimal number of latent variables to construct a model was evaluated using the MCCV approach. The prediction ability of models with an increasing complexity was obtained as an average of the prediction errors for models with a given number of latent variables. In the MCCV course, 114 out of 134 samples were drawn randomly 268 times. For these samples PLS models with an increasing number of latent variables were built and their prediction errors were evaluated for 20 test samples in order to obtain estimates of RMSECV. In Fig. 3a and b, two PLS models for the color and ash content are visualized as the observed vs. predicted response. When modeling sugar color, six latent variables were determined to be the optimal model complexity leading to RMSE equal to 1.678 and RMSEP equal to 1.858. Bearing in mind the calibration range of a response variable, these two errors correspond to 5.08% and 5.63%, respectively.

When modeling the ash content in samples, six latent variables were needed to reach the minimal prediction error of the PLS



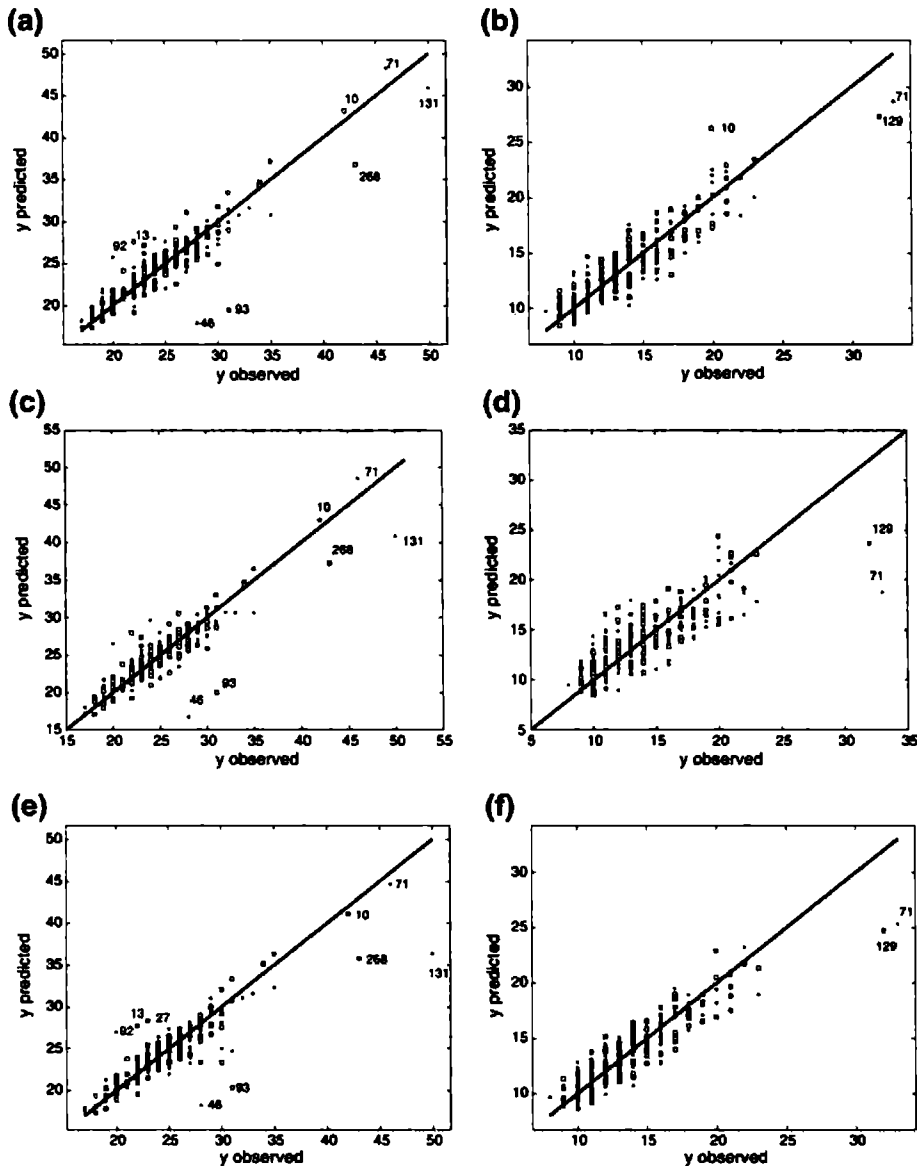


Fig. 3. The classic partial least squares models for: a) color and b) ash content. The N-way partial least squares models for: c) color and d) ash content. The robust partial least squares models for: e) color and f) ash content in sugar samples. Models are presented as  $y$  predicted vs.  $y$  observed for the model (●) and test set samples (○).

model. Its RMSE is equal to 1.508 (5.59%) and RMSEP equals to 1.535 (5.48%).

The fluorescence landscapes of sugar samples are two-way signals and thus for a number of samples they can be arranged as a three-way data array with dimensions  $samples \times emission\ spectra \times excitation\ spectra$ . Due to the specific data structure, one can possibly profit from data trilinearity and enhance modeling results compared to the results of modeling obtained from the classic PLS constructed for unfolded excitation–emission signals. This can be supported using the N-way variant of the PLS method – N-PLS [16].

The N-PLS model for color had complexity seven as was suggested in the course of the LOOCV procedure. The RMS of the N-PLS model was equal to 1.975 (5.98% of the response range) and RMSEP equals 1.784 (5.41% of response range).

The optimal complexity for the N-PLS model for ash content was four. The RMSE and RMSEP values obtained for that model were equal to 2.395 (8.55%) and 1.996 (7.13%). Both N-PLS models are presented in Fig. 3 as the predicted versus the observed response values for color and ash content. Comparing Fig. 3c and d, which illustrate the fit and prediction errors of the models, it can be concluded that

Table 1

Root mean square error of calibration (RMSE) and root mean square error of prediction (RMSEP) for partial least squares regression (PLS), partial robust M-regression (PRM) with 5% trimming and N-way partial least squares regression (N-PLS) built for color and ash content. Error values written in italics correspond to the calibration models obtained after removing seven objects with the highest residuals observed in the PRM models from the calibration and test sets. Errors expressed as a percentage were computed with respect to the calibration range of a given response variables.

Model	Color			Ash content		
	f	RMSE	RMSEP	f	RMSE	RMSEP
PLS	6	1.678 (5.08%)	1.858 (5.63%)	6	1.508 (5.59%)	1.535 (5.48%)
		<i>1.214 (3.68%)</i>	<i>1.172 (3.55%)</i>		<i>1.302 (4.65%)</i>	<i>1.254 (4.48%)</i>
N-PLS	7	1.975 (5.98%)	1.784 (5.41%)	4	2.395 (8.55%)	1.996 (7.13%)
		<i>1.241 (3.76%)</i>	<i>1.183 (3.58%)</i>		<i>1.839 (6.67%)</i>	<i>1.605 (5.73%)</i>
PRM	5	1.147 (3.48%)	1.071 (3.24%)	9	1.056 (3.77%)	1.225 (4.37%)

for the data being studied the N-PLS technique does not outperform classic PLS models.

On the other hand, in Fig. 3a–d a few suspicious objects (located far away from the majority of the calibration samples) can be spotted. Such objects can affect the construction of classic calibration models. Robust modeling was carried out in order to verify their influence on the calibration model and to potentially improve the models. In this study, the PRM technique was used as a robust modeling approach. The PRM models for color and ash content are presented in Fig. 3e and f. In order to obtain robust estimates of error (estimates not affected by the presence of outlying objects in calibration data),

the trimming procedure was applied, assuming that at most 5% of outliers are expected in data. The optimal number of latent factors for the PRM models was determined using the MCCV procedure. Five latent factors were required to model color and to lead to the best fit and prediction properties. The fit and prediction errors are equal to 1.147 and 1.071, respectively. They are substantially lower than the errors obtained for the corresponding PLS and N-PLS models. According to the calibration range of a response variable, they account for 3.48% and 3.24% of error (see Table 1). Nine latent factors were found to be the optimal complexity of the PRM model for calibrating the ash content in sugar samples. The fit and prediction errors obtained for that model were equal to 1.056 (3.77%) and 1.225 (4.37%), respectively. As is shown in Table 1, the fit and prediction errors of the robust model are significantly lower than those obtained from PLS and N-PLS.

In order to compare the results of modeling for the classic and robust techniques, PLS and N-PLS models were again constructed for the model and test sets without the 7 objects with the largest residuals from the PRM model (the PLS and N-PLS models have the same complexity as before). The fit and prediction errors of the models are reported in Table 1 in italics. The PLS and N-PLS models for calibrating color are comparable to the PRM model. The PLS model constructed for the ash content, which is similar to PRM and N-PLS, has the highest error from all models. Therefore, for the data being studied taking into account their trilinear structure does not considerably improve the modeling result.

The PRM approach allows for the straightforward identification of outlying objects in a calibration set. It is performed using the

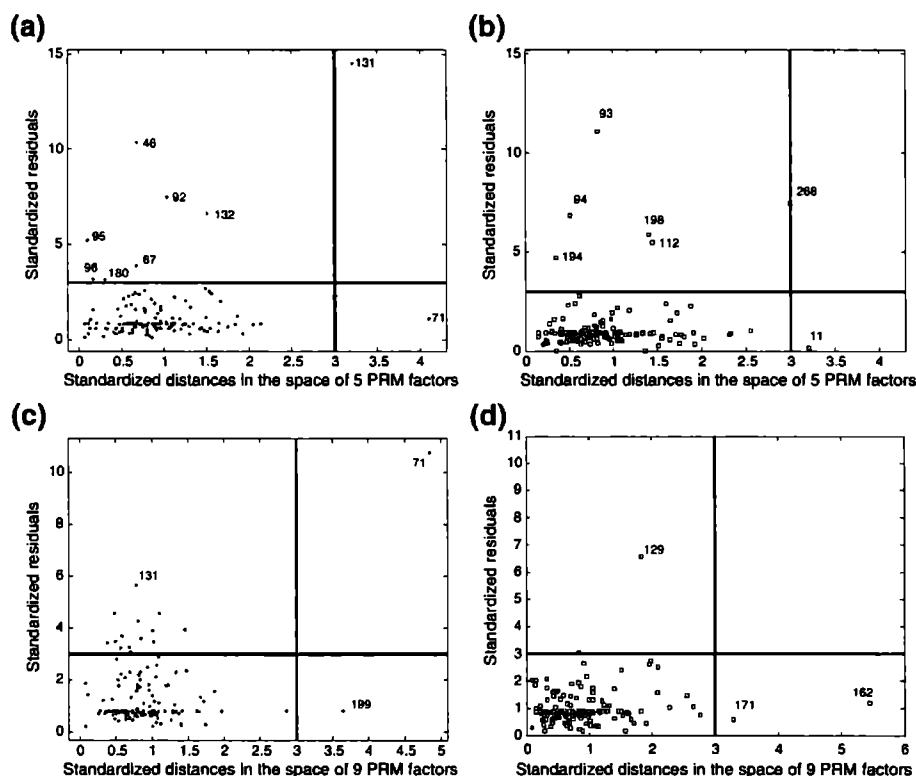


Fig. 4. Distance–distance plots presenting the standardized residual and leverage distances for the model and test sets objects, respectively, for the robust partial least squares model for: a) and b) color (model constructed with five latent factors), c) and d) ash content (model constructed with nine latent factors).

Table 2

Root mean square error of calibration (RMSE), root mean square error of prediction (RMSEP) and errors expressed as a percentage of the response variable range for partial least squares regression (PLS) created for color and ash content on the basis of excitation–emission fluorescence landscapes and selected emission spectra for 230 and 290 nm excitation wavelengths for color and ash content, respectively.

Modeled property	Color		Ash content	
	Landscape	230 nm	Landscape	290 nm
r	6	5	6	4
RMSE	1.678	1.698	1.508	2.012
	5.08%	5.14%	5.59%	8.04%
RMSEP	1.858	2.194	1.535	1.879
	5.63%	6.65%	5.48%	7.51%

distance–distance plot, which presents the standardized leverage and residual distances computed using parameters of a robust model. In this study, the cutoff values for both distances were set to three as was explained in the Theory section.

The PRM model with five latent factors built for color was used to detect outlying objects in the calibration set. In Fig. 4a and b, the distance–distance plots for the model and test set samples are presented, respectively. As is indicated in the figure one bad leverage object is revealed in the model set (object no. 131). Seven objects nos. 46, 67, 92, 95, 96, 132 and 180 are high residual objects for the model set. There is also one good leverage object because its residual distance is small; however, its leverage distance is high. It is object no. 71. It potentially extends the calibration range but has no influence on the construction of classic models such as PLS and N-PLS. As can be seen in Fig. 4b, the test set also contained some objects that exceeded the cutoff values.

The distance–distance plots obtained from the PRM model for ash content (with nine latent factors) reveal one bad leverage object in the model set (object no. 71). This has a significant influence on classic calibration models (see Fig. 4c). Moreover, one can spot a few good leverage objects in the model set. The distance–distance plot constructed for the test set exhibits three outlying objects nos. 129, 162 and 171 (see Fig. 4d).

In order to identify the characteristic excitation wavelength that can provide a suitable spectrum for calibration purposes, PLS models for each emission spectrum observed at a given excitation wavelength for the color and ash content were constructed. Characteristic excitation wavelengths were selected according to the optimal RMSE and RMSEP values obtained from a model for a calibrated property. A satisfactory PLS model for color was built using the emission spectrum corresponding to the excitation wavelength at 230 nm, whereas for the ash content it was the emission spectrum observed at the excitation wavelength 290 nm. As can be observed by the results presented in Table 2, both PLS models (with two and four latent variables, respectively), are comparable to the PLS models constructed for the unfolded fluorescence landscapes.

## 6. Conclusions

The aim of this paper was to highlight the advantages of the robust calibration technique (PRM) used for monitoring sugar production. To achieve this goal, different calibration models – classic and robust – were built and compared for excitation–emission fluorescence landscapes. The prediction errors of the constructed models, given in percentages relative to the total range of response variable, were equal to 3.24% and 4.37% for the color and ash content, respectively. It was confirmed that the sugar production process can be monitored using multivariate calibration models and fluorescence signals and that robust modeling has great potential in the control of real processes. The results obtained from PLS, N-PLS and PRM indicate that robust PLS models are stable against new sources of variation and have better prediction properties than PLS and N-PLS. The PRM

models allow for the direct processing of data containing outlying objects. Therefore, identification of outliers prior to the construction of a PRM model is unnecessary. In addition, it is possible to select characteristic excitation wavelengths and use the corresponding emission spectra for modeling for the purpose of monitoring.

## Acknowledgment

MD wishes to express his gratitude to the Minister of Science and Higher Education of the Polish Republic for funding the scholarship.

## References

- [1] T. Woodcock, C. O'Donnell, G. Downey, Better quality food and beverages: the role of near infrared spectroscopy, *Journal of Near Infrared Spectroscopy* 16 (2008) 1–29.
- [2] J.C. Menezes, P. Felizardo, M.J. Neiva-Correira, The use of process analytical technology (PAT) tools in biofuels production, *Spectroscopy* 23 (2008).
- [3] J. Munson, C.F. Stanfield, B. Cujral, A review of process analytical technology (PAT) in the US pharmaceutical industry, *Current Pharmaceutical Analysis* 2 (2006) 405–414.
- [4] A.S. Rathore, R. Bhambure, V. Ghare, Process analytical technology (PAT) for biopharmaceutical product, *Analytical and Bioanalytical Chemistry* 398 (2010) 137–154.
- [5] T. Kourti, Process analytical technology beyond real-time analyzers: the role of multivariate analysis, *CRC Critical Reviews in Analytical Chemistry* 36 (2006) 257–278.
- [6] D. Kim, S.R. Kim, D.F. Day, M.J. Oh, H.Y. Kim, A dipstick method to measure dextran in raw sugar process stream, *International Sugar Journal* 111 (2009) 119–124.
- [7] S.M. Walford, K.J. Schättler, P.G. Morel Du Boli, Chromatographic solutions for sugar processing, *International Sugar Journal* 107 (2005) 90–102.
- [8] D.L. Massart, B.G.M. Vandeginste, L.M. Budgens, S. de Jong, P.J. Lewi, J. Smeyers-Verbeke, *Handbook of Chemometrics and Qualimetrics Part A*, Elsevier, London, 1992.
- [9] E.M. Becker, J. Christensen, C.S. Frederiksen, V.K. Haugaard, Front-face fluorescence spectroscopy and chemometrics in analysis of yogurt: rapid analysis of riboflavin, *Journal of Dairy Science* 86 (2003) 2508–2515.
- [10] R. Cheikhousman, M. Zude, D. Bouveresse, C.L. Léger, D.N. Rutledge, I. Birlouez-Aragon, Fluorescence spectroscopy for monitoring deterioration of extra virgin olive oil during heating, *Analytical and Bioanalytical Chemistry* 382 (2005) 1438–1443.
- [11] D. Baunsgaard, L. Norgaard, M.A. Godshall, Fluorescence of raw cane sugars evaluated by chemometrics, *Journal of Agricultural and Food Chemistry* 48 (2000) 4955–4962.
- [12] L. Norgaard, Classification and prediction of quality and process parameters of thick juice and beet sugar by fluorescence spectroscopy and chemometrics, *Zuckerind* 120 (1995) 970–981.
- [13] R. Bro, Exploratory study of sugar production using fluorescence spectroscopy and multi-way analysis, *Chemometrics and Intelligent Laboratory Systems* 46 (1999) 133–147.
- [14] I. Stanimirova, S. Serneels, P.J. Van Espen, B. Walczak, How to construct a multiple regression model for data with missing elements and outlying objects, *Analytica Chimica Acta* 581 (2007) 324–332.
- [15] M.C. Ortiz, L.A. Sarabia, A. Herrero, Robust regression techniques. A useful alternative for the detection of outlier data in chemical analysis, *Talanta* 70 (2006) 499–512.
- [16] R. Bro, Multiway calibration, multilinear PLS, *Journal of Chemometrics* 10 (1996) 47–62.
- [17] S. Serneels, C. Croux, P. Filzmoser, P.J. Van Espen, Partial robust M-regression, *Chemometrics and Intelligent Laboratory Systems* 79 (2005) 55–64.
- [18] H. Fritz, P. Filzmoser, C. Croux, A comparison of algorithms for multivariate L1-median, *Comput. Stat.* (in press) 1–18, doi:10.1007/s00180-011-0262-4.
- [19] O. Hössjer, C. Croux, Generalizing univariate signed rank statistics for testing and estimating a multivariate location parameter, *Journal of Nonparametric Statistics* 4 (1995) 293–308.
- [20] D.J. Cummins, C. Andrews, Iteratively reweighted partial least squares: a performance analysis by Monte Carlo simulation, *Journal of Chemometrics* 9 (1995) 489–507.
- [21] M. Hubert, K. Vanden Branden, Robust methods for partial least squares regression, *Journal of Chemometrics* 17 (2003) 537–549.
- [22] M. Hubert, *Robust Calibration in Comprehensive Chemometrics*, vol. 3, Elsevier, Amsterdam, 2010, pp. 315–343.
- [23] S. Frosch Moller, J. von Frese, R. Bro, Robust methods for multivariate data analysis, *Journal of Chemometrics* 19 (2005) 549–563.
- [24] M. Daszykowski, B. Walczak, D.L. Massart, Representative subset selection, *Analytica Chimica Acta* 468 (2002) 91–103.
- [25] R.D. Snee, Validation of regression models: methods and examples, *Technometrics* 19 (1977) 415–428.
- [26] R.W. Kennard, L.A. Stone, Computer aided design of experiments, *Technometrics* 11 (1969) 137–148.
- [27] T. Naes, T. Isaksson, T. Fearn, T. Davies, *A User-friendly Guide to Multivariate Calibration and Classification*, NIR Publications, Chichester, 2002.

- [28] Q.S. Xu, Y.Z. Liang, Uniform design and its applications in chemistry and chemical engineering, *Chemometrics and Intelligent Laboratory Systems* 56 (2001) 1–11.
- [29] K. Baumann, H. Albert, M. Von Korff, A systematic evaluation of the benefits and hazards of variable selection in latent variable regression. Part II. Practical applications, *Journal of Chemometrics* 16 (2002) 339–350.
- [30] [http://www.models.kvl.dk/Sugar\\_Process](http://www.models.kvl.dk/Sugar_Process)(last accessed on the 10th of July 2011).
- [31] M. Daszykowski, S. Serneels, K. Kaczmarek, P. Van Epsen, C. Croux, B. Walczak, TOMCAT: a MATLAB toolbox for multivariate calibration techniques, *Chemometrics and Intelligent Laboratory Systems* 85 (2007) 269–277.
- [32] <http://www.models.kvl.dk/nwaytoolbox>(last accessed on the 10th of July 2011).

Katowice, 03.03.2015

dr hab. Michał Daszykowski, prof. UŚ  
Instytut Chemii  
Uniwersytet Śląski  
ul. Szkolna 9  
40-006 Katowice

Oświadczam, że w artykule pt. *Controlling sugar quality on the basis of fluorescence fingerprints using robust calibration*, opublikowanym w czasopiśmie *Chemometrics and Intelligent Laboratory Systems* 110 (2012) 89-93 mój udział polegał na:

- współtworzeniu hipotezy badawczej i ogólnej koncepcji badań,
- zapewnieniu nadzoru merytorycznego w tym pomocy w interpretacji wyników i weryfikacji wyciągniętych przez doktorantkę wniosków z przeprowadzonych badań,
- opiece i merytorycznym nadzorze podczas przygotowania manuskryptu zgodnie z wymaganiami czasopisma *Chemometrics and Intelligent Laboratory Systems*,
- pomocy w procedurze redakcyjnej (tj. pomocy w złożeniu artykułu do recenzji oraz przygotowaniu odpowiedzi na uzyskane recenzje),
- dokonaniu ostatecznej korekty artykułu.

Michał Daszykowski

17.11.2014. Katowice

prof. dr hab. Beata Walczak  
Instytut Chemii  
Uniwersytet Śląski  
ul. Szkolna 9  
40-006 Katowice

Oświadczam, że w artykule pt. *Controlling sugar quality on the basis of fluorescence fingerprints using robust calibration*, opublikowanym w czasopiśmie *Chemometrics and Intelligent Laboratory Systems* 110 (2012) 89-93 mój udział polegał na merytorycznej dyskusji hipotez badawczych i uzyskanych wyników.

Beata Walczak

## 7.2. Publikacja II

### *Metoda równoczesnego oznaczania Solvent Yellow 124 i Solvent Red 19 w oleju napędowym z wykorzystaniem spektroskopii fluorescencyjnej i chemometrii*

---

Tytuł oryginalny: *Simultaneous determination of Solvent Yellow 124 and Solvent Red 19 in diesel oil using fluorescence spectroscopy and chemometrics*

Autorzy: Joanna Orzeł  
Michał Daszykowski  
Ireneusz Grabowski  
Grzegorz Zaleszczyk  
Mirosław Sznajder  
Beata Walczak

Czasopismo: Talanta

Rok wydania: 2012

Tom: 101

Strony: 78 – 84

Wartość  
współczynnika  
*Impact Factor* : 3,498

---

W publikacji II opisałam opracowaną przeze mnie metodykę analityczną umożliwiającą ocenę zawartości komponentów fiskalnych (znanika – Solvent Yellow 124 i barwnika Solvent Red 19) wprowadzanych do oleju napędowego o obniżonym podatku akcyzowym. Fluorescencyjne obrazy próbek oleju napędowego zawierających zróżnicowane zawartości komponentów fiskalnych wykorzystałam do konstrukcji modeli kalibracyjnych (wybrane metody kalibracji: PLS i N-PLS) umożliwiających ocenę zawartości wybranych składników bez wcześniejszego przygotowywania próbki do analizy. Metodyka charakteryzuje się satysfakcjonującymi wartościami granic wykrywalności dla obydwu komponentów (odpowiednio  $0,042 \text{ mg} \cdot \text{L}^{-1}$  dla znanika oraz  $0,048 \text{ mg} \cdot \text{L}^{-1}$  dla barwnika) oraz wartościami względnego odchylenia standardowego dla obydwu analitów poniżej 3,4%.

Ze względu na potrzebę tego typu rozwiązań i wysokie prawdopodobieństwo aplikacji proponowanego rozwiązania w badaniach rutynowych, zostało ono zgłoszone do ochrony patentowej – zgłoszenie nr P.339194, data zgłoszenia 16.05.2012 r., tytuł zgłoszenia: *Sposób równoczesnego lub pojedynczego oznaczania zawartości znanika Solvent Yellow 124 i barwnika Solvent Red 19 lub Solvent Red 164 lub Solvent Blue 35 w oleju napędowym.*



## Simultaneous determination of Solvent Yellow 124 and Solvent Red 19 in diesel oil using fluorescence spectroscopy and chemometrics<sup>☆</sup>

J. Orzel<sup>a</sup>, M. Daszykowski<sup>a,\*</sup>, I. Grabowski<sup>b</sup>, G. Zaleszczyk<sup>b</sup>, M. Sznajder<sup>b</sup>, B. Walczak<sup>a</sup>

<sup>a</sup> Department of Analytical Chemistry, Chemometric Research Group, Institute of Chemistry, The University of Silesia, 9 Szkolna Street, 40-006 Katowice, Poland

<sup>b</sup> Customs House of Customs Laboratory in Biala Podlaska, 21 Celników Pańskich Street, 21-500 Biala Podlaska, Poland

### ARTICLE INFO

#### Article history:

Received 16 May 2012

Received in revised form

16 August 2012

Accepted 22 August 2012

Available online 30 August 2012

#### Keywords:

Euromarker

Sudan Red 7B

EEM

Fluorescence landscapes

N-PLS

### ABSTRACT

Differences in tax levels for diesel oil stimulate the illegal removal of characteristic diazo compounds purposely added to designate its possible usage. In order to reduce the losses in the national income, there is a strong need to develop a sensitive and cost-effective analytical procedure for the detection of this illegal action. In this study, we describe a novel analytical approach for a qualitative and quantitative determination of two diazo compounds (Solvent Yellow 124 and Solvent Red 19) that are usually added to diesel oil. The methodology proposed combines the use of excitation–emission matrix fluorescence spectroscopy as an analytical technique and partial least squares regression as a multiple modeling tool. With this new methodology, relatively low root mean square errors of prediction (for independent set of test samples) that are equal to 0.223 for Solvent Red 19 and 0.263 for Solvent Yellow 124, were obtained and the results were stable, which were indicated by an analysis performed after 48 and 96 h. The methodology is also nondestructive and allows for (i) simultaneous detection of diesel oil additives, (ii) determination of satisfactory limits of detection (0.048 and 0.042 mg L<sup>-1</sup> for Solvent Red 19 and Solvent Yellow 124, respectively), and (iii) obtaining of considerably low relative standard deviations of 2.33% for Solvent Yellow 124 and of 3.23% for Solvent Red 19 in comparison with the existing norm level.

© 2012 Elsevier B.V. All rights reserved.

### 1. Introduction

Diesel oil is a commonly used fuel for transport, heating, and agricultural machinery drive purposes. Depending on its usage, the tax levels are different in many European and American countries. A low tax fuel (used for heating and agricultural machinery drive purposes) is spiked with additives which change its color from yellow to red. The type of additives varies from country to country, but everywhere a marker and a dye are added at the stage of oil production. In the European Union countries, Solvent Yellow 124 (SY124) is a common marker added to low tax fuels in concentration levels strictly defined in the range of 6.0 mg L<sup>-1</sup>–9.0 mg L<sup>-1</sup> [1]. Various dyes like Solvent Red 164 (SR164), Solvent Red 19 (SR19), and Solvent Red 26 (SR26) can be used to ensure the red color of the fuel, but their specific use and concentration levels are legally regulated in every country. In Poland, SR164 and SR19 dyes are added interchangeably in diesel oil and their concentration levels must be higher than 6.6 mg L<sup>-1</sup>

and 6.3 mg L<sup>-1</sup>, respectively, while the concentration of Solvent Yellow 124 is regulated by the European Norm [2].

Several analytical procedures for the determination of the dye and marker in different types of oils have been described in other literature. In 2004, a validated procedure for determination of SY124 in gas oil and kerosene was introduced as the EU reference method [3]. The method is based on high performance liquid chromatography (HPLC) determination of SY124 spiked with different dyes. Another procedure for simultaneous quantitative determination of both SY124 and SR19 (a dye used in Poland and other European countries) in fuel has been developed [4]. However, it requires the separation of the reagents of interest with the HPLC technique before their qualitative or quantitative determination with UV-vis or diode array (DAD) detector. Recently, a method for detection of SR164 in vehicle exhaust has also been described in [5].

SY124 and diazo compounds SR19, SR164, and SR26, also known as Sudan dyes, have fluorescence properties since their molecules contain aromatic rings and coupled double bonds. Chen et al. have proposed [6] the use of conventional fluorescence spectroscopy for the determination of the Sudan IV dye in food samples. Traditionally, either the maximum intensity of the emission spectrum or the emission spectrum in a selected range of wavelengths is recorded at a single excitation wavelength

<sup>☆</sup> Patent pending.

\* Corresponding author.

E-mail address: [mdaszyk@us.edu.pl](mailto:mdaszyk@us.edu.pl) (M. Daszykowski).



characteristic for the analyzed fluorophor. This method of analysis may be preferred when the concentration of one fluorophor is to be determined or when the mixture being analyzed contains only a limited number of compounds with fluorescence properties and a quenching effect is not to be expected. In this context, the simultaneous determination of the dye and marker introduced in diesel oil, which may contain many other fluorophors, by the conventional fluorescence spectroscopy can be a difficult task. This task is further complicated by the fact that the fluorophors of interest have different characteristic excitation and emission ranges.

Excitation–emission fluorescence spectroscopy allows for the simultaneous collecting of excitation and emission spectra of samples. A sample is described by a two-dimensional signal that contains peaks from all excited fluorophors. All two-dimensional signals collected for a set of samples form a complex three-way data (third-order tensor) of dimensions: *samples* × *excitation wavelengths* × *emission wavelengths*. Such a type of data follows the trilinear or parallel factor analysis (PARAFAC) model. Construction of multivariate or multi-way calibration models [7–10] offers the possibility to simultaneously determine the dye and marker that were added to diesel oil when the samples contain many constituents that are not of interest. This is the so-called second-order advantage [10–12]. In general, the calibration models using the unfolded second-order data for a sample are to be preferred when the analyte–background interactions or changes in spectral properties of samples are examined [11].

The aim of this work is to develop a new analytical approach for the simultaneous determination of SY124 and SR19 in diesel oil without any sample preparation. For this purpose, the excitation–emission fluorescence spectroscopy is used as the analytical technique and the results of two calibration methods, e.g. partial least squares regression and N-way partial least squares regression, are compared [13] for the studied problem.

## 2. Materials and methods

### 2.1. Samples preparation

Diesel oil was purchased from a local gas station, while SY124 (98.0% purity) was obtained from Sigma-Aldrich. A stock solution was prepared by dissolving 5 mg of SY124 in 50 mL of diesel oil, while 5 mg of SR19 (obtained from IBPO Poland, 92.3% purity) was dissolved in 50 mL diesel oil in order to obtain a stock solution of SR19. The SR19 and SY124 stock solutions were mixed so that the concentration of each reagent in the mixture was varied in the range of 0–10 mg L<sup>-1</sup>. A total of 20 mixture combinations were considered (see Fig. 1). Three samples were prepared (three laboratory replicates) for each mixture combination (for example 8 mg L<sup>-1</sup> of SR19 and 4 mg L<sup>-1</sup> of SY124, see Fig. 1), and for each of them the EEM fluorescence spectroscopy measurements were repeated three times (technical replicates). Thus, a total of 180 EEM fluorescence images were registered. The scheme of the experimental design is shown in Fig. 1.

To determine the limits of detection and quantification, nine laboratory replicates containing SY124 at a concentration level of 2 × 10<sup>-3</sup> mg L<sup>-1</sup> and nine laboratory replicates containing SR19 at a level of 2 × 10<sup>-3</sup> mg L<sup>-1</sup> were additionally prepared. Repeatability was evaluated using a set of 18 samples at three concentration levels (4, 5, and 6 mg L<sup>-1</sup>). Three solutions at each concentration level were prepared for SY124 (a total of 9 samples) and the same number of solutions at the same concentration levels were prepared for SR19 (9 samples). The measurements were performed immediately after sample preparation. In order to evaluate the stability of the measurement results over time, 18 samples (used for testing the repeatability) were analyzed after 48 and 96 h.

### 2.2. Fluorescence measurements

A Carry Eclipse Varian FL0811M000 spectrofluorometer with right angle geometry was used to perform the measurements. The emission spectra were registered in a 2 nm interval from 350 to 800 nm (226 wavelengths) at 46 excitation wavelengths selected in a 10 nm interval in the range of 250–700 nm. Detector sensitivity was set to 500 V and the excitation and emission slits were set to 5 nm. Raw spectral data were subjected to a further chemometric analysis and modeling.

### 2.3. Preprocessing of fluorescence signals

One of the most important preprocessing steps when working with fluorescence signals is the correction of Rayleigh scattering, which is chemically irrelevant. Different approaches for scattering correction have been described in the literature. The matrix elements corresponding to the spectral regions with the Rayleigh scattering can be (i) replaced with zeros [14,15], (ii) treated as missing values [16], or (iii) removed and the missing elements of the signal can be interpolated in different possible ways [17,18]. In this paper the scattering effect was removed and the signals were interpolated by the Delaunay triangulation [17].

### 2.4. Modeling of excitation–emission data

#### 2.4.1. Partial least squares regression

Partial least squares regression, PLS, is a popular chemometric tool used to construct multivariate calibration models. The aim of the multiple PLS regression is to describe the relationship between a set of explanatory variables,  $\mathbf{X}$  (a set of matrixized EEM images as: *samples* × (*excitation wavelengths* × *emission wavelengths*)) and a response variable,  $\mathbf{y}$  (a marker or a dye concentration) [13]. In other words, the original strongly correlated spectral signals are replaced by a small number of latent factors (new variables),  $\mathbf{T}$ , for which the maximum covariance with the modeled property,  $\mathbf{y}$ , is observed. The PLS-1 model can be described in the following way:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E} \quad (1)$$

$$\mathbf{y} = \mathbf{T}\mathbf{q} + \mathbf{r} \quad (2)$$

where  $\mathbf{X}(I \times JK)$  is the matrixized form of the data, vector  $\mathbf{q}(f \times 1)$  holds the regression coefficients for  $f$  PLS factors,  $\mathbf{T}(I \times f)$ , the elements of  $\mathbf{P}(JK \times f)$  matrix the PLS loading values, matrix  $\mathbf{E}(I \times JK)$  holds the differences between the observed and predicted  $\mathbf{X}$  with  $f$  PLS factors and the residual vector  $\mathbf{r}(1 \times 1)$  contains the differences between the observed and predicted  $\mathbf{y}(1 \times 1)$  values.

#### 2.4.2. N-way partial least squares

The N-way partial least squares, N-PLS, is a generalization of the classic PLS regression for higher order data arrays [19]. In the course of the model construction, three-way data array,  $\underline{\mathbf{X}}$ , of size  $I \times J \times K$  (e.g. *samples* × *excitation wavelengths* × *emission wavelengths*), is decomposed into new variables (the so-called triads). The number of the new variables constituting the triads depends on the dimensionality of modeled data.

Prior to modeling, the three-way array is metricized into a data matrix. For the data matrix,  $\mathbf{X}^{I \times JK}$ , of size  $I \times JK$ , where  $I$  is the number of samples,  $J$  is the number excitation wavelengths, and  $K$  is the number of emission wavelengths, the decomposition can be presented as follows:

$$\mathbf{X} = \mathbf{H}(\mathbf{W}^J) \otimes \mathbf{W}^K + \mathbf{E} \quad (3)$$

where  $\mathbf{X}(I \times JK)$  is the matrixized form of the data,  $\mathbf{H}(I \times f)$  is the score matrix,  $\mathbf{W}^J$  and  $\mathbf{W}^K$  are the respective loading matrices for

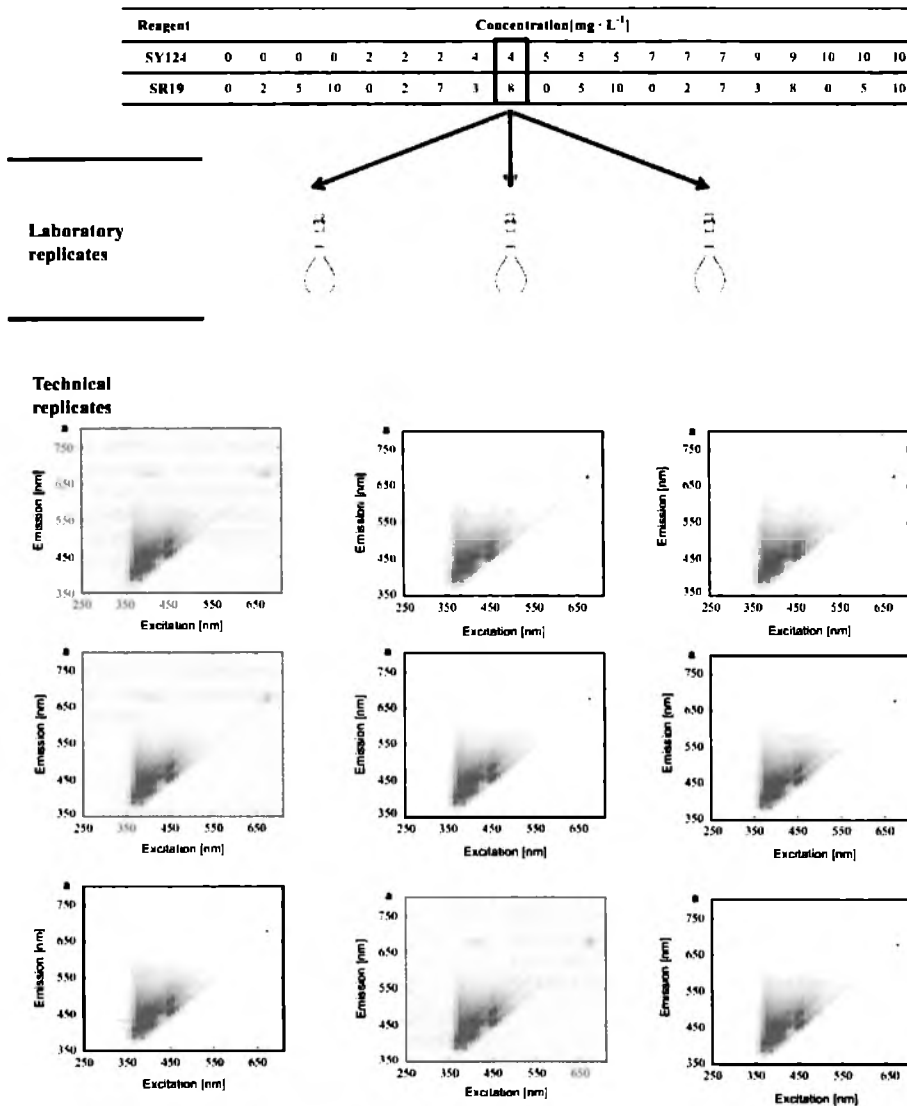


Fig. 1. Design of the experiment with respect to concentrations of a marker compound (Solvent Yellow 124, SY124) and a dye compound (Solvent Red 19, SR19).

the excitation mode and the emission mode,  $E(I \times JK)$  is the error matrix. The symbol  $\otimes$  denotes the Khatri Rao product [20].

The construction of triads is optimized in order to maximize the covariance between  $H$  and  $y$ . Calibration models are constructed using the new variables. A more detailed description of the N-PLS method can be found in [19].

### 2.5. The complexity of regression models

The number of new variables (factors) used for the model's construction is called the complexity of the model,  $f$ . To determine the optimal complexity a cross-validation procedure is usually used [21]. At each step of the validation procedure either a sample or a subset of  $p$  samples (a validation set) is removed from the

data and PLS models with increasing complexity are built for the remaining samples (a model set). Then, a prediction is performed for the removed samples based on the model set. The procedure is repeated for the next subset of  $p$  objects removed from the data, while all possible subsets are not considered when validating the models with an increasing complexity. The root mean square error of cross-validation, RMSECV, is calculated as a measure of the model's performance using the following equation:

$$\text{RMSECV}(f) = \sqrt{\frac{1}{m} \sum_{i=1}^m (y_{-i} - \hat{y}_{-if})^2} \quad (4)$$

where  $y_{-i}$  is the  $i$ -th experimental value of the response variable removed during the cross-validation procedure,  $\hat{y}_{-if}$  is the  $i$ -th

predicted value of the removed response variable using the PLS model with  $f$  latent factors and  $m$  is the number of objects in the model set.

The optimal PLS model is the one characterized by the lowest or acceptable RMSCV. The choice of  $p$  in the leave- $p$ -out cross-validation scheme depends on the user and data dimensionality. The Monte Carlo cross-validation procedure is a cross-validation scheme where the validation set of samples is selected randomly from the model set [22]. In our study, during the cross-validation procedure all of the technical replicates were included in the same validation set of samples (a cancellation group). Once a calibration model of a definite complexity is constructed, its fit is scored by the root mean square error (RMSE):

$$\text{RMSE} = \sqrt{\frac{1}{m} \sum_{i=1}^m (y_i - \hat{y}_i)^2} \quad (5)$$

where  $y_i$  is the  $i$ -th experimental response value,  $\hat{y}_i$  is the  $i$ -th predicted response value using the model of definite factors and  $m$  is the number of samples in the model set.

The prediction properties of the constructed model are evaluated on the basis of an independent test set (dataset that was not used for the model's construction) and expressed as root mean square error of prediction (RMSEP):

$$\text{RMSEP} = \sqrt{\frac{1}{m_t} \sum_{i=1}^{m_t} (y_i^t - \hat{y}_i^t)^2} \quad (6)$$

where  $y_i^t$  and  $\hat{y}_i^t$  are the  $i$ -th experimental and  $i$ -th predicted response value for the  $i$ -th independent test sample using the calibration model of a definite complexity and  $m_t$  is the number of independent test samples.

## 2.6. Software

All calculations were performed within the MATLAB environment. PLS models were constructed using the freely available TOMCAT toolbox [23]. The algorithm for scattering correction described in [17] was used. The N-way Toolbox (version 3.1) [24] was adopted to construct N-PLS models.

## 3. Results and discussion

The raw spectra collected were characterized by a relatively high signal-to-noise ratio and therefore, neither a noise correction nor baseline elimination was performed.

The characteristic Rayleigh scattering was observed for the samples examined, e.g. samples of a clean diesel oil without additives, an oil with the SY124 marker, an oil with SR19 dye and an oil spiked with both the SY124 marker and the SR19 dye. Fig. 2a displays the scattering effect as a diagonal line of peaks starting at the wavelength of 350 nm. As was mentioned earlier, the Rayleigh scattering is a chemically irrelevant component and should be removed prior to the models' construction (the first step of the raw data preprocessing). The correction procedure used in this article consists of detecting the maxima of the Rayleigh scattering peaks followed by counting the number of sampling points with the scattering effect and replacing the values of the sampling points with a scattering effect with values interpolated using the Delaunay triangulation [25]. For our dataset, it was found that seven sampling points on the left side and seven sampling points on the right side of the detected maxima were sufficient to handle the scattering effect in all EEM fluorescence images. A contour map of the EEM fluorescence image obtained from a pure diesel oil sample after removing the

Rayleigh scattering is presented in Fig. 2b. It should be emphasized that only meaningful spectra were considered, e.g. the spectra for each sample were recorded so that the lower emission wavelength level was at least equal to the lower excitation wavelength level.

### 3.1. Construction of PLS models for SR19 and SY124

Prior to the construction of the multiple PLS models, the EEM data containing 180 excitation–emission spectra were unfolded in the form of *spectra* × (*emission wavelengths* × *excitation wavelengths*). Then, model and test sets were selected. The PLS model calibrating Solvent Yellow 124 was built using 117 spectra (model set) describing samples with analyte concentration at levels of 0, 5, 7, and 10 mg L<sup>-1</sup>. The remaining 63 spectra (i.e. samples with SY124 concentrations equal to 2, 4, and 9 mg L<sup>-1</sup>) formed the test set. To construct model for calibration of Solvent Red 19 spectra were divided into a model set containing 108 spectra (EEMs of samples with concentrations of SR19 equal to 0, 3, 5, 7, and 10 mg L<sup>-1</sup>, respectively) and a test set formed by 72 spectra (containing 2, 5, and 8 × 10<sup>-3</sup> mg L<sup>-1</sup> of SR19).

Two individual PLS-1 models were constructed to predict the concentration of SY124 or SR19, respectively. The Monte Carlo cross-validation scheme with a validation set of 20 samples ( $p=20$ ) was adopted to evaluate the optimal complexity of the models. At each cross-validation step, 20 samples out of 500 were randomly selected for the validation set.

Firstly, the PLS-1 model for the prediction of the concentration of SR19 was constructed. The model with eight latent variables, presented in Fig. 3a, was considered as the optimal. Its RMSCV was equal to 0.174. The fit and the prediction properties, expressed as RMSE and RMSEP, were found to be 0.153 and 0.223, respectively. As indicated in Fig. 3a, the model has similar variances for the predictions of the model and test set samples regardless of their concentration levels.

The optimal PLS model obtained for SY124 also contained eight latent variables. RMSE and RMSEP were equal to 0.229 and 0.263, respectively. Compared to the PLS model constructed for SR19, a larger scatter of the values predicted for the model and test set samples was observed. The variances of concentrations predicted for the model and test set samples are comparable at different concentration levels (see Fig. 3b).

### 3.2. Construction of N-PLS models for SR19 and SY124

A possible improvement of the calibration models can be expected when using N-PLS because of the well-defined three-linear structure of the excitation–emission data. Therefore, the N-PLS models were built for the same model and test set samples (see Section 3.1) as those used to construct the conventional PLS-1 models on the metricized data, but the data sets were arranged as *spectra* × *excitation wavelengths* × *emission wavelengths*. Orthogonality constraint was not considered for any of the modes. Two N-PLS models, each of which had 10 latent variables, were used for the prediction of the concentrations for SR19 and SY124, respectively.

As indicated in Table 1 and the graphical representation in Fig. 3d, the N-PLS model and the classic PLS model constructed for SY124 have virtually the same performance (compare the values for fit and prediction shown in Table 1). On the other hand, when modeling SR19, PLS outperforms the N-PLS model (see Table 1 and Fig. 3b and d) and requires a smaller number of latent variables compared to the N-PLS model.

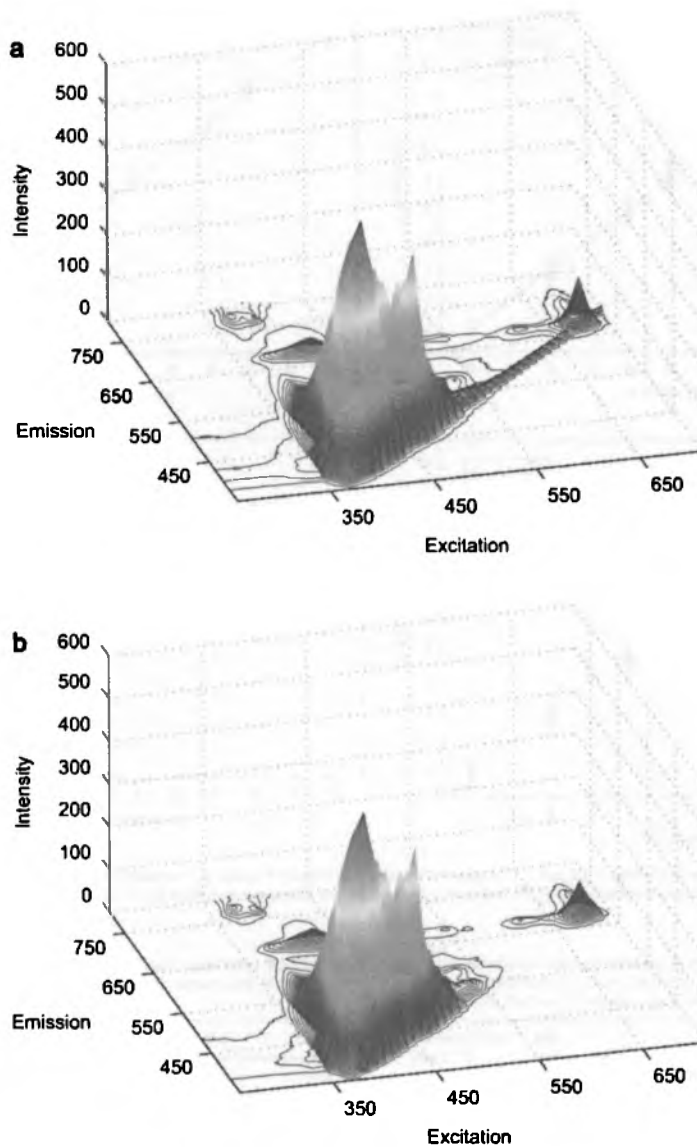


Fig. 2. Landscapes of excitation-emission matrices for pure diesel oil: (a) with the Rayleigh scattering effect and (b) without the Rayleigh scattering effect.

### 3.3. Figures of merit of PLS models

Considering a good performance of classic PLS models, figures of merit were calculated as described in [26,27]. The sensitivity was calculated as the inverse of Euclidean norm of vector containing regression coefficients, and was equal to 13.67 and 11.49 for models describing concentration of SR19 and SY124, respectively. Selectivity of PLS models was equal to 0.015 for SR19 and 0.036 for SY124.

Next, the ratio between instrumental noise and sensitivity was calculated for both calibration models. On the basis of the ratios, limits of detection (LODs) were calculated as the ratio times 3.3 for both analytes (a dye and a marker).

The LOD value of  $0.048 \text{ mg L}^{-1}$  and  $0.042 \text{ mg L}^{-1}$  was obtained for dye and for marker, respectively. The limits of quantification (LOQ) were calculated as 10 times the ratio values for both analytes. The LOQ values were equal  $0.144$  and  $0.126 \text{ mg L}^{-1}$  for SR19 and SY124, respectively.

### 3.4. Analytical validation

Next, the repeatability of the proposed analytical approach was tested. Predicted mean concentration values, relative standard deviations (RSD), and the uncertainty for the three concentration levels that were obtained by conventional PLS are shown in Table 2.

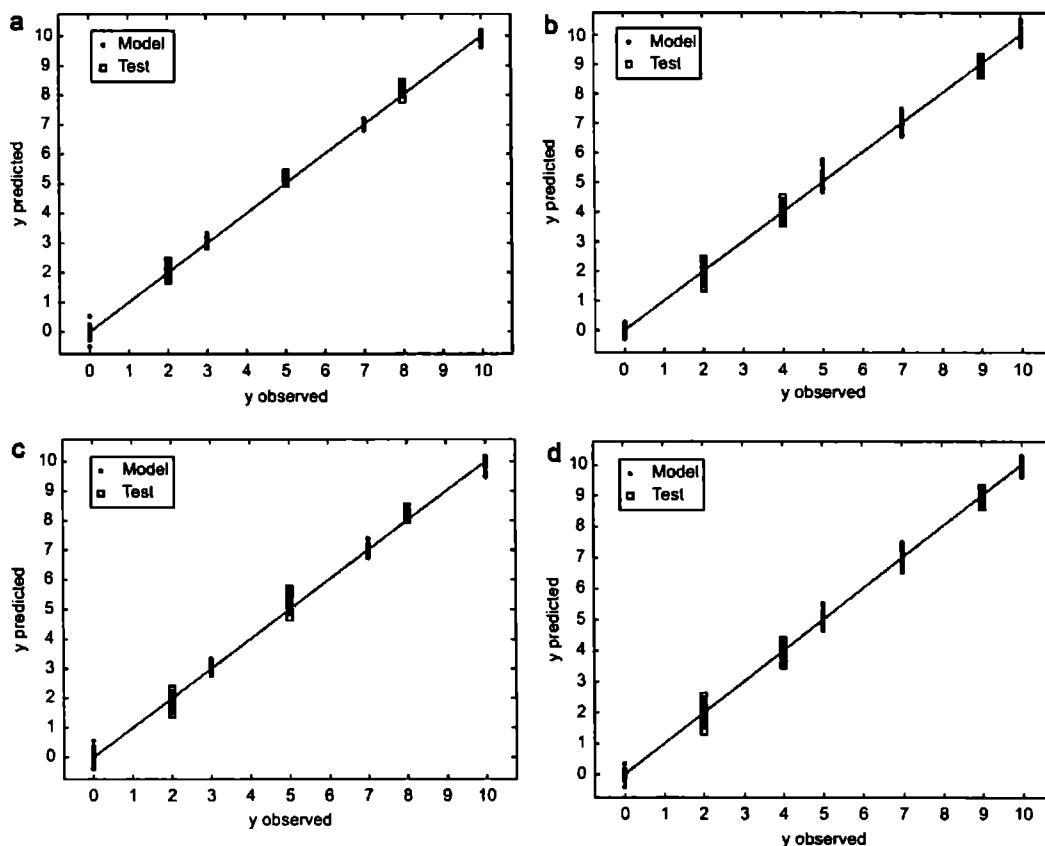


Fig. 3. Calibration models are presented as  $y$  predicted vs.  $y$  observed for model (●) and test set samples (□). The PLS models of complexity eight for: (a) dye (Solvent Red 19) and (b) marker (Solvent Yellow 124). The N-PLS models of complexity ten for: (c) dye (Solvent Red 19) and (d) marker (Solvent Yellow 124).

Table 1

Fit and prediction properties of calibration models constructed for SR19 and SY124. PLS models were constructed with eight factors, for N-PLS models construction 10 factors were used.

Additive	PLS		N-PLS	
	RMSE	RMSEP	RMSE	RMSEP
SR19	0.153	0.223	0.193	0.305
SY124	0.229	0.263	0.212	0.260

Table 2

Repeatability of proposed method obtained from three concentration levels. Mean values were calculated for three laboratory replicates, uncertainty calculated for 95% confidence interval. PLS models were constructed with 8 factors.

Concentration [mg L <sup>-1</sup> ]	PLS					
	SY124			SR19		
	Mean [mg L <sup>-1</sup> ]	RSD [%]	± [mg L <sup>-1</sup> ]	Mean [mg L <sup>-1</sup> ]	RSD [%]	± [mg L <sup>-1</sup> ]
4	3.559	3.14	0.341	4.049	1.33	0.164
5	4.394	3.25	0.435	5.129	1.24	0.194
6	6.027	3.31	0.607	6.282	4.42	0.846

From the RSD values presented in Table 2, one can conclude that the validated approach for SR19 performs better than the one for SY124. In the literature [3], the analytical validation of the HPLC method for Solvent Yellow determination is presented. RSD (calculated at the level of 6 mg mL<sup>-1</sup>) and LOD are equal 0.68% and 0.02 mg mL<sup>-1</sup>, respectively. The corresponding values for presented method are 3.14% and 0.048 mg L<sup>-1</sup> for RSD (at the same concentration level) and LOD, respectively. The HPLC method for the dye determination presented in [4] is not validated thus comparison of the results cannot be compared.

The stability of the results obtained from PLS method was tested statistically using the measurements performed immediately after preparation (set as a conditional level of 0 h) and after

48 and 96 h. The results of the analysis in three 48 h intervals are presented in Table 3.

Two statistical hypotheses were checked for the dye and marker concentrations predicted by the conventional PLS method using F-test. The first hypothesis is that there is no difference in the variances of the predicted values obtained from PLS for samples analyzed at a conditional time of 0 h and after 48 h. For oil samples spiked with a dye the F-value was equal to 1.107, while a F-value of 0.899 was calculated for the oil samples mixed with a marker. The p-values calculated for the two-sided F-test

Table 3

Robustness of proposed method calculated for three concentration levels tested after 0 h, 48 h, and 96 h. Mean values were calculated for three laboratory replicates, uncertainty calculated for 95% confidence interval. PLS models were constructed with eight factors.

Concentration [ $\text{mg L}^{-1}$ ]	Time [h]	PLS SY124			SR19		
		Mean [ $\text{mg L}^{-1}$ ]	RSD [%]	$\pm$ [ $\text{mg L}^{-1}$ ]	Mean [ $\text{mg L}^{-1}$ ]	RSD [%]	$\pm$ [ $\text{mg L}^{-1}$ ]
4.000	cond. 0	3.559					0.164
	48	3.233	3.65	0.360	3.941	1.92	0.350
	96	3.410	2.33	0.242	3.917	1.65	0.197
5.000	cond. 0	4.394	3.25	0.435	5.129	1.24	0.194
	48	4.727	3.27	0.471	5.076	3.29	0.663
	96	4.518	4.37	0.505	5.017	3.69	0.564
6.000	cond. 0	6.027	3.31	0.607	6.282	4.42	0.846
	48	5.647	2.26	0.388	6.100	4.21	0.967
	96	5.787	4.82	0.849	5.996	3.52	0.644

are 0.9492 and 1.053, respectively. The null hypothesis is rejected when the p-values are smaller than a definite significant level value,  $\alpha$ , of 0.01 or 0.05. For the studied case, the null hypothesis of no significant differences in the two variances calculated for the PLS predicted values can be accepted for both types of samples at a level of significance of 0.05. The second null hypothesis is that there is no difference in the variances of the predicted values obtained from PLS for the samples analyzed at a conditional time of 0 h and after 96 h. The values of the estimated F ratios were 1.154 (p-value of 0.9284) for the samples with the dye and 0.899 (p-value of 1.053) for the samples mixed with the marker. Again the null hypothesis can be accepted at a significant level of 0.05.

Assuming a significant level of 0.05 it can be concluded that the proposed analytical approach is stable over time (there are no significant differences in models' performance) and constructed calibration models allow for the determination of analytes in samples after 48 and 96 h with comparable and acceptable error levels.

#### 4. Conclusions

In this paper, a novel approach combining an excitation–emission matrix fluorescence spectroscopy as an analytical technique and partial least squares regression as a multiple modeling tool was developed to quantitatively and qualitatively determine Solvent Red 19 and Solvent Yellow 124 in diesel oil. It is a nondestructive procedure which does not require expensive reagents and laborious sample preparation prior to analysis. The results obtained from the validation procedure give evidence that the approach is stable over time. These attractive features make the proposed methodology a potential screening technique that can be used directly at the place where samples are collected and thus to support the actions of the customs office and related agencies. Two calibration methods were evaluated and compared, namely PLS and N-PLS. Calibration models constructed to predict concentration of SY124 in samples have comparable fit and prediction properties. Compared to PLS, the N-PLS model describing content of SR19 was characterized with higher RMSE and RMSEP values. In our application, the PLS model is preferred due to its conceptual simplicity. In terms of validation parameters (RSD and LOD), the HPLC method for SY124 determination performs better than the proposed approach. On the other hand, simplicity and low cost of fluorescence spectroscopy encourage its use.

#### Acknowledgment

M.D. wishes to express his gratitude to the Minister of Science and Higher Education of the Polish Republic for funding the scholarship.

#### References

- [1] European Council Directive 95/60/EC.
- [2] European Commission Decision 2003/900/EC.
- [3] T. Lisinger, G. Koomen, H. Emteborg, G. Roebben, G. Kramer, A. Lamberty, *Energy Fuels* 18 (2004) 1851–1854.
- [4] DIN 51430.
- [5] H.D. Scott, B.W. Wright, *Talanta* 86 (2011) 148–156.
- [6] G.-Q. Chen, C.-Q. Ma, Y.-M. Wu, H.-J. Liu, S.-M. Gao, T. Zhu, Determination and identification of Sudan IV using fluorescence spectrometry and artificial neural networks, in: *The 4th International Conference on Information and Computing*, Phuket Island: 25–27 April 2011.
- [7] R. Bro, *Anal. Chim. Acta* 500 (2003) 185–194.
- [8] A.C. Olivieri, *Anal. Chem.* 80 (2008) 5713–5720.
- [9] M. Meira, C.M. Quintella, A.D.S. Tanajura, H.R.G. Da Silva, J.D.S. Fernando, P.R. Da Costa Neto, I.M. Pepe, M.A. Santos, L.L. Nascimento, *Talanta* 85 (2011) 430–434.
- [10] A.C. Olivieri, G.M. Escandar, A. Munoz de la Pena, *Trends Anal. Chem.* 30 (2011) 607–617.
- [11] M.J. Culzoni, H.C. Goicoechea, A.P. Pagani, M.A. Cabezon, A.C. Olivieri, *Analyst* 131 (2006) 718–723.
- [12] O. Divya, A.K. Mishra, *Appl. Spectrosc.* 62 (2008) 753–758.
- [13] D.L. Massart, B.G.M. Vandeginste, L.M.C. Buydens, S. de Jong, P.J. Lewi, J. Smeyers-Verbeke, *Handbook of Chemometrics and Qualimetrics Part A*, Elsevier, London, 1992.
- [14] J. Orzel, M. Daszykowski, B. Walczak, *Chemom. Intell. Lab. Syst.* 110 (2012) 89–96.
- [15] L.G. Thygesen, A. Rinnan, S. Barsberg, J.K.S. Moller, *Chemom. Intell. Lab. Syst.* 71 (2004) 97–106.
- [16] G. Tomasi, R. Bro, *Chemom. Intell. Lab. Syst.* 75 (2005) 163–180.
- [17] R.G. Zepp, W.M. Sheldon, M.A. Moran, *Mar. Chem.* 89 (2004) 15–36.
- [18] M. Bahram, R. Bro, C. Stedmon, A. Afkhami, *J. Chemometr.* 20 (2006) 99–105.
- [19] R. Bro, *J. Chemometr.* 10 (1996) 47–62.
- [20] R. Bro, *Multivariate Analysis in the Food Industry. Models, Algorithms, and Applications*, Royal Veterinary and Agricultural University Denmark, 1998. Available from:.
- [21] T. Naes, T. Isaksson, T. Fearn, T. Davies, *A User-friendly Guide to Multivariate Calibration and Classification*, NIR Publications, Chichester, 2002.
- [22] Q.-S. Xu, Y.-Z. Liang, *Chemom. Intell. Lab. Syst.* 56 (2001) 1–11.
- [23] M. Daszykowski, S. Semeels, K. Kaczmarek, P. Van Espen, C. Croux, B. Walczak, *Chemom. Intell. Lab. Syst.* 85 (2007) 269–277.
- [24] <http://www.models.kvl.dk/nwaytoolbox>.
- [25] C.B. Barber, D.P. Dobkin, H.T. Huhdanpaa, *ACM Trans. Math. Softw.* 22 (1996) 469–483.
- [26] A. Lorber, *Anal. Chem.* 58 (1986) 1167–1172.
- [27] A.C. Olivieri, N.M. Faber, J. Ferre, R. Boque, J.H. Kalivas, H. Mark, *Pure Appl. Chem.* 78 (2006) 633–661.

Katowice, 03.03.2015

dr hab. Michał Daszykowski, prof. UŚ  
Instytut Chemii  
Uniwersytet Śląski  
ul. Szkolna 9  
40-006 Katowice

Oświadczam, że w artykule pt. *Simultaneous determination of Solvent Yellow 124 and Solvent Red 19 in diesel oil using fluorescence spectroscopy and chemometrics*, opublikowanym w czasopiśmie *Talanta* 101 (2012) 78-84 mój udział polegał na:

- współpracy podczas formułowania hipotezy badawczej i ogólnej koncepcji badań,
- zapewnieniu nadzoru merytorycznego w tym pomocy w interpretacji wyników i weryfikacji wyciągniętych przez doktorantkę wniosków z przeprowadzonych badań,
- opiece i merytorycznym nadzorze podczas przygotowania manuskryptu,
- uczestnictwie w procedurze redakcyjnej (tj. pomocy w przygotowaniu odpowiedzi na recenzje)
- dokonaniu ostatecznej korekty artykułu.

Michał Daszykowski

Biała Podlaska, 29.09.2014

dr Ireneusz Grabowski  
Laboratorium Celne  
Izba Celna w Białej Podlaskiej  
Terminal Samochodowy w Koroszczynie  
21-550 Terespol

Oświadczam, że w artykule pt. *Simultaneous determination of Solvent Yellow 124 and Solvent Red 19 in diesel oil using fluorescence spectroscopy and chemometrics*, opublikowanym w czasopiśmie *Talanta* 101 (2012) 78-84 mój udział polegał na:

- postawieniu hipotez badawczych i pomocy w interpretacji uzyskanych wyników.



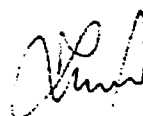


Biała Podlaska, 29.09.2014

mgr Grzegorz Zaleszczyk  
Laboratorium Celne  
Izba Celna w Białej Podlaskiej  
Terminal Samochodowy w Koroszczynie  
21-550 Terespol

Oświadczam, że w artykule pt. *Simultaneous determination of Solvent Yellow 124 and Solvent Red 19 in diesel oil using fluorescence spectroscopy and chemometrics*, opublikowanym w czasopiśmie *Talanta* 101 (2012) 78-84 mój udział polegał na:

- pomocy w postawieniu hipotez badawczych i w interpretacji uzyskanych wyników.



Biała Podlaska, 29.09.2014

mgr inż. Mirosław Sznajder  
Laboratorium Celne  
Izba Celna w Białej Podlaskiej  
Terminal Samochodowy w Koroszczynie  
21-550 Terespol

Oświadczam, że w artykule pt. *Simultaneous determination of Solvent Yellow 124 and Solvent Red 19 in diesel oil using fluorescence spectroscopy and chemometrics*, opublikowanym w czasopiśmie *Talanta* 101 (2012) 78-84 mój udział polegał na:

- pomocy w postawieniu hipotez badawczych i w interpretacji uzyskanych wyników.

.. - -



Katowice, 17.11.2014

prof. dr hab. Beata Walczak  
Instytut Chemii  
Uniwersytet Śląski  
ul. Szkolna 9  
40-006 Katowice

Oświadczam, że w artykule pt. *Simultaneous determination of Solvent Yellow 124 and Solvent Red 19 in diesel oil using fluorescence spectroscopy and chemometrics*, opublikowanym w czasopiśmie *Talanta* 101 (2012) 78-84 mój wkład polegał na merytorycznej dyskusji hipotez badawczych oraz uzyskanych wyników.

Beata Walczak

### 7.3. Publikacja III

#### *Identyfikacja nielegalnego procederu usuwania komponentów fiskalnych z oleju napędowego o obniżonej akcyzie*

---

Tytuł oryginalny: *Identifying the illegal removal from diesel oil of certain chemical markers that designate excise duty*

Autorzy: Joanna Orzeł  
Michał Daszykowski  
Ireneusz Grabowski  
Grzegorz Zaleszczyk  
Mirosław Sznajder  
Beata Walczak

Czasopismo: Fuel

Rok wydania: 2014

Tom: 117

Strony: 224 – 229

Wartość  
współczynnika  
*Impact Factor* : 3,406

---

W publikacji III zaproponowałam metodę badania legalności oleju napędowego ze względu na uwarunkowania prawne wynikające z wysokości nałożonego podatku akcyzowego. Próbki oleju napędowego zawierające komponenty fiskalne w ściśle określonym stężeniu poddałam symulacji ich nielegalnego usunięcia na drodze sorpcji. Każda próbka została scharakteryzowana pod kątem właściwości fluorescencyjnych. Na podstawie zestawu fluorescencyjnych obrazów próbek skonstruowałam modele dyskryminacyjne używając dyskryminacyjnego wariantu regresji częściowych najmniejszych kwadratów, D-PLS. Stosując różne kryteria dyskryminacyjne (zdefiniowane w oparciu o przepisy prawne obowiązujące w Rzeczypospolitej Polskiej oraz w Unii Europejskiej) uzyskałam modele umożliwiające poprawną dyskryminację próbek zawartych w niezależnych zbiorach testowych z dokładnością od 77% do 100% (w zależności od przyjętego kryterium).



ELSEVIER

Contents lists available at ScienceDirect

Fuel

journal homepage: [www.elsevier.com/locate/fuel](http://www.elsevier.com/locate/fuel)

## Identifying the illegal removal from diesel oil of certain chemical markers that designate excise duty



J. Orzel<sup>a</sup>, M. Daszykowski<sup>a,\*</sup>, I. Grabowski<sup>b</sup>, G. Zaleszczyk<sup>b</sup>, M. Sznajder<sup>b</sup>

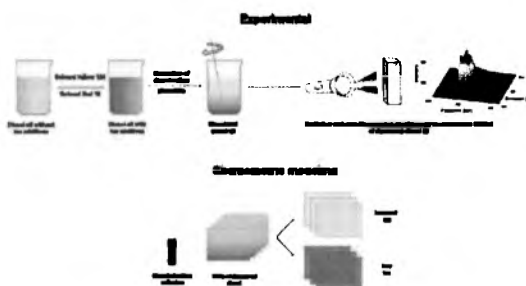
<sup>a</sup> Institute of Chemistry, The University of Silesia, 9 Szkolna Street, 40-006 Katowice, Poland

<sup>b</sup> Customs Chamber of Customs Laboratory in Biala Podlaska, 21 Celnikow Polskich Street, 21-500 Biala Podlaska, Poland

### HIGHLIGHTS

- Procedure of rebated tax diesel oil discoloration was performed in a laboratory.
- Fluorescence properties of discolored rebated tax diesel oil were studied.
- On the basis of 3D fluorescence signals discrimination models were constructed.
- A tool for discolored diesel oil discrimination is proposed.

### GRAPHICAL ABSTRACT



### ARTICLE INFO

#### Article history:

Received 10 July 2013

Received in revised form 6 September 2013

Accepted 10 September 2013

Available online 23 September 2013

#### Keywords:

Solvent Yellow 124

Solvent Red 19

Euromarker

Sudan Red 7B

EEM

### ABSTRACT

The fluorescent fingerprints of diesel oil samples were investigated in order to develop a fast and cost-effective method to facilitate the discrimination of rebated tax diesel fuel from oil that is illegally processed by the sorption process. In the experiment oil samples were spiked with a fiscal marker (Solvent Yellow 124) and a dye (Solvent Red 19) and then these were removed using a simulated sorption process. The excitation–emission fluorescence fingerprints were recorded for each sample. Discriminant models were constructed on the basis of fluorescence spectra to distinguish oil samples with respect to four possible discrimination schemes (corresponding to the concentrations of chemical additives). Using discriminant partial least squares models, in all of the discrimination cases that were considered, 100% of the samples from the model set were discriminated correctly. Prediction results for the test set samples were encouraging and varied between 77% and 100% of correctly discriminated samples.

© 2013 Elsevier Ltd. All rights reserved.

### 1. Introduction

Diesel oil is a complex mixture of different chemical compounds. The major components of diesel are saturated and aromatic hydrocarbon compounds (usually containing from 8 to 21 carbon atoms per molecule with a boiling point between 290 and 330 °C) [1]. Their amounts in commercially available diesel oil vary within a wide range. Different additives are also used to modify and improve the properties of a fuel and this considerably in-

creases the chemical variability of samples. Among the additives used, there are some that increase the cetane number, depressants, corrosion inhibitors, antioxidants, biocides, anti-foaming additives, dyes, markers, bio components (e.g. fame acid methyl esters – FAME), etc. Chemical additives are incorporated into the oil in various concentrations and thus diesel oil properties are deliberately modified. Consequently, oil can be characterized according to different criteria, for example, quality, which can be described by the cetane number, density, viscosity or sulfur content of the fuel [2]. The proposed use of diesel oil can be regarded as yet another classification criterion. It is commonly used for vehicle transport, heating, driving agricultural machinery and shipping goods. Diesel

\* Corresponding author. Tel. +48 32 359 1568; fax: +48 32 259 9978.  
E-mail address: [mdaszyk@us.edu.pl](mailto:mdaszyk@us.edu.pl) (M. Daszykowski).

oil designated for purposes other than a regular transport is spiked with chemical additives that change its physical properties (e.g. changing the color from yellow to red). This is done to prevent the illegal use and distribution of rebated excise tax diesel oil from being introduced on the market in order to reduce the costs of heating, agricultural production and marine transportation. Thus, another criterion for diesel oil characteristics is also available, i.e. tax rate criterion. Two excise tax components have been introduced in all countries that use such regulations. However, their type and concentrations are strictly defined by local laws. The first excise tax component is a marker. The same marker is used in all member states of the European Union – Solvent Yellow 124 (SY124). The second excise tax component is a dye that changes the color of diesel oil. Oil that is designed to be used for heating and agricultural machinery purposes usually contains a red dye and oil for marine transport is usually spiked with a blue dye. In Poland, two red dyes are used interchangeably: Solvent Red 19 (SR19) and Solvent Red 164 (SR164). Solvent Blue 35 (SB35) is the only blue dye used in Poland [3]. Diesel oil that is meant for vehicle transport has no tax components.

In order to distinguish diesel oil that has a rebated or full rate of duty, discrimination can be performed based on the chemical content of samples. It is a relatively easy task when samples of diesel oil are marked properly. Some typical methods that are used to describe the chemical content of samples, like UV-Vis spectrophotometry, have the potential to reveal differences between spiked and unspiked samples (i.e. samples are characterized by different colors and absorption properties). A problem occurs when the samples being examined are not genuine. In such a case, excise tax additives are removed from diesel oil and a 'cleaned' variant is sold at a higher price. Some efforts have been undertaken in order to detect the illegal process of the sorption of excise tax components. In 2007, Zadora and in 2011 Grabowski et al. investigated the chemical composition of sediment samples that were suspected to be the remains of the sorption process [4,5]. However, to the best of our knowledge, no procedure to distinguish illegally processed diesel oil samples from those with a full rate of duty or for the detection of residuals that remain after the sorption process has been described in the literature. The goal of our study was to develop a simple and cost-effective method to facilitate the discrimination of diesel oil after the removal of tax additives.

Because of the fluorescence properties of the diazo compounds that are introduced into diesel oil (i.e. marker and dyes), fluorescence spectroscopy is considered to be a sensitive analytical technique with the potential to detect the residue of tax compounds or changes in the chemical composition that are due to the discoloration process. In the literature, fluorescence spectroscopy has been acknowledged as a valuable method to characterize the quality of fuel [6–10].

Bearing in mind the different spectral characteristics of markers and dyes (i.e. different characteristic excitation and emission ranges) and possible analyte-background interactions (i.e. fluorescent molecules that are present in the matrix can contribute to analytical signal that is registered), the excitation-emission (EEM) fluorescence spectroscopy was the fingerprinting technique that was used to provide the comprehensive chemical information for complex mixtures in our study. Using the EEM spectroscopy, a sample is excited with a number of excitation wavelengths and the emission spectra for every excitation wavelength are registered. Thus, during a single measurement one can simultaneously observe the fluorescence of fluorophores that have different spectral properties. The complexity of the data that is obtained with EEM fluorescence spectroscopy is great and often chemical information that explains the phenomenon that is being studied requires chemometric data mining. Therefore, in order to extract the relevant

chemical information advanced chemometric methods are necessary [11–13].

In this study a set of oil samples with and without SY124 and SR19 was characterized by its EEM fingerprints. Then, these were preprocessed and a chemometric analysis was performed. Principal component analysis (PCA) [14] was carried out first in order to explore the structure of the data. Next, a partial least squares discriminant model (PLS-DA) [15] was built to explain any differences in the chemical composition of the oil samples with respect to the concentrations of SY124 and SR19 in order to detect any levels below the concentration limits that are acceptable under Polish law.

## 2. Materials and methods

### 2.1. Experimental

#### 2.1.1. Preparation of samples

Diesel oil, which was purchased at a local gas station, was spiked with SY124 (obtained from Sigma-Aldrich, 98.0% purity) and dyed with SR19 (obtained from IBPO, 92.3% purity). The concentrations of the additives in the oil samples were kept within the range of between 0 and 10 mg L<sup>-1</sup>. Then, the marker and dye were removed from the samples using the sorption process simulated in a laboratory environment. For each sample, the same discoloration procedure was performed. 10 mL of a freshly prepared sample was introduced into a vessel with 1.50 g of a sorption agent and mixed. After 10 min the sample was filtered using a paper filter. Then, the EEMs of the filtered and discolored samples were collected. In total, 20 different concentration levels of a marker and dye were tested. Three laboratory replicates for a particular concentration level were prepared (independent samples were prepared individually). For each laboratory replicate three EEMs were registered (the so-called technical replicates), which led to 180 EEMs.

#### 2.1.2. Fluorescence measurements of diesel oil samples

All diesel oil samples were measured spectrofluorometrically in a 10 × 10 mm quartz cuvette using a Carry Eclipse Varian FLO811M000 spectrofluorometer with the right-angled geometry. The emission spectra from 350 nm to 800 nm were measured in 2 nm intervals (226 wavelengths) at 46 excitation wavelengths from 250 nm to 700 nm (10 nm interval) with the detector sensitivity of the detector set to 500 V.

### 2.2. Chemometric analysis of EEMs data

#### 2.2.1. Data exploration – principal component analysis

Data exploration is one of the first steps in data analysis. It helps to uncover any similarities among samples and/or variables. The most popular method used for that purpose is principal component analysis (PCA) [14,16,17]. During PCA new features, called principal components (PCs), are constructed. They aim to maximize the description of the total data variance. When data variables are strongly correlated, a small number of PCs sufficiently summarize the data structure and data dimensionality is greatly reduced. An efficient data compression leads to a few PCs that can be used to visualize the data structure using the so-called score and loading plots. A score plot is a projection of samples onto a new coordinate system that is defined by selected pairs of PCs that explain a large portion of the data variability. Projection(s) often reveal a clustering tendency in the data (if any exists). Furthermore, coding samples on a score plot according to additional information, e.g. the concentration of the chemicals of interests, groups of samples,

etc., facilitates the analysis of any trends in the data. When loading plots are interpreted, one can study the contribution of explanatory variables into the construction of each PC and explain the specific data structure that is observed on a corresponding score plot [18].

### 2.2.2. Construction of discrimination models – partial least squares discriminant analysis

The purpose of discrimination methods is to divide a set of samples into a number of groups that are characterized according to their physico-chemical properties, e.g., the discrimination of fuel samples with an increased or decreased tax rate. The discrimination based on multivariate chemical data can be performed using the partial least squares discriminant model (PLS-DA) [19]. It is a useful technique when the data contain a large number of correlated variables because only a few latent factors are necessary to efficiently model the relationship between a set of explanatory variables (EEMs) and a dependent variable. For a two group problem, a dependent variable is represented by a binary vector that indicates membership for each sample group (a tax level – increased or decreased). A PLS-DA model is constructed for a limited number of latent variables and their number is called model complexity,  $f$ . The construction of a model with an acceptable discrimination power requires the optimization of the number of latent variables. This is usually done using the so-called cross-validation procedure [20].

## 2.3. Data set and data pretreatment

### 2.3.1. Data preprocessing

The first preprocessing step of excitation–emission fluorescence matrices is to handle the Rayleigh scattering. This is a chemically irrelevant component of the EEM spectra and thus can negatively influence their modeling. In the literature, there are several proposals as to how it can be corrected. For instance, the Rayleigh scattering can be removed from the EEM data by replacing its corresponding values with zeros [21,22], treating them as a missing values [23] or interpolation using different approaches [24,25]. In this paper the Rayleigh scattering was removed from the EEMs and interpolated with the Delaunay triangulation as described in Ref. [24].

Next, the EEMs were unfolded. A single fluorescence matrix was rearranged into a row vector containing 46 augmented emission spectra (each consisting of 226 sampling points), see Fig. 1. The unfolded 180 EEMs produced a data matrix of size  $180 \times 10396$ .

The last step in the data preprocessing workflow consisted of averaging the intensities of the fluorescence measurements that are observed for three technical replicates that are recorded for each laboratory sample. The mean spectra formed the final data

set – a matrix of size  $60 \times 10396$  (samples  $\times$  (emission wavelengths  $\times$  excitation wavelengths)).

### 2.3.2. Discrimination schemes

Before the construction of a dependent variable, a careful study of any possible fluctuations in the chemical composition of the samples should be carried out. For example, if a diesel oil of a full rate tax value has been transported in a container that had previously been used to transport a fuel with another tax level, it may be a source of residues from a spiked oil that will mix with a diesel oil that does not have excise tax additives. Therefore, these samples can be incorrectly recognized as samples after the sorption process. Thus, in order to draw a definite conclusion about the presence of tax additives, it would be an insufficient criterion to discriminate oil samples. A discrimination model as a discrimination criterion should also take into account the concentration levels of the excise tax additives (described by law). According to Polish law [3], Solvent Yellow 124 is introduced to diesel oil in an amount that varies between  $6.0 \text{ mg L}^{-1}$  and  $9.0 \text{ mg L}^{-1}$ . The concentration level of SY19 must be at least equal to  $6.3 \text{ mg L}^{-1}$  or larger. Taking into account the concentration criterion and experimental settings, some of the available samples can be labeled as either a full rate of duty or as rebated tax samples. However, in a data set there may be samples with ambiguous, according to the law, concentrations of a marker and dye that cannot support any definite discrimination. For example, a sample with concentration of a marker equal to  $7 \text{ mg L}^{-1}$  (that defines a rebated tax oil) and a concentration of dye  $2 \text{ mg L}^{-1}$  (oil with a full rate of duty). Therefore, in this study four different discrimination problems were considered. Each discrimination problem is described by a dependent binary variable. In all of the discrimination cases, the available samples were divided into two groups with respect to the concentration of an additive(s). Samples included in the full rate of duty group of oils are denoted as '1' and the remaining samples (form the rebated tax group) as '-1'. The design of the experiment and the discrimination schemes that were investigated are presented in Fig. 2.

The first discrimination problem, n1, is based on a concentration of a marker compound (samples characterized with SY124 concentration below the level described in the law were included in the full rate of duty group, and the rest of the samples formed the rebated tax group). The concentration of a dye was used as the second criterion. The remaining two discrimination problems focus on two additives. For n3, simultaneous concentrations of both a marker and dye should exceed the levels described by the law to assign samples to the group of oils with a full rate of duty. For n4, the concentration of at least one of the tax additives should be equal to or larger than the specified concentration in order to recognize a sample as a rebated tax oil.

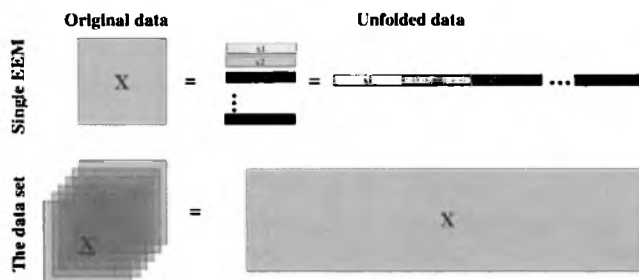


Fig. 1. Different possibilities of EEM data representation – the unfolding of the EEM fluorescence fingerprints.

	Concentration [ $\text{mg} \cdot \text{l}^{-1}$ ]																			
SY124	0	0	0	0	2	2	2	4	4	5	5	5	7	7	7	9	9	10	10	10
SM19	0	2	5	10	0	2	7	3	8	0	5	10	0	2	7	3	8	0	5	10
n1	1	1	1	1	1	1	1	1	1	1	1	1	-1	-1	-1	-1	-1	-1	-1	-1
n2	1	1	1	-1	1	1	-1	1	-1	1	1	-1	1	1	-1	1	-1	1	1	-1
n3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	-1	1	-1	1	1
n4	1	1	1	-1	1	1	-1	1	-1	1	1	-1	1	1	-1	1	-1	1	1	-1

Fig. 2. Design of the experiment indicating four discrimination problems.

### 3. Results and discussion

The data that was studied were first preprocessed according to the steps described in Section 2.3. In order to efficiently handle the Rayleigh scattering, the location of the peaks related to the scattering were detected (in all of the emission spectra registered). Then, fifteen sampling points corresponding to the Rayleigh scattering peaks were removed and interpolated using the procedure described in Ref. [24]. The corrected data were unfolded, averaged (as described in Section 2.3) and centered column-wise prior to further analysis. The first two principal components described over 99.05% of the total data variance. The PCA score plot, which displays 60 diesel oil spectra on the PC1 and PC2 plane, is

presented in Fig. 3. In order to investigate the scatter of the samples included in the tax groups being considered, they were marked according to four possible discrimination problems (see Section 2.3).

Exploring the PC1 vs. PC2 score plot, two clusters of samples are observed. However, they do not represent any of the discrimination problems being considered. The score plot presented in Fig. 3a helps to spot separate clusters that match discrimination problem n1. Thus, the concentration of SY124 is a possible marker that can be used to indicate the process of illegal sorption. Further interpretation of the score plots with the other three discrimination criteria indicated (Fig. 3b–d) could not be associated with any particular cluster of samples.

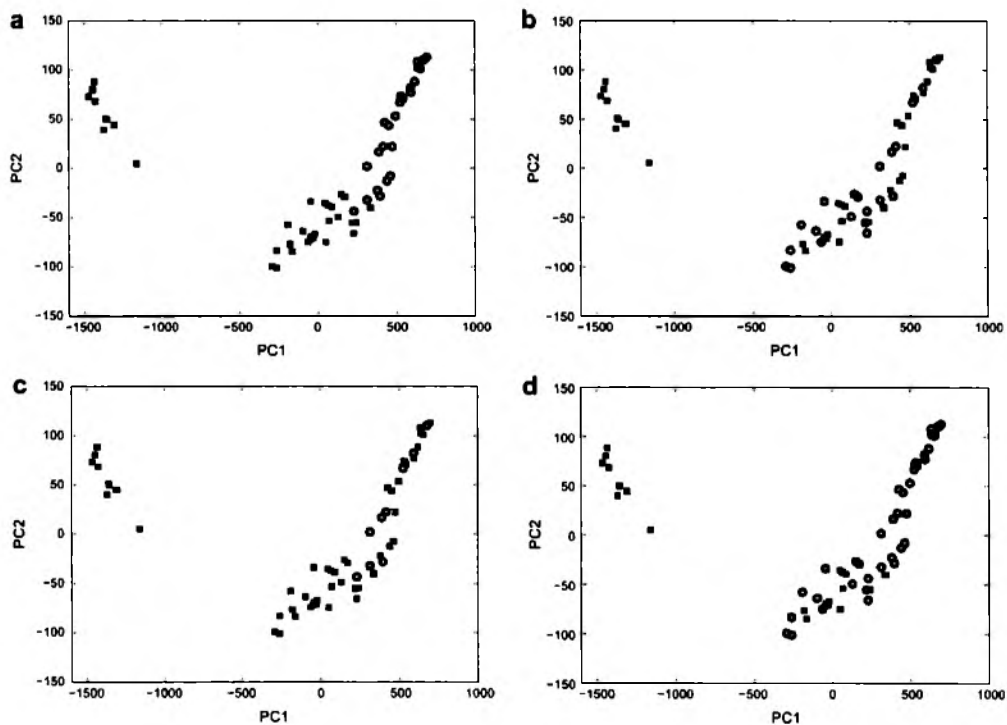


Fig. 3. PC1 vs. PC2 score plots of oil samples marked according to four discrimination schemes (a) n1, (b) n2, (c) n3 and (d) n4. Samples included in the rebated tax group are marked with 'L' and samples included in the full rate of duty group are marked with 'O'.



Table 1

Discrimination results obtained from the PLS-DA models with the optimal number of factors. *F* and *R* indicate the number of samples from the full rate of duty and the rebated tax oil groups, respectively.

Binary dependent variable	The considered compound	Samples in the model set	Samples in the test set	<i>f</i>	Number of correctly discriminated samples in model set	Number of correctly discriminated samples in test set
n1	SY124	30 ( <i>F</i> = 15 <i>R</i> = 15)	30 ( <i>F</i> = 9 <i>R</i> = 21)	3	30 (100%)	30 (100%)
n2	SR19	30 ( <i>F</i> = 15 <i>R</i> = 15)	30 ( <i>F</i> = 6 <i>R</i> = 24)	10	30 (100%)	23 (77%)
n3	SY124 and SR19	14 ( <i>F</i> = 7 <i>R</i> = 7)	46 ( <i>F</i> = 2 <i>R</i> = 44)	6	14 (100%)	35 (78%)
n4	SY124 or SR19	30 ( <i>F</i> = 15 <i>R</i> = 15)	30 ( <i>F</i> = 9 <i>R</i> = 21)	10	30 (100%)	27 (90%)

After the PCA analysis, multivariate discriminant models were constructed. For each of the four discrimination problems, samples were split into model and test sets, according to the following procedure:

- (i) samples were divided into two subsets, i.e. denoted as '–1' and as '1'.
- (ii) an equal number of samples was selected from each subset using the Kennard and Stone algorithm [26] (selecting samples with unique characteristics) and included into the model set, while the remaining samples formed the test set.

The number of samples defined as the full rate of duty and rebated tax oil that were included in each model and test set is different (see Table 1).

The model and test sets were then used for the construction and validation of discriminant models. In the case of discriminant problem n1, three latent factors were found to be sufficient to construct a model with an acceptable performance, i.e. all model and test sets samples are correctly recognized.

For the second discrimination problem (n2), the optimal discrimination results were achieved with a model containing ten latent factors. However, its performance is worse compared to the previous model, i.e. all samples from the model set are recognized correctly but seven samples (30%) from the test set are incorrectly classified. A better performance is observed when the samples are discriminated with respect to the concentrations of a marker and dye. 86% and 93% of the test set samples are correctly classified using the discrimination models constructed for the n3 and n4 discrimination criteria, respectively. The discrimination results obtained from all of the models constructed are summarized in Table 1.

An analysis of the results obtained from the optimal models allows the conclusion to be drawn that discrimination criterion n1 is the best, i.e. the concentration of SY124 is a marker that indicates a possible sorption process with a high probability.

Misclassified samples according to discrimination criterion n2, are characterized by concentration of SR19 in the range from 0 to 8 mg L<sup>-1</sup>. However, concentration of SY124 in these samples is high, i.e. 9 (two samples) and 10 (five samples) mg L<sup>-1</sup>. It can be concluded that discrimination criterion based on the SR19 concentration is affected to some extent by presence of SY124. Thus, criterion n2 can be expected to yield the worst model's performance (from all possible discrimination schemes).

According to n3 criterion samples with a high SY124 concentration (i.e. 9 mg L<sup>-1</sup> and above) are incorrectly discriminated. For four wrongly discriminated samples using criterion n4 it is impossible to explain the fact of their incorrect discrimination using concentrations of SY124 and SR19.

EEM fingerprints can be modeled using the *n*-way approach, e.g. *n*-PLS [27]. The *n*-PLS method can be regarded as an extension of PLS to handle data with tri-linear structure, and by including this property into modeling, compared with classic PLS, may have a better performance. Initially, we have modeled our data using the *n*-PLS approach. However, for our data, in general, *n*-PLS did not outperformed than standard PLS in terms of discrimination

error. In case of criterion n1, the PLS model (with three factors only) performed better than more complex *n*-PLS (with four factors). The complexity of the optimal *n*-PLS model for criterion n2 contains eleven factors (one more than in PLS) and the performance of this model is worse (in case of PLS correct classification rate approaches 77%, whereas for *n*-PLS 70%). The *n*-PLS model constructed for criterion n3 requires six factors and performs virtually the same as the six-factor PLS model. Criterion n4 can be modeled using *n*-PLS with the same effectiveness using one factor less than for the optimal PLS model constructed. Since the *n*-PLS models are more complex, we have decided to discuss results only using a classic PLS modeling.

#### 4. Conclusions

In this paper, a method to facilitate the discrimination of a rebated tax diesel oil from samples after the sorption process was introduced. This goal was achieved using the excitation-emission fluorescence spectroscopic signals that were preprocessed and then modeled using chemometric tools. The data exploration performed using PCA supported the hypothesis that Solvent Yellow 124 can be regarded as a possible marker for the detection of an illegal sorption process. The best discrimination results based on the PLS-DA model, i.e. 100% correctly discriminated samples (both in the model and test sets) were obtained by focusing on the marker concentration (discrimination criterion n1). Other discrimination models had a performance that was similar to the first model. The model with the worst predictive performance, only 77% of correctly classified samples, was constructed assuming the concentration of a dye, Solvent Red 19, as the discriminant criterion. The results that were obtained support the hypothesis that residues of a marker indicate a possible sorption process. The dye was practically completely removed during the sorption procedure, and thus the discrimination based on its concentration is the least effective. The method presented here can be considered as a solution for the detection of illegally preprocessed oil samples. However, in order to obtain a general solution, a larger pool of samples of commercially available diesel oil should be included in the data set that is used for the model construction and validation of the model. To adopt our procedure to comply with regulations of other countries, their law need to be studied. A set of samples with marker and a particular dye (used in a considered country) need to be prepared and the cut-off value corresponding to assumed in the country law concentration level have to be used.

Simulated sorption process used in our experiment is one of the possible activities that influence properties of rebated tax diesel oil. To extend the proposed method to be used in a forensic context (useful in court trial) other activities (e.g. transport and storage conditions) need to be considered. It can be performed after adopting, e.g. the likelihood ratio approach [28].

#### Acknowledgements

M.D. wishes to express his gratitude to the Minister of Science and Higher Education of the Polish Republic for funding the scholarship.

J.O. appreciates the support of National Science Center Grant No. 2011/03/N/ST4/00713 and the UPGOW fellowship.

## References

- [1] Agency for toxic substances and disease registry, toxicological profile for fuel oils. Atlanta: U.S. Department of Health and Human Services; 1995. [cited 07.02.12]. <http://www.atsdr.cdc.gov/toxprofiles/tp75.pdf>.
- [2] Dabelstein W, Reglitzky A, Schütze A, Reders K. Automotive fuels. In: Ullmann F, editor. Ullmann's encyclopedia of industrial chemistry. Weinheim: Wiley; 2007.
- [3] Rozporządzenie ministra finansów w sprawie znakowania i barwienia wyrobów energetycznych [Disposition of Polish Minister of Finance concerning marking and dyeing of energy products] (Dz. U. 2010, No. 157, item 1054).
- [4] Zadora G. Laundering of "Illegal" fuels – a forensic chemistry perspective. Acta Chim Slov 2007;54:110–3.
- [5] Grabowski I, Zaleszczyk G. Identification of sediments from discoloration of energetic fuels with soil performed with FTIR and ED-XRF spectroscopy. In: Proceedings of the 15th polish symposium on applications of spectroscopic methods for investigation of materials and chemical compounds. Poznan, Poland; 2011 May 25–27.
- [6] Alexander J, Mashak G, Kapitan N, Siegel JA. Fluorescence of petroleum products. II. Three-dimensional fluorescence plots of gasolines. J Forensic Sci 1987;32:72–86.
- [7] Sheff M, Siegel J. Fluorescence of petroleum products V: three-dimensional fluorescence spectroscopy and capillary gas chromatography of neat and evaporated gasoline samples. J Forensic Sci 1994;39:1201–14.
- [8] Siegel J. Solving crimes with 3-D fluorescence spectroscopy. Anal Chem 1985;57:934–40.
- [9] Christensen JH, Hansen AB, Mortensen J, Andersen O. Characterization and matching of oil samples using fluorescence spectroscopy and parallel factor analysis. Anal Chem 2005;77:2210–7.
- [10] Divya D, Mishra AK. Multivariate methods on the excitation emission matrix fluorescence spectroscopic data of diesel–kerosene mixtures: a comparative study. Anal Chim Acta 2007;592:82–90.
- [11] Guimet F, Ferre J, Boque R, Rius FX. Application of unfold principal component analysis and parallel factor analysis to the exploratory analysis of olive oils by means of excitation–emission matrix fluorescence spectroscopy. Anal Chim Acta 2004;515:75–85.
- [12] Persson T, Wedborg M. Multivariate evaluation of the fluorescence of aquatic organic matter. Anal Chim Acta 2001;434:179–92.
- [13] Piccirilli GN, Escandar GM. Second-order advantage with excitation–emission fluorescence spectroscopy and a flow-through opensing device. Simultaneous determination of thiabendazole and fuberidazole in the presence of uncalibrated interferences. Analyst 2010;135:1299–308.
- [14] Wold S, Esbensen K, Geladi P. Principal component analysis. Chemom Intell Lab Syst 1987;2:37–52.
- [15] Barker M, Rayens W. Partial least squares for discrimination. J Chemom 2003;17:166–73.
- [16] Daszykowski M, Walczak B, Massart DL. Projection methods in chemistry. Chemom Intell Lab Syst 2003;65:97–112.
- [17] Honorato FA, Neto BdB, Pimentel MF, Stragevitch L, Galvão RKH. Using principal component analysis to find the best calibration settings for simultaneous spectroscopic determination of several gasoline properties. Fuel 2008;87:3706–9.
- [18] Brown SD, Tauter R, Walczak B. Comprehensive chemometrics. London: Elsevier; 2009.
- [19] Massart DL, Vandeginste BGM, Budgens LM, de Jong S, Lewi PJ, Smeyers-Verbeke J. Handbook of chemometrics and qualimetrics part A. London: Elsevier; 1992.
- [20] Naes T, Isaksson T, Fearn T, Davies T. A user-friendly guide to multivariate calibration and classification. Chichester: NIR Publications; 2002.
- [21] Orzel J, Daszykowski M, Walczak B. Controlling sugar quality on the basis of fluorescence fingerprints using robust calibration. Chemom Intell Lab Syst 2012;110:89–96.
- [22] Thygesen LG, Rinnan A, Barsberg S, Moller JKS. Stabilizing the PARAFAC decomposition of fluorescence spectra by insertion of zeros outside the data area. Chemom Intell Lab Syst 2004;71:97–106.
- [23] Tomasi G, Bro R. PARAFAC and missing values. Chemom Intell Lab Syst 2005;75:163–80.
- [24] Zepp RG, Sheldon WM, Moran MA. Dissolved organic fluorophores in southeastern US coastal waters: correction method for eliminating rayleigh and raman scattering peaks in excitation–emission matrices. Mar Chem 2004;89:15–36.
- [25] Bahram M, Bro R, Stedmon C, Afkhami A. Handling of rayleigh and raman scatter for PARAFAC modeling of fluorescence data using interpolation. J Chemom 2006;20:99–105.
- [26] Kennard RW, Stone LA. Computer aided design of experiments. Technometrics 1969;11:137–48.
- [27] Bro R. Multiway calibration, multilinear PLS. J Chemom 1996;10:47–62.
- [28] Aitken CGG, Lucy D. Evaluation of trace evidence in the form of multivariate data. J R Stat Soc 2004;53:109–22.

Katowice, 03.03.2015

dr hab. Michał Daszykowski, prof. UŚ  
Instytut Chemii  
Uniwersytet Śląski  
ul. Szkolna 9  
40-006 Katowice

Oświadczam, że w artykule pt. *Identifying the illegal removal from diesel oil of certain chemical markers that designate excise duty*, opublikowanym w czasopiśmie Fuel 117 (2013) 224-229 mój udział polegał na:

- współtworzeniu hipotezy badawczej i koncepcji badań,
- zapewnieniu nadzoru merytorycznego w tym pomocy w interpretacji wyników i weryfikacji wyciągniętych przez doktorantkę wniosków z przeprowadzonych badań,
- opiece i merytorycznym nadzorze podczas przygotowania manuskryptu,
- pomocy w procedurze redakcyjnej (tj. pomocy w przygotowaniu odpowiedzi na recenzje),
- dokonaniu ostatecznej korekty artykułu.

Michał Daszykowski

Biała Podlaska, 29.09.2014

dr Ireneusz Grabowski  
Laboratorium Celne  
Izba Celna w Białej Podlaskiej  
Terminal Samochodowy w Koroszczynie  
21-550 Terespol

Oświadczam, że w artykule pt. *Identifying the illegal removal from diesel oil of certain chemical markers that designate excise duty*, opublikowanym w czasopiśmie Fuel, 117 (2013) 224-229 mój udział polegał na:

- pomocy w postawieniu hipotez badawczych i w interpretacji uzyskanych wyników.



Biała Podlaska, 29.09.2014

mgr Grzegorz Zaleszczyk  
Laboratorium Celne  
Izba Celna w Białej Podlaskiej  
Terminal Samochodowy w Koroszczynie  
21-550 Terespol

Oświadczam, że w artykule pt. *Identifying the illegal removal from diesel oil of certain chemical markers that designate excise duty*, opublikowanym w czasopiśmie Fuel 117 (2013) 224-229 mój udział polegał na:

- pomocy w postawieniu hipotez badawczych i w interpretacji uzyskanych wyników.

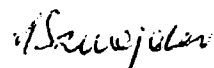


Biała Podlaska, 29.09.2014

mgr inż. Mirosław Sznajder  
Laboratorium Celne  
Izba Celna w Białej Podlaskiej  
Terminal Samochodowy w Koroszczynie  
21-550 Terespol

Oświadczam, że w artykule pt. *Identifying the illegal removal from diesel oil of certain chemical markers that designate excise duty*, opublikowanym w czasopiśmie Fuel 117 (2013) 224-229 mój udział polegał na:

- pomocy w postawieniu hipotez badawczych i w interpretacji uzyskanych wyników.



#### 7.4. Publikacja IV

*Szybka metoda oceny zdolności antyoksydacyjnej żywności używając fluorescencyjnych obrazów na przykładzie ekstraktów kawy i mięty pieprzowej*

---

Tytuł oryginalny: *A rapid validation of the antioxidant capacity of food commodities based on their fluorescence excitation emission spectra as applicable to coffee and peppermint extracts*

Autorzy: Joanna Orzeł  
Michał Daszykowski

Czasopismo: Chemometrics and Intelligent Laboratory Systems

Rok wydania: 2014

Tom: 137

Strony: 74 – 81

Wartość  
współczynnika  
*Impact Factor\**: 2,381

---

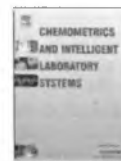
W publikacji IV opisuję opracowaną przeze mnie metodykę analityczną, opierającą się na pomiarach fluorescencyjnych obrazów, która umożliwia w sposób szybki i niewymagający odczynników chemicznych ocenić parametr jakości produktów spożywczych jakim jest ich całkowita zdolność antyoksydacyjna. Parametr wyznaczałam przy użyciu metody referencyjnej, dedykowanej analizie produktów spożywczych – test ORAC (ang. *oxygen radical absorbance capacity*). Równocześnie przeprowadziłam eksperyment, w którym oceniłam całkowitą zawartość polifenoli wykorzystując metodę z użyciem odczynnika Folina-Ciocalteau. Wyzaczyłam zawartość antyoksydantów (i związków fenolowych) w wodnych ekstraktach kawy i mięty pieprzowej. Scharakteryzowałam próbki ze względu na ich właściwości spektralne, rejestrując fluorescencyjne obrazy. Uzyskane dane podałam kalibracji używając regresji częściowych najmniejszych kwadratów lub jej N-modalnej wersji. Skonstruowane modle umożliwiają ocenę parametru ORAC dla wybranych ekstraktów z błędem niższym niż 9,10% (dla niezależnych zbiorów testowych). Efektywna kalibracja zawartości związków fenolowych była możliwa dla ekstraktów kawy (błąd przewidywania 7,02%). Niestety, dla ekstraktów mięty pieprzowej konstrukcja modelu o satysfakcjonujących parametrach, na podstawie uzyskanych danych była nieosiągalna.

Badania zamieszczone w publikacji realizowałam w ramach projektu badawczego Preludium, finansowanego przez Narodowe Centrum Nauki pt. *Fluorescencyjne krajobrazy w połączeniu z metodami chemometrycznymi jako potencjalne narzędzie do wyznaczania całkowitej zdolności antyoksydacyjnej*, nr 2011/03/N/ST4/00713, którego byłam kierownikiem.



Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: [www.elsevier.com/locate/chemolab](http://www.elsevier.com/locate/chemolab)

## A rapid validation of the antioxidant capacity of food commodities based on their fluorescence excitation emission spectra as applicable to coffee and peppermint extracts



Joanna Orzel, Michal Daszykowski\*

Institute of Chemistry, The University of Silesia, 9 Szkolna Street, 40-006 Katowice, Poland

### ARTICLE INFO

#### Article history:

Received 15 April 2014

Received in revised form 17 June 2014

Accepted 23 June 2014

Available online 28 June 2014

#### Keywords:

EEM

Non-selective analytical signals

Multivariate calibration

Chemometrics

### ABSTRACT

A novel and rapid approach to screening antioxidant content is proposed. Its usefulness is illustrated for coffee and peppermint extracts. The antioxidant properties of extracts were predicted from their excitation–emission fluorescence matrices using the partial least squares and the N-way partial least squares regression. The total antioxidant capacity and the total phenolic content were estimated using oxygen radical absorbance capacity (ORAC) and Folin–Ciocalteu assays, respectively. Prediction errors of calibration models describing the total antioxidant content evaluated using the ORAC assay were equal to 448.33 (6.29%) and 647.74 (9.04%) of the Trolox equivalents for coffee and peppermint extracts, respectively. The prediction error of the total phenolic content for coffee extracts was about 0.270 (7.02%) expressed as gallic acid equivalents. The limits of detection were below 85.43 of the Trolox equivalents and 0.038 of the gallic acid equivalents, and the relative standard deviation was in a range of between 1.61% and 7.13%.

© 2014 Elsevier B.V. All rights reserved.

### 1. Introduction

A lifestyle that ensures a long and healthy life requires introducing into one's diet food commodities that contain vital components necessary to maintain our body in good health (vitamins, omega-3 fatty acids, microelements, etc.). Potentially, these can help to reduce the chance of developing diseases (such as Alzheimer's, cancer and atherosclerosis) or inhibit the aging process. Due to their ability to deactivate free radicals, antioxidants are claimed to be one of the most important vital food components. The antioxidant properties of food products as well as their content of chemical compounds that are recognized as antioxidants vary considerably [1]. The total amount of antioxidants contained in a sample, which has the ability to scavenge free radicals, is described by the total antioxidant capacity (TAC) parameter. It is evaluated using antioxidant capacity assays [2]. The assays that have been developed to date are based on the chemical reaction between antioxidants and artificially generated free radicals. The reaction is monitored using an instrumental technique, for instance, spectrophotometry, fluorimetry or electron paramagnetic resonance spectroscopy. Another approach that is used to estimate the TAC parameter aims to evaluate the antioxidant content in samples by the quantification of some known antioxidants (e.g., using chromatographic methods). In this case, the TAC value is expressed as the sum of quantified antioxidant compounds [3], however, the result cannot be compared with the TAC

value of active antioxidants that are obtained from the chemical-based assays. Moreover, often concentrations of certain antioxidants do not resemble their true TAC. These phenomena can be explained by a possible synergic effect and/or redox interactions between chemical components [2]. The evaluation of the TAC for different food commodities has become quite popular these days. Standard assays, which are based on a chemical reaction, e.g., the Trolox equivalent antioxidant capacity assay (TEAC) [4], the DPPH assay [2,2-diphenyl-1-picrylhydrazyl] [5], the ferric reducing ability of plasma (FRAP) [6] or the oxygen radical antioxidant capacity assay (ORAC) [7] are frequently used for the evaluation of the TAC of vegetables [1], fruits [8], alcohol free beverages (e.g., coffee [9]) and alcoholic beverages that is beers and wines [10,11]. Although all of the methods that have been developed to estimate the TAC value are routinely applied, they are time or reagent consuming. A routine and rapid analysis with available assays that is applicable, e.g., during the production process or monitoring the TAC of food products is difficult to perform in an on-line setting.

On the other hand, many antioxidant compounds exhibit fluorescent properties and thus they can be screened using a fast, sensitive and non-destructive analytical technique – fluorescence spectroscopy. To date, fluorescence spectroscopy has been successfully applied for screening the properties of various products (e.g., riboflavin in dairy products [12]) or some global properties including the shelf life of yoghurt or beer [13,14]. The antioxidants that are contained in a complex food sample have different spectral characteristics (i.e. characteristic excitation and emission wavelengths). The basic emission spectrum that is collected for a single excitation wavelength is insufficient to describe

\* Corresponding author. Tel.: + 48 32591569; fax: + 48 32 259 99 78.  
E-mail address: [mdaszyk@us.edu.pl](mailto:mdaszyk@us.edu.pl) (M. Daszykowski).



all of the compounds of interest. Therefore, in order to obtain information about a wide range of fluorophores, a number of emission spectra are collected by the sequential excitation of a sample at certain wavelengths. These emission spectra are then arranged into a so-called excitation–emission fluorescence matrix (EEM), which represents a 2-D analytical signal containing emission spectra collected for particular excitation wavelengths in rows. Considering the relatively high information content, an EEM can be treated as a sample fingerprint that potentially resembles information about its TAC value. For instance, EEMs were used as fluorescence fingerprints of diesel oil samples in order to determine the concentrations of excise tax components [15] or to detect the illegal process of fuel laundering (the removal of excise duty components) [16]. They have also been considered to be useful analytical signals to monitor the quality of food products such as sugar [17], yoghurt [13] and edible oils [18]. Bearing in mind the advantages of fluorescence fingerprints, a novel approach to evaluate the TAC of food samples was developed based on their EEMs. Multivariate calibration methods were applied in order to extract useful chemical information from EEMs that described the TAC value. The TAC values were obtained from two reference analytical assays – the ORAC assay [7] and the total phenol content of samples assay, TPC [19]. To illustrate the performance of the proposed method, two food commodities such as coffee and peppermint, both of which have been relatively well described in the literature, were selected. ORAC and TPC values for each sample were expressed as the Trolox (TE) and gallic acid (GA) equivalents, respectively. The EEMs were registered in selected excitation and emission ranges. Then, the collected fluorescence fingerprints were related to their corresponding response TAC values using the partial least squares regression [20], PLS and the N-way partial least squares [21], N-PLS. The proposed method is characterized by its figures of merit (limits of detection and quantitation) as calculated for the optimal calibration models.

## 2. Materials and methods

### 2.1. Theory

#### 2.1.1. Determination of the total phenolic content

The total polyphenol content of samples was determined using the Folin–Ciocalteu method as described in detail in reference [19]. The calibration curve, which served as the standard, was constructed for gallic acid and the absorbance was recorded at 760 nm.

#### 2.1.2. The oxygen radical absorbance capacity assay

The response of the ORAC assay relies on the damage to a fluorescent agent – fluorescein, which is caused by free radicals and is monitored as the decrease of the fluorescence intensity over a certain period of time. The antioxidants that are present in a sample inhibit damage to the fluorescent agent and thus the fluorescence signal is sustained. Simultaneously, the fluorescence intensity of a blank probe (without antioxidants, but with the same amounts of a free radical generator and a fluorescence agent as in the original sample) is examined. These two chemical reactions are driven to completion. The TAC is expressed as the area between the two curves that described the decrease of fluorescence intensity over time for the blank and tested sample, respectively. In our study, a calibration curve was built for a set of standards with known TAC values. The water soluble vitamin E analog, which is called Trolox (6-hydroxy-2,5,7,8-tetramethylchroman-2-carboxylic acid), served as the standard antioxidant for the ORAC assay.

#### 2.1.3. Data modeling

Excitation–emission fluorescence matrices contain multiple fluorescence measurements, which result in 2-D spectra. For a set of samples, a collection of EEMs can be organized as a 3-D data array [22] with the dimensions *excitation wavelength* × *emission wavelength* × *samples*. To extract the relevant chemical information from the EEMs, advanced

chemometric preprocessing and modeling methods are necessary. One of the most important preprocessing steps for the EEMs consists of the detection and elimination of the Rayleigh and Raman scattering. The scattering is manifested on the EEMs as diagonal lines of peaks (see Fig. 1a). It is a chemically irrelevant component and can significantly influence the shape of the characteristic of the peaks for key fluorophores. Therefore, prior to the construction of the model, the scattering must be corrected. In this study, correction of scattering was done using the approach described earlier [23], although there are other methods also performing well [24]. Peaks corresponding to scattering were detected, removed and missing elements were interpolated using the Delaunay triangulation. The exemplary fluorescence signals (EEMs) of peppermint extract before and after scatter correction are shown in Fig. 1.

Prior to the construction of a calibration model, it is important to analyze the data structure to foresee any possible difficulties at the modeling stage. In particular, data homogeneity, including the detection of outlying samples and possible measurement errors that can negatively influence construction of a least squares model, is evaluated [25]. Data multivariate exploratory methods are required in order to uncover the structure of the EEMs. Principal component analysis, PCA [26], is frequently used to compress and visualize multivariate data. In PCA, the original data variables are described by a set of new variables, which are called principal components, PCs. They maximize the description of the data variance. Data structure is visualized on score and loading plots that reveal similarities among the samples and variables, respectively.

To predict the TAC values of samples from their EEMs, it is necessary to construct a multivariate calibration model. Assuming a linear relationship between the TAC value and fluorescence spectra, the partial least squares regression, PLS, is a straightforward choice [20]. Another

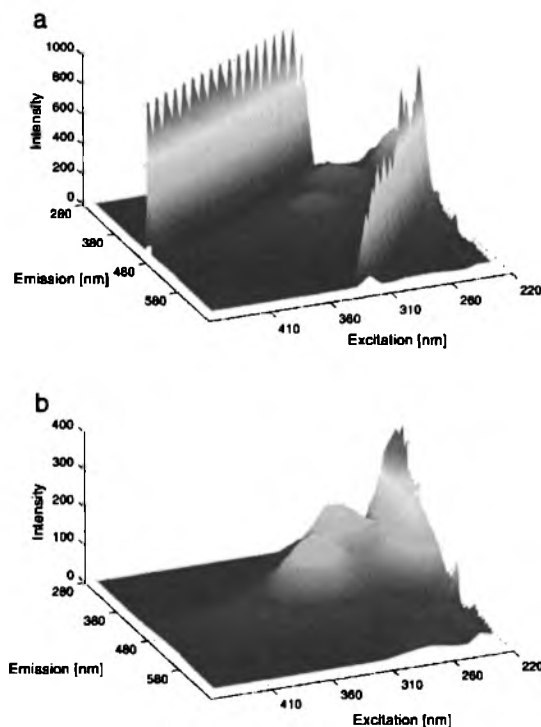


Fig. 1. Exemplary fluorescence signals (EEMs) for a peppermint extract: a) raw signal and b) signal after scatter correction.

more advanced option is the N-way partial least squares regression, N-PLS, [21] which takes into account the trilinear structure of the fluorescence data, which can possibly reinforce the calibration results [22] (improving the fit and prediction properties of how a N-PLS model compares with the PLS). The relationship between the dependent variable (TAC) and the explanatory variables (EEM) is described by a few latent factors in both calibration models. In PLS, the original 3-D data are unfolded into a matrix of the dimensions *samples* × (*excitation wavelength* × *emission wavelengths*). The unfolded data are then represented by a few latent factors that maximize the covariance between the modeled property and the explanatory variables.

In N-PLS, the 3-D data of the dimensions *samples* × *excitation wavelengths* × *emission wavelengths* are decomposed into the so-called triads that maximize the cost function and are used for the construction of the model.

In order to characterize the model fit and its prediction properties, the available set of samples is divided into training and test sets. The model set is used to construct a calibration model and samples from the test set are used to estimate its prediction properties. The fit and prediction of a model are described by the root mean squared error, RMSE, and the root mean squares error of prediction, RMSEP [20], respectively.

To evaluate the number of latent factors (complexity of a model), which is necessary to construct the optimal model, the cross-validation procedure is often applied [20]. In the cross-validation procedure, error estimates are derived based on the iterative deletion of a certain number of samples that serve as validation samples. At one step, a number of models are constructed with an increasing number of latent factors. Error estimates are obtained from each model for the group of validation samples. After repeating the cross-validation procedure a number of times, error estimates are averaged for each model complexity and expressed as the root mean square error of cross-validation, RMSECV.

## 2.2. Experimental

### 2.2.1. Reagents

The fluorescence probe – fluorescein disodium salt ( $C_{20}H_{10}Na_2O_3$ , 98.5–100.5% purity) was purchased from Fluka. 2,2'-Azobis (2-methylpropionamide) dihydrochloride, AAPH, ( $C_8H_{18}N_2$  2HCl, 97% purity grade) and ( $\pm$ )-6-hydroxy-2,5,7,8-tetramethylchromane-2-carboxylic acid, Trolox, ( $C_{14}H_{18}O_4$ , 97% purity grade) were obtained from Sigma-Aldrich. The phosphate buffer ingredients (sodium dihydrogen phosphate ( $NaH_2PO_4$ , pure grade p.a.) and di-sodium hydrogen phosphate dodecahydrate ( $Na_2HPO_4 \cdot 12H_2O$ , pure grade p.a.), Folin–Ciocalteu reagent (2.0 M) and sodium carbonate ( $Na_2CO_3$ , pure grade p.a.) were purchased from POCH Poland.

### 2.2.2. Preparation of coffee extracts

In our experiment, 15 types of 'arabica' whole bean coffee samples that had been collected from different geographical locations and that had been differently roasted were considered. In general, the coffee was produced in three continents: Asia, South America and Africa. Firstly, the coffee beans were ground in an electric coffee grinder and sieved (the size of the sieve was equal 0.43 mm). Then, 5 g of the sieved coffee sample were brewed in 50 mL of distilled water at 95 °C. After 5 min of brewing, the coffee extracts were filtered using a standard coffee filter and stored in a freezer at –20 °C until a further examination.

### 2.2.3. Preparation of peppermint extracts

Ten peppermint samples were collected from seven different producers (samples obtained from the same producer varied with respect to production batch). Two grams of dried peppermint were extracted for 15 min using 50 mL of hot (95 °C) distilled water, filtered and stored in a freezer at –20 °C until further analysis.

### 2.2.4. Determination of the total phenolic content in the samples

The total phenolic content of the samples was evaluated using the Folin–Ciocalteu reagent (FCR). A 20  $\mu$ L of the investigated sample was added to 250  $\mu$ L of FCR and 1000  $\mu$ L of saturated  $Na_2CO_3$ . Then, the mixture was diluted up to a volume of 10 mL and incubated for 30 min at 50 °C. Absorption of the samples was measured at 760 nm. The TPC values were expressed as gallic acid equivalents, i.e.  $\mu$ mol of gallic acid per 1 g of a sample and predicted from the calibration model ( $R^2 = 0.9989$ ) that was constructed within a calibration range between 0.088 and 0.294  $\mu$ mol of gallic acid.

### 2.2.5. The ORAC assay

In the ORAC assay, which was executed simultaneously in a 96-well plate, fluorescein was used as the fluorescence probe [27]. 50  $\mu$ L of fluorescein (4.75  $\mu$ M) was added to 50  $\mu$ L of a diluted sample, 50  $\mu$ L of a standard (80  $\mu$ M) or 50  $\mu$ L of a phosphate buffer (pH = 7.4) and then incubated for 20 min at 37 °C. Next, a portion of 25  $\mu$ L of AAPH (0.0029 M) was introduced into each sample. The initiated reaction was driven to completion (for approximately 300 min). Excitation and emission wavelengths were set to 492 nm and 530 nm, respectively. For each sample, the ORAC value was calculated according to the following equation:

$$ORAC = \frac{C_{standard} AUC_{sample,blank}}{AUC_{standard/blank}} d \quad (1)$$

where,  $C_{standard}$  is the Trolox concentration,  $AUC_{sample,blank}$  and  $AUC_{standard/blank}$  are the areas under the curve that described a decrease in the fluorescence intensity observed over time for a sample or standard and a blank sample and  $d$  is the dilution factor. All of the ORAC values were expressed as Trolox equivalents, i.e.  $\mu$ mol of Trolox per 1 g of dry sample matter (coffee or peppermint).

### 2.2.6. Measurements of the excitation–emission fluorescence spectra

In order to eliminate fluorescence quenching, a portion of 200  $\mu$ L of the coffee extracts or 50  $\mu$ L of the peppermint extracts were diluted in distilled water up to the final volume of 10 mL. The fluorophores contained in a sample were characterized by the EEMs that were collected in spectral ranges, which were characteristic for a particular set of samples. For the coffee extracts, emission spectra from 330 nm to 700 nm were measured at 2 nm intervals (185 wavelength channels) at 42 excitation wavelengths from 230 nm to 650 nm every 10 nm. For the peppermint extracts, the EEMs were composed of 185 emission wavelength channels (measured from 280 to 650 nm every 2 nm) and excitation within the range of 220 and 450 nm with a 10 nm interval.

In order to derive the validation parameters of analytical method, the EEMs of additional samples were registered. Two 2  $\mu$ L of coffee and peppermint extracts diluted in up to 10 mL of distilled water were prepared as nine replicates in order to evaluate the limits of the detection and quantitation (LOD and LOQ) of a method. To estimate the repeatability of a method, six replicates of coffee and six replicates of peppermint extracts (50  $\mu$ L of extracts diluted in up to 10 mL of distilled water) were measured immediately after preparation.

## 3. Results and discussion

The structure of the unfolded EEMs data after scattering correction was explored using PCA. The first two PCs explained 98.2% and 68.9% of the total data variance for coffee and peppermint extracts, respectively. On the score plot of the first two principal components, Fig. 2a, two coffee extracts are located relatively far from the data majority along the second principal component. They represent different coffee producers and different batches of replicate measurements. The score plot of peppermint extracts, see Fig. 2b, uncovered a group of three distant samples, which are indicated with \*. They all represent spectra of one sample batch, for which extracts were prepared in triplicate. It is

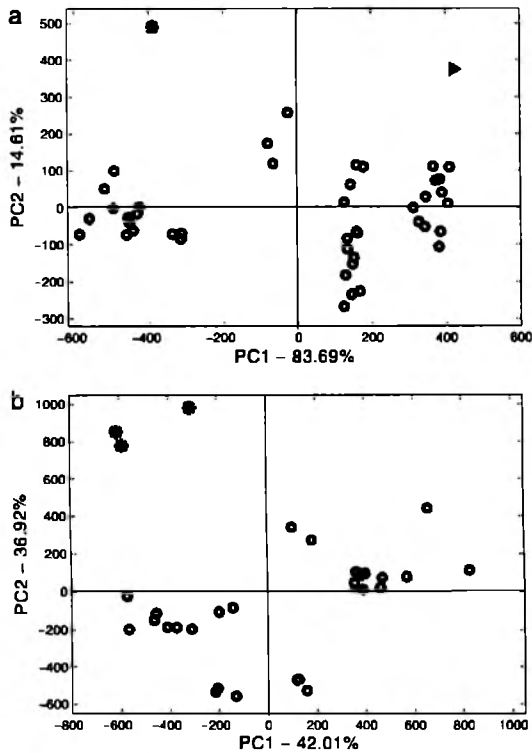


Fig. 2. Score plots of the first two principal components, PC 1 and PC 2, of unfolded EEMs that describe a collection of: a) coffee and b) peppermint extracts. Extreme samples are marked as  $\blacktriangle$  and  $\blacktriangleright$ .

known that, depending on an outlying character of samples, they can influence least squares models. For this reason, samples with an outlying character were included in test sets.

For coffee extracts, since its geographical origin is known, it is possible to spot a tendency to differentiate between coffee samples collected at different continents on the PC 1 and PC 4 score plot. Along the PC 1, the samples from Africa can be separated from the remaining continents (see Fig. 3a). PC 4 models the differences between the Asian coffees and the other ones. The corresponding loading plots, which are presented in the form of a color map, indicate spectral regions that explain the chemical differences between the samples (see Fig. 3b and c). Each pixel of a map has a color intensity that is proportional to a loading value. The interpretation of a color map of the loadings of PC 1 leads to the conclusion that coffee samples from Africa are characterized by a relatively high content of fluorophores that induces a relatively low fluorescence intensity, which can be observed in the dark regions of a color map. On the other hand, samples from Asia contain fluorophores that are characterized by a relatively large fluorescence in the dark regions of a color map.

The ORAC and TPC values of the same samples were also analyzed. In the case of the peppermint extracts, a relatively strong linear correlation between the results from two antioxidant assays was observed with a correlation coefficient equal to 0.880. For the coffee extracts, the agreement between the results obtained from two assays was rather weak (the correlation coefficient value was equal to 0.462); however, the ORAC assay is discussed in the literature as having weak agreement with other assays [28,29]. A more detailed analysis of the total antioxidant capacity estimated using either the ORAC or the TPC assay did not reveal extreme samples.

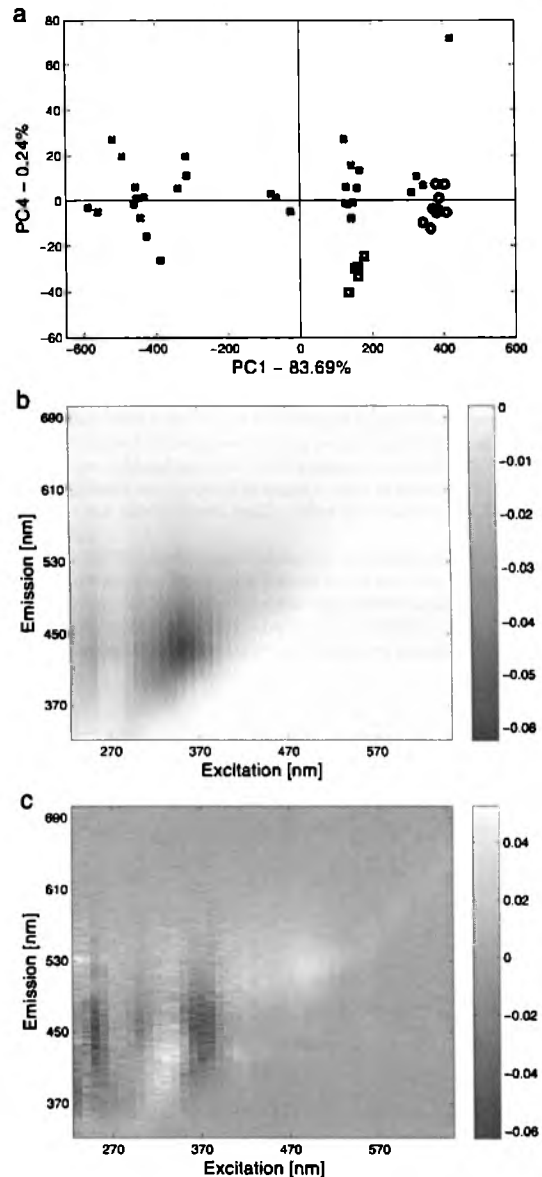


Fig. 3. PC 1 and PC 4 score plot of unfolded EEMs for coffee extracts with the geographical origin indicated ( $\square$  – Asia,  $\times$  – South America and  $\circ$  – Africa). Color maps of the loadings presenting the contribution of particular variables to the construction of: b) PC 1 and c) PC 2.

To construct the least squares models, training sets were constructed for the coffee and for the peppermint extracts to cover the range of the TAC values that incorporated all of the laboratory replicates within the same subset. Thirty samples of coffee extracts and eighteen samples of peppermint extracts were selected for model sets. The remaining fifteen and twelve samples formed test sets for the coffee and peppermint extracts, respectively.

To construct a PLS model with a satisfactory fit and prediction properties to describe the ORAC value of coffee extracts, four latent factors were selected based on the leave-one-out cross-validation procedure. On the ORAC predicted vs. ORAC observed plot that was constructed for the optimal PLS model, there are two extreme samples from the test set (see Fig. 4a). Both samples, marked as \* and ►, were also found to be extreme based on the score plot. The ORAC parameters of these two outlying objects are not extreme, thus they can be considered as outlying in the space of fluorescence spectra. Their outlying character can be explained as a consequence of instrumental fluctuations or unexpected laboratory error because their laboratory replicates have small residuals from the model.

Next, the N-PLS model was constructed for the coffee extracts. Five factors were found to be sufficient to construct the optimal model. In Fig. 4b, the ORAC predicted vs. ORAC observed plot, which contains one extreme object belonging to test set and marked as \*, is presented. It is interesting to note that now for the sample marked in Fig. 4a as ►, the N-PLS model offers a good prediction (a small residual error).

The fits of the optimal PLS and N-PLS models were equal to 331.14 and 445.90 TE, respectively. The prediction errors obtained for the independent test set extracts (without outlying objects) were equal to 605.09 and 448.33 TE, respectively.

Additionally, PLS and N-PLS models that described the TPC of the coffee extracts were constructed. The PLS model contained five latent factors and the N-PLS model had six latent factors. The same sample, indicated with \*, appeared to be outlying and had large residuals from the PLS and N-PLS models (compare Fig. 4c and d). The prediction errors for the models were calculated without this sample. On Fig. 4d, which was constructed for the six-factor N-PLS model, object ► from the test set

was also classified as an extreme object with large residuals from the model. It was not taken into account in calculating the RMSE error. In Table 1, the fit and prediction properties of the constructed models are presented.

In order to describe the ORAC and TPC of the peppermint extracts, additional PLS and N-PLS models were built. The optimal PLS and N-PLS models to predict the ORAC values required five latent factors and offered a prediction error for calibration samples equal to 181.08 TE and 413.21 TE, respectively. On the ORAC predicted vs. ORAC observed plots, see Fig. 5a and b, no outlying objects were observed. The RMSEP error for the PLS model, as compared with RMSEP for N-PLS model (see Table 1), is better. A similar conclusion can be drawn when the fit and prediction properties are evaluated for the optimal PLS and N-PLS that was constructed to predict the TPC. It is interesting to note that for one batch of peppermint for all of the laboratory replicates the TPC values were underestimated using the PLS and N-PLS models, thus they considerably influence the RMSEPs of the constructed models. This may be a consequence of the unique character of a sample. Compared with the coffee samples, the modeling of peppermint seems to be more challenging because it was difficult to maintain the same structure of a sample. A visual inspection of the samples gave evidence that they were different with respect to quality. Some of them contained an increased number of leaves, while others had stalks and a small number of leaves.

A more detailed analysis of the multivariate models permitted a conclusion to be drawn that for the coffee extracts, the N-PLS for the ORAC and the PLS for the TPC had the most satisfactory performance, whereas for the peppermint only the PLS constructed for the ORAC is considered as having an acceptable prediction error. This can be explained by the

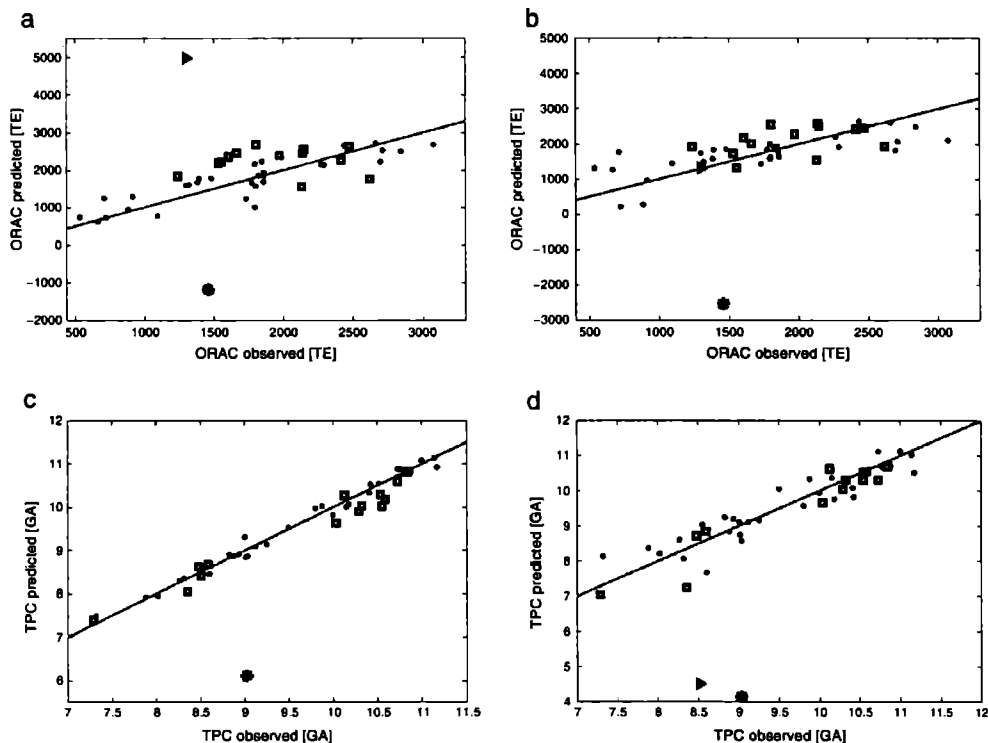


Fig. 4. The PLS and N-PLS models presented as parameter predicted vs. parameter observed plots for model (●) and test set samples (◻), constructed for: a) and b) the ORAC assay results; c) and d) the TPC assay results of coffee extracts. Two extreme samples, included in the test sets, are marked with \* and ►.

Table 1

The fit and prediction properties of the PLS and N-PLS models that described the ORAC and TPC parameters of coffee and peppermint extracts, respectively. The parameters are expressed as  $\mu\text{mol}$  of the standard, i.e. Trolox and gallic acid for ORAC and TPC, respectively. The error expressed as percentage was calculated with respect to the range of the ORAC or TPC values that were observed for the calibration samples.

Method of model construction	Coffee extracts				Peppermint					
	Number of factors	Fit properties		Prediction properties		Number of factors	Fit properties		Prediction properties	
		$\mu\text{mol}$ of standard	%	$\mu\text{mol}$ of standard	%		$\mu\text{mol}$ of standard	%	$\mu\text{mol}$ of standard	%
	ORAC									
PLS	4	331.14	4.64	605.09	8.48	5	181.08	2.54	647.74	9.08
N-PLS	5	445.90	6.25	448.33	6.29	5	413.21	5.79	878.91	12.32
	TPC									
PLS	5	0.13	3.37	0.27	7.02	3	2.69	5.31	10.90	21.54
N-PLS	6	0.39	10.26	0.40	10.53	3	3.98	7.86	13.08	25.85

different sample quality and homogeneity. In the case of the coffee samples, their homogeneity and loss of quality over time was better controlled than for the peppermint samples. It is very likely that some of the peppermint samples were stored for a longer period of time, thus they had a longer contact with oxygen and were affected by the storage conditions.

### 3.1. Figures of the merit of multivariate models

The figures of the merit and analytical validation parameters were calculated for the best models. The sensitivities of the constructed models were calculated according to the description provided in [30,

31] as the inverse of the Euclidean norm of the regression coefficients that are obtained for a given model.

The ratio between the instrumental noise and sensitivity was calculated and multiplied by 3.3 in order to obtain the LOD for the considered models, whereas the limits of quantification (LOQ) were calculated as ten times the ratio values for both models. The figures of merit for selected models are presented in Table 2.

### 3.2. Analytical validation parameters

The repeatability of the proposed methods was expressed as the relative standard deviation, RSD, of the results obtained for six laboratory

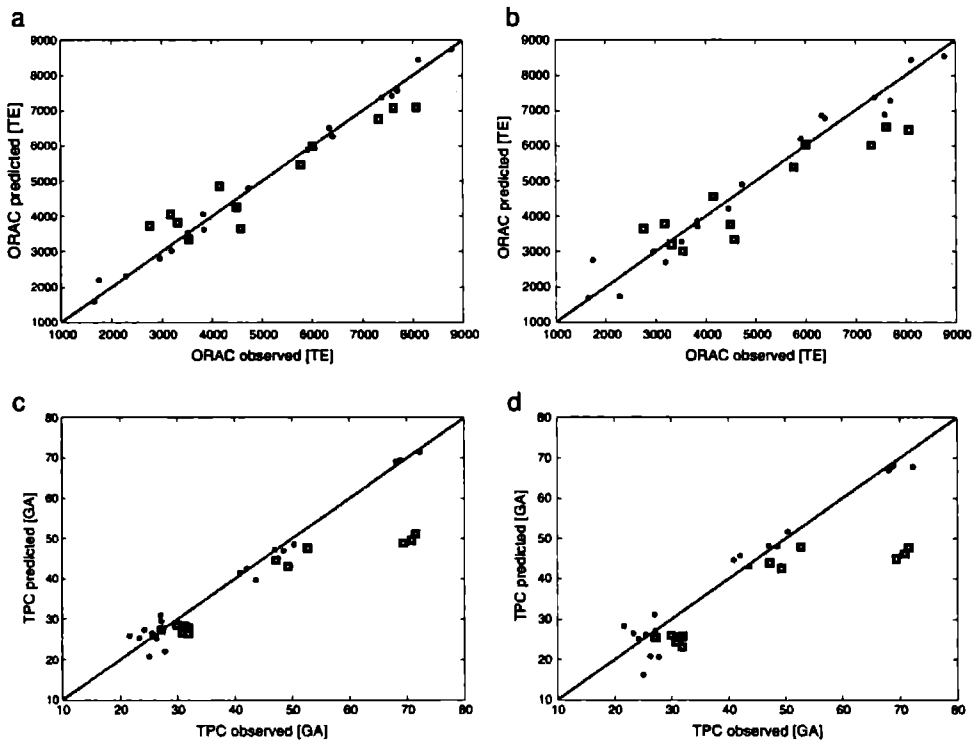


Fig. 5. PLS and N-PLS models presented as the parameter predicted vs. parameter observed for model (●) and test set samples (□), constructed for: a) and b) the ORAC assay results; c) and d) the TPC assay results of peppermint extracts.

Table 2

Figures of merit for the N-PLS model constructed for the ORAC and PLS models that were constructed for the TPC of the coffee extracts and the PLS model that was constructed for the ORAC of the peppermint extracts. The figures of merit for the ORAC are expressed as TE and for the TPC are expressed as GA.

Parameter	Coffee extracts		Pepper mint extracts
	ORAC N-PLS, 5 factors	TPC PLS, 5 factors	ORAC PLS, 5 factors
Sensitivity	8.33	10.75	5.55
Selectivity	$5.57 \times 10^{-1}$	$6.88 \times 10^{-1}$	$1.29 \times 10^{-2}$
LOD	85.42	0.037	12.58
LOQ	854.2	0.378	125.8

replicates measured immediately after preparation. The RSD values are presented in Table 3.

The analytical limit of detection, LOD, was calculated as the standard deviation of the concentrations that had been predicted for a set of nine samples times the critical value of the Student's *t*-distribution assuming a confidence level of 95% and eight degrees of freedom. For the coffee extracts, the LODs were equal to 498.55 TE and 0.285 GA for the ORAC and TPC, respectively. The LOD calculated of the ORAC model describing the antioxidant properties of the peppermint extracts was equal to 488.94 TE. In the literature, different models that express the total antioxidant capacity as evaluated from preferable analytical techniques have been proposed. Their performance seems to be comparable to our approach. For instance, in [32] a robust variant of the PLS method to evaluate the TAC of green tea using HPLC fingerprints was described. The possibility of using infrared spectra supported by multivariate modeling to evaluate the TPC or TAC (examined using different antioxidant assays) of onion and shallot [33], salvia [34] and other food commodities [35] was investigated. Unfortunately, in these articles the LOD, LOQ or RSD of the discussed methods were not reported.

In our study, in order to characterize the non-selective analytical signals of samples – the EEMs were registered and the total antioxidant capacity expressed as the ORAC and TPC values were examined. The collected data were used to construct calibration models to predict the antioxidant properties of coffee and peppermint extracts directly from their EEMs. The calibration models that were obtained have a relatively good fit and prediction properties, bearing in mind the chemical complexity of the selected food samples. On the basis of the EEMs of water extracts, one can evaluate the ORAC parameter within an error of prediction equal to 6.29% or 9.08% for the coffee and peppermint extracts, respectively. Increased levels of RMSEPs may be a consequence of a known drawback of the ORAC assay, which has been widely discussed in the literature [2], i.e. the low repeatability of the method. Nevertheless, for comparative purposes, the ORAC assay results were also included due to its popularity in evaluating the TAC of various food commodities. The TPC of the examined coffee samples was evaluated using the five-factor PLS model and had a satisfactory performance (prediction error equal to 7.02%). The considerably high prediction errors that were obtained for the models that were constructed for the TPC indicates that some antioxidant compounds, which were contained in the analyzed samples, do not exhibit fluorescence properties.

Table 3

Repeatability of proposed methods. Mean values were calculated for six laboratory replicates and uncertainty was calculated assuming a 95% confidence interval. Results for the ORAC are expressed as TE and for the TPC are expressed as GA.

Coffee extracts			Pepper mint extracts		
ORAC N-PLS, 5 factors	TPC PLS, 5 factors		ORAC PLS, 5 factors		
Mean	RSD	±	Mean	RSD	±
447.46	7.13%	26.198	6.802	1.61%	0.089
			3173.39	4.04%	105.18

## Conflict of interest

We wish to confirm that there are no known conflicts of interest associated with this publication.

## Acknowledgment

JO appreciates the support of the National Science Center grant no. 2011/03/N/ST4/00713.

## References

- N. Pellegrini, M. Serafini, B. Colombi, D.D. Rio, S. Salvatore, M. Bianchi, et al., Total antioxidant capacity of plant foods, beverages and oils consumed in Italy assessed by three different in vitro assays, *J. Nutr.* 133 (2003) 2812–2819.
- R. Apak, S. Gorinstein, V. Böhm, K.M. Schaich, M. Özyürek, K. Güçlü, Methods of measurement and evaluation of natural antioxidant capacity/activity (IUPAC Technical Report), *Pure Appl. Chem.* 85 (2013) 957–998, <http://dx.doi.org/10.1351/PAC-REP-12-07-15>.
- E. Joubert, T. Beelders, D. de Beer, C.J. Malherbe, A.J. de Villiers, G.D. Sigge, Variation in phenolic content and antioxidant activity of fermented rooibos herbal tea infusions: role of production season and quality grade, *J. Agric. Food Chem.* 60 (2012) 9171–9179, <http://dx.doi.org/10.1021/jf302583r>.
- N.J. Miller, C. Rice-Evans, M.J. Davies, V. Gopinathan, A. Milner, A novel method for measuring antioxidant capacity and its application to monitoring the antioxidant status in premature neonates, *Clin. Sci. (Lond.)* 1979 (84) (1993) 407–412.
- A. Jiménez-Escrig, I. Jiménez-Jiménez, C. Sánchez-Moreno, F. Saura-Calixto, Evaluation of free radical scavenging of dietary carotenoids by the stable radical 2,2-diphenyl-1-picrylhydrazyl, *J. Sci. Food Agric.* 80 (2000) 1686–1690, [http://dx.doi.org/10.1002/1097-0010\(20000901\)80:11<1686::AID-JSFA694>3.0.CO;2-Y](http://dx.doi.org/10.1002/1097-0010(20000901)80:11<1686::AID-JSFA694>3.0.CO;2-Y).
- J.F.F. Benzie, J.J. Strain, The Ferric Reducing Ability of Plasma (FRAP) as a measure of "antioxidant power": the FRAP assay, *Anal. Biochem.* 239 (1996) 70–76, <http://dx.doi.org/10.1006/abio.1996.0292>.
- B. Ou, M. Hampsch-Woodill, R.L. Prior, Development and validation of an improved oxygen radical absorbance capacity assay using fluorescein as the fluorescent probe, *J. Agric. Food Chem.* 49 (2001) 4619–4626, <http://dx.doi.org/10.1021/jf010586o>.
- A. Zulueta, M.J. Esteve, A. Frigola, ORAC and TEAC assays comparison to measure the antioxidant capacity of food products, *Food Chem.* 114 (2009) 310–316, <http://dx.doi.org/10.1016/j.foodchem.2008.09.033>.
- I.A. Ludwig, L. Sanchez, B. Caemmerer, L.W. Kroh, M.P. De Peña, C. Cid, Extraction of coffee antioxidants: impact of brewing time and method, *Food Res. Int.* 48 (2012) 57–64, <http://dx.doi.org/10.1016/j.foodres.2012.02.023>.
- J. Polak, M. Bartoszek, I. Stanimirova, A study of the antioxidant properties of beers using electron paramagnetic resonance, *Food Chem.* 141 (2013) 3042–3049, <http://dx.doi.org/10.1016/j.foodchem.2013.05.133>.
- M. Bartoszek, J. Polak, An electron paramagnetic resonance study of antioxidant properties of alcoholic beverages, *Food Chem.* 132 (2012) 2089–2093, <http://dx.doi.org/10.1016/j.foodchem.2011.12.060>.
- R. Karoui, J. De Baerdemaeker, A review of the analytical methods coupled with chemometric tools for the determination of the quality and identity of dairy products, *Food Chem.* 102 (2007) 621–640, <http://dx.doi.org/10.1016/j.foodchem.2006.05.042>.
- J. Christensen, E.M. Becker, C.S. Frederiksen, Fluorescence spectroscopy and PARAFAC in the analysis of yogurt, *Chemom. Intell. Lab. Syst. 75* (2005) 201–208, <http://dx.doi.org/10.1016/j.chemolab.2004.07.007>.
- E. Sikorska, A. Gliszczynska-Swiglo, M. Insińska-Rak, I. Khmelinskii, D. De Keuleleire, M. Sikorski, Simultaneous analysis of riboflavin and aromatic amino acids in beer using fluorescence and multivariate calibration methods, *Anal. Chim. Acta.* 613 (2008) 207–217, <http://dx.doi.org/10.1016/j.aca.2008.02.063>.
- J. Orzel, M. Daszykowski, I. Grabowski, G. Zaleszczyk, M. Sznajder, B. Walczak, Simultaneous determination of Solvent Yellow 124 and Solvent Red 19 in diesel oil using fluorescence spectroscopy and chemometrics, *Talanta* 101 (2012) 78–84, <http://dx.doi.org/10.1016/j.talanta.2012.08.031>.
- J. Orzel, M. Daszykowski, I. Grabowski, G. Zaleszczyk, M. Sznajder, Identifying the illegal removal from diesel oil of certain chemical markers that designate excise duty, *Fuel* 117 (Part A) (2014) 224–229, <http://dx.doi.org/10.1016/j.fuel.2013.09.029>.
- J. Orzel, M. Daszykowski, B. Walczak, Controlling sugar quality on the basis of fluorescence fingerprints using robust calibration, *Chemom. Intell. Lab. Syst.* 110 (2012) 89–96.
- E. Sikorska, A. Gliszczynska-Swiglo, I. Khmelinskii, M. Sikorski, Synchronous fluorescence spectroscopy of edible vegetable oils. Quantification of tocopherols, *J. Agric. Food Chem.* 53 (2005) 6988–6994.
- V.L. Singleton, R. Orthofer, R.M. Lamuela-Raventós, Analysis of total phenols and other oxidation substrates and antioxidants by means of folin-ciocalteu reagent, in: Lester Packer (Ed.), *Methods Enzymol.* Academic Press, 1999, pp. 152–178.
- T. Naes, T. Isaksson, T. Fearn, T. Davies, A User Friendly Guide to Multivariate Calibration and Classification, NIR Publications, Chichester, 2002.
- R. Bro, Multiway calibration. Multilinear PLS, *J. Chemom.* 10 (1996) 47–61.
- J.A. Arancibia, P.C. Damiani, G.M. Escandar, G.A. Ibañez, A.C. Olivieri, A review on second- and third-order multivariate calibration applied to chromatographic data, *J. Chromatogr. B* 910 (2012) 22–30, <http://dx.doi.org/10.1016/j.jchromb.2012.02.004>.

- [23] R.C. Zepp, W.M. Sheldon, M.A. Moran, Dissolved organic fluorophores in southeastern US coastal waters: correction method for eliminating Rayleigh and Raman scattering peaks in excitation–emission matrices. *Mar. Chem.* 89 (2004) 15–36, <http://dx.doi.org/10.1016/j.marchem.2004.02.006>.
- [24] P.H.C. Eilers, P.M. Kroonenberg, Modeling and correction of Raman and Rayleigh scatter in fluorescence landscapes. *Chemom. Intell. Lab. Syst.* 130 (2014) 1–5, <http://dx.doi.org/10.1016/j.chemolab.2013.09.002>.
- [25] M. Daszykowski, K. Kaczmarek, Y. Vander Heyden, B. Walczak, Robust statistics in data analysis – a review: basic concepts. *Chemom. Intell. Lab. Syst.* 85 (2007) 203–219, <http://dx.doi.org/10.1016/j.chemolab.2006.06.016>.
- [26] S. Wold, K. Esbensen, P. Geladi, Principal component analysis. *Chemom. Intell. Lab. Syst.* 2 (1987) 37–52, [http://dx.doi.org/10.1016/0169-7439\(87\)80084-9](http://dx.doi.org/10.1016/0169-7439(87)80084-9).
- [27] D. Huang, B. Ou, M. Hampsch-Woodill, J.A. Flanagan, R.L. Prior, High-throughput assay of Oxygen Radical Absorbance Capacity (ORAC) using a multichannel liquid handling system coupled with a microplate fluorescence reader in 96-well format. *J. Agric. Food Chem.* 50 (2002) 4437–4444, <http://dx.doi.org/10.1021/jf0201529>.
- [28] S. Kohri, H. Fujii, S. Oowada, N. Endoh, Y. Sueishi, M. Kusakabe, et al., An oxygen radical absorbance capacity-like assay that directly quantifies the antioxidant's scavenging capacity against AAPH-derived free radicals. *Anal. Biochem.* 386 (2009) 167–171, <http://dx.doi.org/10.1016/j.ab.2008.12.022>.
- [29] R.L. Prior, X. Wu, K. Schaich, Standardized methods for the determination of antioxidant capacity and phenolics in foods and dietary supplements. *J. Agric. Food Chem.* 53 (2005) 4290–4302, <http://dx.doi.org/10.1021/jf0502698>.
- [30] A. Lorber, Error propagation and figures of merit for quantification by solving matrix equations. *Anal. Chem.* 58 (1986) 1167–1172, <http://dx.doi.org/10.1021/ac00297a042>.
- [31] A.C. Olivieri, N.M. Faber, J. Ferré, R. Boqué, J.H. Kalivas, H. Mark, Uncertainty estimation and figures of merit for multivariate calibration (IUPAC Technical Report). *Pure Appl. Chem.* 78 (2006) 633–661, <http://dx.doi.org/10.1351/pac200678030633>.
- [32] M. Daszykowski, Y. Vander Heyden, C. Boucon, B. Walczak, Automated alignment of one-dimensional chromatographic fingerprints. *J. Chromatogr. A* 1217 (2010) 6127–6133, <http://dx.doi.org/10.1016/j.chroma.2010.08.008>.
- [33] J.W. Xiaonan Lu, Determination of total phenolic content and antioxidant capacity of onion (*Allium cepa*) and shallot (*Allium oschaninii*) using infrared spectroscopy. *Food Chem.* 129 (2011), <http://dx.doi.org/10.1016/j.foodchem.2011.04.105>.
- [34] X. Duan, D. Zhang, L. Nie, H. Zang, Rapid discrimination of geographical origin and evaluation of antioxidant activity of *Salvia miltiorrhiza* var. alba by Fourier transform near infrared spectroscopy. *Spectrochim. Acta A Mol. Biomol. Spectrosc.* 122 (2014) 751–757.
- [35] X. Lu, B.A. Rasco, Determination of antioxidant content and antioxidant activity in foods using infrared spectroscopy and chemometrics: a review. *Crit. Rev. Food Sci. Nutr.* 52 (2012) 853–875, <http://dx.doi.org/10.1080/10408398.2010.511322>.

Katowice, 03.03.2015

dr hab. Michał Daszykowski, prof. UŚ  
Instytut Chemii  
Uniwersytet Śląski  
ul. Szkolna 9  
40-006 Katowice

Oświadczam, że w artykule pt. *A rapid validation of the antioxidant capacity of food commodities based on their fluorescence excitation emission spectra as applicable to coffee and peppermint extracts*, *Chemometrics and Intelligent Laboratory Systems*, 137 (2014) 74-81 mój udział polegał na:

- współtworzeniu hipotezy badawczej i koncepcji badań,
- zapewnieniu nadzoru merytorycznego w tym pomocy w interpretacji wyników i weryfikacji wyciągniętych przez doktorantkę wniosków z przeprowadzonych badań,
- merytorycznym nadzorze podczas przygotowania manuskryptu,
- uczestnictwie w procedurze redakcyjnej (tj. pomocy w przygotowaniu odpowiedzi na recenzje)
- pomocy w trakcie ostatecznej korekty artykułu.

Michał Daszykowski



## 7.5. Publikacja V

**Metoda przewidywania zawartości hydrofilowych antyoksydantów zawartych w paście pomidorowej na podstawie widm z zakresu podczerwieni oraz fluorescencyjnych obrazów zarejestrowanych dla próbek surowych i ekstraktów**

---

Tytuł oryginalny:	<i>Prediction of hydrophilic antioxidant capacity of tomato pastes from the IR and fluorescence excitation-emission spectra of extracts and intact samples</i>
Autorzy:	Joanna Orzeł Ivana Stanimirova Bogusława Czarnik-Matusiewicz Michał Daszykowski
Czasopismo:	Talanta
Rok wydania:	2015
Tom:	138
Strony:	64 – 70
Wartość współczynnika <i>Impact Factor</i> *	3,511

---

W publikacji V przedstawiam kontynuację badań opisanych w publikacji IV. Zaproponowaną metodę oceny zawartości antyoksydantów na podstawie fluorescencyjnych obrazów porównałam z metodą opierającą się na pomiarze widm z zakresu podczerwieni zarejestrowanych dla próbek past pomidorowych oraz ich wodnych ekstraktów. Optymalne warunki ekstrakcji (ekstrakcja z użyciem ultradźwięków, przez trzy minuty w temperaturze 80 °C) wyznaczono z użyciem centralnego planu kompozycyjnego. Informację o zawartości antyoksydantów w próbkach zbadałam z pomocą metody ORAC, zawartość polifenoli wyznaczyłam używając odczynnika Folin-Ciocalteu. Widma oraz informacje o zawartości antyoksydantów użyłam do konstrukcji modeli PLS oraz N-PLS. Modele skonstruowane dla przewidywania parametru ORAC charakteryzowały się wartościami średnich błędów kwadratowych (wyrażonych jako ekwiwalenty Troloxu) na poziomie 2,011 (14,21%) dla fluorescencyjnych obrazów oraz 2,426 (17,15%) dla widm IR. Błędy przewidywania zawartości polifenoli wyznaczonej w ekwiwalentach kwasu galusowego wynosiły 0,067 (10,78%) oraz 0,033 (5,36%) odpowiednio dla obrazów i widm IR.

Badania zamieszczone w publikacji realizowałam w ramach projektu badawczego Preludium, finansowanego przez Narodowe Centrum Nauki pt. *Fluorescencyjne krajobrazy w połączeniu z metodami chemometrycznymi jako potencjalne narzędzie do wyznaczania całkowitej zdolności antyoksydacyjnej*, nr 2011/03/N/ST4/00713, którego byłam kierownikiem.



Contents lists available at ScienceDirect

Talanta

journal homepage: [www.elsevier.com/locate/talanta](http://www.elsevier.com/locate/talanta)

## Prediction of the hydrophilic antioxidant capacity of tomato pastes from the IR and fluorescence excitation–emission spectra of extracts and intact samples



Joanna Orzel<sup>a</sup>, Ivana Stanimirova<sup>a</sup>, Bogusława Czarnik-Matusewicz<sup>b</sup>,  
Michał Daszykowski<sup>a,\*</sup>

<sup>a</sup> Institute of Chemistry, University of Silesia, 9 Szkalna Street, 40-006 Katowice, Poland

<sup>b</sup> Faculty of Chemistry, University of Wrocław, 14 Joliot Curie Street, 50-383 Wrocław, Poland

### ARTICLE INFO

#### Article history:

Received 28 November 2014

Received in revised form

12 January 2015

Accepted 19 January 2015

Available online 12 February 2015

#### Keywords:

TAC

ORAC

Fiber optics probe

IR-ATR

PLS

N-PLS

### ABSTRACT

The performance of the recently proposed excitation–emission fluorescence method was compared to the method using infrared measurements for the evaluation of the antioxidant properties of intact samples and extracts that had been obtained from tomato pastes. The oxygen radical absorbance capacity assay (ORAC) was applied in order to estimate the antioxidant capacity of the extracts, while the Folin–Ciocalteu reagent was adopted for the evaluation of the total phenolic content. The optimal extraction conditions for tomato pastes (three minutes of sonication under 80 °C) were determined using the central composite design. Chemometric models such as the partial least squares regression and its N-way variant were further constructed in order to predict the antioxidant capacity or total phenolic content of the samples using either the IR or fluorescence spectra. The prediction errors that were obtained for the total antioxidant content were evaluated as the Trolox equivalents from the ORAC assay and were found to be equal to 2.011 (14.21%) for the fluorescence and 2.426 (17.15%) for the IR spectra, respectively. The prediction errors of the total phenolic content expressed as gallic acid equivalents were 0.067 (10.78%) for the fluorescence and 0.033 (5.36%) for the IR spectra, which were used as independent variables in the regression models.

© 2015 Elsevier B.V. All rights reserved.

### 1. Introduction

Antioxidants are vital compounds and their content, which is expressed as the total antioxidant capacity (TAC) in various food commodities, is a subject of a great interest for researchers and consumers. The number of scientific papers that deal with the evaluation of the TAC in different food matrices has been growing steadily for the past several years. Simultaneously, consumer awareness of the TAC and its importance as a quality parameter for food commodities has also increased. Therefore, it is not surprising that the main goal of organizations like ORAC watch (established in USA) is to introduce information about the TAC levels on product labels. A major difficulty in making this information available to the consumer is the inconsistency of the procedure(s) that are used for evaluating the TAC. All of the procedures are based on the chemical reaction between free radicals that are deliberately generated and the antioxidants that are

contained in the sample being analyzed. This reaction is usually monitored using a spectrophotometer or fluorometer and the result is expressed with respect to a selected standard antioxidant (e.g., Trolox, vitamin C or gallic acid). Several generators of free radicals, such as e.g., AAPH (2,2'-Azobis(2-amidinopropane) dihydrochloride), ABTS (2,2'-azino-bis(3-ethylbenzothiazoline-6-sulfonic acid)) or DPPH (2,2-diphenyl-1-picrylhydrazyl), are available [1]. Furthermore, methods based on the reduction of cupric [2] or ferric [3] ions are also used. Differences in the reaction mechanisms between various free radicals and the antioxidants that are contained in a food sample result in the determination of different TAC values. Moreover, the large variety of available TAC assays and their inconsistencies are not the only concern. Some authors argue in the literature that the TAC parameter of a given food commodity is affected by several external factors including its geographical origin, the cultivation approach that was used, the time of harvesting and the type of food processing [4–11]. Therefore, the TAC of each production batch should be carefully monitored. Unfortunately, in practice, standard assays require a substantial amount of solvents and reagents and they are also expensive, time-consuming and laborious. Therefore, new

\* Corresponding author. Tel.: +48 32 359 1568; fax: +48 32 259 9978.

E-mail address: [mdaszyk@us.edu.pl](mailto:mdaszyk@us.edu.pl) (M. Daszykowski).

methods that are based on easy-to-perform on-line measurements that have no requirements for a special laboratory pretreatment with reagents are greatly desired.

A good candidate is a method based on the multivariate calibration of IR spectra (non-selective signals with respect to the particular antioxidant compounds). Lu and Rasco (2012) reported on near- and mid-IR (12,800–400  $\text{cm}^{-1}$ ) spectroscopy as a suitable technique to evaluate the TAC of selected food commodities [12]. That is why the IR measurements for vegetables [13], fruits [14], beverages [15], etc. have been recently proposed as a replacement for conventional assays. This approach is claimed to be accurate, non-destructive and to substantially reduce the amount of preparative work that is required to analyze samples on one hand, while it is sensitive to the presence of water in the samples on the other. Therefore, the analysis of either extracts or vegetable and fruit pulps still presents a challenge. Moreover, the IR radiation is relatively low [16] for fruits or vegetables with a thick skin or occlusive flesh penetration.

Another promising methodology that uses non-selective fluorescence spectra and multivariate modeling has been proposed to evaluate the TAC of coffee and peppermint extracts [17]. Each sample was excited using radiation at different energy levels (various excitation wavelengths) and the emission spectrum was collected for each excitation wavelength. The excitation–emission fluorescence matrices (EEMs) that were collected are examples of second order signals and their use can facilitate the calibration of the constituents of samples (or some specific features like color, ash content, and shelf life) when interfering compounds are present in the sample matrix [18,19]. The relatively low prediction errors for the TAC of the multivariate calibration models that use EEMs (equal to 6.25% and 9.08% for coffee and peppermint extracts, respectively) indicate the potential of the proposed methodology. The preliminary studies were focused only on the water extracts of samples [17].

The aim of this study is to develop a suitable approach for the determination of the TAC levels of intact food samples that would allow for the possibility of routine on-line testing. Tomato paste, which is an integral part of the human diet worldwide, was chosen as the subject of this study. For the collection of intact samples, EEMs were registered using a fiber optic probe in a selected UV–vis range. In order to compare the results that were obtained from EEMs with those that were obtained from the methodology proposed in the literature for the IR measurements, the mid-IR spectra (4000–400  $\text{cm}^{-1}$ ) for the same collection of samples were also registered. Antioxidant water extraction conditions (time and temperature) were optimized using the central composite design [20] in order to guarantee the optimal TAC determination. The TAC parameter in this study was evaluated using the oxygen radical absorbance capacity assay, the ORAC assay [21]. Even though, there are no strict criteria for the selection of an assay, the ORAC assay seems to be selected the most among the assays [22,23]. It should be emphasized the TAC of water extracts discussed in this study refer to the total content of hydrophilic antioxidants extracted from selected samples.

The total phenolic content (TPC) is another parameter that refers to the antioxidant content in samples. In this study, the TPC was determined using the Folin–Ciocalteu reagent [24]. Partial least squares regression, PLSR [25] was used to model the TAC (or TPC) as a function of the spectra for respective intact food samples. In addition, N-way partial least squares regression models (extension of PLS), N-PLS [26], were developed for the collection of the EEMs spectra.

## 2. Experimental

### 2.1. Origin of the samples

The tomato pastes that are available on the Polish market were purchased in local stores. They were produced by different producers

that were located in different production places (nine tomato pastes originated from Poland, three from Italy and one from Croatia). The packaging materials that were used were also different (jars, cans and cardboard boxes). A single sample was prepared using the water extraction of 0.5 g tomato paste under optimal extraction conditions. Each paste was analyzed in triplicate resulting in a total of 39 samples.

### 2.2. Determining the optimal conditions for the extraction procedure

To reduce the amount of hazardous solvents for the environment, distilled water was used for the extractions. Optimal extraction conditions were determined using the central composite design, CCD, with two factors (time of sonication and temperature). Five levels were considered for each factor. The ranges and coded levels of investigated factors are presented in Table 1.

A total of 20 experiments were carried out. In order to evaluate the pure experimental error rate, experiments with nos. 1–8 were repeated twice and the experiment at the central point was repeated four times. The ORAC parameter was the response variable,  $y$ , that was determined in each experiment. In general, the following quadratic model was considered:

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_1^2 + b_4x_2^2 + b_5x_1x_2 + \varepsilon \quad (1)$$

Analysis of variance, ANOVA, with the type III sum of squares was adopted to determine the significance of each regression coefficient of the quadratic equation. ANOVA was also used to check the adequacy of the final model (the significance of lack-of-fit), since the estimation of the pure measurement error was available.

### 2.3. Evaluation of the total antioxidant capacity using the ORAC assay

The oxygen radical absorbance capacity, the ORAC assay [21], is measured as the damage to a fluorescence probe that is caused by the free radicals generated from AAPH. Reactions for a sample with the analytes (antioxidants) and for a sample without analytes are monitored in parallel and the TAC is evaluated on the basis of the area between the two curves that are registered.

The ORAC assay was performed using a 96-well plate. Fluorescence was measured at fixed excitation and emission wavelengths

Table 1  
Design of experiment.

Run (random order)	Time (min)	Temperature (°C)	Time coded ( $x_1$ )	Temperature coded ( $x_2$ )
6	5	34	–1	–1
3	5	34	–1	–1
16	13	34	+1	–1
11	13	34	+1	–1
7	5	72	–1	+1
17	5	72	–1	+1
14	13	72	+1	+1
8	13	72	+1	+1
5	3	53	–1.414	0
19	3	53	–1.414	0
15	15	53	+1.414	0
1	15	53	+1.414	0
2	9	26	0	–1.414
4	9	26	0	–1.414
18	9	80	0	1.414
13	9	80	0	1.414
9	9	53	0	0
20	9	53	0	0
10	9	53	0	0
12	9	53	0	0

(492 nm and 530 nm, respectively). Fluorescein (4.75  $\mu\text{M}$ ) was used as the fluorescence probe. A probe of 50  $\mu\text{L}$  was mixed with either 50  $\mu\text{L}$  of a diluted extract or with a phosphate buffer (pH=7.4) of 50  $\mu\text{L}$  as a blank sample or with a standard sample of 50  $\mu\text{L}$  containing 80  $\mu\text{M}$  Trolox (6-hydroxy-2,5,7,8-tetramethylchroman-2-carboxylic acid). After 20 min of incubation at 37  $^{\circ}\text{C}$ , 25  $\mu\text{L}$  of AAPH (0.0029 M) was introduced into each sample. The reaction was driven to completion (for 6 h). The water soluble analog of vitamin E—Trolox (6-hydroxy-2,5,7,8-tetramethylchroman-2-carboxylic acid) was used as the antioxidant standard and the TAC value was calculated as  $\mu\text{mole}$  of Trolox per g of a sample. Trolox equivalent (TE).

#### 2.4. Evaluation of the total phenolic content using the Folin–Ciocalteu method

The total phenolic content was determined using the Folin–Ciocalteu method, which was described in detail in [24]. The absorbance of a sample is measured after the reaction between the reagent and polyphenols is completed.

Each 10 mL sample was prepared by diluting a mixture of 500  $\mu\text{L}$  tomato paste extract, 250  $\mu\text{L}$  of Folin–Ciocalteu reagent and 1 mL of saturated  $\text{Na}_2\text{CO}_3$ . After 30 min of incubation at 50  $^{\circ}\text{C}$ , the sample was allowed to equilibrate and its absorbance was measured at 760 nm. A calibration model (with a correlation coefficient,  $r$ , of 0.9999) was constructed for gallic acid (3,4,5-trihydroxybenzoic acid) in the range of 0.088–0.294  $\mu\text{mol}$  and the TPC values for samples were expressed as the gallic acid equivalents, GAE, ( $\mu\text{mol}$  of gallic acid per g of a sample). Gallic acid is the recommended standard for this reaction.

#### 2.5. Registration of the fluorescence spectra

In our experiment, a Varian Cary Eclipse fluorescence spectrophotometer was used to record the spectra of intact tomato pastes and extracts. The EEMs were recorded at 161 emission wavelengths from 280 to 600 nm in a 2 nm interval and 26 excitation wavelengths in the range of 250 and 500 nm in a 10 nm interval. The emission intensities of the intact tomato paste samples were measured with a detector sensitivity of 800 V using a fiber optic probe that was adjusted to measure solid samples. The excitation and emission slits were set to 5 nm. The measurements for each sample were performed in triplicate.

The EEMs for 39 tomato extracts were registered with a detector sensitivity of 650 V using a standard (10  $\times$  10 mm) quartz cuvette. Excitation and emission slits were set to 5 nm and right angle geometry was set.

#### 2.6. Registration of the infrared spectra

The IR spectra were recorded for the intact tomato pastes using a Nicolet iS50 FTIR spectrometer equipped with a ceramic source ( $\text{Si}_3\text{N}_4$ ) and DLATGS detector. The attenuated total reflection (ATR) accessory was equipped with a single bounce diamond (Golden Gate<sup>™</sup>, Specac Ltd., UK). The crystal was cleaned with distilled water and its surface was dried with nitrogen after each analysis. The background spectrum was registered before each measurement. Spectra were collected in the spectral range of 4000–400  $\text{cm}^{-1}$  at a resolution of 4  $\text{cm}^{-1}$ . A total of 64 replicate scans for every sample were collected and averaged in order to obtain the final IR spectrum. Three replicate measurements were performed in a random order for each sample.

### 3. Multivariate modeling of the spectral data

Preprocessing of fluorescence and IR signals is required before the application of any multivariate method in order to extract relevant information from the data. The quality expressed as the signal-to-noise ratio for both the IR and fluorescence spectra should be evaluated first. If this ratio is relatively low, the removal of instrumental noise is necessary. Additionally, EEMs require the removal of the Rayleigh and Raman scattering signals, which are irrelevant from the chemical point of view and can negatively influence further chemometric modeling. The matricizing of the excitation–emission spectra is a necessary step when they are modeled using a classic exploratory or calibration approach (e.g., principal component analysis, PCA or partial least squares regression, PLSR). For this purpose, each EEM signal in a matrix form with the dimensions *excitation wavelengths*  $\times$  *emission wavelengths* is rearranged in a vector form with the dimensions  $1 \times (\text{excitation wavelengths} \times \text{emission wavelengths})$ . Thus, the final data are organized in a matrix with the dimensions *samples*  $\times$  (*excitation wavelengths*  $\times$  *emission wavelengths*) by putting all of the sample row vectors together. These multivariate data can then be analyzed using principal component analysis [27], which facilitates the exploration and visualization of the data. In the course of PCA, new variables, which are called principal components, are constructed by maximizing the variance of the projected data. Projections of objects on a principal component are called scores, while the projections of variables on a component are called loadings. Plotting the scores of one component against another (a scores plot) allows for the analysis of any patterns of the objects (if present), while the respective loadings plot can provide information about the importance of a particular variable in the trends that are observed. In our study, PCA was specifically used in order to see any similarity in the tomato pastes of different origins (score plots) that were described by their excitation–emission spectra and to identify the substances responsible for the trends that were observed in the score plots. The excitation–emission loadings for each principal component are presented as an image (see for example Fig. 2b) in order to enhance their interpretability.

Partial least squares regression was used to construct a model of the TAC or TPC values,  $y$ , that was obtained for the water extracts of the tomato pastes as a function of the excitation–emission spectra (in a matricized form) and the IR spectra for the tomato pastes. A PLSR model is built using a set of a few new variables that maximize the covariance between the dependent variable (TAC or TPC) and the explanatory variables (e.g., a collection of spectra) [25]. As was mentioned earlier, the explanatory variables can be arranged in a matrix form whether they represent a simple IR or UV–vis spectra or unfolded EEMs. The  $N$ -way partial least squares regression can be regarded as an extension of two-way PLSR to model three- or higher-way data. The  $N$ -way models often account for the second order advantage [28]. That is why such a model is more stable in the presence of new sources of variation.

To construct the PLSR model and to evaluate its performance, the available data should be divided into the model and test sets [29]. The model set is used to construct the calibration model and to evaluate its fit, which is expressed as the root mean square error, RMSE (see Eq. (2)), while the evaluation of the predictive properties of a model, which is expressed as the root mean square error of prediction, RMSEP, is performed using the test set:

$$\text{RMSE}(P) = \sqrt{\frac{1}{p} \sum_{i=1}^p (y_i - \hat{y}_i)^2} \quad (2)$$

In this equation,  $P$  is the number of latent PLSR factors,  $p$  is the total number of objects that are included into the model set (or the test set),  $y_i$  is the  $i$ th ( $i=1, 2, \dots, p$ ) experimental value of the

response variable, and  $\hat{y}_i$  is the  $i$ th value of the response variable as predicted using the constructed model.

Another parameter that characterizes the explained model variance is the coefficient of determination,  $D^2$ :

$$D^2 = 1 - \frac{\sum_{i=1}^p (y_i - \hat{y}_i)^2}{\sum_{i=1}^p (y_i - \bar{y})^2} \quad (3)$$

where  $\bar{y}$  is the mean value of dependent variable.

#### 4. Results and discussion

Firstly, the central composite design was used in order to find the optimal extraction conditions. As was mentioned earlier, the ranges of the temperature and sonication time were considered to be in agreement with the limitations of the sonication bath. The second-order model as described by Eq. (1) was fitted to the data and the significance of the regression coefficients was verified using the analysis of variance, ANOVA, with the type III sum of squares (see Table 2). The assumed level of significance was 0.05 ( $P$ -value < 0.05).

From the  $P$ -values that are presented in Table 2, one can conclude that the regression coefficient of variable  $x_1$ , which is associated with the time effect, is insignificant ( $P$ -value > 0.05), but the regression coefficient that describes its quadratic form is found to be significant ( $P$ -value < 0.05). This suggests a nonlinear dependency of the ORAC parameter and the extraction sonication time. The final regression model can be expressed with the following equation:

$$y = 119.70 + 10.73x_2 + 7.22x_1^2 - 5.75x_2^2 - 5.61x_1x_2 + \epsilon \quad (4)$$

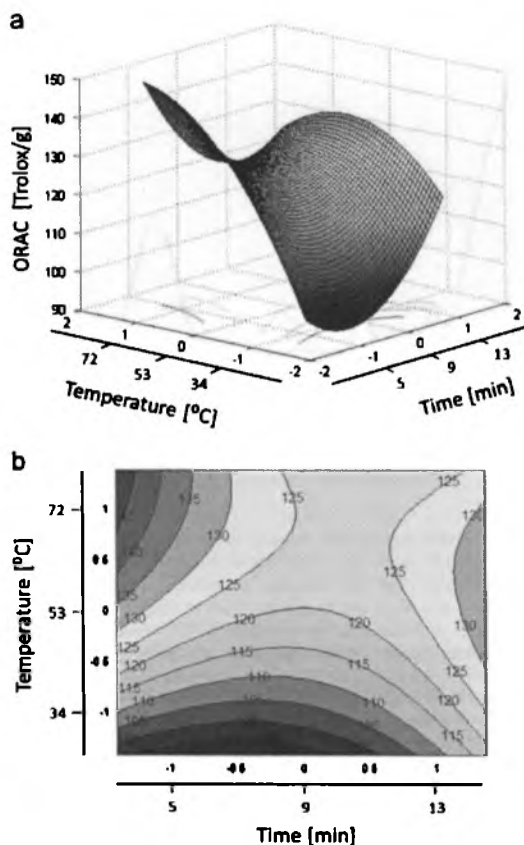
The coefficient of determination,  $D^2$ , was equal to 0.835. Because repeated measurements were performed, the residual variance can be partitioned further into the variance related to the pure measurement error and the variance of lack-of-fit with ANOVA. The model that is constructed is adequate and useful because the lack-of-fit was insignificant ( $P$ -value > 0.05). The response surface depicted in Fig. 1 shows one saddle point.

The largest value of the ORAC parameter was found for the highest temperature of 80 °C and the shortest time of sonication (3 min). This finding is in agreement with the suggestion of some authors that sonication and a long time of high temperature exposure may lead to a degradation of some antioxidant substances [30,31] and consequently to a decrease in the antioxidant capacity values as measured by the ORAC assay. The optimal conditions were further validated with additional measurements. The RMSE value was equal to 13.14 TE (7.97%).

The extractions of antioxidants that were present in the tomato pastes were carried out under optimal conditions. The ORAC parameter was evaluated for each extract and the EEM spectra were registered. The fluorescence and IR spectra for the intact samples were collected at the same time. Raw instrumental signals were then preprocessed prior to multivariate modeling. A visual inspection of the signals allowed it to be concluded that the signal-to-noise ratio

**Table 2**  
Results of the analysis of variance with the type III sum of squares.

Source of variation	Sum of squares	Degrees of freedom	Mean squares	F-value	P-value
$x_1$	3.40	1	3.40	0.070	0.795
$x_2$	$1.84 \times 10^3$	1	$1.84 \times 10^3$	37.98	< 0.001
$x_1^2$	476.37	1	476.37	9.822	0.007
$x_2^2$	302.16	1	302.16	6.230	0.026
$x_1x_2$	251.84	1	251.84	5.193	0.039
Model	$3.44 \times 10^3$	4	860.00	18.91	< 0.001
Residuals	682.36	15	45.49		



**Fig. 1.** Response surface that was obtained from the central composite design. A description of the coded parameters is presented in Table 1.

was at a satisfactory level for EEMs and IR spectra and that the noise reduction step was unnecessary. The preprocessing of EEMs consisted of removing the Rayleigh and Raman scattering using the routine that was described in detail in [32]. Scattering peaks were detected and removed. Then, the corresponding regions of signals were interpolated using the Delaunay triangulation. Next, each excitation–emission matrix with the dimensions  $26 \times 161$  was unfolded into a vector with the dimensions  $1 \times 4186$ . The final set of EEMs was presented into a matrix with the dimensions  $39 \times 4186$ .

The PCA score plots of the unfolded EEMs for the extracts of tomato pastes revealed a clustering tendency. Namely, the samples from Italy and Croatia (south of Europe), the projections of which are found along PC 1 axis (see Fig. 2a), can be distinguished.

An analysis of PC 1 loadings was performed after their refolding to an image that corresponded to the excitation–emission ranges of the EEMs (see Fig. 2b). The intensity of the color of each pixel is proportional to its loading value. Cold colors correspond to low values while hot colors indicate high values. An analysis of the image allowed for the conclusion that the mixtures that were studied contain chemical components that exhibit fluorescence of around 320 nm (excitation around 300 nm). These compounds are found at relatively high concentrations in the samples of tomato paste from Poland. A similar fluorescence characteristic is typical for catechin and epicatechin [33]. Both substances are common antioxidants that are present in tomatoes [34]. A further analysis

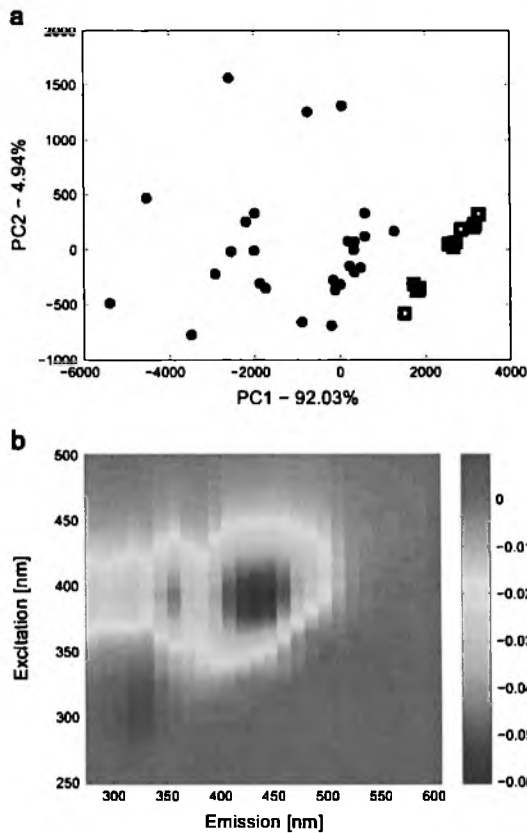


Fig. 2. (a) Projections of the intact tomato pastes that were described by the EEMs in the space spanned by PC 1 and PC 2. The samples from southern Europe are marked with 'o' and the samples from Poland are marked with 's' and (b) a color image of PC 1 reloaded loadings. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

of the image of the PC 1 loadings revealed the activity of fluorophores in the range of 350–500 nm with an excitation of around 360 nm. These fluorophores may be caffeic and ferulic acids. High levels of those fluorophores were found in the samples that originated from southern Europe. This observation is in the agreement with the observations that tomato products that were cultivated under influence of increased UV radiation had 20% higher concentrations of caffeic and ferulic acids [35]. Differences in UV radiation conditions are also common for the samples in our study (i.e. tomato products from southern Europe compared with those from Poland) [36]. The PC analysis of the IR spectra did not reveal any interesting data trends.

The IR spectra hold information about the chemical content of samples. Depending on the band that is analyzed, the water, amino acids or phenolic content can be evaluated. For the study of the total antioxidant capacity, the spectral region between 2000 and 900  $\text{cm}^{-1}$  is recommended in the literature [12,37]. This spectral range includes the stretch of phenyl groups and an asymmetric deformation of methyl and carbonyl groups that are prerequisite to the chemical structure of antioxidants. That is why, in our study, the PLSR models for TAC and TPC were constructed using this particular spectral range. Moreover, these models were compared to the models that were obtained for the samples that were described by their entire spectra (see Fig. 3).

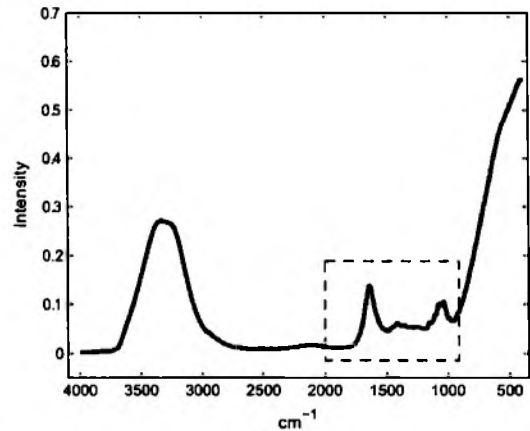


Fig. 3. Mean IR spectrum of the intact tomato samples. The spectral region between 2000 and 900  $\text{cm}^{-1}$  is indicated with a dashed line.

In order to include all of the samples with unique characteristics into the model set, the Kennard and Stone algorithm [38] was used. Evaluation of the uniqueness of samples that were analyzed was performed on the basis of mean spectra of the replicates. Thus, the model set contained 70% of available samples (each described by three replicates), while the remaining 30% of samples formed the test set. The same routine was applied when modeling the TAC as a function of ether EEMs (of the intact samples and the extracts of tomato paste) or the IR spectra (in full and reduced spectral ranges). The spectra of the same samples were selected for the model set in order to perform the comparison.

Both PLSR and N-PLS models were constructed for the different sets of spectra: 1) EEMs registered for the extracts from tomato pastes, 2) EEMs registered for the intact tomato pastes using a fiber optics probe, 3) IR spectra of the intact samples that were collected in the range of 4000–400  $\text{cm}^{-1}$ , and 4) IR spectra of the intact samples reduced to the spectral range of 2000–900  $\text{cm}^{-1}$ . The complexity of the models that were constructed (evaluated using the leave-one cross validation procedure [29]) and the figures of merit that were obtained (RMSE, RMSEP and  $D^2$ ) for TAC and TPC, respectively, are listed in Table 3.

The analysis of the figures of merit that are listed led to the conclusion that the PLSR models that were obtained from the tomato paste samples in the reduced spectral range of IR (2000–900  $\text{cm}^{-1}$ ) are similar (or better) to the models that were built for the model set that was described with the entire IR spectra.

The best result was obtained for the model of the ORAC parameter as a function of the EEMs (fit and prediction of the model were equal to 5.87% and 8.52%, respectively) for the extracts. Higher errors were obtained for the TAC prediction that was evaluated from the spectra of the intact samples than from spectra of extracts. The N-PLS model that was built to predict the TAC value from the EEMs had much better prediction properties than any model built using the IR spectra (in the entire or the shortened range). In contrast, the PLSR models that were built to predict the TPC from IR had lower errors than both the PLSR and N-PLS that were built using the EEMs. This was also confirmed by the improved performance of the models that were constructed for the IR spectra of the intact samples. These IR models were characterized by two-fold larger values of fit and prediction than the models that were built for the other type of spectra. In the Supplementary material section Tables S1 and S2 include observed and predicted TAC and TPC contents from PLS and N-PLS models for the studied model and test set samples.

**Table 3**  
Parameters of the optimal PLS and N-PLS models constructed to evaluate TAC and TPC.

	Spectra	EEM (extract)		EEM (paste)		IR (paste)	
		PLS	N-PLS	PLS	N-PLS	4000–400 cm <sup>-1</sup>	2000–900 cm <sup>-1</sup>
Total antioxidant capacity	Method	PLS	N-PLS	PLS	N-PLS	PLS	PLS
	Complexity	9	10	2	3	6	6
	RMSE <sup>a</sup>	0.830 (5.07%)	1.938 (13.70%)	2.471 (17.46%)	2.485 (17.56%)	1.948 (13.77%)	1.401 (9.90%)
	D <sup>2</sup>	0.960	0.864	0.650	0.537	0.782	0.887
	RMSEP <sup>b</sup>	1.206 (8.52%)	1.536 (10.86%)	2.575 (18.20%)	2.011 (14.21%)	3.398 (24.01%)	2.426 (17.15%)
Total phenolic content	Complexity	7	3	3	6	5	4
	RMSE <sup>b</sup>	0.037 (5.93%)	0.194 (31.31%)	0.069 (11.10%)	0.049 (8.06%)	0.063 (10.15%)	0.056 (9.08%)
	D <sup>2</sup>	0.950	0.805	0.823	0.911	0.853	0.882
	RMSEP <sup>b</sup>	0.038 (6.12%)	0.153 (24.81%)	0.067 (10.78%)	0.069 (11.08%)	0.033 (5.36%)	0.033 (5.36%)

<sup>a</sup> RMSE and RMSEP are expressed as Trolox equivalents (μmole of Trolox per g of sample), the percentage errors are computed according to the range of obtained TAC values.

<sup>b</sup> RMSE and RMSEP are expressed as gallic acid equivalents (μmole of gallic acid per g of sample), the percentage errors are computed according to the range of obtained TPC values.

## 5. Conclusions

In this study, a novel procedure that is based on the EEM measurements was proposed to evaluate the total antioxidant properties of intact tomato pastes. The N-PLS model that was built to predict the TAC values that were measured with the ORAC assay using the EEMs presented an acceptable performance in comparison to the calibration models that were obtained for a collection of the IR spectra. Even though the collection of the EEM signals requires twice as much time (around 5 min per sample), the measurement is simpler to perform in comparison to the IR measurement in which the crystal must be cleaned and then dried with nitrogen before each measurement. Due to the simplicity of the measurement, the procedure proposed can be routinely performed as a quality screening procedure with respect to the TAC parameter for intact samples. The EEM-based procedure does not outperform the IR-based method when the total polyphenol content is evaluated. We have confirmed the observation that has been reported in the other literature that the IR spectra that are registered in the range of 2000–900 cm<sup>-1</sup> contain useful information about the TPC of samples.

## Acknowledgment

JO is grateful for the financial support within the framework of the DoktorIS program –scholarship program for innov Silesia co-financed by the European Union under the European Social Fund (London) and the financial support of National Science Center (Poland) (Grant no. 2011/03/N/ST4/00713).

## Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at <http://dx.doi.org/10.1016/j.talanta.2015.01.026>.

## References

[1] G. Cao, R.L. Prior, *Clin. Chem.* 44 (1998) 1309–1315.  
 [2] M. Özyürek, K. Güçlü, E. Tötem, K.S. Başkan, E. Erçağ, S.E. Çelik, et al., *Anal. Methods* 3 (2011) 2439–2453. <http://dx.doi.org/10.1039/C1AY05320E>.  
 [3] I.F.F. Benzie, J.J. Strain, *Anal. Biochem.* 239 (1996) 70–76. <http://dx.doi.org/10.1006/abio.1996.0292>.  
 [4] E. Capanoglu, J. Beehilder, D. Boyacioglu, R. Hall, R. de Vos, *J. Agric. Food Chem.* 56 (2008) 964–973. <http://dx.doi.org/10.1021/jf072990e>.  
 [5] B. Cämmerer, L.W. Kroh, *Eur. Food Res. Technol.* 223 (2006) 469–474. <http://dx.doi.org/10.1007/s00217-005-0226-4>.

[6] M. Pérez-Martínez, B. Caemmerer, M.P. De Peña, C. Gd, L.W. Kroh, *J. Agric. Food Chem.* 58 (2010) 2958–2965. <http://dx.doi.org/10.1021/jf9037375>.  
 [7] C. Anesini, G.E. Ferraro, R. Filup, *J. Agric. Food Chem.* 56 (2008) 9225–9229. <http://dx.doi.org/10.1021/jf8022782>.  
 [8] M. Bartoszek, J. Polak, *Food Chem.* 132 (2012) 2089–2093. <http://dx.doi.org/10.1016/j.foodchem.2011.12.060>.  
 [9] J. Polak, M. Bartoszek, I. Stanimirova, *Food Chem.* 141 (2013) 3042–3049. <http://dx.doi.org/10.1016/j.foodchem.2013.05.133>.  
 [10] M.B. Hossain, C. Barry-Ryan, A.B. Martin-Diana, N.P. Brunton, *Food Chem.* 123 (2010) 85–91. <http://dx.doi.org/10.1016/j.foodchem.2010.04.003>.  
 [11] K. Ravichandran, N.M.M.T. Saw, A.A.A. Mohdaly, A.M.M. Gabr, A. Kastell, H. Riedel, et al., *Food Res. Int.* 50 (2013) 670–675. <http://dx.doi.org/10.1016/j.foodres.2011.07.002>.  
 [12] X. Lu, B.A. Rasco, *Crit. Rev. Food Sci. Nutr.* 52 (2012) 853–875. <http://dx.doi.org/10.1080/10408398.2010.511322>.  
 [13] L. Bittner, S. Schonbichler, G. Bonn, C. Huck, *Curr. Anal. Chem.* 9 (2013) 417–423.  
 [14] H.S. Lam, A. Proctor, L. Howard, M.J. Cho, *J. Food Sci.* 70 (2005) C545–C549. <http://dx.doi.org/10.1111/j.1365-2621.2005.tb08303.x>.  
 [15] M.J. Martelo-Vidal, M. Vázquez, *Food Chem.* 158 (2014) 28–34. <http://dx.doi.org/10.1016/j.foodchem.2014.02.080>.  
 [16] F.J. Rodríguez-Pulido, J.M. Hernández-Hierro, J. Nogales-Buena, B. Gordillo, M. L. González-Miret, F.J. Heredia, *Talanta* 122 (2014) 145–150. <http://dx.doi.org/10.1016/j.talanta.2014.01.044>.  
 [17] J. Orzeł, M. Daszykowski, *Chemom. Intell. Lab. Syst.* 137 (2014) 74–81. <http://dx.doi.org/10.1016/j.chemolab.2014.06.014>.  
 [18] J. Orzeł, M. Daszykowski, B. Walczak, *Chemom. Intell. Lab. Syst.* 110 (2012) 89–96.  
 [19] J. Christensen, E.M. Becker, C.S. Frederiksen, *Chemom. Intell. Lab. Syst.* 75 (2005) 201–208. <http://dx.doi.org/10.1016/j.chemolab.2004.07.007>.  
 [20] R.H. Myers, D.C. Montgomery, C.M. Anderson-Cook, *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*, John Wiley & Sons, London, 2009.  
 [21] D. Huang, B. Ou, M. Hampsch-Woodill, J.A. Flanagan, R.L. Prior, *J. Agric. Food Chem.* 50 (2002) 4437–4444. <http://dx.doi.org/10.1021/jf0201529>.  
 [22] A. Zulueta, M.J. Esteve, A. Frigola, *Food Chem.* 114 (2009) 310–316. <http://dx.doi.org/10.1016/j.foodchem.2008.09.033>.  
 [23] R. Apak, S. Gorinstein, V. Böhm, K.M. Schaich, M. Özyürek, K. Güçlü, *Pure Appl. Chem.* 85 (2013) 957–998. <http://dx.doi.org/10.1031/PAC-REP-12-07-15>.  
 [24] V.L. Singleton, R. Orthofer, R.M. Lamuela-Raventós, in: Lester Packer (Ed.), *Methods in Enzymology*, Academic Press, 1999, pp. 152–178 (accessed February 13, 2014).  
 [25] D.L. Massart, B.G.M. Vandeginste, L.M. Buydens, S. de Jong, P.J. Lewi, J. Smeyers-Verbeke, *Handbook of Chemometrics and Qualometrics: A Elsevier*, London, 1997.  
 [26] R. Bro, *J. Chemom.* 10 (1996) 47–61 (doi:10.1002/(SICI)1099-128X(199601)10:1<47::AID-CEM400>3.0.CO;2-C).  
 [27] S. Wold, K. Esbensen, P. Geladi, *Chemom. Intell. Lab. Syst.* 2 (1987) 37–52. [http://dx.doi.org/10.1016/0169-7439\(87\)80084-9](http://dx.doi.org/10.1016/0169-7439(87)80084-9).  
 [28] A.C. Olivieri, G.M. Escandar, A.M. de la Peña, *TrAC Trends Anal. Chem.* 30 (2011) 607–617. <http://dx.doi.org/10.1016/j.trac.2010.11.018>.  
 [29] T. Naes, T. Isaksson, T. Fearn, T. Davies, *A User Friendly Guide to Multivariate Calibration and Classification*, NIR Publications, Chichester, 2002.  
 [30] H.V. Annegowda, L.N. Anwar, M.N. Mordt, S. Ramanathan, S.M. Mansor, *Pharmacogn. Res.* 2 (2010) 368–373. <http://dx.doi.org/10.4103/0974-8490.75457>.  
 [31] A. Rawson, A. Patras, B.K. Tiwari, F. Noci, T. Koutchma, N. Brunton, *Food Res. Int.* 44 (2011) 1875–1887. <http://dx.doi.org/10.1016/j.foodres.2011.02.053>.  
 [32] R.G. Zepp, W.M. Sheldon, M.A. Moran, *Mar. Chem.* 89 (2004) 15–36. <http://dx.doi.org/10.1016/j.marchem.2004.02.006>.

- [33] I. Ramirez-Sanchez, L. Maya, G. Ceballos, F. Villarreal, J. Food Compos. Anal. 23 (2010) 790–793. <http://dx.doi.org/10.1016/j.jfca.2010.03.014>.
- [34] M.S. Lenucci, D. Cadinu, M. Taurino, G. Piro, G. Dalessandro, J. Agric. Food Chem. 54 (2006) 2606–2613. <http://dx.doi.org/10.1021/jf052920c>.
- [35] D.L. Luthria, S. Mukhopadhyay, D.T. Krizek, J. Food Compos. Anal. 19 (2006) 771–777. <http://dx.doi.org/10.1016/j.jfca.2006.04.005>.
- [36] B. Liley, R. McKenzie, Where on Earth has the Highest UV7 (2006). ([https://www.niwa.co.nz/sites/default/files/import/attachments/Liley\\_2.pdf](https://www.niwa.co.nz/sites/default/files/import/attachments/Liley_2.pdf)).
- [37] X. Lu, C.F. Ross, J.R. Powers, D.E. Aston, B.A. Rasco, J. Agric. Food Chem. 59 (2011) 5215–5221. <http://dx.doi.org/10.1021/j201254f>.
- [38] R.W. Kennard, L.A. Stone, Technometrics 11 (1969) 137–148. <http://dx.doi.org/10.2307/1266770>.



Katowice, 03.03.2015

dr Ivana Stanimirova-Daszykowska  
Instytut Chemii  
Uniwersytet Śląski  
ul. Szkolna 9  
40-006 Katowice

Oświadczam, że w artykule pt. *Prediction of the hydrophilic antioxidant capacity of tomato pastes from the IR and fluorescence excitation-emission spectra of extracts and intact samples*, opublikowanym w czasopiśmie *Talanta* 138 (2015) 64-70 mój udział polegał na:

- przygotowaniu planu eksperymentu i statystycznej ocenie uzyskanych wyników,
- pomocy w interpretacji uzyskanych wyników,
- opiece merytorycznej w trakcie przygotowywania manuskryptu zgodnie z wymaganiami czasopisma *Talanta*.



Wrocław, 03.03.2015

prof. dr hab. Bogusława Czarnik-Matusewicz  
Wydział Chemii  
Uniwersytetu Wrocławskiego  
Ul. Joliot Curie 14  
50-383 Wrocław

Oświadczam, że w artykule pt. *Prediction of the hydrophilic antioxidant capacity of tomato pastes from the IR and fluorescence excitation-emission spectra of extracts and intact samples*, opublikowanym w czasopiśmie *Talanta* 138 (2015) 64-70 mój udział polegał na pomocy w rejestracji widin z zakresu podczerwieni oraz interpretacji uzyskanych wyników.

B. Czarnik-Matusewicz

Katowice, 03.03.2015

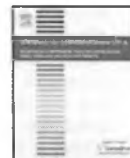
dr hab. Michał Daszykowski, prof. UŚ  
Instytut Chemii  
Uniwersytet Śląski  
ul. Szkolna 9  
40-006 Katowice

Oświadczam, że w artykule pt. *Prediction of the hydrophilic antioxidant capacity of tomato pastes from the IR and fluorescence excitation-emission spectra of extracts and intact samples*, opublikowanym w czasopiśmie *Talanta* 138 (2015) 64-70 mój udział polegał na:

- współtworzeniu i weryfikacji hipotezy badawczej oraz ogólnej koncepcji badań,
- merytorycznej opiece w trakcie realizacji eksperymentu oraz podczas interpretacji uzyskanych wyników,
- merytorycznym nadzorze przygotowania manuskryptu przez doktorantkę,
- udziale w procedurze redakcyjnej (tj. udział w przygotowaniu odpowiedzi na recenzje),
- pomocy w dokonaniu ostatecznej korekty artykułu.

Michał Daszykowski





## Modeling of the total antioxidant capacity of rooibos (*Aspalathus linearis*) tea infusions from chromatographic fingerprints and identification of potential antioxidant markers



Joanna Orzel<sup>a</sup>, Michal Daszykowski<sup>a,\*</sup>, Malgorzata Kazura<sup>a</sup>, Dalene de Beer<sup>b</sup>, Elizabeth Joubert<sup>b,c</sup>, Alexandra E. Schulze<sup>c</sup>, Theresa Beelders<sup>c</sup>, André de Villiers<sup>d</sup>, Christiaan J. Malherbe<sup>b</sup>, Beata Walczak<sup>a</sup>

<sup>a</sup> Institute of Chemistry, The University of Silesia, 9 Szkolna Street, 40-006 Katowice, Poland

<sup>b</sup> Post-Harvest and Wine Technology Division, Agricultural Research Council (ARC), Infruitec-Nietvaarbij, Private Bag X5026, Stellenbosch 7599, South Africa

<sup>c</sup> Department of Food Science, Stellenbosch University, Private Bag X1, Matieland (Stellenbosch) 7602, South Africa

<sup>d</sup> Department of Chemistry and Polymer Science, Stellenbosch University, Private Bag X1, Matieland (Stellenbosch) 7602, South Africa

### ARTICLE INFO

#### Article history:

Received 27 June 2014

Received in revised form

12 September 2014

Accepted 12 September 2014

Available online 22 September 2014

#### Keywords:

Antioxidant markers

*Aspalathus linearis*

Polyphenols

Calibration

Chemometrics

Quality control

### ABSTRACT

Models to predict the total antioxidant capacity (TAC) of rooibos tea infusions from their chromatographic fingerprints and peak table data (content of individual phenolic compounds), obtained using HPLC with diode array detection, were developed in order to identify potential antioxidant markers. Peak table data included the content of 12 compounds, namely phenylpyruvic acid-2-O-glucoside, aspalathin, nothofagin, isoorientin, orientin, ferulic acid, quercetin-3-O-robinobioside, vitexin, hyperoside, rutin, isovitexin and isoquercitrin. The TAC values, measured using the oxygen radical absorbance capacity (ORAC) and DPPH radical scavenging assays, could be predicted from the peak table data or the chromatographic fingerprints (prediction errors 9–12%) using partial least squares (PLS) regression. Prediction models created from samples of only two production years could additionally be used to predict the TAC of samples from another production year (prediction errors <13%) indicating the robustness of the models in a quality control environment. Furthermore, the uninformative variable elimination (UVE)-PLS method was used to identify potential antioxidant markers for rooibos infusions. All individual phenolic compounds that were quantified were selected as informative variables, except vitexin, while UVE-PLS models developed from chromatographic fingerprints indicated additional antioxidant markers, namely (S)-eriodictyol-6-C-glucoside, (R)-eriodictyol-6-C-glucoside, aspalalinin and two unidentified compounds. The potential antioxidant markers should be validated prior to use in quality control of rooibos tea.

© 2014 Elsevier B.V. All rights reserved.

### 1. Introduction

Total antioxidant capacity (TAC) has become a popular quality parameter of food products and plant extracts, providing scientists and manufacturers with a convenient tool to compare products. Studies establishing a link between dietary TAC, as a measure of antioxidant intake, and health [1,2] serve to support its wide-spread use. Notwithstanding such a link, knowledge about the individual antioxidants, contributing to TAC, is required as the in vivo impact of products may vastly differ due to compositional differences. Once identified, the antioxidants could be

further characterized in terms of specific antioxidant properties and bioavailability, amongst others.

The classical approach to identify bioactive compounds in complex mixtures is by bio-assay-guided fractionation, a long and tedious process. Recent papers demonstrated that advanced chromatographic fingerprint analysis, combined with chemometric data analysis, could provide information about potentially active compounds [3]. This approach has been employed for identifying antioxidants in plant extracts [4–6].

In the present paper the focus falls on rooibos, a popular herbal tea [7]. Its inclusion in antioxidant tables of various beverages and food products [8], combined with evidence that the beverage improves the plasma antioxidant status of healthy subjects [9], merits greater insight into compounds contributing to its antioxidant capacity. Previous studies, investigating the antioxidant

\* Corresponding author. Tel.: +48 32 359 1568; fax: +48 32 259 9978.  
E-mail address: [mdaszyk@us.edu.pl](mailto:mdaszyk@us.edu.pl) (M. Daszykowski).

activity of rooibos polyphenols, focused largely on the major compounds, and in particular aspalathin, a novel dihydrochalcone C-glycoside [10–13]. Modeling of TAC, based on chromatographic fingerprint analysis, could offer a new approach for identifying new antioxidant markers.

A prerequisite of such fingerprint analysis is good separation of peaks as co-elution will decrease information. Chromatographic data generated with a new validated, high-resolution HPLC-DAD method [14] and TAC values determined with the oxygen radical scavenging (ORAC) and 2,2-diphenyl-1-picrylhydrazyl (DPPH) radical scavenging assays, were used in the present study. Our first aim was to verify the possibility of predicting the TAC of rooibos tea infusions directly from the HPLC-DAD chromatographic data as both quantitative data on selected bioactive marker compounds and TAC are used by industry as quality control parameters. Secondly, prediction models using UVE-PLS (uninformative variable elimination-partial least squares) regression [15] were used to identify new potential antioxidant marker compounds.

In the current study, both a targeted and non-targeted approach for analysis of chromatographic data was used. For the targeted approach, the TAC values of rooibos tea infusions were evaluated from peak table data (content of individual phenolic compounds). This requires the identification, detection and quantification of compounds selected using published information. For the non-targeted approach, TAC was modeled based on the chromatographic fingerprints, representing a comprehensive phenolic profile, of rooibos tea infusions. Chromatographic methods have the potential to provide information about a relatively large number of chemical compounds (assuming that their separation during a chromatographic run is possible). Thus, a large number of potential antioxidant compounds can be screened at one time. Such a data representation contains a larger amount of potentially relevant chemical information than a peak table, while also eliminating the need for expensive chromatographic standards. However, raw chromatographic fingerprints of complex samples need careful preprocessing prior to further comparative analysis and/or modeling, i.e. improvement of the signal-to-noise ratio, baseline elimination and alignment [16]. Prior to the construction of calibration models, a detailed exploration of the multivariate chromatographic data using principal component analysis (PCA) [17] was performed in order to uncover their structure. Calibration models based on chromatographic data to predict the TAC values of rooibos tea infusions were developed using PLS [18]. Relevant variables that pinpoint the potential antioxidant marker compounds were selected using UVE-PLS [15] for both peak table and chromatographic fingerprint data.

## 2. Materials and methods

### 2.1. Sample preparation and chemical analysis

HPLC-DAD and TAC data using the microplate ORAC and DPPH radical scavenging assays (data designated TAC<sub>DPPH</sub> and TAC<sub>ORAC</sub>, respectively) for 114 rooibos samples were obtained from a study described in detail by Joubert et al. [14]. Briefly: A total of 114 production batches of “fermented” (oxidized) rooibos plant material representing four quality grades (A, B, C and D, where A and D is the highest and lowest quality grade, respectively) were sampled during three different production years (2009, 2010 and 2011). The pasteurization treatment, a processing step in the production of all commercially available rooibos, was performed on a sub-sample of each of the 114 samples. Duplicate infusions were prepared for pasteurized samples and the infusions were analyzed to determine their TAC and phenolic composition. Duplicate HPLC-DAD

chromatograms were registered for each infusion. Twelve monomeric compounds, i.e. the phenylpropenoic glycoside, phenylpyruvic acid-2-O-glucoside (PPAG), the flavonoid glycosides, aspalathin, nothofagin, isoorientin, orientin, isovitexin, vitexin, quercetin-3-O-robinobioside, hyperoside, rutin and isoquercitrin, as well as ferulic acid, a phenolic acid, were quantified using the chromatographic data (described in Joubert et al. (2012) [14]). HPLC-DAD, TAC<sub>DPPH</sub> and TAC<sub>ORAC</sub> data for the unpasteurized samples (data not previously reported) were also obtained in the same manner to create a larger data set.

Precision of the TAC<sub>DPPH</sub> assay was determined using three randomly chosen samples. Each sample was analyzed six times on one microplate, with each value representing the mean for three wells. The relative standard deviation (%RSD) was calculated for the six repetitions, and represented the intra-plate precision. This procedure was repeated for six different plates and %RSD calculated between the six plates, representing the inter-plate precision.

### 2.2. Data preprocessing

The HPLC-DAD data from 298 to 500 nm at 2 nm intervals (102 wavelengths) for retention time points from 1.7 to 40 min (12,000 sampling points) were exported from OpenLab ChemStation software (Agilent Technologies, Waldbronn, Germany) in CSV format and imported to Matlab (version R2009b, Mathworks, Natick, MA, USA) for further data processing. A single chromatogram was constructed for each injection by calculating the mean absorbance intensity over the wavelengths obtained. The data set consisted of 912 chromatograms, namely 114 samples each of pasteurized and unpasteurized rooibos for which duplicate infusions were prepared and duplicate injections performed.

Prior to data analysis, the chromatograms required signal preprocessing such as alignment and baseline correction. For alignment, a target chromatogram was selected as the one with the highest mean Pearson's correlation coefficient with respect to all of the chromatographic signals [19]. In the first step, all the chromatograms were aligned with the target using the automated alignment approach (AA) described in detail by Daszykowski et al. [20] in order to eliminate peak shifts. For each alignment run, the number of spline functions was optimized (between 10 and 20) in order to ensure the satisfactory alignment of most chromatograms. The chromatographic signals that required an improvement of the alignment were additionally corrected using correlation optimized warping (COW) [21].

Elimination of the background component from chromatographic signals was achieved by applying the asymmetric penalized least squares algorithm [22]. In our study, the first derivative of the signals was used to trace signal smoothness and the penalty parameter ( $\lambda$ ) was set to 10,000. After the alignment of the signals and baseline correction, the data size was  $912 \times 12,000$ . Aligned and baseline corrected chromatograms were then averaged (two chromatograms registered for duplicate infusions were averaged according to the sampling points) in order to obtain 228 chromatograms that corresponded to different samples (unpasteurized and pasteurized) (the size of the data set was  $228 \times 12,000$  (objects  $\times$  sampling points)). For further analysis, sampling points from 1033 to 12,000 were selected to exclude the ascorbic acid peak (ascorbic acid was added to prevent oxidation of the compounds during HPLC analysis [14]), giving a final data matrix with dimensionality of  $228 \times 10,967$ . The content of the twelve quantified phenolic compounds in infusions of pasteurized rooibos was published in Joubert et al. [14]. Additionally, the content of these compounds in infusions of unpasteurized rooibos was added to the data set to construct a peak table with a dimensionality of  $228 \times 12$  (samples  $\times$  concentrations of compounds).

### 2.3. Data exploration

Exploration of multivariate chromatographic data is one of the first steps in chemometric analysis [23]. Its major goal is to uncover similarities among samples and variables and to detect samples that are located far from the majority of the data (the so-called outliers). The studied data, regardless their representation, were multivariate and complex. Thus, their exploration required multivariate exploration and modeling methods. PCA [24] is one of the most popular multivariate methods that is used to explore data structure. Multivariate data are decomposed into a set of new variables called principal components (PCs) using PCA. They are constructed in order to maximize the description of data variance, and therefore can be used for visualization of data structure. This is achieved by the projection of objects and variables onto planes that are defined by the selected pairs of PCs (the so-called score and loading plots). When additional information about samples is available, for instance the corresponding grade, production year, etc., one can visualize the possible effect of such factors on the score projections and uncover the contribution of the original variables to the data structure that is observed on a score plot by analyzing the corresponding loadings.

### 2.4. Multivariate calibration using partial least squares regression (PLS)

The major goal of this study was to model the relationship between the chromatographic data and the TAC parameter ( $TAC = f(\text{chromatographic data})$ ). During multivariate calibration, regression coefficients,  $b$ , which describe the relation between response variable,  $y$  (TAC) and a set of explanatory variables,  $X$  (chromatographic data: a peak table or collection of chromatographic signals) [25], are determined. In general, the calibration model can be described as:

$$y_{(m,1)} = X_{(m,n)}b_{(n,1)} + e_{(m,1)} \quad (1)$$

where  $e$  are residuals from the calibration model,  $m$  is the number of calibration samples and  $n$  is the number of variables.

PLS [18] is a standard multivariate calibration method. The PLS model is constructed using a few new orthogonal variables (also called the PLS factors) that maximize the covariance between the PLS factors and response variable(s).

The model is built using calibration samples (also called a model set). The number of factors included in the PLS model (called the model complexity,  $f$ ) is evaluated on the basis of the so-called cross-validation procedure, CV [26].

### 2.5. Uninformative variable elimination-partial least squares (UVE-PLS)

The purpose of the uninformative variable elimination-partial least squares method, UVE-PLS, is to detect uninformative variables [15], i.e. the variables that have a high variance but little covariance with a dependent variable. Such variables, when considered for modeling, increase the complexity of a PLS model and make its interpretation difficult. In UVE-PLS, the detection of relevant variables for modeling is based on the stability of the regression coefficients that are evaluated under model perturbation for real and artificially added sets of irrelevant variables (drawn from normal distribution with a very small amplitude). For  $m$  calibration objects,  $m$  PLS models of a given complexity are built resulting in  $m$  sets of regression coefficients. Then, for each regression coefficient its stability is expressed as the mean to standard deviation ratio. In order to distinguish relevant variables, the stabilities of their regression coefficients are compared with the stabilities of the regression coefficients that represent the irrelevant variables.

Variables with absolute values of stability larger than the absolute value of stability that is observed for irrelevant variables can be regarded as informative and thus important in explaining the modeled property. A detailed description of the UVE-PLS method can be found in Centner et al. [15].

## 3. Results and discussion

### 3.1. Exploring the quality of TAC data

The TAC of rooibos tea infusions was determined using the DPPH radical scavenging and ORAC assays as these are two popular antioxidant assays with different mechanisms [27]. The DPPH radical scavenging assay is used routinely as a quality control parameter for rooibos extracts in South Africa, while the ORAC assay is popular for nutraceutical products in the USA. As illustrated by the histograms in Fig. 1a and b, the  $TAC_{DPPH}$  and  $TAC_{ORAC}$  values followed a nearly normal distribution. Ranges for  $TAC_{DPPH}$  and  $TAC_{ORAC}$  values were ca. 1000–3600  $\mu\text{mol Trolox equivalents (TE)}/L$  and ca. 5500–14600  $\mu\text{mol TE}/L$ , respectively. Most samples had average TAC values (regardless of the assay used) and there were no evident outlying objects.

Further exploration of  $TAC_{DPPH}$  and  $TAC_{ORAC}$  values showed that the samples were normally distributed according to their TAC properties, regardless of production year, grade or pasteurization treatment.

In the literature, weak to strong correlations among the TAC values obtained from different assays are reported [28–31]. In our study, a relatively good agreement between the two assays was observed (see Fig. 1c;  $r = 0.835$ ;  $p = 1.167 \times 10^{-6}$ ).

### 3.2. Exploratory analysis of chromatographic data

Exploration of the chromatographic data was performed using PCA. Prior to PCA, both sets of chromatographic data were pre-processed as described in Section 2.2. For the peak table data, the first two PCs accounted for 78.30% of the total data variance. PCA compression of the chromatographic fingerprints was similar in its efficiency. The first two PCs explained 79.24% of the total data variance. The score plot for the first two principal components of the chromatographic fingerprints revealed six objects (indicated with solid and dashed red ellipses) located far away from the majority of the data (Fig. 2a). Four out of the six samples, indicated with a dashed red ellipse, can be regarded as bad leverage samples and they may affect the performance of a model [32,33], since their TAC values were not extreme. These samples were, therefore, excluded from the calibration data set. The score plot of the third and the fourth PCs indicated a differentiation of the samples according to production year. A differentiation of samples collected in 2009 and those produced during the following years (2010 and 2011) was observed along PC4 (Fig. 2b). A detailed analysis of the corresponding loading plot (Fig. 2c) allowed regions on the elution time axis of the chromatograms where compounds sensitive to production year elute, to be detected. In this manner, orientin, isoorientin, hyperoside and quercetin-3-O-robinobioside were found to be year-dependent compounds, which differentiated the samples produced during 2009 from those produced in 2010 and 2011. Previously, the production year was found to significantly ( $p < 0.05$ ) affect the isoorientin, hyperoside, isovitexin, nothofagin and PPAG content of pasteurized rooibos tea infusions [14]. The major phenolic constituents of rooibos tea infusions are isoorientin and orientin ( $>10 \text{ mg}/L$ ), with quercetin-3-O-robinobioside, PPAG, and aspalathin present at  $>5 \text{ mg}/L$  [14]. Isovitexin, vitexin, and hyperoside are present at  $<3 \text{ mg}/L$ , while rutin, ferulic acid, and isoquercitrin are present at  $<2 \text{ mg}/L$  and nothofagin is present

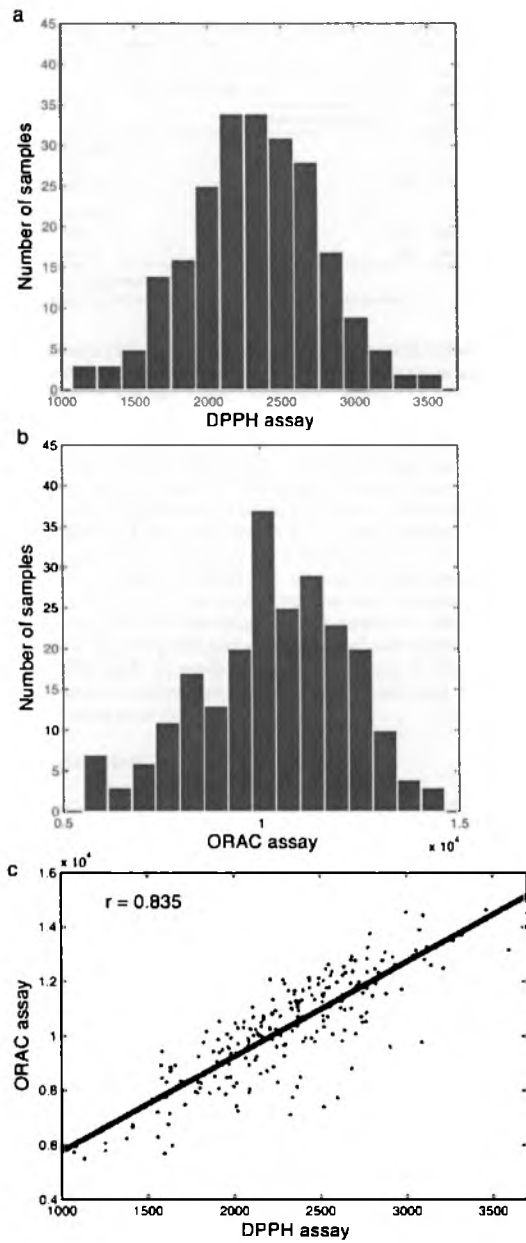


Fig. 1. Distribution of the (a)  $TAC_{DPPH}$  ( $\mu\text{mol Trolox equivalents/L}$ ) and (b)  $TAC_{ORAC}$  ( $\mu\text{mol Trolox equivalents/L}$ ) values for rooibos tea infusions (histograms contain 15 bins), and (c) agreement between  $TAC_{DPPH}$  and  $TAC_{ORAC}$  values of rooibos tea infusions.

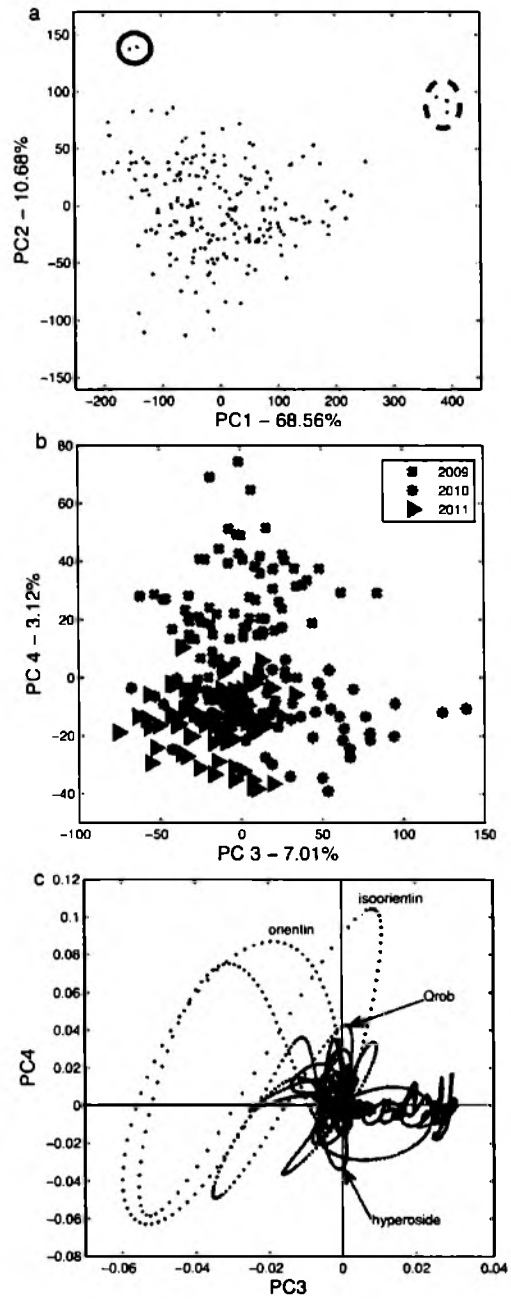


Fig. 2. PCA projections for chromatographic fingerprints: (a) PC1 vs. PC2 score plot, (b) PC2 vs. PC3 score plot and (c) loading plot presenting the contribution of particular variables to the construction of the selected PCs. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)



**Table 1**  
Fit and prediction errors obtained from PLS<sup>a</sup> and UVE-PLS<sup>b</sup> models.

Data	PLS <sup>a</sup>				UVE-PLS <sup>b</sup>					
	Model complexity	Fit		Prediction		Model complexity	Fit		Prediction	
		Trolox equivalents	%	Trolox equivalents	%		Trolox equivalents	%	Trolox equivalents	%
<i>DPPH models</i>										
Peak table	6	241	9.54	274	9.78	4	244	9.67	280	11.1
Fingerprints	6	178	7.05	237	9.38	4	225	8.91	235	9.31
<i>ORAC models</i>										
Peak table	6	1016	11.1	1079	11.8	3	1062	11.3	1057	11.6
Fingerprints	6	828	9.06	1020	11.2	3	990	10.8	1021	11.2

<sup>a</sup> Partial least squares.

<sup>b</sup> Uninformative variable elimination-partial least squares.

at <1 mg/L [14]. A detailed analysis of additional score plots did not uncover any trends with respect to quality grade or pasteurization process. Previously, the pasteurization process [34] and quality grade [14] were shown to affect the phenolic composition of rooibos tea infusions. However, it is important to emphasize that PCA is an unsupervised technique that reveals differences among groups of samples only when these differences are associated with the axes that express large data variability. Therefore, a certain effect will be uncovered on a score plot only if it is directly associated with the variance that is explained by the selected PCs.

Neither outlying objects nor obvious trends were observed (data not shown) on the score plots that were constructed for the peak table data. The significance of the treatment effects (production year, quality grade and pasteurization) was tested previously using the ASCA approach (analysis of variance combined with a simultaneous component analysis) and is discussed in detail by Stanimirova et al. [35].

### 3.3. Modeling of TAC

The next step in the study was to model the TAC using chromatographic data. Before the construction of multivariate calibration models, samples were divided into calibration and test sets using the Duplex algorithm [36]. Such a strategy ensures the uniform distribution of objects over the model space and allows samples with unique characteristics to be included in both the model and test sets. Thus, the fit and prediction errors obtained on the basis of the sets selected with the Duplex algorithm are reliable error estimates that characterize the constructed calibration model.

For modeling based on the peak table data, 178 and 50 samples were included in the model and test sets, respectively (see Supplementary Table S1 for characteristics of the model and test sets). For chromatographic fingerprint modeling, the model set contained 174 samples and the test set 50 samples.

Supplementary Table S1 related to this article can be found, in the online version, at <http://dx.doi.org/10.1016/j.chroma.2014.09.03>.

The TAC<sub>DPPH</sub> and TAC<sub>ORAC</sub> values were modeled separately using the peak table data and chromatographic fingerprints. The optimal complexities of four constructed PLS models were determined using the leave-one-out cross validation procedure. The performance of the optimal models is expressed as root mean square error (RMSE) and root mean square error of prediction (RMSEP) and they are summarized in Table 1. The fit and prediction properties of the PLS models that were constructed for chromatographic fingerprints were comparable with those obtained for the peak table data. The fit for all models were between 7.05 and 11.1%, while prediction errors were between 9.38 and 11.8%. The calibration models for TAC<sub>DPPH</sub> values (prediction error < 10%) showed superior performances to those of the models that were constructed for TAC<sub>ORAC</sub> values (prediction error < 12%). This is reflected in less scatter of the samples

(included in the calibration and test sets) as is shown for the *y* predicted vs. *y* observed plots (Fig. 3). A possible explanation is that the ORAC assay is more sensitive, but at the same time less selective than the DPPH radical scavenging assay as observed for the on-line HPLC-antioxidant assays [37]. Additionally, the TAC<sub>DPPH</sub> of rooibos food ingredient extracts showed a significant correlation ( $p < 0.001$ ) with the total quantified phenolic compound content, while the same was not observed for TAC<sub>ORAC</sub> ( $p \geq 0.05$ ) [38]. The mean intra- and inter-day precision of the ORAC protocol used in this study was reported as %RSD (relative standard deviation)  $\leq 15\%$  [39], while the DPPH assay had a higher precision with %RSD < 5%. This indicates that the prediction errors obtained using PLS were in the same range as experimental error for the respective assays.

The regression coefficients of the constructed calibration model represent the contribution of particular variables to the modeled TAC (see Supplementary Fig. S1 for regression coefficients of constructed models), however, due to problems associated with interpretability of regression coefficients for PLS models as described by Brown and Green [40], they were not used to detect potential antioxidant markers.

Supplementary Fig. S1 related to this article can be found, in the online version, at <http://dx.doi.org/10.1016/j.chroma.2014.09.030>.

The UVE-PLS approach was used in order to identify compounds with a large contribution to the modeled TAC (Table 2), which may be further investigated as potential antioxidant markers. Informative variables were retained from the peak table data and chromatographic fingerprints with respect to the TAC values obtained from the two TAC assays (assuming 100% of confidence). Nine compounds were detected as important variables to explain TAC<sub>DPPH</sub> using the peak table data, namely PPAG, aspalathin, nothofagin, isoorientin, orientin, hyperoside, rutin, isovitexin and isoquercitrin (Table 2). Restricting further PLS modeling only to informative variables resulted in a decrease in the initial model complexity (from six to four PLS factors) without a substantial loss of model performance (model fit and prediction properties) (Table 1). A similar tendency was observed for the UVE-PLS model that was constructed for the peak table data and TAC<sub>ORAC</sub> values. In this case, the relevant compounds selected were nothofagin, quercetin-3-O-robinobioside and orientin (Table 2). Comparing the two subsets of selected compounds, two compounds were simultaneously indicated as uninformative. Regardless of the specific assay that was used to determine the TAC, ferulic acid and vitexin did not appear to contribute to the prediction of the TAC values when using the peak table data.

The variables that were retained from the UVE-PLS models constructed using chromatographic fingerprints and TAC<sub>DPPH</sub> or TAC<sub>ORAC</sub> data are marked as red circles in Fig. 4a and b, respectively. A large number of informative variables were detected for the DPPH (213) and ORAC (126) models using the chromatographic fingerprints. The optimal number of factors for the new PLS models

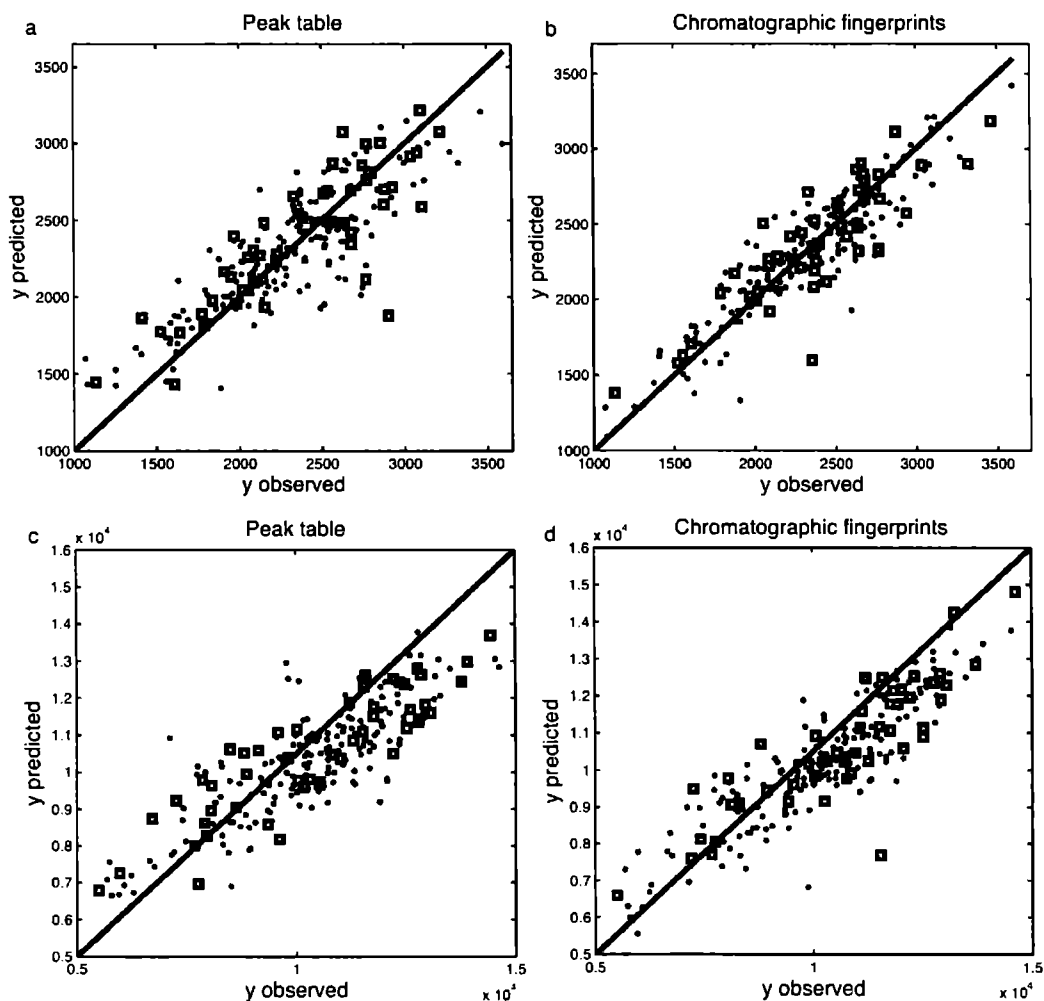


Fig. 3. The PLS models presented as  $y$  predicted vs.  $y$  observed for model (●) and test set samples (□), constructed for (a) and (c) peak table data and (b) and (d) chromatographic fingerprints.  $TAC_{DPPH}$  [(a) and (b)] and  $TAC_{ORAC}$  [(c) and (d)] were dependent variables.

was smaller compared to that of the initial PLS models. The performances of the PLS models that were constructed for a limited number of variables are presented in Table 1. The variables detected as informative for the DPPH radical scavenging and ORAC assays when using chromatographic fingerprints were compared with a published chromatogram of a rooibos tea infusion and corresponding compounds as identified or tentatively identified by Beelders et al. [41]. For the DPPH radical scavenging assay, informative variables corresponded to peaks of (*S*)-eriodictyol-6-*C*-glucoside, (*R*)-eriodictyol-6-*C*-glucoside, aspalathin, ferulic acid, aspalalinin, rutin, and isovitexin, as well as some regions of the chromatogram presumed to correspond to unidentified polymeric compounds (22–35 min) (Fig. 4a; Table 2). For the ORAC assay, on the other hand, informative variables corresponded to peaks of two unidentified compounds (b and f), aspalathin, aspalalinin, and isoquercitrin, as well as some regions of the chromatogram presumed to correspond to unidentified polymeric compounds (22–35 min) (Fig. 4b; Table 2). Informative variables selected for both assays using

chromatographic fingerprints were aspalathin, aspalalinin and regions corresponding to unknown polymeric compounds.

The compounds identified as important for predicting the  $TAC_{DPPH}$  and  $TAC_{ORAC}$  using the peak table and chromatographic fingerprints (Table 2) can be considered as potential antioxidant markers. Among the 12 major compounds in the peak table, only vitexin was not selected as an informative variable in any of the UVE-PLS models. Additional potential antioxidant markers could be identified using the chromatographic fingerprints, namely (*S*)-eriodictyol-6-*C*-glucoside, (*R*)-eriodictyol-6-*C*-glucoside, aspalalinin and two unidentified compounds. Among the potential antioxidant markers identified in this manner, aspalathin, ferulic acid, rutin, isovitexin, isoquercitrin, (*S*)-eriodictyol-6-*C*-glucoside and (*R*)-eriodictyol-6-*C*-glucoside, have been shown to be antioxidants [10–13,42]. Although the nature of the polymeric fraction remains unknown to date, it has been shown to inhibit lipid peroxidation in lipid peroxidation assays [43]. However, no information on the antioxidant activity of

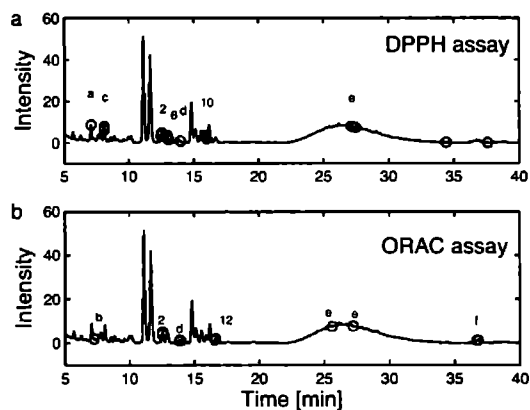
**Table 2**  
Informative variables selected (x) using UVE-PLS<sup>a</sup> for models built from peak table data and chromatographic fingerprints for TAC<sub>DPPH</sub><sup>b</sup> and TAC<sub>ORAC</sub><sup>b</sup> of rooibos tea infusions.

Compound nr	Compound	TAC <sub>DPPH</sub> <sup>b</sup>		TAC <sub>ORAC</sub> <sup>b</sup>	
		Peak table	CF <sup>c</sup>	Peak table	CF <sup>c</sup>
1	Phenylpyruvic acid-2-O-glucoside	x			
2	Aspalathin	x	x		x
3	Nothofagin	x		x	
4	Isoorientin	x			
5	Orientin	x		x	
6	Ferulic acid		x		
7	Quercetin-3-O-robinobioside			x	
8	Vitexin				
9	Hyperoside	x			
10	Rutin	x	x		
11	Isovitexin	x			
12	Isoquercitrin	x			x
a	(S)-Eriodictyol-6-C-glucoside		x		
b	Unidentified compound				x
c	(R)-Eriodictyol-6-C-glucoside		x		
d	Aspalalinin		x		x
e	Unidentified polymers		x		x
f	Unidentified compound		x		

<sup>a</sup> Uninformative variable elimination-partial least squares.

<sup>b</sup> Total antioxidant capacity determined using the DPPH radical scavenging and ORAC assays.

<sup>c</sup> Chromatographic fingerprints.



**Fig. 4.** A mean chromatographic fingerprint with regions containing the relevant variables selected using the UVE-PLS model constructed for TAC<sub>DPPH</sub> and TAC<sub>ORAC</sub> values indicated [peak nrs correspond to those in Table 2]. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

**Table 4**

Fit and prediction properties of PLS<sup>a</sup> models constructed for predicting TAC<sub>DPPH</sub><sup>b</sup> values of rooibos tea infusions from chromatographic fingerprints using samples collected during two out of three production years.

Samples included in the model set	Samples included in the test set	Model complexity	Fit		Prediction	
			RMSE <sup>c</sup>	%	RMSEP <sup>d</sup>	%
2009 and 2010	2011	4	223	8.83	255	10.9
2009 and 2011	2010	5	200	7.92	279	11.0
2010 and 2011	2009	4	203	8.04	326	12.9

<sup>a</sup> Partial least squares.

<sup>b</sup> Total antioxidant capacity determined using the DPPH radical scavenging assay.

<sup>c</sup> Root mean square error.

<sup>d</sup> Root mean square error of prediction.

**Table 3**  
Performance of PLS<sup>a</sup> models for predicting TAC<sub>DPPH</sub><sup>b</sup> of rooibos tea infusions from chromatographic fingerprints constructed for samples obtained during three consecutive production years (2009, 2010, and 2011).

Production year	Model complexity	Fit		Prediction	
		RMSE <sup>c</sup>	%	RMSEP <sup>d</sup>	%
2009	5	152	9.01	198	11.7
2010	5	219	8.71	231	9.19
2011	5	179	7.48	171	7.14

<sup>a</sup> Partial least squares.

<sup>b</sup> Total antioxidant capacity determined using the DPPH radical scavenging assay.

<sup>c</sup> Root mean square error.

<sup>d</sup> Root mean square error of prediction.

PPAG, quercetin-3-O-robinobioside, aspalalinin and the unidentified compounds is available. PPAG is not expected to be an active antioxidant as it lacks the phenolic structural features that are required for free radical scavenging ability. Its contribution to the UVE-PLS model is thus likely solely because its concentration showed a similar trend as TAC<sub>DPPH</sub>. Quercetin-3-O-robinobioside, on the other hand, could be expected to display free radical activity as it fulfills several of the structural requirements of antioxidants shown for other quercetin-3-O-glycosides such as rutin, isoquercitrin and hyperoside [13]. Validation of the selected potential antioxidant markers will be required.

Bearing in mind the satisfactory results that were obtained with modeling using only informative variables of chromatographic fingerprints, additional models were constructed in order to evaluate the possibility of their application. Only TAC<sub>DPPH</sub> data were used for modeling because they offered a performance comparable to modeling of TAC<sub>ORAC</sub>.

Firstly, the prediction properties of PLS models that were constructed using samples produced during a specific production year were evaluated. The number of samples collected in 2009, 2010 and 2011 is equal to 39, 38 and 35, respectively, thereby giving 78, 79 and 70 samples (representing pasteurized and unpasteurized sub-samples in each case). PLS models were constructed using ca. 70% of the samples (from a given production year). The prediction properties of the models were evaluated using the remaining samples from the same production year (Table 3).

By analyzing the results obtained from the calibration models, it can be concluded that the TAC of the samples produced during one production year can be evaluated with an acceptable prediction error (at most 11.7% for a test set).

Next, the possibility of the routine application of the calibration model that was constructed for the samples produced during one production year to predict the TAC of the samples collected during another production year was examined (Table 4), considering that the chemical composition of plant material strongly depends on environmental factors that vary from year to year, e.g., weather conditions, light intensity, etc., as reviewed for *Camellia sinensis* by Tounekti et al. [44]. In such a case, all year-dependent sources of variance must be represented by samples that are included in

the calibration set in order to construct a calibration model with optimal properties.

In all three modeling scenarios, PLS models were well fitted and their prediction properties were slightly higher compared with the PLS models that were constructed for the samples collected during a particular production year only. It can be concluded that the prediction of the TAC can be performed for samples from the same production year with satisfactory results. Accurate prediction of the TAC of samples from a given production year based on the model constructed using samples produced during another production year is also possible.

#### 4. Conclusions

UVE-PLS modeling of chromatographic data and TAC of rooibos infusions indicated a number of potential antioxidant markers. All quantified compounds, except vitexin, were indicated as informative, while analysis of chromatographic fingerprints afforded additional antioxidant markers, namely (5)-eriodictyol-5-C-glucoside, (R)-eriodictyol-6-C-glucoside, aspalalinin and two unidentified compounds. Antioxidant markers selected in this manner should be validated by determining their antioxidant activity in the DPPH and ORAC assays. Such validation will entail purification of several compounds that are not available commercially and identification of unidentified compounds using mass spectrometry and nuclear magnetic resonance. After validation, the antioxidant markers will be invaluable for quality control of rooibos tea and nutraceutical and food ingredient extracts from rooibos tea. HPLC fingerprint data of rooibos infusions will not normally be used to predict TAC values due to the effort and expense involved. However, in cases where levels of specific compounds in rooibos infusions or products are needed for quality control, the available chromatographic data could be used to predict their TAC, as less chemical waste will be generated.

#### Acknowledgements

Rooibos samples were provided by Rooibos Ltd., Clanwilliam, South Africa. This work is based on the research supported in part by the South African Rooibos Council and the National Research Foundation of South Africa (grant 72065 to EJ). The Grantholder acknowledges that the opinions, findings and conclusions or recommendations expressed in any publication generated by NRF supported research are those of the author(s) and that the NRF accepts no liability whatsoever in this regard. A NRF-DST Scholarship to CJM is acknowledged. AS was awarded a post-graduate bursary by the NRF under grant 70525. MD wishes to express his gratitude to the Minister of Science and Higher Education of the Polish Republic for funding the scholarship for outstanding young scientists. JO appreciates the support of National Science Center grant no. 2011/03/N/ST4/00713, the UPGOW and DoktorIS fellowships.

#### References

- [1] S. Valtueña, N. Pellegrini, L. Franzini, M.A. Bianchi, D. Ardigò, D.D. Rio, P. Piatti, F. Scazzina, I. Zavaroni, F. Brighenti, Food selection based on total antioxidant capacity can modify antioxidant intake, systemic inflammation, and liver function without altering markers of oxidative stress, *Am. J. Clin. Nutr.* 87 (2008) 1290–1297.
- [2] B. Puchau, M.A. Zulet, A.C. de Echavari, H.H.M. Hermisdorff, J.A. Martinez, Dietary total antioxidant capacity is negatively associated with some metabolic syndrome features in healthy young adults, *Nutrition* 26 (2010) 534–541.
- [3] G. Alaerts, B. Dejaeger, J. Smeyers-Verbeke, Y. Vander Heyden, Recent developments in chromatographic fingerprints from herbal products: set-up and data analysis, *Comb. Chem. High Throughput Screen.* 13 (2010) 900–922.
- [4] J.R. Lucio-Gutiérrez, A. Garza-Juárez, J. Coello, S. Maspocho, M.L. Salazar-Cavazos, R. Salazar-Aranda, N. Waksman de Torres, Multi-wavelength high-performance liquid chromatographic fingerprints and chemometrics to predict the antioxidant activity of *Turnera diffusa* as part of its quality control, *J. Chromatogr. A* 1235 (2012) 68–76.
- [5] C. Tistaert, B. Dejaeger, G. Chataigné, C. Van Minh, J. Quetin-Leclercq, Y. Vander Heyden, Dissimilar chromatographic systems to indicate and identify antioxidants from *Mallotus* species, *Talanta* 83 (2011) 1198–1208.
- [6] M. Dumarey, I. Smets, Y. Vander Heyden, Prediction and interpretation of the antioxidant capacity of green tea from dissimilar chromatographic fingerprints, *J. Chromatogr. B: Anal. Technol. Biomed. Life Sci.* 878 (2010) 2733–2740.
- [7] E. Joubert, D. de Beer, Rooibos (*Aspalathus linearis*) beyond the farm gate: from herbal tea to potential phytopharmaceutical, *S. Afr. J. Bot.* 77 (2011) 869–886.
- [8] M.H. Carlsen, B.L. Halvorsen, K. Holte, S.K. Bøhn, S. Dragland, L. Sampson, C. Willey, H. Senoo, Y. Umezono, C. Sanada, I. Barikmo, N. Berhe, W.C. Willett, K.M. Phillips, D.R. Jacobs, R. Blomhoff, The total antioxidant content of more than 3100 foods, beverages, spices, herbs and supplements used worldwide, *Nutr. J.* 9 (2010) 3.
- [9] D. Villano, M. Pecorari, M.F. Testa, A. Raguzzini, A. Stalmach, A. Crozier, C. Tubili, M. Serafini, Unfermented and fermented rooibos teas (*Aspalathus linearis*) increase plasma total antioxidant capacity in healthy humans, *Food Chem.* 123 (2010) 679–683.
- [10] M.J. Simpson, D. Hjeltnes, C. López-Alarcón, N. Karamehmedovic, T.G. Minehan, A. Yepremyan, B. Salehani, E. Lissi, E. Joubert, K.J. Udekwu, E.I. Alarcón, Anti-peroxyl radical quality and antibacterial properties of rooibos infusions and their pure glycosylated polyphenolic constituents, *Molecules* 18 (2013) 11264–11280.
- [11] N. Krafczyk, F. Woyand, M.A. Glomb, Structure-antioxidant relationship of flavonoids from fermented rooibos, *Mol. Nutr. Food Res.* 53 (2009) 635–642.
- [12] A. von Gadow, E. Joubert, C.F. Hansmann, Comparison of the antioxidant activity of aspalathin with that of other plant phenols of rooibos tea (*Aspalathus linearis*),  $\alpha$ -tocopherol, BHT, and BHA, *J. Agric. Food Chem.* 45 (1997) 632–638.
- [13] P.W. Snijman, E. Joubert, D. Ferreira, X.-C. Li, Y. Ding, I.R. Green, W.C.A. Gelderblom, Antioxidant activity of the dihydrochalcones Aspalathin and notoifagin and their corresponding flavones in relation to other rooibos (*Aspalathus linearis*) flavonoids, epigallocatechin gallate, and Trolox, *J. Agric. Food Chem.* 57 (2009) 6678–6684.
- [14] E. Joubert, T. Beelders, D. de Beer, C.J. Malherbe, A.J. de Villiers, G.D. Sigge, Variation in phenolic content and antioxidant activity of fermented rooibos herbal tea infusions: role of production season and quality grade, *J. Agric. Food Chem.* 60 (2012) 9171–9179.
- [15] V. Centner, D.-L. Massart, O.E. de Noord, S. de Jong, B.M. Vandeginste, C. Sterna, Elimination of uninformative variables for multivariate calibration, *Anal. Chem.* 68 (1996) 3851–3858.
- [16] Y.V. Heyden, Extracting information from chromatographic herbal fingerprints, *LCGC Eur.* 21 (2008) 438–443.
- [17] M. Daszykowski, B. Walczak, Use and abuse of chemometrics in chromatography, *TrAC - Trends Anal. Chem.* 25 (2006) 1081–1096.
- [18] D.L. Massart, B.G.M. Vandeginste, L.M. Buydens, S. de Jong, P.J. Lewi, J. Smeyers-Verbeke, *Handbook of Chemometrics and Qualimetrics Volume A*, Elsevier, London, 1987.
- [19] M. Daszykowski, B. Walczak, Target selection for alignment of chromatographic signals obtained using monochannel detectors, *J. Chromatogr. A* 1176 (2007) 1–11.
- [20] M. Daszykowski, Y. Vander Heyden, C. Boucon, B. Walczak, Automated alignment of one-dimensional chromatographic fingerprints, *J. Chromatogr. A* 1217 (2010) 6127–6133.
- [21] N.-P.V. Nielsen, J.M. Carstensen, J. Smedsgaard, Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping, *J. Chromatogr. A* 805 (1998) 17–35.
- [22] P.H.C. Eilers, A perfect smoother, *Anal. Chem.* 75 (2003) 3631–3636.
- [23] M. Daszykowski, M. Sajewicz, J. Rzepa, M. Hajnos, D. Staszek, I. Wojtal, T. Kowalska, M. Waksmundzka-Hajnos, B. Walczak, Comparative analysis of the chromatographic fingerprints of twenty different sage species, *Acta Chromatogr.* 21 (2009) 513–530.
- [24] S. Wold, K. Esbensen, P. Geladi, Principal component analysis, *Chemom. Intell. Lab. Syst.* 2 (1987) 37–52.
- [25] A. Smilde, R. Bro, P. Geladi, *Multivariate Analysis with Applications in the Chemical Sciences*, John Wiley & Sons, Ltd., 2005.
- [26] T. Naes, T. Isaksson, T. Fearn, T. Davies, *A User-Friendly Guide to Multivariate Calibration and Classification*, NIR Publications, Chichester, 2002.
- [27] D. Huang, B. Ou, R.L. Prior, The chemistry behind antioxidant capacity assays, *J. Agric. Food Chem.* 53 (2005) 1841–1856.
- [28] A. Zulueta, M.J. Esteve, A. Frigola, ORAC and TEAC assays comparison to measure the antioxidant capacity of food products, *Food Chem.* 114 (2009) 310–316.
- [29] D. Martysiak-Zurowska, W. Went, A comparison of ABTS and DPPH methods for assessing the total antioxidant capacity of human milk, *Acta Sci. Pol. Technol. Aliment.* 11 (2012) 83–89.
- [30] K. Thaipong, U. Boonprakob, K. Crosby, L. Cisneros-Zevallos, D. Hawkins Byrne, Comparison of ABTS, DPPH, FRAP, and ORAC assays for estimating antioxidant activity from guava fruit extracts, *J. Food Compos. Anal.* 19 (2006) 669–675.
- [31] A. Floegel, D.-O. Kim, S.-J. Chung, S.I. Koo, O.K. Chun, Comparison of ABTS/DPPH assays to measure antioxidant capacity in popular antioxidant-rich US foods, *J. Food Compos. Anal.* 24 (2011) 1043–1048.
- [32] I. Stanimirova, M. Daszykowski, B. Walczak, Dealing with missing values and outliers in principal component analysis, *Talanta* 72 (2007) 172–178.
- [33] P.J. Rousseeuw, A.M. Leroy, *Robust Regression and Outlier Detection*, John Wiley & Sons, Inc., New York, NY, USA, 1987.

- [34] I.S. Koch, N. Müller, D. de Beer, T. Næs, E. Joubert, Impact of steam pasteurization on the sensory profile and phenolic composition of rooibos (*Aspalathus linearis*) herbal tea infusions. *Food Res. Int.* 53 (2013) 704–712.
- [35] I. Stanimirova, M. Kazura, D. de Beer, E. Joubert, A.E. Schulze, T. Beelders, A. de Villiers, B. Walczak, High-dimensional nested analysis of variance to assess the effect of production season, quality grade and steam pasteurization on the phenolic composition of fermented rooibos herbal tea. *Talanta* 115 (2013) 590–599.
- [36] R.D. Snee, Validation of regression models: methods and examples. *Technometrics* (1977) 415–428.
- [37] C.J. Malherbe, E. Willenburg, D. de Beer, S.L. Bonnet, J.H. van der Westhuizen, E. Joubert, Iriflophenone-3-C-glucoside from *Cyclopia genistoides*: isolation and quantitative comparison of antioxidant capacity with mangiferin and isomangiferin using on-line HPLC antioxidant assays. *J. Chromatogr. B: Anal. Technol. Biomed. Life Sci.* 951–952 (2014) 164–171.
- [38] E. Joubert, D. de Beer, Phenolic content and antioxidant activity of rooibos food ingredient extracts. *J. Food Compos. Anal.* 27 (2012) 45–51.
- [39] D. Huang, B. Ou, M. Hampsch-Woodill, J.A. Flanagan, R.L. Prior, High-throughput assay of oxygen radical absorbance capacity (ORAC) using a multichannel liquid handling system coupled with a microplate fluorescence reader in 96-well format. *J. Agric. Food Chem.* 50 (2002) 4437–4444.
- [40] C.D. Brown, R.L. Green, Critical factors limiting the interpretation of regression vectors in multivariate calibration. *Trends Anal. Chem.* 28 (2009) 506–514.
- [41] T. Beelders, G.O. Sigge, E. Joubert, D. de Beer, A. de Villiers, Kinetic optimisation of the reversed phase liquid chromatographic separation of rooibos tea (*Aspalathus linearis*) phenolics on conventional high performance liquid chromatographic instrumentation. *J. Chromatogr. A* 1219 (2012) 128–139.
- [42] H. Kikuzaki, M. Hisamoto, K. Hirose, K. Akiyama, H. Taniguchi, Antioxidant properties of ferulic acid and its related compounds. *J. Agric. Food Chem.* 50 (2002) 2161–2168.
- [43] E. Joubert, P. Winterton, T.J. Britz, W.C.A. Gelderblom, Antioxidant and pro-oxidant activities of aqueous extracts and crude polyphenolic fractions of rooibos (*Aspalathus linearis*). *J. Agric. Food Chem.* 53 (2005) 10260–10267.
- [44] T. Tounekti, E. Joubert, I. Hernández, S. Munné-Bosch, Improving the polyphenol content of tea. *Crit. Rev. Plant Sci.* 32 (2013) 192–215.

Katowice, 03.03.2015

dr hab. Michał Daszykowski, prof. UŚ  
Instytut Chemii  
Uniwersytet Śląski  
ul. Szkolna 9  
40-006 Katowice

Oświadczam, że w artykule pt. „*Modeling of the total antioxidant capacity of rooibos (Aspalathus linearis) tea infusions from chromatographic fingerprints and identification of potential antioxidant markers*”, opublikowanym w czasopiśmie *Journal of Chromatography A* 1366 (2014) 101-109, mój udział polegał na:

- współtworzeniu hipotezy badawczej i ogólnej koncepcji badań,
- pomocy w interpretacji wyników i weryfikacji wyciągniętych przez doktorantkę wniosków z przeprowadzonych badań,
- opiece i merytorycznym nadzorze podczas przygotowania manuskryptu zgodnie z wymaganiami czasopisma *Journal of Chromatography A*,
- pomocy w procedurze redakcyjnej (tj. pomocy w przygotowaniu odpowiedzi na pytania recenzentów),
- współpracy podczas ostatecznej korekty artykułu.

Michał Daszykowski

Katowice, 29.09.2014

mgr Małgorzata Kazura  
Instytut Chemii  
Uniwersytet Śląski  
ul. Szkolna 9  
40-006 Katowice

Oświadczam, że w artykule pt. „*Modeling of the total antioxidant capacity of rooibos (Aspalathus linearis) tea infusions from chromatographic fingerprints and identification of potential antioxidant markers*”, przyjętym do druku w czasopiśmie *Journal of Chromatography A*, mój udział polegał na:

- dokonaniu wstępnej organizacji danych,
- eliminacji linii bazowej w zbiorze chromatogramów z wykorzystaniem algorytmu PALS,
- korekcji przesunięć pików w zbiorze chromatogramów z wykorzystaniem algorytmu COW.

Małgorzata Kazura

Stellenbosch, South Africa, 25 September 25, 2014

Dalene de Beer

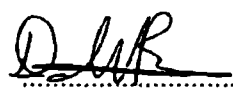
Post-Harvest and Wine Technology Division

Agricultural Research Council, Infruitec-Nietvoorbij

Stellenbosch, South Africa

Hereby, I declare that my contribution in scientific paper entitled: *Modeling of the total antioxidant capacity of rooibos (Aspalathus linearis) tea infusions from chromatographic fingerprints and identification of potential antioxidant markers*, accepted for publication in Journal of Chromatography A mostly concerned:

- Assistance in postulating research hypothesis
- Assistance in interpretation of results obtained from data exploration
- Providing comments to manuscript
- Providing answers to reviewers' questions regarding experiment



Author's signature



Stellenbosch, South Africa, 25 September 25, 2014

Elizabeth Joubert

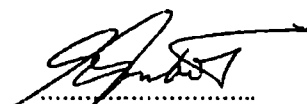
Post-Harvest and Wine Technology Division

Agricultural Research Council, Infruitec-Nietvoorbij

Stellenbosch, South Africa

Hereby, I declare that my contribution in scientific paper entitled: *Modeling of the total antioxidant capacity of rooibos (Aspalathus linearis) tea infusions from chromatographic fingerprints and identification of potential antioxidant markers*, accepted for publication in Journal of Chromatography A mostly concerned:

- Experimental planning
- Collecting samples
- Assistance in postulating research hypothesis
- Assistance in interpretation of results obtained from data exploration
- Providing comments to manuscript
- Providing answers to reviewers' questions regarding experiment



.....  
Author's signature

Stellenbosch, South Africa, 25 September 2014

Alexandra E. Schulze

Department of Food Science

Stellenbosch University

Stellenbosch, South Africa

Hereby, I declare that my contribution in scientific paper entitled: *Modeling of the total antioxidant capacity of rooibos (Aspalathus linearis) tea infusions from chromatographic fingerprints and identification of potential antioxidant markers*, accepted for publication in Journal of Chromatography A mostly concerned:

- Transferring of chromatographic signals from Agilent software to MatLab

  
.....  
Author's signature

Stellenbosch, South Africa, 25 September 2014

**Theresa Beelders**

**Department of Food Science**

**Stellenbosch University**

**Stellenbosch, South Africa**

Hereby, I declare that my contribution in scientific paper entitled: *Modeling of the total antioxidant capacity of rooibos (Aspalathus linearis) tea infusions from chromatographic fingerprints and identification of potential antioxidant markers*, accepted for publication in *Journal of Chromatography A* mostly concerned:

- Optimization of chromatographic conditions
- Samples preparation
- Chemical analysis of samples
- Analysis of chromatographic signals to obtain quantitative data



.....  
Author's signature

Stellenbosch, 25 September 2014

André de Villiers

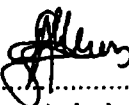
Department of Chemistry and Polymer Science

Stellenbosch University

Stellenbosch, South Africa

Hereby, I declare that my contribution in scientific paper entitled: *Modeling of the total antioxidant capacity of rooibos (Aspalathus linearis) tea infusions from chromatographic fingerprints and identification of potential antioxidant markers*, accepted for publication in Journal of Chromatography A mostly concerned:

- Optimization of chromatographic conditions

  
.....  
Author's signature

Stellenbosch, South Africa, 25 September 2014

Christiaan J. Malherbe

Post-Harvest and Wine Technology Division

Agricultural Research Council, Infruitec-Nietvoorbij

Stellenbosch, South Africa

Hereby, I declare that my contribution in scientific paper entitled: *Modeling of the total antioxidant capacity of rooibos (Aspalathus linearis) tea infusions from chromatographic fingerprints and identification of potential antioxidant markers*, accepted for publication in *Journal of Chromatography A* mostly concerned:

- Samples preparation
- Chemical analysis of samples



.....  
Author's signature

Katowice, 17.11.2014

prof. dr hab. Beata Walczak  
Instytut Chemii  
Uniwersytet Śląski  
ul. Szkolna 9  
40-006 Katowice

Oświadczam, że w artykule pt. *Modeling of the total antioxidant capacity of rooibos (Aspalathus linearis) tea infusions from chromatographic fingerprints and identification of potential antioxidant markers*, opublikowanym w czasopiśmie *Journal of Chromatography A*, 1366 (2014) 101-109 mój wkład polegał na przygotowaniu danych chromatograficznych do analizy oraz merytorycznej dyskusji hipotez badawczych i uzyskanych wyników.

Beata Walczak

## 8. Badania własne nieopublikowane

Antyoksydanty to związki chemiczne wykazujące zdolność do opóźnienia działania wolnych rodników (niezwykle aktywnych form chemicznych, powodujących degradację wielu materiałów w tym podstawowych budulców organizmów). Właściwość ta wyrażona dla próbki w sposób globalny – wykazywana przez wszystkie antyoksydanty w niej zawarte – nazywana jest całkowitą zdolnością antyoksydacyjną próbki (ang. *total antioxidant capacity*, TAC). Najpopularniejsze metody używane do oceny tego parametru to tzw. testy antyoksydacyjne. Są one oparte na reakcji chemicznej pomiędzy wolnymi rodnikami (generowanymi w sposób sztuczny i kontrolowany) i antyoksydantami zawartymi w próbkach. Reakcja jest kontrolowana z użyciem prostych technik instrumentalnych (spektrometrie UV-Vis i fluorescencyjna są najbardziej popularne), a zawartość antyoksydantów wyrażana jest w gramach wybranej substancji wzorcowej (najpopularniejszymi wzorcami są Trolox, witamina C i kwas galusowy). Stosowane są różne generatory wolnych rodników np. AAPH (2,2'-diazobis(2-amidinopropano) dihydrochlorok), ABTS (kwas 2,2'-azo-bis(3-etylobenzotiazolino-6-sulfonowy)) lub DPPH (2,2-diphenylo-1-pikrylohydrazyl) [122], dostępne są także metody oparte na redukcji jonów miedzi lub żelaza [123,124]. Różnorodność mechanizmów reakcji pomiędzy wolnymi rodnikami i antyoksydantami powoduje, że wartości TAC wyznaczone dla tego samego produktu, różnią się w zależności od użytego testu. Parametr może być zróżnicowany ze względu na pochodzenie geograficzne, sposób uprawy czy sposób przetwarzania. Niestety dostępne metody oceny TAC są czasochłonne i wymagają użycia licznych odczynników chemicznych przez co są trudne do zastosowania rutynowego, np. w badaniach typu on-line. Dlatego też potrzebne są nowe metody, które w sposób efektywny i automatyczny umożliwią ocenę zawartości antyoksydantów.

Nieselektywne sygnały analityczne takie jak widma rejestrowane w zakresie podczerwieni, fluorescencyjne obrazy czy chromatogramy umożliwiają rejestrację odpowiedzi detektora związaną z obecnością w próbkach substancji o właściwościach antyoksydacyjnych. Opracowałam metodykę analityczną opartą na nieselektywnych sygnałach analitycznych z pomocą której można dokonać w sposób efektywny oceny tego parametru. Szczegółowy opis proponowanego rozwiązania przedstawiam w **publikacjach IV i V**. Oprócz naparów kawy, naparów mięty pieprzowej i past pomidorowych badania przeprowadziłam również dla innych rodzajów ziół – bazylii i oregano, a także dla suplementów diety w postaci toników witaminowych. Poniżej przedstawiam uzyskane dla tych produktów wyniki. Nie zostały one opublikowane, nie mniej jednak są integralną częścią zrealizowanych badań i wnoszą istotną wiedzę na temat proponowanego rozwiązania.

Badania realizowałam w ramach projektu badawczego Preludium, finansowanego przez Narodowe Centrum Nauki pt. *Fluorescencyjne krajobrazy w połączeniu z metodami chemometrycznymi jako potencjalne narzędzie do wyznaczania całkowitej zdolności antyoksydacyjnej*, nr 2011/03/N/ST4/00713, którego byłam kierownikiem.

### ***Efektywność oceny całkowitej zdolności antyoksydacyjnej na podstawie fluorescencyjnych obrazów dla próbek bazylii, oregano oraz suplementów diety***

Próbki suszonych ziół: bazylii (11 rodzajów od różnych producentów) i oregano (12 rodzajów od różnych producentów), zakupiłam w lokalnych sklepach. Ekstrakcję antyoksydantów do fazy wodnej przeprowadziłam podobnie jak w przypadku mięty pieprzowej (szczegółowy opis w publikacji V). Dwa gramy suszu zalewałam 50 mL wody destylowanej o temperaturze 95 °C i poddawałam ekstrakcji przez 15 minut. Przefiltrowany ekstrakt przechowywałam w temperaturze -20 °C aż do dalszej analizy.

Kolejne próbki zbadane pod kątem zdolności antyoksydacyjnych to 8 rodzajów suplementów diety zróżnicowanych pod względem zaleceń stosowania. Ze względu na płynną postać nie wymagały one przeprowadzenia dodatkowej ekstrakcji antyoksydantów do fazy ciekłej.

Postępując zgodnie z opracowaną procedurą dla każdego ekstraktu (lub toniku witaminowego) wykonałam trzy replikaty laboratoryjne. Zarejestrowałam fluorescencyjne obrazy w zoptymalizowanych dla badanych próbek zakresach spektralnych (Tabela 1). Każdy replikat został oceniony pod względem zawartości antyoksydantów przy użyciu testu ORAC.

**Tabela 1** Zakresy spektralne w których rejestrowane były fluorescencyjne obrazy próbek

Rodzaj próbki	Wzbudzenie		Emisja	
	Zakres [nm]	$\Delta$ [nm]	Zakres [nm]	$\Delta$ [nm]
Ekstrakty bazylii	220 – 450	10	280 – 650	2
Ekstrakty oregano	220 – 450	10	280 – 650	2
Suplementy diety	230 – 550	10	300 – 650	2

Fluorescencyjne obrazy przygotowałam do analizy usuwając linie rozproszenia promieniowania elektromagnetycznego oraz rozwijając sygnały w postać wektorową (do konstrukcji modeli PLS). Dane podzieliłam na zbiory modelowe i testowe używając algorytmu Kennarda i Stonea [22]. Na podstawie uzyskanego zestawu danych skonstruowałam modele kalibracyjne, w których jako zmienną zależną używałam wartości parametru ORAC wyznaczonego dla każdej próbki.

Dla próbek ziół lepsze wyniki modelowania parametru ORAC uzyskałam używając do konstrukcji modeli metody PLS (w porównaniu z modelami N-PLS). Optymalna, ustalona przy użyciu metody walidacji krzyżowej kompleksowość modelu dla próbek oregano wynosi



pięć czynników i pozwala skonstruować model, w którym średni kwadratowy błąd dopasowania modelu wynosi 4,30% natomiast średni kwadratowy błąd przewidywania jest równy 9,60%. Cztery czynniki są wystarczające aby skonstruować model umożliwiający kalibrację parametru ORAC dla próbek bazylii o błędzie dopasowania równym 7,00% oraz błędzie przewidywania równym 18,50%. W przypadku modelowania zdolności antyoksydacyjnej suplementów diety uzyskane błędy dopasowania i przewidywania dla modelu PLS z siedmioma czynnikami (liczba czynników wybrana tą samą metodą walidacji) były wyższe niż 20%.

Parametr dopasowania charakteryzujący model skonstruowany dla próbek oregano jest na zadowalająco niskim poziomie, podobnie jak jego właściwości przewidywania. Nie mniej jednak, dwukrotnie wyższa wartość błędu przewidywania w stosunku do błędu dopasowania wskazuje na konieczność zwiększenia liczby próbek, a co za tym idzie możliwych źródeł wariacji, jeżeli rozpatrywany model miałby być stosowany rutynowo. Ze względu na ograniczoną liczbę dostępnych komercyjnie rodzajów oregano (różnych pod względem producenta) powiększenie liczby analizowanych próbek było niemożliwe.

Niestety, modele skonstruowane dla próbek bazylii i suplementów diety, mimo wyboru optymalnej liczby czynników, charakteryzują się stosunkowo wysokimi wartościami błędów przewidywania (powyżej 18%). Są one zbyt wysokie aby stwierdzić zadowalający efekt modelowania parametru ORAC. Przyczyną takich obserwacji może być zbyt duże zróżnicowanie próbek użytych w eksperymencie. Suszona bazylija charakteryzowała się skrajnymi różnicami pod względem zawartości części roślin w dostępnych komercyjnie produktach (od próbek składających się wyłącznie z suszonych liści do takich które zawierały większość łodyg). Niestety ze względu na ograniczoną dostępność próbek nie można było zawęzić eksperymentu do tych, które charakteryzowały się podobnym składem jakościowym (liczba próbek które stanowiłyby podstawę konstrukcji i oceny modelu byłaby nieakceptowanie niska). Podobnie, dostępne komercyjnie próbki toników witaminowych są wysoce zróżnicowane zarówno pod względem składu chemicznego (ilościowego i jakościowego) jak i właściwości fizycznych (różny kolor i stopień zmętnienia). Błędy powyżej 20%, które charakteryzują zbudowany model, są efektem tak dużych różnic w badanych próbkach.

Pomimo negatywnej oceny modeli dla oregano, bazylii i suplementów diety poczynione obserwacje nie dyskredytują proponowanej metodyki w kontekście badania tego typu próbek. Jeżeli rozpatrywane jest zastosowanie rutynowe można regulować jednorodność próbek, zwłaszcza jeżeli miałyby to być badania prowadzone w sposób ciągły w jednym zakładzie produkcyjnym.

## 9. Podsumowanie i wnioski

Wyniki zrealizowanych przeze mnie badań dowodzą, że nieselektywne sygnały analityczne są dobrym nośnikiem chemicznej informacji. Informacja ta może w efektywny sposób być wykorzystywana do oceny zróżnicowanych parametrów będących wyznacznikiem jakości wielu produktów, zwłaszcza jeżeli parametr opisujący jakość produktu jest zależny od zawartości różnych komponentów chemicznych.

Uzyskanie istotnej informacji z sygnałów nieselektywnych jest możliwe dzięki zastosowaniu odpowiednio dobranych do problemu analitycznego metod chemometrycznego przygotowania, analizy i interpretacji danych wieloparametrowych.

W niniejszej pracy przedstawiłam studium trzech problemów oceny jakości produktów, w których wykorzystałam nieselektywne sygnały analityczne i narzędzia chemometryczne.

### Problem I

Ocena jakości cukru ze względu na jego zabarwienie i zawartość popiołu.

W zrealizowanych badaniach wykorzystałam sygnały i informacje o parametrach jakości opisujące próbki uzyskane przez grupę prof. Rasmusa Bro z Uniwersytetu w Kopenhadze.

Opracowana metoda została oparta na pomiarach nieselektywnych fluorescencyjnych obrazów próbek. Stabilny model PRM umożliwił efektywną kalibrację wybranego parametru charakteryzującego jakość cukru na podstawie widm fluorescencyjnych z błędami przewidywania dla próbek testowych równymi 3,24% oraz 4,37% odpowiednio dla parametru opisującego barwę oraz zawartości popiołu.

### Problem II

Ocena jakości oleju napędowego o obniżonej aktywności ze względu na:

- a) zgodność poziomu stężenia dodatków aktywnych (Solvent Red 19 i Solvent Yellow 124) z wymaganiami stawianymi przez rozporządzenia Ministerstwa Finansów Rzeczypospolitej Polskiej,
- b) możliwość poddania próbki oleju procedurze nielegalnego usuwania dodatków aktywnych na drodze sorpcji.

Zaprojektowałam i przeprowadziłam modelowy eksperyment, w którym do oleju napędowego wprowadzałam dodatki aktywne w zróżnicowanym stężeniu. Następnie próbki poddałam symulacji nielegalnego procesu. Dla każdej z próbek (przed i po usunięciu komponentów) zbadałam właściwości fluorescencyjne w postaci fluorescencyjnych obrazów. Widma próbek barwionych użyłam do konstrukcji modeli kalibracyjnych (PLS), które charakteryzują się satysfakcjonującymi wartościami parametrów przewidywania i dopasowania. Ze względu na aspekt analityczny opracowanej metody wyznaczyłam jej podstawowe parametry walidacyjne.

Fluorescencyjne obrazy próbek odbarwionych użyłam do zbudowania modeli dyskryminacyjnych (D-PLS). Model, w którym kryterium dyskryminacyjnym jest zawartość markera – Solvent Yellow 124 umożliwiła poprawną w 100% dyskryminację próbek zawartych w niezależnym zbiorze testowym.

Proponowane rozwiązanie zostało zgłoszone do ochrony patentowej – zgłoszenie nr P.339194, data zgłoszenia 16.05.2012 r., tytuł zgłoszenia: *Sposób równoczesnego lub pojedynczego oznaczania zawartości znacznika Solvent Yellow 124 i barwnika Solvent Red 19 lub Solvent Red 164 lub Solvent Blue 35 w oleju napędowym.*

### **Problem III**

Ocena jakości produktów spożywczych, która została zdefiniowana jako ich całkowita zawartość antyoksydantów. Badania wykonałam dla następujących typów próbek:

- a) wodne ekstrakty kawy, mięty pieprzowej, bazylii i oregano, a także próbek suplementów diety w postaci toników multiwitaminowych,
- b) pasty pomidorowe w postaci wodnych ekstraktów oraz w nienaruszonej formie,
- c) napary herbaty typu rooibos.

Badania realizowane w ramach podpunktów a) oraz b) były przedmiotem projektu badawczego Preludium finansowanego przez Narodowe Centrum Nauki pt. *Fluorescencyjne krajobrazy w połączeniu z metodami chemometrycznymi jako potencjalne narzędzie do wyznaczania całkowitej zdolności antyoksydacyjnej*, nr 2011/03/N/ST4/00713.

Zaprojektowałam i przeprowadziłam eksperymenty, w których dla każdego z wybranych produktów spożywczych przygotowałam ekstrakt wodny. Następnie oznaczyłam zawartość związków antyoksydacyjnych, używając rekomendowanego do tego rodzaju produktów testu ORAC i całkowitą zawartość związków fenolowych używając odczynnika Folina-Ciocalteu. Dla wszystkich ekstraktów zarejestrowałam również fluorescencyjne obrazy, które następnie stanowiły dane użyte do konstrukcji modeli kalibracyjnych PLS i N-PLS.

Badania pilotażowe wykonałam dla wodnych ekstraktów kawy, mięty pieprzowej, bazylii i oregano. Satysfakcjonujące wartości parametrów przewidywania i dopasowania modeli kalibrujących parametr ORAC (a także zawartość polifenoli) uzyskałam modelując dane opisujące próbki kawy oraz mięty pieprzowej i oregano (wartości nie większe niż 10%). Modele skonstruowane dla próbek bazylii oraz suplementów diety miały niezadowalające wartości błędów przewidywania (wartości powyżej 18%). Postulowaną przyczyną uzyskanych wyników jest niski stopień homogeniczności dostępnych próbek. Opracowana metoda analityczna została poddana walidacji analitycznej i chemometrycznej zarówno dla próbek kawy jak i mięty pieprzowej.

Opracowaną metodykę analityczną porównałam z opisanym w literaturze podejściem również opartym na pomiarze nieselektywnych sygnałów analitycznych, którymi są widma rejestrowane w zakresie podczerwieni. Badania przeprowadziłam dla past pomidorowych. Zarejestrowałam widma w zakresie średniej podczerwieni oraz fluorescencyjne obrazy dla próbek past w nienaruszonej postaci (bezpośrednio po wyjęciu z opakowania). Było to

możliwe dzięki użyciu specjalistycznej fluorescencyjnej sondy światłowodowej, przystosowanej do badań próbek stałych. Oceeniłam całkowitą zawartość antyoksydantów i związków fenolowych wykorzystując te same metody co w badaniach przeprowadzonych dla próbek kawy i mięty. Użyłam metod chemometrycznych do konstrukcji modeli kalibracyjnych. Przeprowadzone studium porównania efektywności modelowania całkowitej zdolności antyoksydacyjnej w oparciu o sygnały absorpcyjne lub emisyjne wykazały, że fluorescencyjne obrazy (modelowane przy użyciu modelu N-PLS) umożliwiają uzyskanie efektywniejszej informacji niż sygnały IR (modelowane metodą PLS). W przypadku oceny całkowitej zawartości polifenoli lepsze parametry przewidywania uzyskałam dla modeli skonstruowanych używając widm zarejestrowanych w zakresie podczerwieni.

Badania przeprowadzone w ramach podpunktu c) realizowałam w oparciu o dane eksperymentalne otrzymane przez współpracowników z Uniwersytetu w Stellenbosch.

Wodne napary herbaty rooibos zostały scharakteryzowane przez: chromatograficzne odciski palca uzyskane przy użyciu wysokosprawnej chromatografii cieczowej, sprzężonej z detektorem DAD, ilościową informację o zawartości 12 związków polifenolowych oraz zawartość związków antyoksydacyjnych wyznaczoną niezależnymi testami ORAC i DPPH. Dane poddałam analizie chemometrycznej. Wykazałam, że nieselektywne chromatogramy umożliwiają ocenę całkowitej zdolności antyoksydacyjnej w sposób efektywny. Błędy przewidywania modeli PLS skonstruowanych dla chromatogramów w porównaniu z modelami zbudowanymi dla informacji selektywnych (informacja ilościowa) są niższe o około 2%. W badaniach wykorzystałam również wariant metody PLS umożliwiającą eliminację zmiennych nieistotnych (UVE-PLS), który umożliwił wskazanie tych polifenoli, które w sposób istotny wpływają na zawartość antyoksydantów.

Podsumowując:

- Wszystkie zaproponowane przeze mnie rozwiązania i podejścia analityczne wykorzystane do oceny jakości produktów opierają się na pomiarach nieselektywnych sygnałów instrumentalnych.
- Proponowane rozwiązania eliminują konieczność przeprowadzania kosztownych i czasochłonnych analiz dzięki zastosowaniu sygnałów nieselektywnych.
- Użycie stabilnego modelu kalibracyjnego PRM do kontroli jakości cukru potwierdziło przydatność stabilnych metod w modelowaniu problemów analityki procesowej.
- Wszystkie eksperymenty projektowane były tak, aby ograniczyć produkcję szkodliwych odpadów chemicznych – zgodnie z zasadami zielonej chemii. Dlatego też, jedynym używanym rozpuszczalnikiem była woda destylowana.
- Wysoka efektywność proponowanych rozwiązań sprawia, że mogą one być rekomendowane do zastosowań praktycznych w laboratoryjnych badaniach rutynowych oraz typu on-line, do kontroli jakości produktów. Taka adaptacja wymaga jednak przeprowadzenia dodatkowych badań, zwłaszcza z uwzględnieniem możliwych zmian składu chemicznego próbek.

- Zastosowanie światłowodowej sondy fluorescencyjnej przystosowanej do pomiarów ciał stałych usprawnia wykonanie pomiarów oraz potwierdza możliwość stosowania proponowanego rozwiązania analitycznego dla próbek bez wcześniejszego przygotowania do analizy. Przyczynia się to również do większej atrakcyjności metodyki w kontekście badań typu on-line.
- Wykorzystanie tej samej sondy z odpowiednią nakładką umożliwiającą pomiary substancji ciekłych, bez konieczności stosowania specjalnych kuwet pomiarowych, również zwiększa atrakcyjność opracowanych metodyk, zwłaszcza w kontekście możliwości prowadzenia badań poza wyposażonym w niezbędny sprzęt laboratorium.
- Rozwiązania analityczne o mniejszej efektywności (opracowane dla próbek bazylii, oregano i suplementów diety), zgodnie z definicją badań podstawowych [125] wnoszą nową wiedzę na temat wykorzystania nieselektywnych sygnałów do oceny jakości produktów. Stanowią dobry punkt wyjściowy do kontynuacji lub planowania nowych badań z zakresu oceny jakości żywności wyrażonej jako ich całkowita zdolność antyoksydacyjna.

## 10. Finansowanie

Badania będące przedmiotem niniejszej pracy doktorskiej zostały zrealizowane dzięki wsparciu finansowemu pochodzącemu z takich źródeł jak:

- **Narodowe Centrum Nauki**



Projekt badawczy PRELUDIUM pt. *Fluorescencyjne krajobrazy w połączeniu z metodami chemometrycznymi jako potencjalne narzędzie do wyznaczania całkowitej zdolności antyoksydacyjnej*, NCN nr 2011/03/N/ST4/00713

- **Unijne programy stypendialne realizowane w ramach Europejskiego Funduszu Społecznego**



- **Dotacji dla młodych naukowców finansowanej przez Ministerstwo Nauki i Szkolnictwa Wyższego**

## 11. Literatura

- [1] Ch 13 - <sup>1</sup>H NMR, <http://www.chem.ucalgary.ca/courses/351/Carey5th/Ch13/ch13-hnmr.html> (ostatni dostęp 24 września 2014).
- [2] Amino acids fluorescence data | Plant Food Science group & Spectroscopy and Chemometrics group, [http://www.models.life.ku.dk/amino\\_acid\\_fluo](http://www.models.life.ku.dk/amino_acid_fluo) (ostatni dostęp 24 września 2014).
- [3] E. Sikorska, A. Gliszczyńska-Świgło, I. Khmelinskii, M. Sikorski, Synchronous fluorescence spectroscopy of edible vegetable oils. Quantification of tocopherols, *J. Agric. Food Chem.* 53 (2005) 6988–6994.
- [4] L. Xu, D.-H. Deng, C.-B. Cai, H.-W. Yang, Automatic discrimination of the geographical origins of milks by excitation-emission fluorescence spectrometry and chemometrics, *J. Autom. Methods Manag. Chem.* 2011 (2011) 1–6.
- [5] O. Divya, A.K. Mishra, Multivariate methods on the excitation emission matrix fluorescence spectroscopic data of diesel-kerosene mixtures: A comparative study, *Anal. Chim. Acta.* 592 (2007) 82–90.
- [6] R. Tauler, B. Walczak, S. Brown, *Comprehensive Chemometrics: Chemical and Biochemical Data Analysis*, Elsevier, Londyn, 2009.
- [7] D. Skoog, D. West, J. Holler, S. Crouch, *Podstawy chemii analitycznej*, PWN, Warszawa, 2000.
- [8] S. Wold, K. Esbensen, P. Geladi, Principal component analysis, *Chemom. Intell. Lab. Syst.* 2 (1987) 37–52.
- [9] B. Walczak, *Wavelets in chemistry*, Elsevier, Londyn, 2000.
- [10] P.H.C. Eilers, A perfect smoother, *Anal. Chem.* 75 (2003) 3631–3636.
- [11] A. de Juan, R. Tauler, Multivariate curve resolution (MCR) from 2000: progress in concepts and applications, *Crit. Rev. Anal. Chem.* 36 (2006) 163–176.
- [12] R. Bro, PARAFAC. Tutorial and applications, *Chemom. Intell. Lab. Syst.* 38 (1997) 149–171.
- [13] J.R. Lakowicz, *Principles of fluorescence spectroscopy*, Springer, Nowy York, 2006.
- [14] L.G. Thygesen, Å. Rinnan, S. Barsberg, J.K.S. Møller, Stabilizing the PARAFAC decomposition of fluorescence spectra by insertion of zeros outside the data area, *Chemom. Intell. Lab. Syst.* 71 (2004) 97–106.
- [15] G. Tomasi, R. Bro, PARAFAC and missing values, *Chemom. Intell. Lab. Syst.* 75 (2005) 163–180.
- [16] R.G. Zepp, W.M. Sheldon, M.A. Moran, Dissolved organic fluorophores in southeastern US coastal waters: Correction method for eliminating Rayleigh and Raman scattering peaks in excitation-emission matrices, *Mar. Chem.* 89 (2004) 15–36.
- [17] X. Lu, C.F. Ross, J.R. Powers, D.E. Aston, B.A. Rasco, Determination of total phenolic content and antioxidant activity of garlic (*Allium sativum*) and elephant garlic (*Allium ampeloprasum*) by attenuated total reflectance–Fourier transformed infrared spectroscopy, *J. Agric. Food Chem.* 59 (2011) 5215–5221.
- [18] N.P.V. Nielsen, J.M. Carstensen, J. Smedsgaard, Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping, *J. Chromatogr. A.* 805 (1998) 17–35.
- [19] D.L. Massart, B.G.M. Vandeginste, L.M. Buydens, S. de Jong, P.J. Lewi, J. Smeyers-Verbeke, *Handbook of chemometrics and qualimetrics volume A*, Elsevier, London, 1987.
- [20] R. Bro, Multiway calibration. Multilinear PLS, *J. Chemom.* 10 (1996) 47–61.
- [21] T. Naes, T. Isaksson, T. Fearn, T. Davies, *A user friendly guide to multivariate calibration and classification*, NIR Publications, Chichester, 2002.
- [22] R.W. Kennard, L.A. Stone, Computer aided design of experiments, *Technometrics.* 11 (1969) 137–148.
- [23] R.D. Snee, Validation of regression models: methods and examples, *Technometrics.* 19 (1977) 415–428.

- [24] Q.-S. Xu, Y.-Z. Liang, Monte Carlo cross validation, *Chemom. Intell. Lab. Syst.* 56 (2001) 1–11.
- [25] S. Semeels, C. Croux, P. Filzmoser, P.J. Van Espen, Partial robust M-regression, *Chemom. Intell. Lab. Syst.* 79 (2005) 55–64.
- [26] H. Fritz, P. Filzmoser, C. Croux, A comparison of algorithms for the multivariate L 1-median, *Comput. Stat.* 27 (2012) 393–410.
- [27] O. Hossjer, C. Croux, Generalizing univariate signed rank statistics for testing and estimating a multivariate location parameter, *J. Nonparametric Stat.* 4 (1995) 293–308.
- [28] A.C. Olivieri, G.M. Escandar, A.M. de la Peña, Second-order and higher-order multivariate calibration methods applied to non-multilinear data using different algorithms, *TrAC Trends Anal. Chem.* 30 (2011) 607–617.
- [29] V. Centner, D.L. Massart, O.E. de Noord, S. de Jong, B.M. Vandeginste, C. Sterna, Elimination of Uninformative Variables for Multivariate Calibration, *Anal. Chem.* 68 (1996) 3851–3858.
- [30] A. Smolinska, A.C. Hauschild, R.R.R. Fijten, J.W. Dallinga, J. Baumbach, F.J. van Schooten, Current breathomics a review on data pre-processing techniques and machine learning in metabolomics breath analysis, *J. Breath Res.* 8 (2014) 027105.
- [31] M. Daszykowski, J. Orzel, M.S. Wrobel, H. Czarnik-Matusewicz, B. Walczak, Improvement of classification using robust soft classification rules for near-infrared reflectance spectral data, *Chemom. Intell. Lab. Syst.* 109 (2011) 86–93.
- [32] R.A. Fisher, The use of multiple measurements in taxonomic problems, *Ann. Eugen.* 7 (1936) 179–188.
- [33] M. Barker, W. Rayens, Partial least squares for discrimination, *J. Chemom.* 17 (2003) 166–173.
- [34] L. Breiman, J. Friedman, C.J. Stone, R.A. Olshen, *Classification and regression trees*, Taylor & Francis, 1984.
- [35] S. Wold, M. Sjostrom, *SIMCA: A method for analyzing chemical data in terms of similarity and analogy*, w: *Chemometrics: Theory and Application*, American Chemical Society, Waszyngton, 1977.
- [36] Dynamic Time Warping and Correlation Optimized Warping | Plant Food Science group & Spectroscopy and Chemometrics group, [http://www.models.life.ku.dk/dtw\\_cow](http://www.models.life.ku.dk/dtw_cow) (ostatni dostep 27 listopada 2014).
- [37] M. Daszykowski, S. Semeels, K. Kaczmarek, P. Van Espen, C. Croux, B. Walczak, TOMCAT: A MATLAB toolbox for multivariate calibration techniques, *Chemom. Intell. Lab. Syst.* 85 (2007) 269–277.
- [38] The N-way toolbox for MATLAB | Plant Food Science group & Spectroscopy and Chemometrics group, <http://www.models.kvl.dk/nwaytoolbox> (ostatni dostep 12 marca 2012).
- [39] W. Cheung, I.T. Shadi, Y. Xu, R. Goodacre, Quantitative analysis of the banned food dye sudan-1 using surface enhanced raman scattering with multivariate chemometrics, *J. Phys. Chem. C.* 114 (2010) 7285–7290.
- [40] Y. Ni, Y. Wang, S. Kokot, Simultaneous kinetic spectrophotometric analysis of five synthetic food colorants with the aid of chemometrics, *Talanta.* 78 (2009) 432–441.
- [41] M. Daszykowski, Y. Vander Heyden, B. Walczak, Robust partial least squares model for prediction of green tea antioxidant capacity from chromatograms, *J. Chromatogr. A.* 1176 (2007) 12–18.
- [42] X. Lu, B.A. Rasco, Determination of antioxidant content and antioxidant activity in foods using infrared spectroscopy and chemometrics: a review, *Crit. Rev. Food Sci. Nutr.* 52 (2012) 853–875.
- [43] X. Lu, J. Wang, H.M. Al-Qadiri, C.F. Ross, J.R. Powers, J. Tang, i in., Determination of total phenolic content and antioxidant capacity of onion (*Allium cepa*) and shallot (*Allium oschaninii*) using infrared spectroscopy, *Food Chem.* 129 (2011) 637–644.



- [44] I. Ścibisz, M. Reich, S. Bureau, B. Gouble, M. Causse, D. Bertrand, i in., Mid-infrared spectroscopy as a tool for rapid determination of internal quality parameters in tomato, *Food Chem.* 125 (2011) 1390–1397.
- [45] P.A. da Costa Filho, Developing a rapid and sensitive method for determination of trans-fatty acids in edible oils using middle-infrared spectroscopy, *Food Chem.* 158 (2014) 1–7.
- [46] F.M. Verbi Pereira, A.P. Rebellato, J.A. Lima Pallone, L.A. Colnago, Through-package fat determination in commercial samples of mayonnaise and salad dressing using time-domain nuclear magnetic resonance spectroscopy and chemometrics, *Food Control.* 48 (2015) 62–66.
- [47] E. Sikorska, A. Gliszczyńska-Świgło, M. Insińska-Rak, I. Khmelinskii, D. De Keukeleire, M. Sikorski, Simultaneous analysis of riboflavin and aromatic amino acids in beer using fluorescence and multivariate calibration methods, *Anal. Chim. Acta.* 613 (2008) 207–217.
- [48] R. Karoui, J. De Baerdemaeker, A review of the analytical methods coupled with chemometric tools for the determination of the quality and identity of dairy products, *Food Chem.* 102 (2007) 621–640.
- [49] J.Y. Shim, I.K. Cho, H.K. Khurana, Q.X. Li, S. Jun, Attenuated total reflectance - Fourier transform infrared spectroscopy coupled with multivariate analysis for measurement of acesulfame-K in diet foods, *J. Food Sci.* 73 (2008) C426–C431.
- [50] Y.B. Monakhova, S.P. Mushtakova, S.S. Kolesnikova, S.A. Astakhov, Chemometrics-assisted spectrophotometric method for simultaneous determination of vitamins in complex mixtures, *Anal. Bioanal. Chem.* 397 (2010) 1297–1306.
- [51] D.M. Luykx, S.M. van Ruth, An overview of analytical methods for determining the geographical origin of food products, *Food Chem.* 107 (2008) 897–911.
- [52] Rozporządzenie Rady W.E. nr 509/2006 z dnia 20 marca 2006 r. w sprawie produktów rolnych i środków spożywczych będących gwarantowanymi tradycyjnymi specjalnościami.
- [53] Rozporządzenie Rady W.E. nr 510/2006 z dnia 20 marca 2006 r. w sprawie ochrony oznaczeń geograficznych i nazw pochodzenia produktów rolnych i środków spożywczych.
- [54] Ł. Zieliński, S. Deja, I. Jasicka-Misiak, P. Kafarski, Chemometrics as a tool of Origin determination of polish monofloral and multifloral honeys, *J. Agric. Food Chem.* 62 (2014) 2973–2981.
- [55] I. Stanimirova, B. Üstün, T. Cajka, K. Riddelova, J. Hajslova, L.M. Buydens, i in., Tracing the geographical origin of honeys based on volatile compounds profiles assessment using pattern recognition techniques, *Food Chem.* 118 (2010) 171–176.
- [56] P. Oliveri, G. Downey, Multivariate class modeling for the verification of food-authenticity claims, *TrAC - Trends Anal. Chem.* 35 (2012) 74–86.
- [57] E. Kurniawati, A. Rohman, K. Triyana, Analysis of lard in meatball broth using Fourier transform infrared spectroscopy and chemometrics, *Meat Sci.* 96 (2014) 94–98.
- [58] İ.H. Boyacı, H.T. Temiz, R.S. Uysal, H.M. Velioglu, R.J. Yadegari, M.M. Rishkan, A novel method for discrimination of beef and horsemeat using Raman spectroscopy, *Food Chem.* 148 (2014) 37–41.
- [59] N.A. Fadzillillah, A. Rohman, A. Ismail, S. Mustafa, A. Khatib, Application of FTIR-ATR spectroscopy coupled with multivariate analysis for rapid estimation of butter adulteration, *J. Oleo Sci.* 62 (2013) 555–562.
- [60] R.S. Uysal, İ.H. Boyacı, H.E. Genis, U. Tamer, Determination of butter adulteration with margarine using Raman spectroscopy, *Food Chem.* 141 (2013) 4397–4403.
- [61] P. Jaiswal, S.N. Jha, A. Borah, A. Gautam, M.K. Grewal, G. Jindal, Detection and quantification of soymilk in cow-buffalo milk using Attenuated Total Reflectance Fourier Transform Infrared spectroscopy (ATR-FTIR), *Food Chem.* 168 (2015) 41–47.
- [62] C.E.T. da Silva, V.L. Filardi, I.M. Pepe, M.A. Chaves, C.M.S. Santos, Classification of food vegetable oils by fluorimetry and artificial neural networks, *Food Control.* 47 (2015) 86–91.

- [63] F. Guimet, J. Ferré, R. Boqué, Rapid detection of olive-pomace oil adulteration in extra virgin olive oils from the protected denomination of origin “Siurana” using excitation-emission fluorescence spectroscopy and three-way methods of analysis, *Anal. Chim. Acta.* 544 (2005) 143–152.
- [64] S.A. Ordoudi, M. De Los Mozos Pascual, M.Z. Tsimidou, On the quality control of traded saffron by means of transmission Fourier-transform mid-infrared (FT-MIR) spectroscopy and chemometrics, *Food Chem.* 150 (2014) 414–421.
- [65] N. Reis, A.S. Franca, L.S. Oliveira, Quantitative evaluation of multiple adulterants in roasted coffee by Diffuse Reflectance Infrared Fourier Transform Spectroscopy (DRIFTS) and chemometrics, *Talanta.* 115 (2013) 563–568.
- [66] L.S. Magwaza, S. Landahl, P.J.R. Cronje, H.H. Nieuwoudt, A.M. Mouazen, B.M. Nicolaï, i in., The use of Vis/NIRS and chemometric analysis to predict fruit defects and postharvest behaviour of “Nules Clementine” mandarin fruit, *Food Chem.* 163 (2014) 267–274.
- [67] V. Giovenzana, R. Beghi, S. Buratti, R. Civelli, R. Guidetti, Monitoring of fresh-cut *Valerianella locusta* Laterr. shelf life by electronic nose and VIS-NIR spectroscopy, *Talanta.* 120 (2014) 368–375.
- [68] E.Z. Panagou, O. Papadopoulou, J.M. Carstensen, G.J.E. Nychas, Potential of multispectral imaging technology for rapid and non-destructive determination of the microbiological quality of beef filets during aerobic storage, *Int. J. Food Microbiol.* 174 (2014) 1–11.
- [69] H. Vasconcelos, C. Saraiva, J.M.M. de Almeida, Evaluation of the spoilage of raw chicken breast fillets using Fourier transform infrared spectroscopy in tandem with chemometrics, *Food Bioprocess Technol.* 7 (2014) 2330–2341.
- [70] J.H. Christensen, G. Tomasi, J. Strand, O. Andersen, PARAFAC modeling of fluorescence excitation - emission spectra of fish bile for rapid en route screening of PAC exposure, *Environ. Sci. Technol.* 43 (2009) 4439–4445.
- [71] J. Christensen, E.M. Becker, C.S. Frederiksen, Fluorescence spectroscopy and PARAFAC in the analysis of yogurt, *Chemom. Intell. Lab. Syst.* 75 (2005) 201–208.
- [72] M. Ghasemi-Varnamkhasti, M. Forina, NIR spectroscopy coupled with multivariate computational tools for qualitative characterization of the aging of beer, *Comput. Electron. Agric.* 100 (2014) 34–40.
- [73] A. Aït Kaddour, B. Cuq, Dynamic NIR spectroscopy to monitor bread dough mixing: A short review, *Am. J. Food Technol.* 6 (2011) 173–185.
- [74] C.C. Fagan, C. Everard, C.P. O’Donnell, G. Downey, E.M. Sheehan, C.M. Delahunty, i in., Prediction of processed cheese instrumental texture and meltability by mid-infrared spectroscopy coupled with chemometric tools, *J. Food Eng.* 80 (2007) 1068–1077.
- [75] I. Stanimirova, C. Boucon, B. Walczak, Relating gas chromatographic profiles to sensory measurements describing the end products of the Maillard reaction, *Talanta.* 83 (2011) 1239–1246.
- [76] S.A. Bortolato, J.A. Arancibia, G.M. Escandar, Chemometrics-assisted excitation-emission fluorescence spectroscopy on nylon membranes. Simultaneous determination of benzo[a]pyrene and dibenz[a,h]anthracene at parts-per-trillion levels in the presence of the remaining EPA PAH priority pollutants as interferences, *Anal. Chem.* 80 (2008) 8276–8286.
- [77] M. Bravo, L.F. Aguilar, W. Quiroz, A.C. Olivieri, G.M. Escandar, Determination of tributyltin at parts-per-trillion levels in natural waters by second-order multivariate calibration and fluorescence spectroscopy, *Microchem. J.* 106 (2013) 95–101.
- [78] A.G. García-Reiriz, Parallel factor analysis and multivariate curve resolution as data fusion tools to supervise a stream, *Chemom. Intell. Lab. Syst.* 137 (2014) 120–127.
- [79] K.E. Clow, G.J. Hall, H. Chen, J.E. Kenny, Spectral fingerprinting and classification by location of origin of natural waters by multidimensional fluorescence, w: *Proceedings of SPIE - The International Society for Optical Engineering* 2004

- [80] L. Cécillon, G. Certini, H. Lange, C. Forte, L.T. Strand, Spectral fingerprinting of soil organic matter composition, *Org. Geochem.* 46 (2012) 127–136.
- [81] R. Linker, A. Kenny, A. Shaviv, L. Singher, I. Shmulevich, Fourier transform infrared-attenuated total reflection nitrate determination of soil pastes using principal component regression, partial least squares, and cross-correlation, *Appl. Spectrosc.* 58 (2004) 516–520.
- [82] G.N. Elliott, H. Worgan, D. Broadhurst, J. Draper, J. Scullion, Soil differentiation using fingerprint Fourier transform infrared spectroscopy, chemometrics and genetic algorithm-based feature selection, *Soil Biol. Biochem.* 39 (2007) 2888–2896.
- [83] L. Cristescu, G. Pavelescu, E.M. Cârstea, D. Săvăstru, Evaluation of petroleum contaminants in soil by fluorescence spectroscopy, *Environ. Eng. Manag. J.* 8 (2009) 1269–1273.
- [84] B.R. Stallard, M.J. Garcia, S. Kaushik, Near-IR reflectance spectroscopy for the determination of motor oil contamination in sandy loam, *Appl. Spectrosc.* 50 (1996) 334–338.
- [85] X.D. Qing, H.L. Wu, C.C. Nie, X.F. Yan, Y.N. Li, J.Y. Wang, i in., Simultaneous determination of plant growth regulators in environmental samples using chemometrics-assisted excitation-emission matrix fluorescence: Experimental study on the prediction quality of second-order calibration method, *Talanta.* 103 (2013) 86–94.
- [86] X.D. Qing, H.L. Wu, Y.N. Li, C.C. Nie, J.Y. Wang, S.H. Zhu, i in., Simultaneous determination of pre-emergence herbicides in environmental samples using HPLC-DAD combined with second-order calibration based on self-weighted alternating trilinear decomposition algorithm, *Anal. Methods.* 4 (2012) 685–692.
- [87] R.B. Horton, M. McConico, C. Landry, T. Tran, F. Vogt, Introducing nonlinear, multivariate “Predictor Surfaces” for quantitative modeling of chemical systems with higher-order, coupled predictor variables, *Anal. Chim. Acta.* 746 (2012) 1–14.
- [88] H. Wagner, S. Dunker, Z. Liu, C. Wilhelm, Subcommunity FTIR-spectroscopy to determine physiological cell states, *Curr. Opin. Biotechnol.* 24 (2013) 88–94.
- [89] X.U. Zhanfeng, C.E. Bunker, P.B. De Harrington, Classification of jet fuel properties by near-infrared spectroscopy using fuzzy rule-building expert systems and support vector machines, *Appl. Spectrosc.* 64 (2010) 1251–1258.
- [90] J.C.L. Alves, R.J. Poppi, Simultaneous determination of hydrocarbon renewable diesel, biodiesel and petroleum diesel contents in diesel fuel blends using near infrared (NIR) spectroscopy and chemometrics, *Analyst.* 138 (2013) 6477–6487.
- [91] M.T. Bona, J.M. Andrés, Application of chemometric tools for coal classification and multivariate calibration by transmission and drift mid-infrared spectroscopy, *Anal. Chim. Acta.* 624 (2008) 68–78.
- [92] H. Omrani, J.A. Barnes, A.E. Dudelzak, H.-P. Looek, H. Waechter, Fluorescence excitation-emission matrix (EEM) spectroscopy and cavity ring-down (CRD) absorption spectroscopy of oil-contaminated jet fuel using fiber-optic probes, *Analyst.* 137 (2012) 2782–2790.
- [93] J. Silberring, A. Drabik, A. Kraj, *Proteomika i metabolomika*, Wydawnictwa Uniwersytetu Warszawskiego, Warszawa, 2010.
- [94] M. Daszykowski, A. Bierzynska-Krzysik, J. Silberring, B. Walczak, Avoiding spots detection in analysis of electrophoretic gel images, *Chemom. Intell. Lab. Syst.* 104 (2010) 2–7.
- [95] P. Suder, A. Bodzon-Kulakowska, P. Mak, A. Bierzynska-Krzysik, M. Daszykowski, B. Walczak, i in., The proteomic analysis of primary cortical astrocyte cell culture after morphine administration, *J. Proteome Res.* 8 (2009) 4633–4640.
- [96] J.O. Lay Jr., R. Liyanage, S. Borgmann, C.L. Wilkins, Problems with the “omics,” *TrAC - Trends Anal. Chem.* 25 (2006) 1046–1056.
- [97] E. Ibáñez, A. Cifuentes, Foodomics: Food science and nutrition in the postgenomic era, *Comprehensive Analytical Chemistry*, 64 (2014) 395–440.
- [98] A. Valdés, C. Simó, C. Ibáñez, V. García-Cañas, Foodomics strategies for the analysis of transgenic foods, *TrAC Trends Anal. Chem.* 52 (2013) 2–15.

- [99] M. Herrero, C. Simõ, V. García-Cañas, E. Ibáñez, A. Cifuentes, Foodomics: MS-based strategies in modern food science and nutrition, *Mass Spectrom. Rev.* 31 (2012) 49–69.
- [100] M. Streefland, D.E. Martens, E.C. Beuvery, R.H. Wijffels, Process analytical technology (PAT) tools for the cultivation step in biopharmaceutical production, *Eng. Life Sci.* 13 (2013) 212–223.
- [101] E. Van Gyseghem, S. Van Hemelryck, M. Daszykowski, F. Questier, D.L. Massart, Y. Vander Heyden, Determining orthogonal chromatographic systems prior to the development of methods to characterise impurities in drug substances, *J. Chromatogr. A.* 988 (2003) 77–93.
- [102] J.M.M. Leitão, J.C.G. Esteves Da Silva, A.J. Girón, A. Muñoz De La Peña, Optimization of verapamil drug analysis by excitation-emission fluorescence in combination with second-order multivariate calibration, *J. Fluoresc.* 18 (2008) 1065–1076.
- [103] P. Datta, D. Bhattacharyya, Analysis of fluorescence excitation-emission matrix of multicomponent drugs: A case study with human placental extract used as wound healer, *J. Pharm. Biomed. Anal.* 36 (2004) 211–218.
- [104] Y. Roggo, P. Chalus, L. Maurer, C. Lema-Martinez, A. Edmond, N. Jent, A review of near infrared spectroscopy and chemometrics in pharmaceutical technologies, *J. Pharm. Biomed. Anal.* 44 (2007) 683–700.
- [105] I. Tomuta, L. Rus, R. Iovanov, L.L. Rus, High-throughput NIR-chemometric methods for determination of drug content and pharmaceutical properties of indapamide tablets, *J. Pharm. Biomed. Anal.* 84 (2013) 285–292.
- [106] O. Scheibelhofer, N. Balak, D.M. Koller, J.G. Khinast, Spatially resolved monitoring of powder mixing processes via multiple NIR-probes, *Powder Technol.* 243 (2013) 161–170.
- [107] D. Mainali, J. Li, P. Yehl, N. Chetwyn, Development of a comprehensive near infrared spectroscopy calibration model for rapid measurements of moisture content in multiple pharmaceutical products, *J. Pharm. Biomed. Anal.* 95 (2014) 169–175.
- [108] M.C. Hennigan, A.G. Ryder, Quantitative polymorph contaminant analysis in tablets using Raman and near infra-red spectroscopies, *J. Pharm. Biomed. Anal.* 72 (2013) 163–171.
- [109] M. Khanmohammadi, A.B. Garmarudi, N. Moazzen, K. Ghasemi, Qualitative discrimination between paracetamol tablets made by near infrared spectroscopy and chemometrics with regard to polymorphism, *J. Struct. Chem.* 51 (2010) 663–669.
- [110] Y. Ni, W. Lin, S. Kokot, Classification of *Rhizoma et Radix Notoperygii* by HPLC fingerprints with the aid of chemometrics, *Anal. Lett.* 45 (2012) 1810–1823.
- [111] M. Dumarey, A.M. van Nederkassel, I. Stanimirova, M. Daszykowski, F. Bensaid, M. Lees, *in.*, Recognizing paracetamol formulations with the same synthesis pathway based on their trace-enriched chromatographic impurity profiles, *Anal. Chim. Acta.* 655 (2009) 43–51.
- [112] X. Yin, H. Wu, X. Zhang, H. Gu, R. Yu, Simultaneous determination of irinotecan and its metabolite 7-ethyl-10-hydroxycamptothecin in biological fluids using excitation-emission matrix fluorescence coupled with second-order calibration method, *Acta Chim. Sin.* 71 (2013) 560–566.
- [113] Y.J. Liu, H.L. Wu, J.F. Nie, J.Y. Wang, L.Q. Ouyang, R.Q. Yu, Simultaneous determination of dextromethorphan and its metabolite dextrorphan in plasma samples using second-order calibration coupled with excitation-emission matrix fluorescence, *Anal. Sci.* 27 (2011) 663–666.
- [114] C.D. Bernardes, R.J. Poppi, M.M. Sena, Direct determination of trans-resveratrol in human plasma by spectrofluorimetry and second-order standard addition, *Talanta.* 82 (2010) 640–645.
- [115] S.F. Li, H.L. Wu, L. Huang, Y.N. Li, J.F. Nie, S.R. Zhang, *in.*, Quantitative analysis of fluphenazine hydrochloride in human urine using excitation-emission matrix fluorescence based on oxidation derivatization and combined with second-order calibration methods, *Anal. Methods.* 2 (2010) 1069–1077.

- [116] P.C. Damiani, Determination of atenolol in human urine by emission-excitation fluorescence matrices and unfolded partial least-squares with residual bilinearization, *Talanta*. 85 (2011) 1526–1534.
- [117] T. Madrakian, A. Afkhami, M. Mohammadnejad, Second-order advantage applied to simultaneous spectrofluorimetric determination of paracetamol and mefenamic acid in urine samples, *Anal. Chim. Acta*. 645 (2009) 25–29.
- [118] J. Nie, H. Wu, X. Wang, Y. Zhang, S. Zhu, R. Yu, Determination of testosterone propionate in cosmetics using excitation-emission matrix fluorescence based on oxidation derivatization with the aid of second-order calibration methods, *Anal. Chim. Acta*. 628 (2008) 24–32.
- [119] R. George, M. Michaelides, M.A. Brewer, U. Utzinger, Parallel factor analysis of ovarian autofluorescence as a cancer diagnostic, *Lasers Surg. Med.* 44 (2012) 282–295.
- [120] M. Daszykowski, W. Wu, A.W. Nicholls, R.J. Ball, T. Czekaj, B. Walczak, Identifying potential biomarkers in LC-MS data, *J. Chemom.* 21 (2007) 292–302.
- [121] R. Rajasekaran, P.R. Aruna, D. Koteeswaran, L. Padmanabhan, K. Muthuvelu, R.R. Rai, i in., Characterization and diagnosis of cancer by native fluorescence spectroscopy of human urine, *Photochem. Photobiol.* 89 (2013) 483–491.
- [122] G. Cao, R.L. Prior, Comparison of different analytical methods for assessing total antioxidant capacity of human serum, *Clin. Chem.* 44 (1998) 1309–1315.
- [123] M. Özyürek, K. Güçlü, E. Tütem, K.S. Başkan, E. Erçağ, S.E. Çelik, i in., A comprehensive review of CUPRAC methodology, *Anal. Methods*. 3 (2011) 2439–2453.
- [124] I.F.F. Benzie, J.J. Strain, The ferric reducing ability of plasma (FRAP) as a measure of “antioxidant power”: The FRAP assay, *Anal. Biochem.* 239 (1996) 70–76.
- [125] OECD, *Podręcznik Frascati 2002 Proponowane procedury standardowe dla badań statystycznych w zakresie działalności badawczo-rozwojowej: Proponowane procedury standardowe dla badań statystycznych w zakresie działalności badawczo-rozwojowej*, OECD Publishing, 2010.

## 12. Curriculum Vitae

### Joanna ORZEŁ



data urodzenia:

**24.01.1986**

miejsce zamieszkania:

**Katowice**

telefon:

**506-216-586**

e-mail:

**joanna\_orzel@interia.pl**

www:

**<https://sites.google.com/site/chemomlab/members/joanna-orzel>**

---

#### ..... *Wykształcenie*

- 2010- Uniwersytet Śląski, *Wydział Matematyki, Fizyki i Chemii, Instytut Chemii*, Katowice, Doktorant.
- 2005 – 2010 Uniwersytet Śląski, *Wydział Matematyki, Fizyki i Chemii, Instytut Chemii*, Katowice, Student.  
Pięcioletnie jednolite studia magisterskie na kierunku Chemia, specjalizacja Chemia analityczna, ukończone z wynikiem bardzo dobrym.
- 2002 - 2005 IX Liceum Ogólnokształcące, *im. C. K. Norwida*, Częstochowa  
Klasa o profilu biologiczno-chemicznym.
- 

#### ..... *Zainteresowania naukowe*

- Chemometria Sposoby modelowania danych chemicznych z uwzględnieniem problemów kalibracyjnych i klasyfikacyjnych.
- Fluorescencja Fluorescencyjne obrazy złożonych układów chemicznych.
- Chemia analityczna Nowe sposoby wykorzystania instrumentalnych metod analitycznych w badaniach złożonych układów chemicznych.
- Antyoksydanty Metody oceny całkowitej zdolności antyoksydacyjnej próbek modelowych i spożywczych.
- Analiza próbek kryminalistycznych Opracowywanie nowych metodyk do zastosowań w analizie kryminalistycznej.
- 

#### ..... *Języki*

- Angielski Zaawansowany w mowie i piśmie
- Niemiecki Podstawowy w mowie i piśmie
-

---

..... *Publikacje naukowe*

1. **J. Orzel**, I. Stanimirova, B. Czarnik-Matusewicz, M. Daszykowski, Prediction of hydrophilic antioxidant capacity of tomato pastes from the IR and fluorescence excitation-emission spectra of extracts and intact samples, *Talanta* 138 (2015) 64-70, IF=3,511
  2. B. Krakowska, I. Stanimirova, **J. Orzel**, M. Daszykowski, I. Grabowski, G. Zaleszczyk, M. Sznajder, Detection of discoloration in diesel fuel based on gas chromatographic fingerprints, *Analytical and Bioanalytical Chemistry* 407 (2015) 1157-1170, IF=3,578
  3. **J. Orzel**, M. Daszykowski, M. Kazura, D. de Beer, E. Joubert, A.E. Schulze, T. Beelders, A. de Villiers, C.J. Malherbe, B. Walczak, Modeling of the total antioxidant capacity of rooibos (*Aspalathus linearis*) tea infusions from chromatographic fingerprints and identification of potential antioxidant markers, *Journal of Chromatography A*, 1366 (2014) 101-109, IF=4,258
  4. **J. Orzel**, M. Daszykowski, A rapid validation of the antioxidant capacity of food commodities based on their fluorescence excitation emission spectra as applicable to coffee and peppermint extracts, *Chemometrics and Intelligent Laboratory Systems*, 137 (2014) 74-81, IF=2,381
  5. **J. Orzel**, M. Daszykowski, I. Grabowski, G. Zaleszczyk, M. Sznajder, Identifying the illegal removal from diesel oil of certain chemical markers that designate excise duty, *Fuel*, 117 (2013) 224-229, IF=3,406
  6. **J. Orzel**, M. Daszykowski, I. Grabowski, G. Zaleszczyk, M. Sznajder, B. Walczak, Simultaneous determination of solvent yellow 124 and solvent red 19 using fluorescence spectroscopy and chemometrics, *Talanta*, 101 (2012) 78-84, IF=3,511
  7. **J. Orzel**, M. Daszykowski, B. Walczak, Controlling sugar quality on the basis of fluorescence fingerprints using robust calibration, *Chemometrics and Intelligent Laboratory Systems*, 110 (2012) 89-93, IF=2,381
  8. M. Daszykowski, **J. Orzel**, M.S. Wrobel, H. Czarnik-Matusewicz, B. Walczak, Improvement of classification using robust soft classification rules for near-infrared reflectance spectral data, *Chemometrics and Intelligent Laboratory Systems*, 109 (2011) 86-93, IF=2,381
- 

..... *Komunikaty konferencyjne*

1. **J. Orzel**, M. Daszykowski, PLS and n-PLS for modelling the total antioxidant capacity of food extracts using excitation emission fluorescence spectroscopy, PLS 2014, Paryż (19-21 maja 2014), Francja
  2. **J. Orzel**, M. Daszykowski, Prosta i szybka metoda analizy antyoksydantów w żywności, Wiosenny Zjazd Sekcji Studenckiej PTChem 2014, Zawoja (9-13 kwietnia 2014), Polska
  3. **J. Orzel**, M. Daszykowski, I. Grabowski, G. Zaleszczyk, M. Sznajder, B. Walczak, Metoda równoczesnego oznaczania dodatków akcyzowych do oleju napędowego, V konferencja "Chemometria – metody i zastosowania" 2012, Zakopane (18-21 października 2012), Polska
  4. **J. Orzel**, M. Daszykowski, D. de Beer, E. Joubert, A.E. Schluze, T. Beelders, A.J. de Villiers, B. Walczak, Opracowanie sprawnej metody wyznaczania zdolności antyoksydacyjnej naparów herbaty w oparciu o dane chromatograficzne, 55 Zjazd Polskiego Towarzystwa Chemicznego i Stowarzyszenia Inżynierów i Techników Przemysłu Chemicznego 2012, Białystok (16-20 września 2012), Polska
  5. **J. Orzel**, M. Daszykowski, M. Bartoszek, J. Polak, B. Walczak, Czy piwo fluoryzuje i co z tego wynika?, Wiosenny Zjazd Sekcji Studenckiej PTChem 2012, Łazy (18-22 kwietnia 2012), Polska
  6. **J. Orzel**, M. Daszykowski, B. Walczak, Porównanie metod kalibracyjnych klasycznej i elastycznej w kontekście możliwości ich zastosowań w analizie procesowej, 54 Zjazd Polskiego Towarzystwa Chemicznego i Stowarzyszenia Inżynierów i Techników Przemysłu Chemicznego, Lublin (18-22 września 2011), Polska
- 

*Plakaty konferencyjne*

1. **J. Orzel**, M. Daszykowski, Metoda oznaczania kwasu mlekowego z wykorzystaniem sondy fluorescencyjnej, Zimowy Zjazd Studenckiej Sekcji PTChem 2014, Wrocław (13 grudnia 2014), Polska
2. **J. Orzel**, I. Stanimirova, M. Daszykowski, A novel method for rapid evaluation of total antioxidant capacity of food commodities based on nondestructive fluorescence measurements supported by chemometric modeling, Food Analysis Congress, Barcelona (29-30 października 2014), Hiszpania
3. **J. Orzel**, M. Daszykowski, Porównanie metody oznaczenia kwasu mlekowego z wykorzystaniem technik spektroskopowych, *Pomiędzy naukami – Zjazd fizyków i chemików 2014*, Chorzów (26 września 2014), Polska
4. **J. Orzel**, I. Stanimirova, M. Daszykowski, Modelowanie sygnałów spektroskopowych w kontekście oceny zdolności antyoksydacyjnej past pomidorowych, 57 Zjazd Polskiego Towarzystwa Chemicznego i Stowarzyszenia Inżynierów i Techników Przemysłu Chemicznego 2014, Częstochowa (14-18 września 2014), Polska

5. **J. Orzel**, M. Daszykowski, Identification of geographical origin of Indian opium on the basis of amino acid profiles, XXXVII Sympozjum nt. Chromatograficzne metody badania związków organicznych 2014, Szczyrk (11-13 czerwca 2014), Polska
6. **J. Orzel**, I. Stanimirova, M. Daszykowski, Ocena zdolności antyoksydacyjnej przecierów pomidorowych, VIII seminarium naukowe: "Aktualne problemy chemii analitycznej" 2014, Katowice (16 maja 2014), Polska
7. B. Krakowska, **J. Orzel**, M. Daszykowski, I. Grabowski, G. Zaleszczyk, M. Sznajder, Partial Least Squares for detection of diesel oil adulteration, PLS 2014, Paryż (19-21 maja 2014), Francja
8. **J. Orzel**, M. Daszykowski, I. Grabowski, G. Zaleszczyk, M. Sznajder, Analiza dwuwymiarowych widm fluorescencyjnych odbarwionego oleju napędowego o obniżonej aktywności, Zimowy Zjazd Studenckiej Sekcji PTChem 2013, Łódź (7 grudnia 2013), Polska
9. **J. Orzel**, M. Daszykowski, Fluorescence spectroscopy imaging as a method for total antioxidant capacity evaluation, Recent Advances in Food Analysis, Praga (5-8 listopada 2013), Czechy
10. **J. Orzel**, M. Daszykowski, Studium właściwości spektralnych fluoresceiny w zależności od otoczenia chemicznego, Pomiedzy naukami – Zjazd fizyków i chemików 2013 (27 września 2013), Polska
11. **J. Orzel**, M. Daszykowski, D. de Beer, E. Joubert, A.E. Schluz, T. Beelders, A.J. de Villiers, B. Walczak, Chromatographic fingerprints and chemometrics as a tool for prediction of the total antioxidant capacity of rooibos infusions, XXXVI Sympozjum nt. Chromatograficzne metody badania związków organicznych 2013, Szczyrk (5 czerwca 2013), Polska
12. **J. Orzel**, M. Daszykowski, I. Grabowski, G. Zaleszczyk, M. Sznajder, Badanie właściwości fluorescencyjnych odbarwionego oleju o obniżonym podatku akcyzowym, VII seminarium naukowe: "Aktualne problemy chemii analitycznej" 2013, Katowice (17 maja 2013), Polska
13. B. Krakowska, **J. Orzel**, M. Daszykowski, I. Grabowski, G. Zaleszczyk, M. Sznajder, Wykrywanie procedury fałszowania oleju napędowego przy użyciu metod chemometrycznych, VII seminarium naukowe: "Aktualne problemy chemii analitycznej" 2013, Katowice (17 maja 2013), Polska
14. **J. Orzel**, M. Daszykowski, M. Bartoszek, J. Polak, A. Pędziwiatr, J. Spaltenstein, Wyznaczanie zdolności antyoksydacyjnej nalewek alkoholowych, Wiosenny Zjazd Sekcji Studenckiej PTChem 2013, Augustów (11-18 kwietnia 2013), Polska
15. **J. Orzel**, M. Daszykowski, Chemometria w chemii analitycznej - kalibracja wieloparametrowa, Pomiedzy naukami – Zjazd fizyków i chemików 2012, Chorzów (5 października 2012), Polska
16. **J. Orzel**, M. Daszykowski, B. Walczak, Simultaneous determination of Solvent Yellow 124 and Solvent Red 19 in diesel oil using fluorescence spectroscopy and chemometrics, The 13th International Conference on Chemometrics in Analytical Chemistry, Budapeszt (25-29 czerwca 2012), Węgry
17. **J. Orzel**, M. Daszykowski, B. Walczak, Zaawansowane pomiary instrumentalne i metody ich modelowania na przykładzie fluorescencyjnych obrazów, The VI seminarium naukowe: "Aktualne problemy chemii analitycznej" 2012, Katowice (18 maja 2012), Polska
18. **J. Orzel**, M. Daszykowski, B. Walczak, Porównanie metod kalibracyjnych klasycznej i elastycznej w kontekście możliwości ich zastosowań w analizie procesowej, 54 Zjazd Polskiego Towarzystwa Chemicznego i Stowarzyszenia Inżynierów i Techników Przemysłu Chemicznego 2012, Lublin (18-22 września 2011), Polska
19. **J. Orzel**, M. Daszykowski, B. Walczak, Chemometrics in fluorescence data calibration, XV Sympozjum nt. Zastosowania metod spektroskopowych w badaniu materiałów i związków chemicznych 2011, Poznań (25 maja 2011), Polska
20. **J. Orzel**, M. Daszykowski, B. Walczak, Modelowanie danych fluorescencyjnych, V seminarium naukowe: "Aktualne problemy chemii analitycznej" 2011, Katowice (13 maja 2011), Polska
21. **J. Orzel**, M. Daszykowski, B. Walczak, Chemometric in analysis of fluorescence fingerprints, Wiosenny Zjazd Sekcji Studenckiej PTChem 2011, Murzasichle (13-17 kwietnia 2011), Polska
22. **J. Orzel**, J. Połedniok, Spectrophotometric study of Ga(III) and V(IV) complex formation with ChromeazurolS and benzylododecyldimethylammonium bromide in a presence of surfactants, V seminarium naukowe: "Aktualne problemy chemii analitycznej" 2011, Katowice (11 czerwca 2011), Polska
23. **J. Orzel**, J. Połedniok, Influence of non-ionic surfactants on ternary complexes of Ga(III) and V(IV) with ChromeazurolS and benzylododecyldimethylammonium bromide, Wiosenny Zjazd Sekcji Studenckiej PTChem 2010, Sokolec (7-11 kwietnia 2010), Polska

---

### Zgłoszenie patentowe

*Sposób równoczesnego lub pojedynczego oznaczania zawartości znacznika Solvent Yellow 124 i barwnika Solvent Red 19 lub Solvent Red 164 lub Solvent Blue 35 w oleju napędowym.*  
Zgłoszenie patentowe nr **P.399194**, Data zgłoszenia 16.05.2012

---



---

..... *Doświadczenie zawodowe*

1. Pracownik techniczny Instytutu Chemii Uniwersytetu Śląskiego od 2014
  2. Ćwiczenia z modułu spektroskopia cząsteczkowa, prowadzone dla studentów studiów podyplomowych z analizy instrumentalnej w ramach projektu pt. Zwiększenie udziału osób dorosłych w kształceniu w zakresie narzędzi informatycznych i technologii – NITKA, 2015
  3. Kierownik projektu badawczego PRELUDIUM, finansowanego przez NCN, *Fluorescencyjne krajobrazy w połączeniu z metodami chemometrycznymi jako potencjalne narzędzie do wyznaczania całkowitej zdolności antyoksydacyjnej*, 2012-2014
  4. Zajęcia laboratoryjne z chemii analitycznej, dla studentów I roku chemii pierwszego stopnia, 2010/11, 2011/12, 2012/13, 2013/14
  5. Zajęcia wyrównawcze z chemii, dla studentów I roku chemii pierwszego stopnia w ramach projektu pt. Zwiększenie liczby absolwentów kierunku chemia – ZLAB, 2012/13
  6. Zajęcia z chemii analitycznej dla uczniów szkół średnich biorących udział w olimpiadzie chemicznej, 2011
- 

..... *Stypendia i projekty*

1. Stypendium DoktorIS finansowane z funduszy Unijnych – POKL, 2013/14, 2014/15
  2. Współpraca z firmą GLOKOR Sp. z o.o. w kwestii oceny zawartości kwasu mlekowego w serwatce od 2014
  3. Projekt badawczy PRELUDIUM, *Fluorescencyjne krajobrazy w połączeniu z metodami chemometrycznymi jako potencjalne narzędzie do wyznaczania całkowitej zdolności antyoksydacyjnej*, NCN nr 2011/03/N/ST4/00713, 2012-2014
  4. Współpraca z Laboratorium Izby Celnej w Białej Podlaskiej w kwestii zastosowania metod chemometrycznych do detekcji nielegalnie odbarwionych paliw od 2012
  5. Stypendium doktoranckie finansowane z projektu unijnego UPGOW, 2012/13
  6. Stypendium projakościowe dla najlepszych doktorantów przyznane z dotacji Ministra Nauki i Szkolnictwa Wyższego, 2012/13
  7. Stypendium dla najlepszych doktorantów przyznawane z puli Rektora Uniwersytetu Śląskiego, 2012/13
  8. Stypendium doktoranckie Rektora Uniwersytetu Śląskiego, 2010/11, 2011/12, 2013/14
  9. Stypendium naukowe dla najlepszych studentów Uniwersytetu Śląskiego, 2009/10
- 

..... *Szkolenia i praktyki zawodowe*

1. Szkolenie w zakresie pozyskiwania środków na badania naukowe, 2015
  2. Szkolenie w zakresie prowadzenia własnej działalności gospodarczej, 2014
  3. Szkolenie w zakresie efektywnego prezentowania wyników naukowych, 2014
  4. Warsztaty FORSTAT – ocena dowodów kryminalnych problemy i zastosowania, 2013
  5. Szkolenie w zakresie ochrony własności intelektualnej oraz komercjalizacji wynalazków, 2011, 2014
  6. Praktyki zawodowe, jednostka Wojewódzkiego Inspektoratu Ochrony Środowiska w Katowicach z siedzibą w Częstochowie, laboratorium analiz chemicznych, laboratorium analiz biologicznych, 2009
- 

*Janina Orlik*