



WYKORZYSTANIE DATA MINING DO OCENY PRODUKTÓW NA PODSTAWIE WIDMA NIR

Tomasz Demski, StatSoft Polska Sp. z o.o.

Spektroskopia w bliskiej podczerwieni (ang. *Near Infrared, NIR*) oraz podobne techniki są coraz częściej stosowane w wielu dziedzinach, np. przemyśle farmaceutycznym, spożywczym i petrochemicznym, ale także w kryminalistyce. Dzieje się tak, ponieważ spektroskopia NIR umożliwia szybkie uzyskanie wyników badań i może być stosowana *on-line*, a nawet *in-line*, czyli bezpośrednio na linii produkcyjnej. Jednak do jej praktycznego wykorzystania wymagana jest tzw. kalibracja, umożliwiająca powiązanie widma z właściwościami badanych próbek. Zazwyczaj kalibrację wykonuje się za pomocą metod klasycznej statystyki wielowymiarowej: regresji wielorakiej, analizy składowych głównych (PCA), modeli cząstkowych najmniejszych kwadratów (PLS) i analizy Fouriera (zob. [1], [5]). My wykorzystamy do tego celu zgłębianie danych (data mining), a mówiąc dokładnie - metodę MARSplines [5].

Krótko o NIR

Mówiąc w skrócie, spektroskopia w bliskiej podczerwieni (ang. *Near Infrared Spectroscopy*) polega na badaniu, jak próbka pochłania, odbija lub przepuszcza promieniowanie podczerwone w zależności od jego długości fali. Badany zakres długości fal to 780–2500 nm. Na podstawie tak uzyskanego widma możemy oceniać rozmaite właściwości bardzo różnych produktów.

Najważniejsze zalety spektroskopii NIR to (zob. [1], [3]):

- ◆ uniwersalność,
- ◆ szybkie uzyskiwanie wyników (często pomiar trwa mniej niż sekundę!),
- ◆ jest to metoda nieniszcząca,
- ◆ pomiaru można dokonywać bezpośrednio na linii produkcyjnej,
- ◆ można mierzyć zawartość kilku składników jednocześnie (np. tłuszczu i protein),
- ◆ często nie jest wymagane specjalne przygotowanie próbki do badań.

Głównym utrudnieniem jest to, że pomiar jest pośredni, a widmo NIR jest zazwyczaj skomplikowane, a zależność między widmem a właściwościami próbki jest złożona. Zanim



zaczniemy używać spektroskopii NIR konieczne jest przeprowadzenie kalibracji, innymi słowy - powiązanie widma z właściwościami próbki. Wymaga to dosyć złożonych metod analizy danych, a często również wstępnego przetworzenia widma: wygładzenia, filtrowania itd. Warto jednak zauważyć, że kalibracja jest wykonywana rzadko; jeśli już ją zrobimy, to stosowanie NIR staje się bardzo proste.

Kalibrację widm NIR najczęściej wykonuje się za pomocą regresji wielorakiej, metod PCA i PLS, analizy Fouriera. Przed analizą widmo często jest przekształcane: standaryzowane, różnicowane, wygładzane itp.

Poniżej znajduje się kilka przykładów stosowania NIR (więcej przykładów i dokładniejsze opisy znajdują się w pracach [1], [3] oraz [4]).

- ◆ analiza procesu mieszania (np. w przemyśle farmaceutycznym) – na podstawie widm NIR określamy, czy mieszanina jest już jednorodna i czy można zakończyć mieszanie,
- ◆ przewidywanie twardości tabletek,
- ◆ wykrywanie fałszowania miodu, cygar i innych produktów,
- ◆ określanie składu mięsa,
- ◆ określanie wilgotności i zawartości protein w zbożu,
- ◆ badanie zawartości paliwa lotniczego,
- ◆ monitorowanie parametrów procesu wytwarzanie włókien syntetycznych,
- ◆ badanie *in-situ* zawartości glukozy we krwi,
- ◆ identyfikacja pochodzenia produktów żywnościowych.

Warto podkreślić rolę spektroskopii NIR w przemyśle farmaceutycznym. Otóż zalecane przez amerykańską agencję FDA i jej europejski odpowiednik EMEA podejście PAT (*Process Analytical Technology*) wymaga zapewnienia jakości procesów, a co za tym idzie przeprowadzania pomiarów w toku procesu *on-line* lub *in-line*, czyli bezpośrednio na linii produkcyjnej. W praktyce oznacza to konieczność stosowania spektroskopii NIR do badania produktów i surowców.

Krótko o metodzie MARSplines

W metodzie MARSplines (*Multivariate Adaptive Regression Splines*) bazuje się na podziale obszaru zmienności zmiennych niezależnych na przedziały, w których mają one różny wpływ na badane zjawisko. Punkty, w których zachodzi zmiana rodzaju wpływu nazywamy węzłami (ang. *knots*). Aby uniknąć gwałtownego skoku odpowiedzi modelu, rozwiązanie tworzą tzw. funkcje bazowe, które po jednej stronie węzła są równe zero, a po drugiej są funkcją liniową. Model jest sumą iloczynów funkcji bazowych.

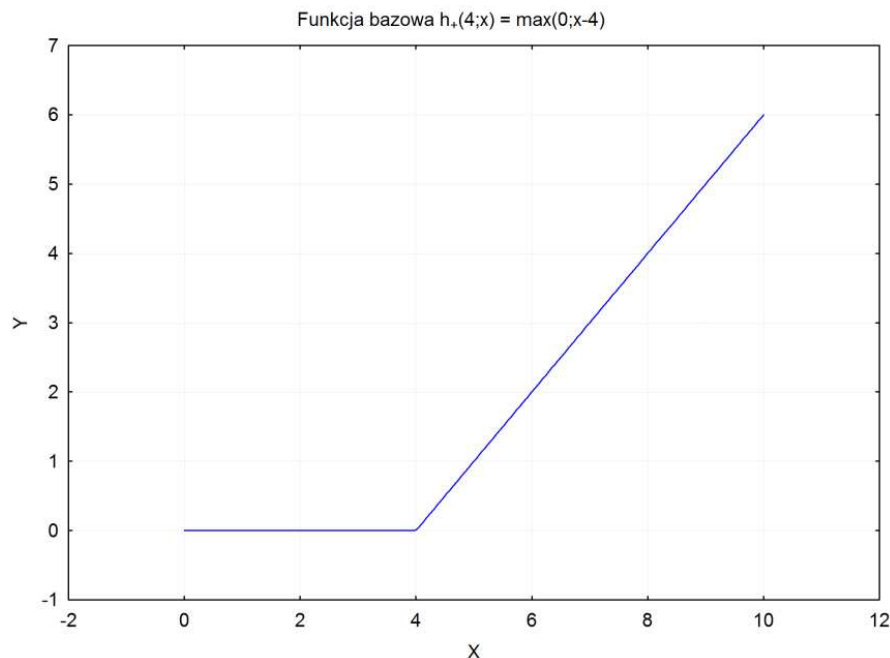
Mówiąc bardziej formalnie, funkcja bazowa h dla predyktora x_p ma postać:

$$h_+(t; x_p) = \begin{cases} 0 & \text{dla } x_p < t \\ (x_p - t) & \text{dla } x_p \geq t \end{cases}$$

lub

$$h_-(t; x_p) = \begin{cases} -x_p + t & \text{dla } x_p < t \\ 0 & \text{dla } x_p \geq t \end{cases}$$

Na rysunku poniżej widzimy funkcję bazową dla węzła w punkcie 4.



Rys. 1. Przykład funkcji bazowej.

Zmienną zależną y obliczamy ze wzoru

$$y = b + \sum_{i=1}^M a_i H_i(x)$$

gdzie b i a_i to współczynniki modelu, a H_i jest iloczynem funkcji bazowych:

$$H_i(x) = \prod_{k=1}^K h(t_k; x_p)$$

Kolejne człony w powyższych wzorach dodajemy jeden po drugim, tak aby uzyskać jak najmniejszy błąd GCV (zob. wzór poniżej). Dzięki temu duża liczba zmiennych niezależnych nie powoduje zasadniczych trudności w uzyskaniu modelu, powoduje jedynie zwiększenie czasu obliczeń.

$$GCV = \frac{1}{N} \frac{\sum_{i=1}^N (y_{\text{obserwowany}} - y_{\text{przewidywany}})^2}{(1 - C/N)^2} = MSE \frac{1}{(1 - C/N)^2}$$

W powyższym wzorze N to liczba przypadków, a C jest miarą złożoności modelu proporcjonalną do liczby jego składowych (zob. [4]).

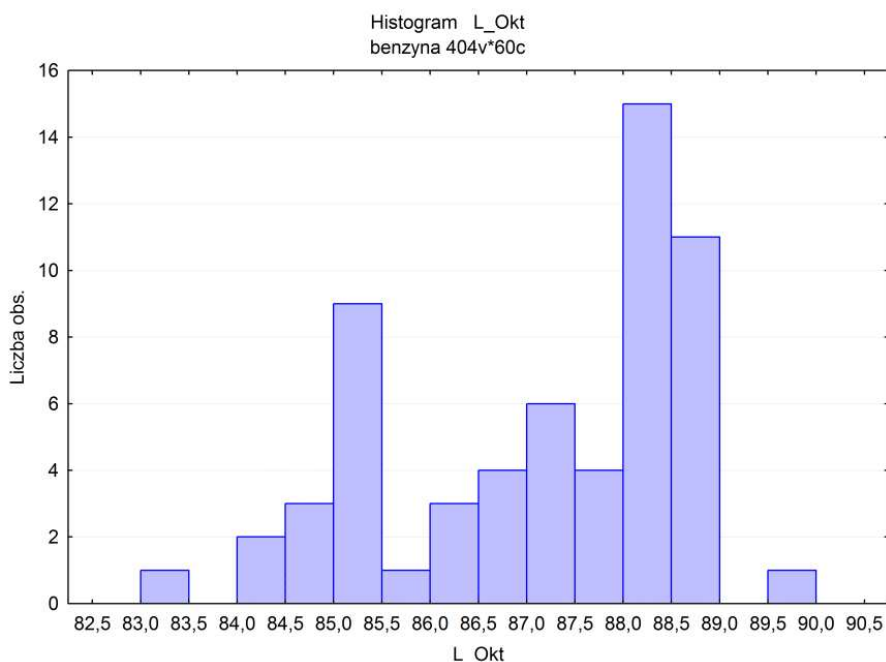
Najważniejszymi zaletami MARSplines są: możliwość odtworzenia złożonej zależności, jawna postać modelu, szybkość obliczeń i stosunkowo łatwe stosowanie modelu.

Przykład

Naszym zadaniem będzie zbudowanie modelu, który na podstawie widma NIR benzyny będzie w stanie określić jej liczbę oktanową. Zastosowanie takie jest opisane w [1]. Skorzystamy z danych wykorzystywanych w artykule [2], jednak do modelowania zamiast PLS użyjemy techniki MARSplines. Zauważmy, że w *STATISTICA* jest dostępny specjalny moduł MSPC, przeznaczony do stosowania PLS i PCA w sterowaniu jakością procesów przemysłowych [7].

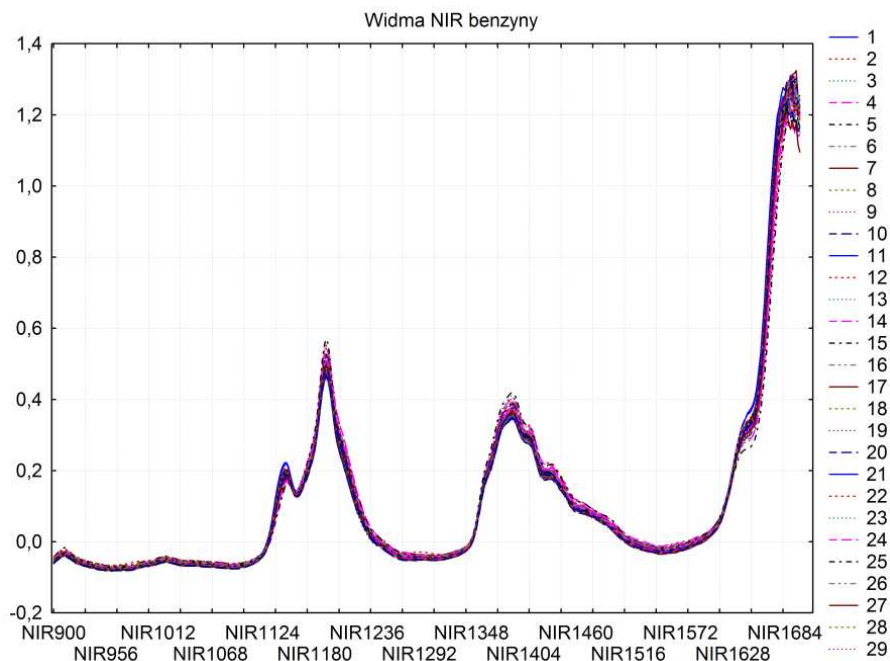
W wykorzystywanym przez nas zbiorze danych zapisano wartości widma dla długości fali od 900 nm do 1700 nm. Jest to łącznie 401 zmiennych objaśniających (predyktorów), na podstawie których będziemy przewidywać liczbę oktanową. Niestety mamy tylko 60 przypadków, a więc nie możemy zastosować zwykłej regresji.

Na początek zobaczymy, jaki jest rozkład zmiennej zależnej L_{okt} . Jak widać na histogramie poniżej, jest on dosyć nieprzyjemny do modelowania: mamy wiele maksimumów, a między nimi są rzadko obsadzone „dziury”.

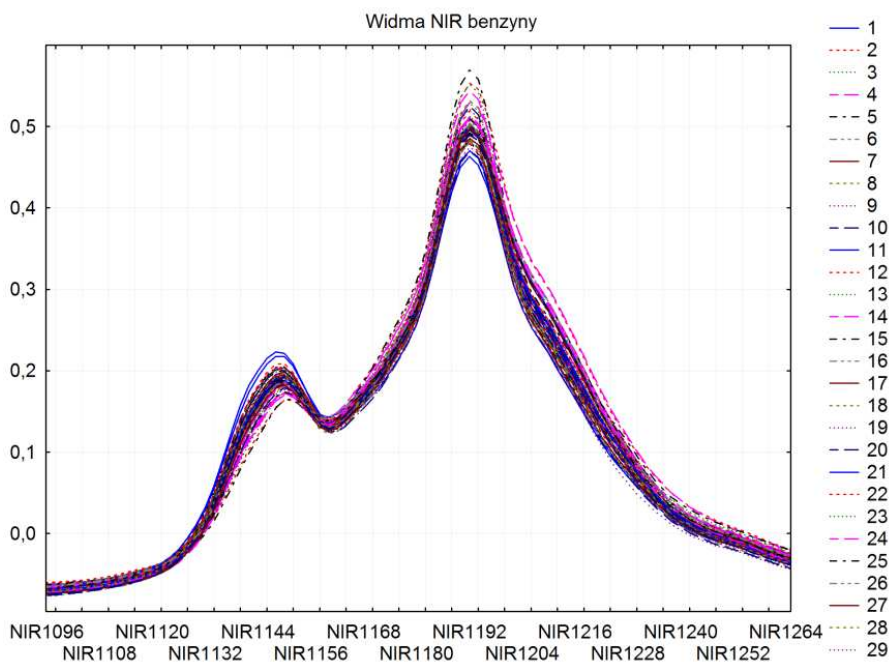


Rys. 2. Rozkład liczby oktanowej w badanym zbiorze.

Następnym krokiem wstępnej analizy jest zbadanie przebiegu widma. Zauważmy, że w tej analizie wartości widm dla kolejnych długości fal nie są już zmiennymi, tylko przypadkami. Podobnie nasze wcześniejsze przypadki, tzn. próbki benzyny stają się zmiennymi dla wykresu. Na rys. 3 widzimy, jak wyglądały widma 60 badanych próbek.



Rys. 3. Widma 60 badanych próbek.

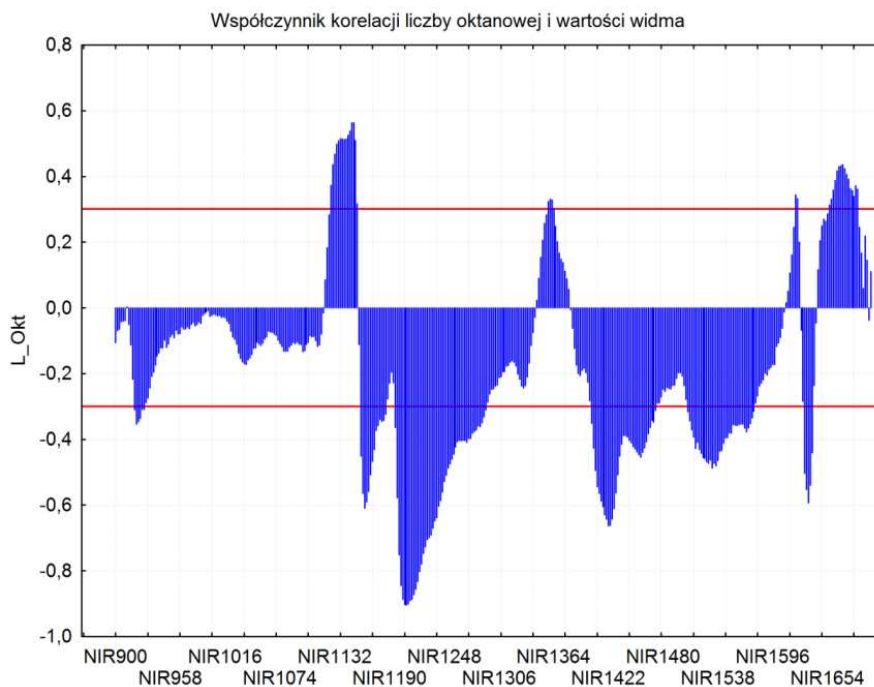


Rys. 4. Powiększony fragment widma.

Widma są dosyć podobne, ale jednak się różnią: daje nam to nadzieję na zbudowanie skutecznego modelu. Na rys. 4 widzimy powiększony obszar pierwszego podwójnego pików widma. Widać na nim wyraźnie różnice między widmami dla poszczególnych próbek.

Zauważmy, że widma nie wykazują gwałtownych skoków i są raczej gładkie. Możemy spróbować zbudować model bez wcześniejszego wygładzania lub filtrowania danych.

Sprawdźmy teraz, jak wyglądają związki między liczbą oktanową a wartościami widma dla poszczególnych częstości. Użyjemy do tego celu współczynników korelacji liniowej Pearsona. Wartości współczynników przedstawia poniższy wykres.



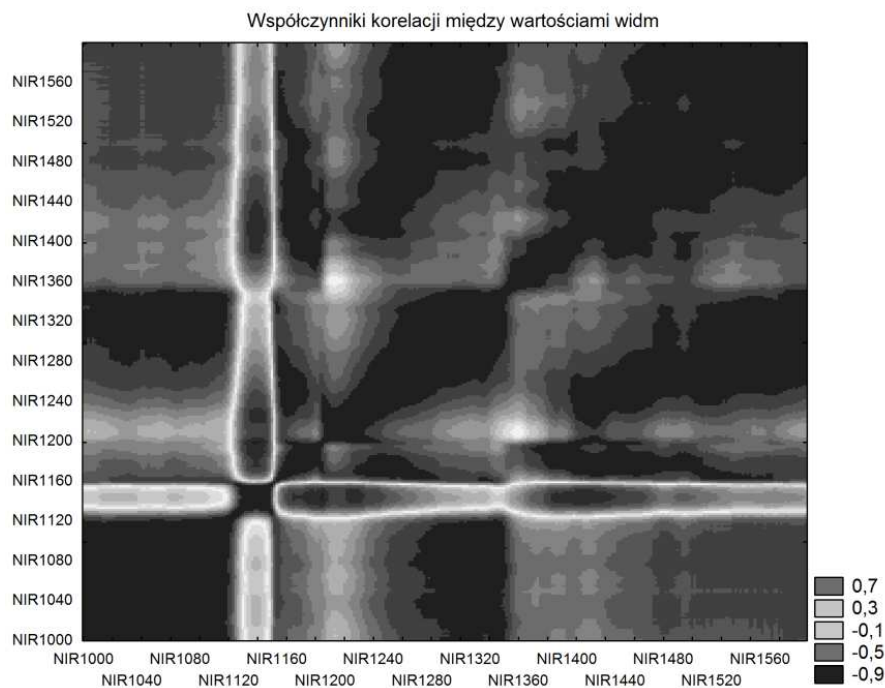
Rys. 5. Korelacja wartości widma z liczbą oktanową.

Jak widać, mamy bardzo dużo silnych związków, o dużej wartości modułu współczynnika korelacji. Mamy aż 59 zmiennych (odpowiadających różnym długościom fal) o module współczynnika korelacji z liczbą oktanową większym od 0,5!

Sprawdzimy jeszcze, jak wyglądają związki między wartościami widma dla poszczególnych długości fali. Macierz korelacji będzie duża: 401 na 401, poniżej widzimy fragment tabeli. Jednak przeglądając ją, można uzyskać ogólny pogląd o zależnościach między różnymi długościami fal. Przede wszystkim każda składowa widma jest silnie, a nawet bardzo silnie skorelowana z innymi. Silne związki występują między sąsiednimi długościami fal, ale również między odległymi, przykładowo współczynnik korelacji między składową dla 900 nm a 1318 nm wynosi 0,89. Ogólnie można powiedzieć, że przebieg zależności jest złożony.

Tabela 1. Fragment macierzy korelacji.

Dł. fali	1100	1102	1104	1106	1108	1110	1112	1114	1116	1118	1120	1122	1124	1126	1128	1130	1132	1134
1100	1,00	1,00	1,00	1,00	1,00	0,99	0,99	0,99	0,98	0,97	0,94	0,87	0,78	0,63	0,47	0,30	0,15	0,03
1102	1,00	1,00	1,00	1,00	1,00	0,99	0,99	0,99	0,98	0,97	0,94	0,88	0,79	0,65	0,48	0,31	0,17	0,05
1104	1,00	1,00	1,00	1,00	1,00	1,00	0,99	0,99	0,99	0,97	0,95	0,89	0,80	0,66	0,50	0,34	0,19	0,07
1106	1,00	1,00	1,00	1,00	1,00	1,00	0,99	0,99	0,99	0,98	0,95	0,90	0,81	0,68	0,52	0,35	0,21	0,09
1108	1,00	1,00	1,00	1,00	1,00	1,00	1,00	0,99	0,99	0,98	0,96	0,91	0,82	0,69	0,53	0,36	0,22	0,10
1110	0,99	0,99	1,00	1,00	1,00	1,00	1,00	1,00	0,99	0,98	0,96	0,92	0,83	0,70	0,55	0,39	0,24	0,12
1112	0,99	0,99	0,99	0,99	1,00	1,00	1,00	1,00	1,00	0,99	0,97	0,92	0,84	0,72	0,56	0,40	0,26	0,14
1114	0,99	0,99	0,99	0,99	0,99	1,00	1,00	1,00	1,00	0,99	0,97	0,93	0,85	0,72	0,57	0,41	0,27	0,15
1116	0,98	0,98	0,99	0,99	0,99	0,99	1,00	1,00	1,00	1,00	0,98	0,94	0,87	0,75	0,60	0,45	0,31	0,19
1118	0,97	0,97	0,97	0,98	0,98	0,98	0,99	0,99	1,00	1,00	0,99	0,96	0,90	0,80	0,66	0,51	0,37	0,26
1120	0,94	0,94	0,95	0,95	0,96	0,96	0,97	0,97	0,98	0,99	1,00	0,99	0,94	0,86	0,73	0,60	0,47	0,36
1122	0,87	0,88	0,89	0,90	0,91	0,92	0,92	0,93	0,94	0,96	0,99	1,00	0,98	0,93	0,83	0,71	0,60	0,50
1124	0,78	0,79	0,80	0,81	0,82	0,83	0,84	0,85	0,87	0,90	0,94	0,98	1,00	0,98	0,92	0,83	0,73	0,64
1126	0,63	0,65	0,66	0,68	0,69	0,70	0,72	0,72	0,75	0,80	0,86	0,93	0,98	1,00	0,98	0,92	0,85	0,79
1128	0,47	0,48	0,50	0,52	0,53	0,55	0,56	0,57	0,60	0,66	0,73	0,83	0,92	0,98	1,00	0,98	0,94	0,89
1130	0,30	0,31	0,34	0,35	0,36	0,39	0,40	0,41	0,45	0,51	0,60	0,71	0,83	0,92	0,98	1,00	0,99	0,96
1132	0,15	0,17	0,19	0,21	0,22	0,24	0,26	0,27	0,31	0,37	0,47	0,60	0,73	0,85	0,94	0,99	1,00	0,99
1134	0,03	0,05	0,07	0,09	0,10	0,12	0,14	0,15	0,19	0,26	0,36	0,50	0,64	0,79	0,89	0,96	0,99	1,00
1136	-0,05	-0,03	-0,01	0,01	0,02	0,04	0,06	0,07	0,11	0,18	0,28	0,42	0,58	0,73	0,86	0,94	0,98	1,00



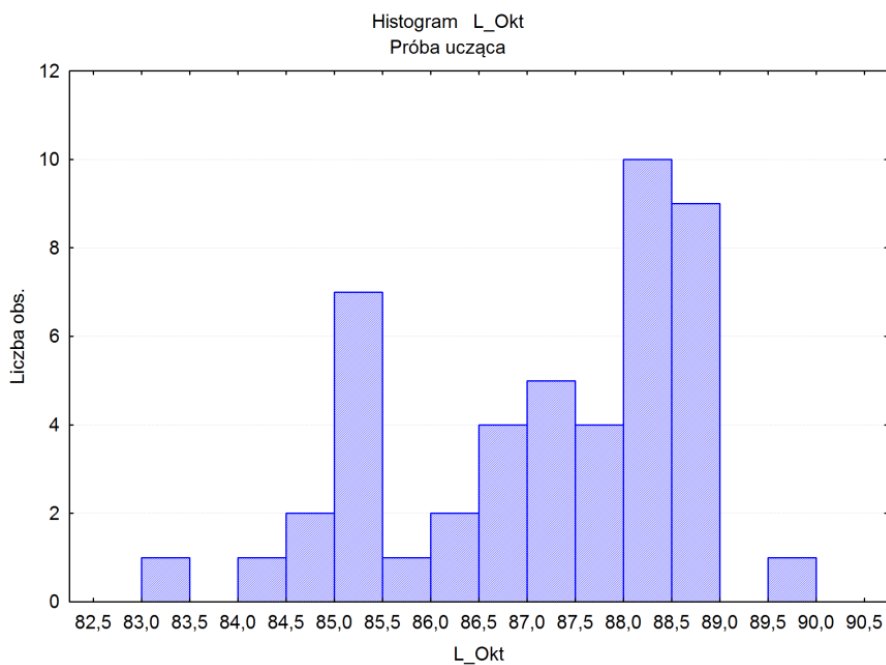
Rys. 6. Wykres wartości współczynników korelacji wartości widma.

Na rys. 6 widzimy macierz korelacji przedstawioną na wykresie: im ciemniejszy kolor, tym większy moduł współczynnika korelacji, a co za tym idzie silniejszy liniowy związek między składowymi widma. Wykres pokazuje, jak skomplikowane są te zależności.

Zwróćmy uwagę, że silne liniowe związki między składowymi dla różnych długości fal umożliwiają skuteczne wykonywanie analizy składowych głównych (PCA) i PLS.

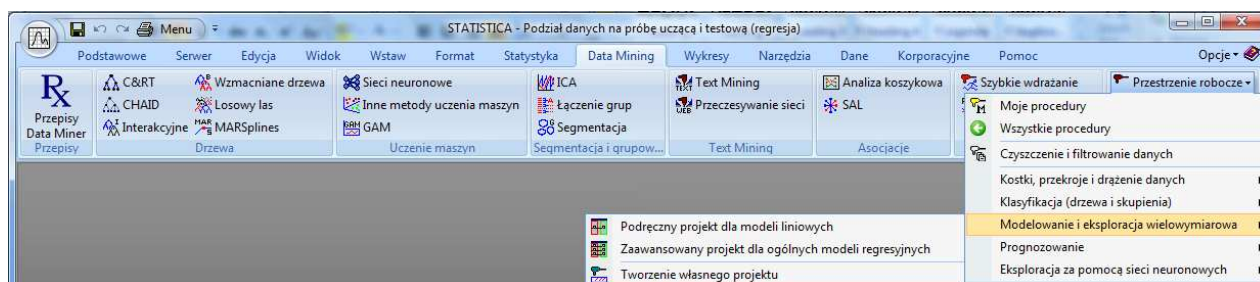
Teraz stworzymy model przewidujący liczbę oktanową na podstawie widma NIR. Zastosujemy metodę MARSplines.

W celu oceny uzyskanego modelu z danych wydzieliśmy 13 przypadków, które nie będą wykorzystywane do dopasowania modelu, lecz do oceny jego działania; innymi słowy przypadki te będą stanowiły próbę testową. Mamy stosunkowo nieduży zbiór danych i musimy tu zachować ostrożność. W szczególności warto sprawdzić, czy próba ucząca pokrywa cały zakres zmiennej zależnej. Poniżej znajduje się histogram liczby oktanowej w próbie uczącej: jak widać znalazły się w niej próbki z całego zakresu danych.



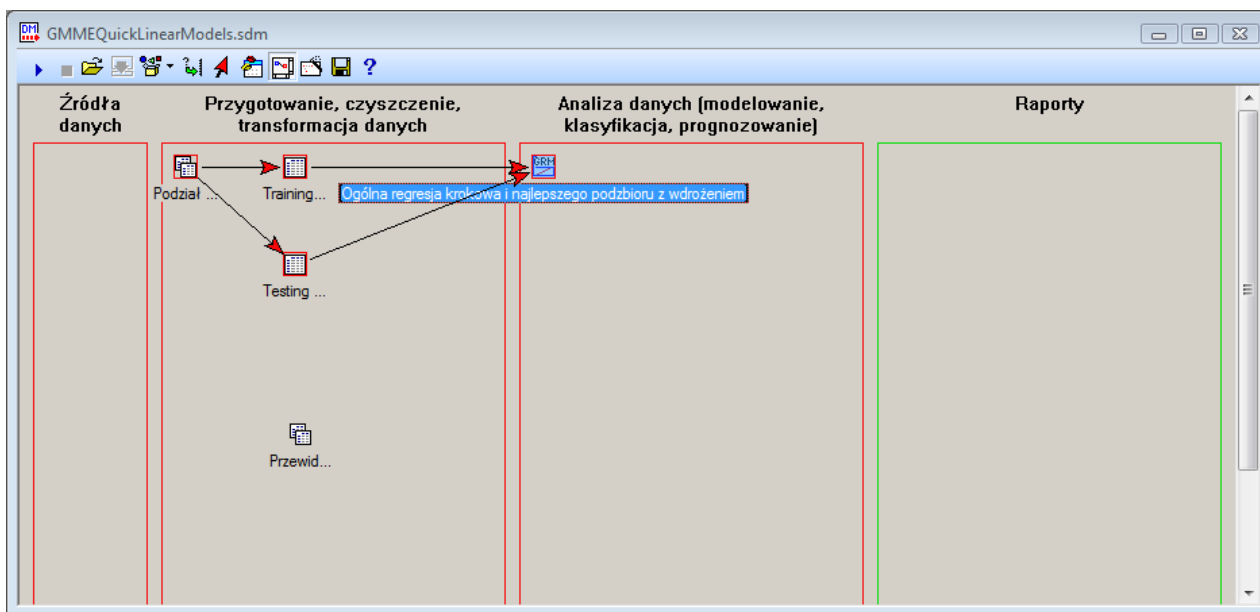
Rys. 7. Histogram dla próby uczącej.

Model zbudujemy w przestrzeni roboczej *STATISTICA Data Miner*. Jako podstawę wykorzystamy gotowy szablon projektu dla regresji. Na karcie *Data mining* wstążki klikamy przycisk *Przestrzeń robocza* i wybieramy polecenie *Modelowanie i eksploracja wielowymiarowa – Podręczny projekt dla modeli liniowych* (zob. rysunek poniżej).




Rys. 8. Wybór szblonu projektu.

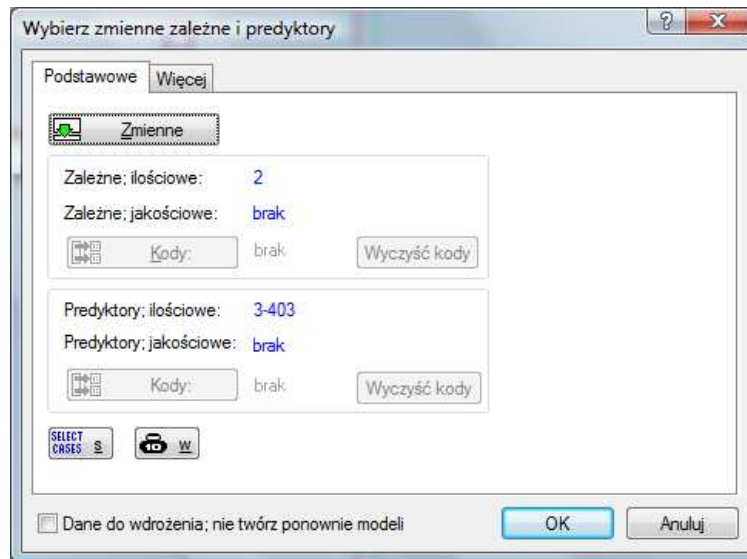
Otwarta zostanie przestrzeń robocza data mining z szablonem projektu pokazana na rysunku poniżej.




Rys. 9. Szablon projektu.

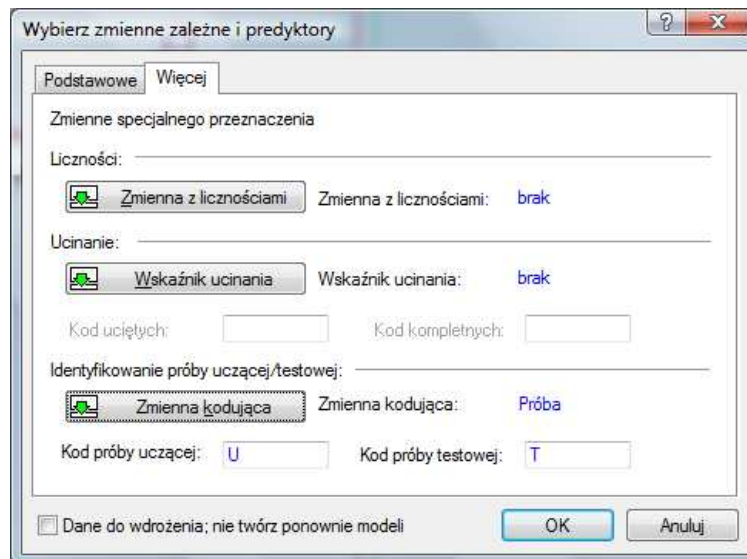
Klikamy przycisk  na pasku narzędzi przestrzeni roboczej, aby umieścić w niej arkusz danych *benzyna* jako źródło danych. Program automatycznie wyświetli na ekranie okno określania ustawień dla źródła danych. Klikamy przycisk *Zmienne* i wskazujemy *L_okt* jako zmienną zależną ilościową, a zmienne od *NIR900* do *NIR1700* jako predyktory ilościowe.

Następnie określamy zmienną identyfikującą dane uczące i testowe. W tym celu przejdziemy na kartę *Więcej* i w grupie *Identyfikowanie próby uczącej/testowej* naciskamy przycisk *Zmienna kodująca*, po czym jako wskaźnik próby wybieramy zmienną *Próba*. Następnie w polu *Kod próby uczącej* wpisujemy *U*, a w polu *Kod próby testowej* wpisujemy *T* (zob. rys. 11).




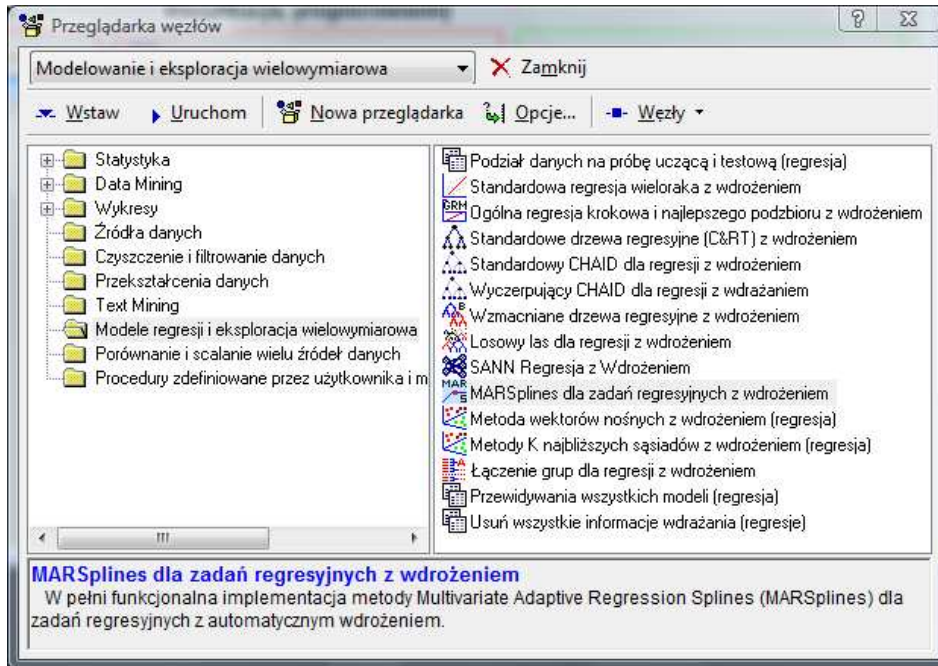
Rys. 10. Wybór zmiennych.

Po określeniu właściwości źródła danych, łączymy je z węzłem *Podział danych na próbę uczącą i testową (regresja)*: zaznaczamy oba te węzły, a potem klikamy przycisk  na pasku narzędzi.




Rys. 11. Wybór identyfikatora prób.

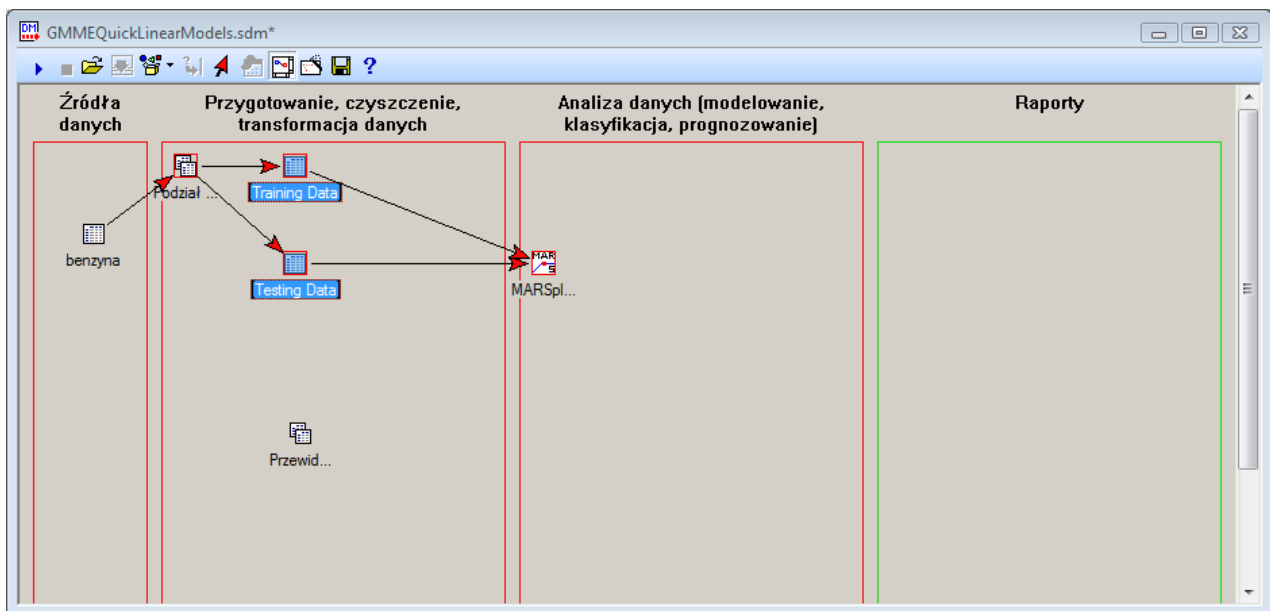
W szablonie projektu model jest tworzony przez węzeł *Ogólna regresja krokowa i najlepszego podzbioru z wdrożeniem*. My chcemy zastosować metodę *MARSplines*, dlatego usuwamy ten węzeł. Następnie zaznaczamy węzły z danymi uczącymi i testowymi, po czym naciskamy przycisk  na pasku narzędzi. Na ekranie pojawi się okno *Przeglądarka węzłów* (zob. poniżej), w którym wybieramy węzeł *MARSplines dla zadań regresyjnych z wdrożeniem* i klikamy *Wstaw*.



Rys. 12. Przeglądarka węzłów.

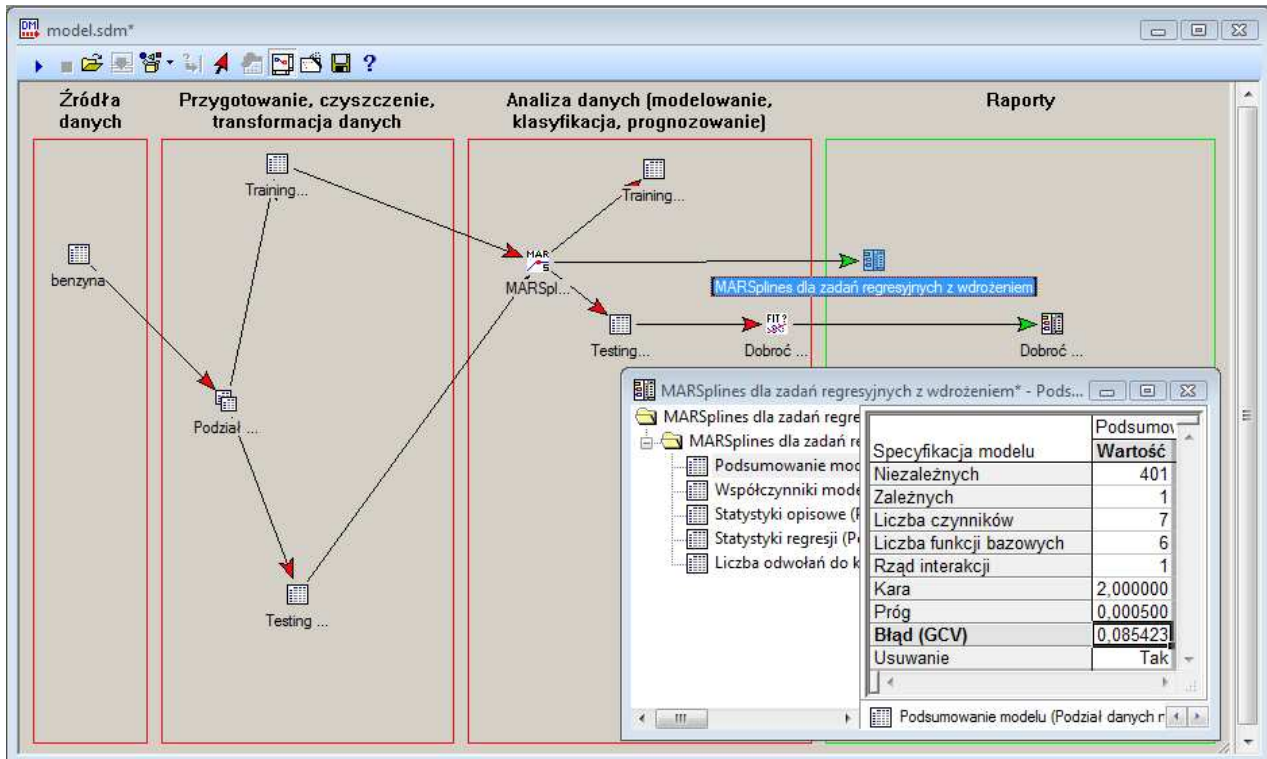
Program automatycznie połączy nowy węzeł z danymi do uczenia i testowania modelu. Projekt przedstawia rysunek poniżej. Aby utworzyć model wystarczy uruchomić projekt naciskając przycisk  na pasku narzędzi. Po zakończeniu obliczeń w przestrzeni roboczej pojawią się nowe węzły:

- ◆ w obszarze *Raporty* z podsumowaniem analizy,
- ◆ w obszarze *Analiza danych* dwa źródła danych z wynikami stosowania modelu dla danych uczących i testowych.



Rys. 13. Projekt przed uruchomieniem.

Kolejny krok analizy to ocena trafności modelu. Aby ją wykonać, dwukrotnie klikamy węzeł *Testing...* w obszarze *Analiza danych*. Następnie jako zmienną zależną ilościową wybieramy *L_okt*, a jako predyktor ilościowy *MARSplineModelPrzew*. Po wybraniu zmiennych podłączamy do węzła *Testing...* w obszarze *Analiza danych* węzeł analityczny *Dobroć dopasowania* z foldera *Data mining – Dobroć dopasowania*. Uruchamiamy węzeł, klikając go prawym klawiszem myszy i wybierając polecenie *Uruchom węzeł*. Poniższy rysunek przedstawia gotowy projekt.

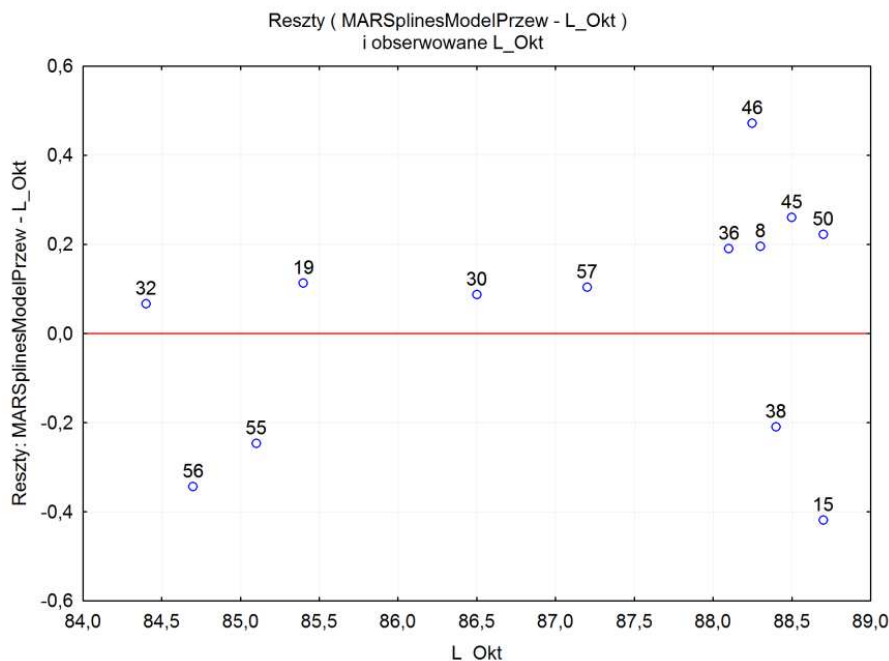


Rys. 14. Projekt po wykonaniu.

Ocenę modelu zaczniemy od przejrzania podsumowania modelu: dwukrotnie klikamy węzeł *MARSplines dla zadań regresyjnych z wdrożeniem* w obszarze *Raporty*. Na ekranie otworzy się skoroszyt z wynikami dopasowania modelu do danych. Zajrzyjmy do pierwszego arkusza *Podsumowanie modelu* (jest on widoczny na rysunku powyżej). Program szacuje błąd przewidywania za pomocą uogólnionego sprawdzianu krzyżowego (ang. *generalized cross validation*, GCV). Podejście to omówione jest w rozdziale 7 podręcznika [6], a jego zastosowanie dla MARSplines w rozdziale 9. Błąd GCV bierze pod uwagę nie tylko różnice między wartościami obserwowanymi a przewidywanym z modelu, ale również złożoność modelu: bardziej złożone modele są „karane”: błąd jest nieco powiększany. Poprawka uwzględnia to, że model zwykle lepiej działa dla danych uczących niż dla nowych danych.

W naszym przypadku błąd GCV wynosi około 0,085. Średnia wartość błędu wyznaczenia liczby oktanowej jest pierwiastkiem z błędu (GCV) i wynosi około 0,3. Jest to całkiem przyzwoity wynik, podobny do uzyskiwanego innymi metodami (zob. [1]).

Węzeł *Dobroć dopasowania* wyznaczył błąd średniokwadratowy *MSE* w próbie testowej; wynosi on w naszym przypadku 0,065 – jest porównywalny z szacunkową miarą trafności modelu błędem *GCV*. Do oceny modelu warto użyć wykresu reszt w zależności od wartości obserwowanej (również utworzonego przez węzeł *Dobroć dopasowania*).



Rys. 15. Wykres reszt.

Wykres nie wykazuje wad modelu: nie ma zależności reszt od zmiennej niezależnej (*L_Okt*), ani odstających reszt wyraźnie większych od innych, czy też wyraźnych, przyczynowych wzorców dla reszt.

Podsumowanie

W podsumowaniu można powiedzieć, że metoda *MARSplines* umożliwiła nam szybkie i stosunkowo łatwe zbudowanie zadowalającego modelu zależności liczby oktanowej od widma NIR. Wystarczyło użyć domyślnych ustawień systemu *STATISTICA Data Miner* i nie było potrzeby wykonywania wstępnych przekształceń danych. Dzięki zastosowaniu przestrzeni roboczej model możemy łatwo oceniać, aktualizować dla nowych danych, stosować do przewidywania i modyfikować. Ponadto model *MARSplines* ma jawną postać i można go przenieść do prawie każdego systemu.

Warto też zwrócić uwagę, że standardowy pomiar liczby oktanowej jest kłopotliwy. Potrzebny jest do tego specjalny wzorcowy silnik (jego cena wynosi ponad 250 000 \$) pracujący w kontrolowanych warunkach laboratoryjnych. Trudno sobie wyobrazić wykonywanie takiego pomiaru w toku produkcji. Natomiast spektrometr NIR i model kalibracyjny mogą być stosowane wręcz na linii produkcyjnej, bez konieczności pobierania i przygotowywania próbek.



Literatura

1. Burns D.A., Ciurczak E. W., *Handbook of Near Infrared Analysis*, Third Ed., CRC Press, 2008.
2. Mevik B-H., Wehrens R., *The pls Package: Principal Component and Partial Least Squares Regression in R*, Journal of Statistical Software, Vol. 18(2).
3. Quinn G. P., Keough M. J., *Near infrared spectroscopy for on /in-line monitoring of quality in foods and beverages*, *Journal of Food Engineering* 87 (2008).
4. Downey G., Kelly D., Petisco Rodriguez C., *Food authentication – Has near infrared spectroscopy a role?*, *Spectroscopy Europe* vol 18 no. 3, 2006.
5. *Elektroniczny Podręcznik Statystyki PL*, StatSoft, 2006, www.statsoft.pl/textbook.
6. Hastie T., Tibshirani R., Friedman J., *The Elements of Statistical Learning*, Springer-Verlag 2002.
7. Demski T., *Jak ugotować dobrą zupę, czyli modelowanie procesów wsadowych z wykorzystaniem MSPC na przykładzie procesu polimeryzacji*, www.statsoft.pl/czytelnia/artykuly/Modelowanie_proc_wsadowych.pdf.