



## DATA MINING W PRZEMYŚLE: PROJEKTOWANIE, UDOSKONALANIE, WYTWARZANIE

**Tomasz Demski**

*StatSoft Polska Sp. z o.o.*

### Wprowadzenie

Współczesne produkty i procesy produkcyjne stają się coraz bardziej skomplikowane, a wymagania odnośnie ich jakości są coraz większe. Zarówno podczas projektowania produktu, jego wytwarzania, jak i korzystania z niego przez użytkownika gromadzone są bardzo duże ilości danych. W danych tych ukryta jest wiedza, którą można wydobyć, korzystając z narzędzi *data mining* (zgłębiania danych).

### Co to jest *data mining*?

Jest wiele różnych definicji *data mining*; kładą one nacisk na różne aspekty i cechy tej metodyki. Z całą pewnością *data mining* jest pewnym sposobem postępowania z danymi, przetwarzania ich, mówiąc dokładniej ich analizy. Rozsądną i wyważoną definicję zaproponowano w 1997 roku w pracy [1]:

„*Data mining* jest procesem badania i analizy dużych ilości danych metodami automatycznymi lub półautomatycznymi w celu odkrycia znaczących wzorców i reguł.”

Warto zwrócić uwagę na następujące cechy *data mining*, które w szczególności odróżniają tę metodę od tradycyjnej statystycznej analizy danych:

1. Analiza dużych zbiorów danych.
2. Nastawienie na praktyczne wyniki i zastosowania, a nie na budowę lub sprawdzanie teorii.
3. Korzystanie z istniejących danych, na których zawartość badacz ma niewielki wpływ
4. Ocena modelu na podstawie próby testowej, a nie na podstawie wskaźników statystycznych

Za „duże” uznajemy takie zbiory danych, których człowiek nie jest w stanie objąć i wykorzystać bez pomocy komputera i specjalistycznego oprogramowania. Bardzo często w praktyce spotykamy się z sytuacją, gdy danych jest „za dużo”, a głównym zadaniem we

wnoskowaniu z danych jest odsianie bezużytecznej informacji (taką sytuację podsumowujemy stwierdzeniem, że „toniemy w danych”).

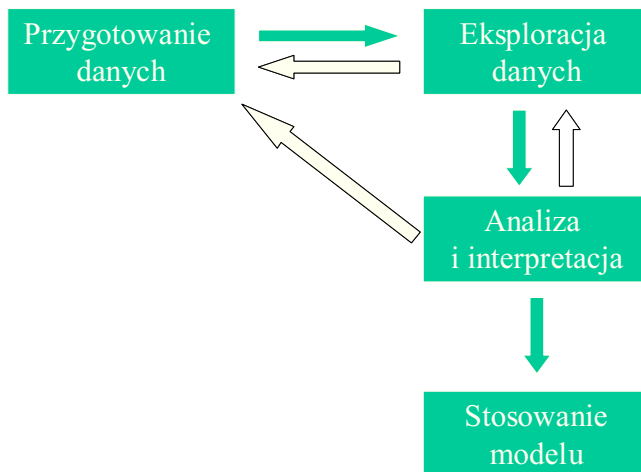
Ważną częścią tradycyjnego badania statystycznego jest zaplanowanie doświadczenia, które da nam informacje podlegające właściwej analizie. W *data mining* mamy do czynienia z inną sytuacją: zazwyczaj analizujemy istniejące dane, gromadzone zwykle do innych celów niż analiza danych, w pewnym „sensie zbierane przy okazji”. Przykładami typowych źródeł danych będą informacje z systemu automatyki przemysłowej (którego głównym celem jest sterowanie produkcją), systemu rejestrującego reklamacje zgłaszane przez klientów (który przecież służy przede wszystkim do wspomagania rozwiązywania problemów klientów) itp.

W *data mining* wykorzystuje się narzędzia pochodzące z trzech dziedzin:

- ◆ Technologii bazodanowej (gromadzenie, udostępnianie i przetwarzanie danych),
- ◆ Statystyki,
- ◆ Uczenia maszyn i sztucznej inteligencji.

W procesie *data mining* możemy wyróżnić cztery zasadnicze etapy:

1. Przygotowanie danych,
2. Eksploracyjna analiza danych,
3. Właściwa analiza danych (budowa i ocena modelu lub odkrywanie wiedzy),
4. Wdrożenie i stosowanie modelu.



Warto zwrócić uwagę, że powyższe etapy nie przebiegają liniowo jeden za drugim. Bardzo często na kolejnym etapie okazuje się, że powinniśmy wrócić do wcześniejszego (por. rysunek powyżej). Poniżej przedstawiamy podstawowe informacje o etapach procesu *data mining* (więcej informacji można znaleźć w pracach [2] i [3]).

Na etapie przygotowania danych decydujemy, z jakich informacji będziemy korzystać w analizie, pobieramy odpowiednie dane, sprawdzamy ich poprawność i dokonujemy



odpowiednich przekształceń, aby zapewnić zgodność danych pochodzących z różnych źródeł.

Celem eksploracji danych jest poznanie ogólnych własności analizowanych danych: rozkładów jedno- i wielowymiarowych cech i podstawowych związków między zmiennymi. Wynikiem takiej wstępnej analizy jest wykrycie nietypowych przypadków. Po wykryciu odstających przypadków powinniśmy podjąć decyzję, jak będziemy z nimi postępować. Podczas eksploracji uzyskujemy również informacje, czy potrzebne i użyteczne będą jakieś przekształcenia oryginalnych danych. Przykładowo, w wyniku eksploracji danych może okazać się, że klasy zmiennej jakościowej występują tak rzadko, iż należy je połączyć z innymi.

Na etapie eksploracji danych bardzo często wykonujemy wstępną selekcję zmiennych, aby w dalszych analizach uwzględniać tylko te właściwości obiektów, które są istotne (np. wpływają na zmienną zależną).

W razie wykrycia niejednorodności danych, możemy pogrupować wszystkie przypadki (obiekty analiz) w jednorodne grupy i właściwą analizę wykonywać osobno dla jednorodnych grup.

Etap właściwej analizy danych rozpoczynamy od wstępnego doboru metod, odpowiednich do rozwiązania naszego problemu. Przy wyborze metody bierzemy pod uwagę rodzaj problemu, wielkość zbioru danych, dopuszczalną złożoność modelu oraz wymagania odnośnie możliwości interpretacji modelu.

Po wykonaniu analiz oceniamy, czy uzyskane wyniki są zadowalające. Kluczową sprawą jest, czy uzyskana informacja jest użyteczna z praktycznego punktu widzenia. Zazwyczaj wykorzystujemy więcej niż jedną technikę analizy danych. Istnieje wiele różnych metod oceny modeli i wyboru najlepszego z nich. Często stosuje się techniki bazujące na porównawczej ocenie modeli (ang. *competitive evaluation of models*) polegającej na stosowaniu poszczególnych metod dla tych samych zbiorów danych, a następnie wybraniu najlepszej z nich lub zbudowaniu modelu złożonego. Techniki oceny i łączenia modeli (uważane często za kluczową część predykcyjnego *data mining*) to m.in.: agregacja modeli (głosowanie i uśrednianie; ang. *bagging*), wzmacnianie (nazywane też losowaniem adaptacyjnym i łączeniem modeli, ang. *boosting*), kontaminacja modeli (ang. *stacking*, *stacked generalizations*) i metauczenie (ang. *meta-learning*). Obszerne omówienie wskaźników jakości modeli i sposobów ich porównania znajduje się w pracy [3].

Wdrożenie i stosowanie modelu jest to końcowy etap, na którym stosujemy dla nowych danych uzyskany model, który został uznany za najlepszy.

## **Data mining w przemyśle**

*Data mining* (zgłębianie danych) w zastosowaniach przemysłowych wykorzystuje te same metody analizy danych i wiąże się z podobnymi problemami jak w innych dziedzinach. Jest jednak kilka specyficznych cech zgłębiania danych w przemyśle.



Pierwsza z nich to bliski związek ze statystycznym sterowaniem jakością procesów (SPC). Bardzo często zgłębianie danych stosujemy jako rozwinięcie i uzupełnienie SPC. Stąd często tego typu zastosowania określa się nazwą *Quality Control data mining* (lub *QC data mining*). Nakłada to na systemy *data mining* stosowane w zastosowaniach przemysłowych wymóg zintegrowania z narzędziami do tradycyjnego SPC, takimi jak: karty kontrolne, analiza zdolności procesu i analiza niezawodności.

Na etapie projektowania użyteczna jest współpraca z aplikacjami wspomagającymi projektowanie (CAD).

Kolejnym wyróżnikiem *data mining* dla przemysłu jest wymóg automatycznego reagowania na zmiany w danych na bieżąco. Jako ilustrację rozważmy system, który przed zakończeniem wieloetapowego procesu technologicznego ma przewidywać, które produkty prawdopodobnie będą wadliwe, aby zaoszczędzić na końcowych etapach procesu. Oczywiście jest, że wyniki działania systemu muszą być dostępne natychmiast, tak abyśmy mieli czas i możliwość skorzystać z wyników analizy.

Dosyć często dane dotyczące procesów technologicznych składają się z bardzo wielu zmiennych: setek lub nawet tysięcy; taka sytuacja jest rzadko spotykana w innych dziedzinach. Bierze się to stąd, że zazwyczaj tworzone są one przez urządzenia automatyki przemysłowej. Zapisują one zazwyczaj całe mnóstwo parametrów, które często nie mają żadnego wpływu na wytwarzany w danej chwili produkt, ale mogą być decydujące dla innego produktu. Duża ilość zmiennych występuje także w przypadku analizy danych o procesach wsadowych (zob. artykuł „Monitorowanie i sterowanie jakością procesów wsadowych” w niniejszej publikacji), w których zmiennymi są wyniki pomiarów parametrów procesów dokonane w różnym czasie.

Ponadto w wielu dziedzinach produkcji zmiany zachodzą bardzo szybko – czas życia produktów i okres stosowania konkretnej technologii ciągle się zmniejsza. W związku z tym bardzo często będziemy potrzebować narzędzia tworzącego modele typu „czarna skrzynka” – na ich zrozumienie nie będziemy mieli po prostu czasu. Modele muszą radzić sobie z dużą liczbą danych nie wpływających w żaden sposób na zmienną wyjściową i łatwo adaptować się do zmienionych technologii i nowych produktów.

W przemyśle *data mining* stosujemy w:

- ◆ Projektowaniu i doskonaleniu produktu,
- ◆ Sterowaniu i optymalizacji procesu produkcyjnego,
- ◆ Analizie reklamacji i niezawodności.

Na etapie projektowania *data mining* może dotyczyć kwestii związanych z klientem, np. identyfikacji potrzeb klientów i prognozowania popytu. Ponadto możemy badać zależności między projektami produktów, portfelem produktów i potrzebami klientów.

Przykłady zastosowań *data mining* na etapie wytwarzania produktu to:

- ◆ Statystyczne sterowania jakością procesu,
- ◆ Przewidywanie problemów z jakością,



- ◆ Wykrywanie przyczyn problemów z jakością,
- ◆ Utrzymanie maszyn (np. planowanie przeglądów i remontów tak, aby uniknąć awarii),
- ◆ Sterowanie przebiegiem procesów,
- ◆ Wykrywanie przyczyn i związków między parametrami procesów,
- ◆ Podsumowanie wielowymiarowych danych,
- ◆ Adaptacyjne interfejsy człowiek – maszyna.

Poniżej prezentujemy dwa przykłady praktycznego i udanego zastosowania *data mining* w przemyśle. Wiele różnorodnych przykładów omówiono w pracy [2].

### ***Sterowanie procesem technologicznym***

Poniższy przykład przedstawiono w podręczniku [4].

W pewnym procesie technologicznym wykorzystywany był duży zbiornik przechowujący surowce dla tego procesu. Proces obserwowany był przez operatorów, którzy korygowali jego ustawienia, tak aby zmniejszyć lub zwiększyć stopień wypełnienia zbiornika. Występują przy tym dwa zagrożenia:

1. Za mała ilość surowca w zbiorniku,
2. Przepelnienie zbiornika.

W pierwszym przypadku cały proces technologiczny musi zostać zatrzymany i uruchomiony ponownie. Procedura taka jest bardzo kosztowna i niebezpieczna.

Natomiast w przypadku przepelnienia i wylania się zawartości zbiornika występuje duże zagrożenie.

Operator ma do dyspozycji wiele ustawień, tj. wartości zmiennych sterujących, ale faktycznie na ilość materiału w zbiorniku wpływa tylko jedna z nich. Celem analizy było przewidzenie wartości zmiennej sterującej, przy której zachowana zostanie bezpieczna ilość surowca w zbiorniku.

Parametry opisujące stan procesu i zbiornika oraz wartość zmiennej sterującej zapisywane są co 30 sekund.

Jako wzorzec odpowiedniego ustawienia zmiennej sterującej przyjęto działania kilku różnych operatorów. Niestety nie udało się teoretycznie wyznaczyć optymalnych ustawień.

Można postawić pytanie: po co modelować działania operatora i potem je naśladować? Otóż formalny, komputerowy model ma wiele zalet: np. w wypadku odejścia doświadczonego pracownika tracimy jego doświadczenia, dopóki nowy operator nie nabierze doświadczenia ryzyko wystąpienia problemów jest znacznie większe, przy sterowaniu automatycznym będzie (a przynajmniej powinno być) mniej losowych wahań, a cały proces powinien być stabilniejszy.



Zmiany dokonywane były przez operatora w oparciu o jego wyczucie i doświadczenie. Większość z nich były to małe i nieistotne zmiany. W związku z tym zdecydowano, że przewidywana będzie zmiana poziomu zmiennej sterującej po 3 minutach.

Jakość uzyskanych wyników zdecydowanie polepszyło stosowanie średnich ruchomych i wartości trendów zamiast surowych wartości parametrów. Pozwoliło to na wyeliminowanie losowych wahań parametrów. Ponadto w analizie uwzględniono tylko te obserwacje, w których zmienna sterująca została znacząco zmodyfikowana – drobne, wykonywane przez człowieka korekty były mylące i niedokładne.

Ważnym elementem modelu było wykrywanie trendów w wielkości strumienia surowca wpływającego i wypływającego ze zbiornika.

Dwa kluczowe parametry wstępnego przekształcenia danych to:

- ◆ Liczba punktów wykorzystywanych przy obliczaniu średniej ruchomej,
- ◆ Wielkość zmiany zmiennej sterującej uznawana za istotną.

Optymalizację tych dwóch parametrów analizy połączono z optymalnym doбором liczności próby uwzględnianej w analizie. Przykładowo za mała wartość progowej zmiany powodowała uzyskiwanie rozwiązań faworyzujących niewykonywanie żadnych zmian.

W wyniku analizy uzyskano rozwiązania o mniejszej i większej złożoności. Pomimo tego, że rozwiązanie o większej złożoności nieco lepiej przewidywało rzeczywiste zmiany, do stosowania wybrano prostsze rozwiązanie, ze względu na jego zgodność ze standardami przedsiębiorstwa.

W wyniku analizy uzyskano zaskakująco dobre wyniki: prosty i skuteczny model. Działanie modeli zbadano dla oryginalnych, zapisywanych co 30 sekund danych, a model spisał się dobrze, pozwalając uniknąć zarówno przepełnienia, jak i opróżnienia zbiornika.

### ***Optymalizacja procesu technologicznego w drukarni R.R. Donneley***

Niniejszy przykład przedstawiony jest w podręczniku [5].

W drukarni R. R. Donneley występował tajemniczy problem polegający na pojawianiu się serii rys na walcu wykorzystywanym przy drukowaniu rotograwiurowym. Na wydrukach problem objawiał się jako kolorowe linie przecinające cały wydruk. Problem zaczął występować przy drukowaniu z szybkością ponad 300 m/s.

Celem *data mining* w tym wypadku było zminimalizowanie częstości występowania problemu.

Stosowana technologia powodowała, że każda przerwa w drukowaniu i ponowne uruchamianie procesu były bardzo kosztowne. Ponadto wystąpienie rysy powodowało marnotrawstwo matryc oraz dużych ilości papieru i farby drukarskiej.

Usunięcie wady walca zajmowało średnio półtorej godziny, a w tym czasie cały proces był zatrzymany. Ponieważ terminy drukowania zazwyczaj są bardzo napięte, każde opóźnienie skutkowało dodatkowymi kosztami nadgodzin.



Przed rozpoczęciem projektu nie gromadzono żadnych danych. Jego pierwszym etapem było zdecydowanie, jakie informacje mają być zbierane i zapisywane w bazie danych. Początkowo zdecydowano, że dla procesów poprawnych i wadliwych gromadzone będą dane m.in. o: wilgotności, temperaturze farby, lepkości farby, odczynie farby, napięciu i rodzaju papieru. Ostatecznie zestaw zbieranych informacji uzyskano na podstawie konsultacji z ekspertami, tworzenia kolejnych modeli i wybierania zmiennych istotnie wpływających na wystąpienie problemu.

Jako metodę modelowania zastosowano różne algorytmy drzew decyzyjnych.

Ostatecznie w wyniku zastosowania *data mining* uzyskano zestaw reguł, które można było zastosować przy ustawianiu procesu produkcyjnego. Reguły te nie wyjaśniły, dlaczego pojawiają się problemy, ale pozwoliły zmniejszyć częstość ich występowania.

Wdrożenie reguł wydobytych z danych za pomocą *data mining* pozwoliło zmniejszyć liczbę wystąpień problemów w ciągu roku z 538 do 21. Łączny czas przestojów przed wprowadzeniem reguł wynosił 800 godzin/rok, a po ich zastosowaniu spadł do 30 godzin/rok.

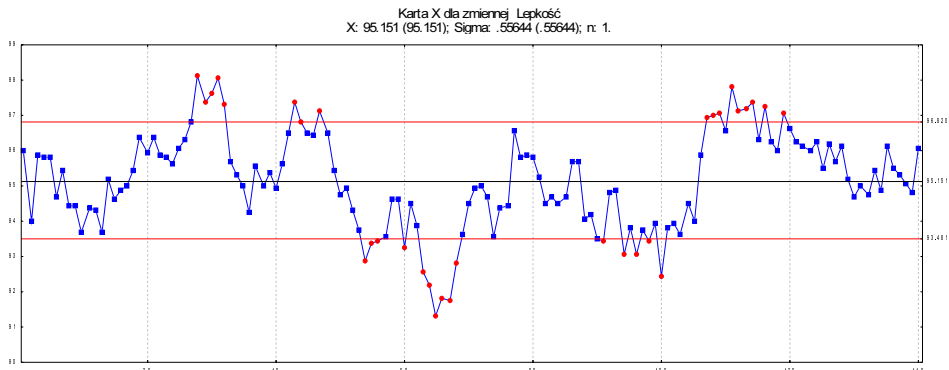
Doświadczenia uzyskane w drukarni, gdzie przeprowadzono oryginalny projekt, zostały przeniesione do innych zakładów. Chociaż same modele należało dostosować do każdej drukarni, to sposób rozwiązania problemu był ten sam.

## Przykłady analiz *data mining* w zastosowaniach przemysłowych

### *Karty kontrolne i predykcja*

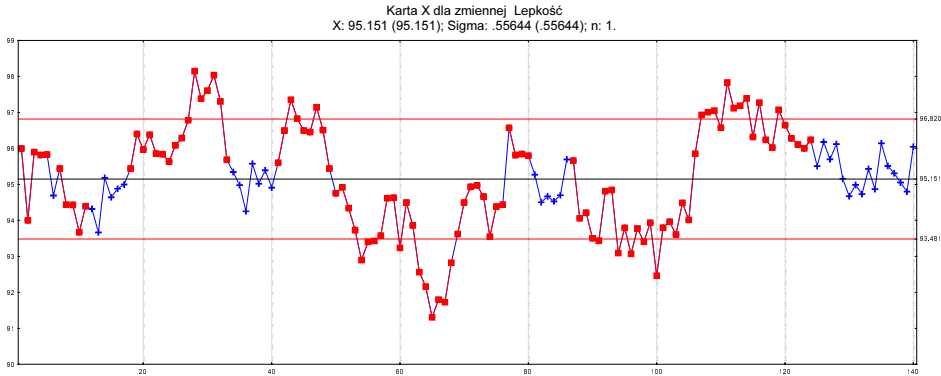
W tym przykładzie przedstawimy, jak *data mining* może uzupełnić i wzbogacić standardowe karty kontrolne.

Na poniższym rysunku widzimy kartę kontrolną pojedynczych obserwacji dla pewnego procesu.



Karta kontrolna sygnalizuje całkiem sporo rozregulowań. Ponadto cały przebieg sprawia wrażenie nielosowego. Jeśli włączymy testy konfiguracji, wykrywające nielosowe

sekwencje wartości, to większość punktów na karcie kontrolnej będzie wskazywało na rozregulowanie.



Ponieważ testy konfiguracji wykazały, że układ kolejnych punktów nie jest losowy, to możemy spróbować zbudować model przewidujący przyszłą wartość na podstawie poprzednich.

Skorzystamy z podejścia *data mining*, a konkretnie procedury *Predictive Quality Control* systemu *STATISTICA QC Miner*. Jest to „maszyna” automatycznie wykonująca:

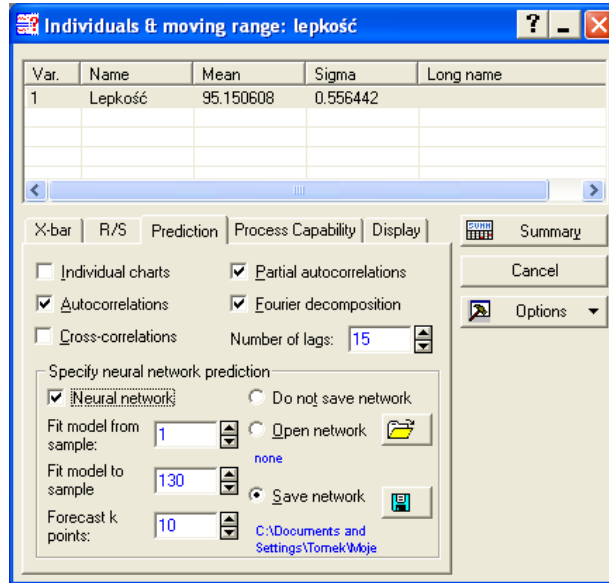
- ◆ badanie związku między bieżącą obserwacją a poprzednimi (tabele i wykresy autokorelacji i autokorelacji cząstkowej),
- ◆ analizę widmową (Fouriera),
- ◆ model sieci neuronowej dla karty X (lub X średnie) i karty rozstępu, z uwzględnieniem ewentualnych dodatkowych predyktorów (np. parametrów surowców).

Najważniejsze parametry procedury *Predictive Quality Control* to:

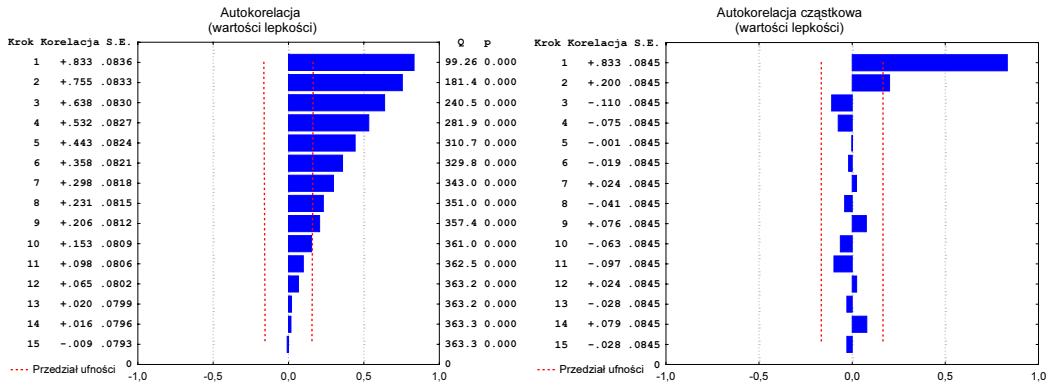
- ◆ długość opóźnienia (liczba obserwacji z przeszłości branych pod uwagę przy badaniu autokorelacji i budowie modelu),
- ◆ horyzont prognozy (ile obserwacji będzie przewidywał model),
- ◆ jaka część obserwacji zostanie wykorzystana do uczenia sieci.

Program automatycznie dobiera najlepszą architekturę sieci neuronowej (pod uwagę brane są: sieć liniowa, perceptron wielowarstwowy i sieć RBF), określa jej złożoność i uczy ją (więcej informacji można znaleźć w opisie Automatycznego projektanta sieci w podręczniku elektronicznym *STATISTICA*).

Do dyspozycji mamy 140 obserwacji. Jako próbę do uczenia modelu przyjmujemy próbki od 1 do 130 (ostanie 10 wykorzystamy do sprawdzenia działania modelu). Długość opóźnienia przyjmujemy równą 15, a prognozować będziemy 10 przyszłych obserwacji.

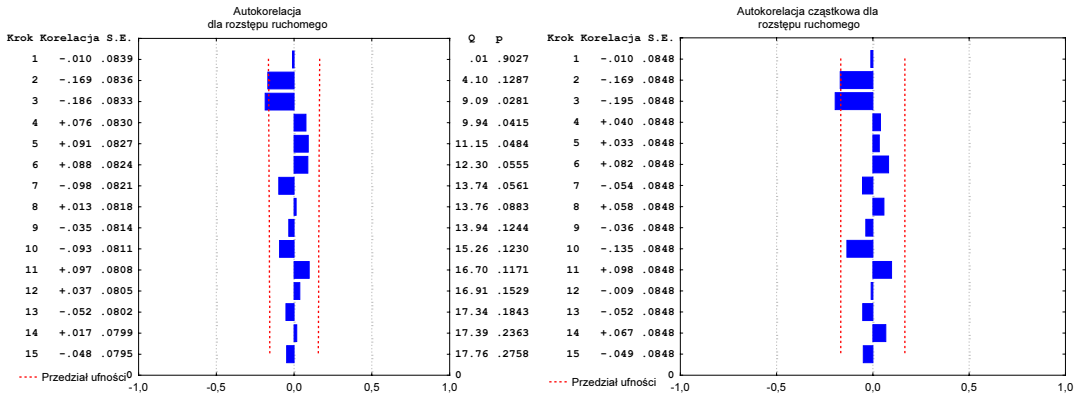


Zobaczmy najpierw, jak wyglądają autokorelacje i autokorelacje cząstkowe dla interesującej nas właściwości. Autokorelacja to współczynnik korelacji bieżącej obserwacji i obserwacji wcześniejszej (przesuniętej o pewien krok). Zauważmy, że jeśli obserwacja jest silnie skorelowana z obserwacją ją poprzedzającą, to będzie również związana z obserwacją wcześniejszą o dwa kroki itd. Taką „przechodnią” korelację eliminuje się przy wyznaczaniu autokorelacji cząstkowej.



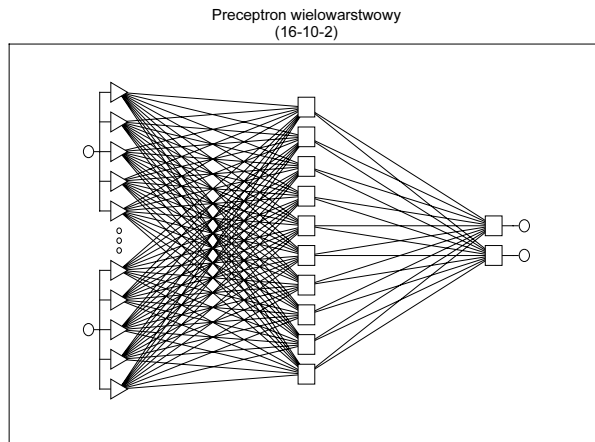
Na powyższych wykresach autokorelacji widzimy jak silny jest związek bieżącej obserwacji z obserwacją poprzedzającą ją o jeden krok. Jest także widoczna zależność z obserwacją wcześniejszą o dwa kroki, a większe opóźnienia wydają się być nieistotne.

W przypadku ruchomych rozstępów związek z wcześniejszymi obserwacjami nie jest już tak silny – można w niego w ogóle powątpiewać.

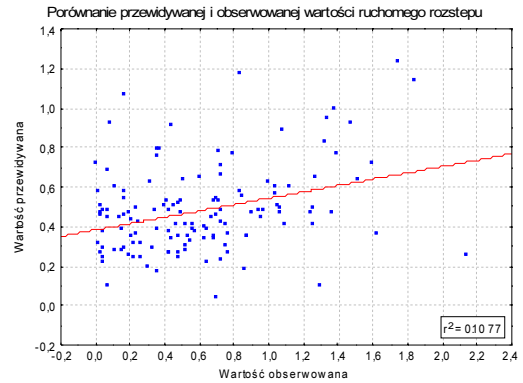
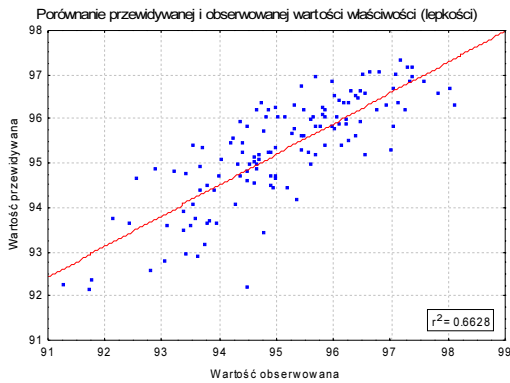


Analiza Fouriera (widmowa) w przypadku naszych danych nie wykryła żadnej interesującej okresowości.

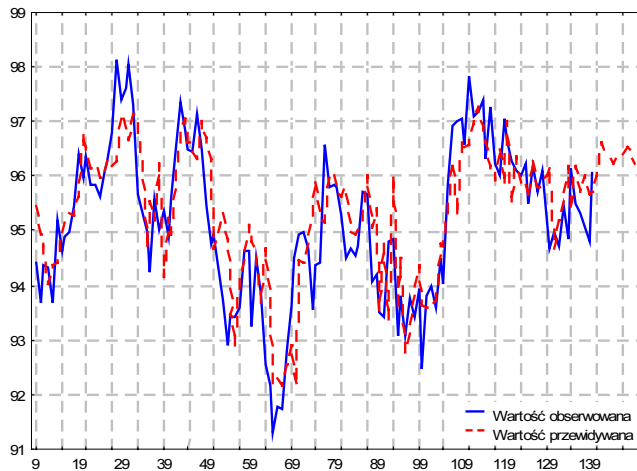
Teraz przejdźmy do modelu sieci neuronowej. Na poniższym rysunku widzimy schemat sieci neuronowej, która przewiduje zarówno wartości monitorowanej właściwości, jak rozstępu ruchomego.



Skuteczność działania modelu możemy sprawdzić za pomocą diagramów korelacyjnych dla wartości przewidywanych i obserwowanych. Z wykresów tych wynika, że dużo trafniejsze prognozy uzyskujemy dla wartości właściwości. Jeśli popatrzymy na wykresy autokorelacji zwykłej i cząstkowej (powyżej), to zrozumiemy, dlaczego model dla ruchomego rozstępu działa gorzej: po prostu związek bieżącego ruchomego rozstępu z wcześniejszymi jest dużo słabszy niż w przypadku wartości właściwości.



Zobaczmy jeszcze, jak wygląda przebieg wartości przewidywanych i obserwowanych badanej właściwości:



Jak widać, prognoza dla wartości badanej właściwości jest zadowalająca, zwłaszcza jeśli weźmiemy pod uwagę, że przy tworzeniu modelu korzystaliśmy tylko z danych o wcześniejszych obserwacjach.

Przewidywania przyszłych wartości możemy wykorzystać do korekty ustawień parametrów procesu. Więcej informacji o takim podejściu zwanym *Engineering Process Control* (EPC) można znaleźć w podręczniku [6].

## Modelowanie

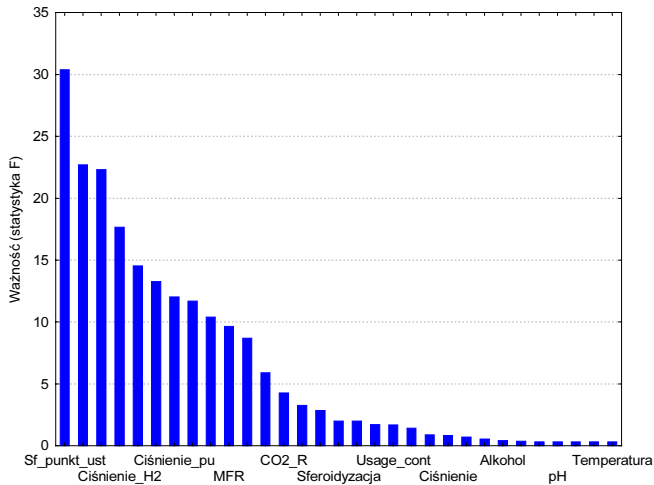
W tym przykładzie zbudujemy model przewidujący właściwość produktu na podstawie parametrów procesu.

Pod uwagę bierzemy 33 parametry procesu: będą one stanowiły predyktory w naszej analizie. Dane te zostały wstępnie przygotowane do analizy i oczyszczone, jednak

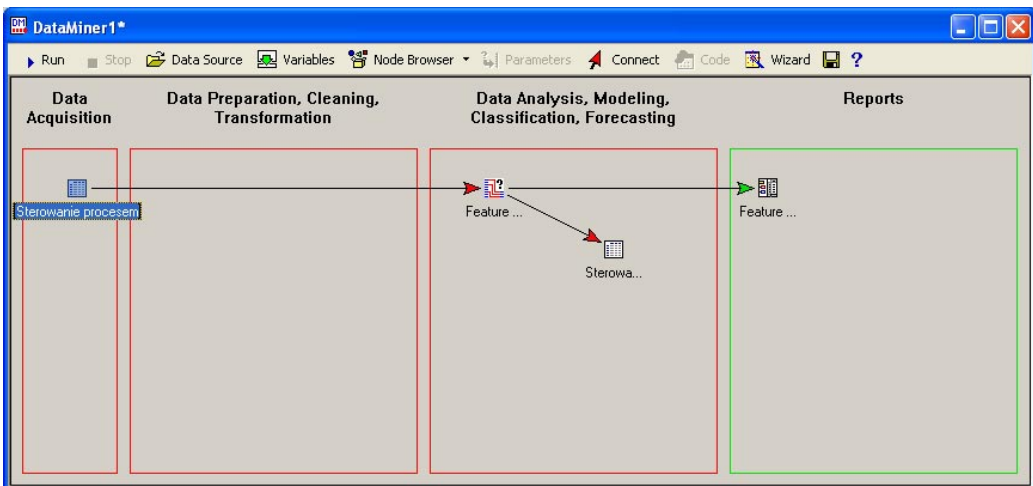


pozostawiliśmy wśród z nich jedną zmienną, która nie przynosi żadnej informacji, aby sprawdzić, jak można poradzić sobie z takimi „pustymi” zmiennymi.

Model zbudujemy w przestrzeni roboczej systemu *STATISTICA Data Miner*. Pierwszym krokiem analizy będzie odrzucenie zmiennych, które nie wpływają na wielkość wyjściową. Użyjemy do tego celu modułu *Feature selection*. Moduł ten dla każdej zmiennej przeprowadza jednowymiarowy test, czy wpływa ona na zmienną wyjściową (w *STATISTICA QC Miner* jest również procedura uwzględniająca interakcje między zmiennymi). Jako kryterium doboru predyktorów wybierzemy poziom  $p$  i ustawimy go jako równy 0,1.



Patrząc na powyższy wykres ważności potencjalnych predyktorów, wydaje się, że procedura selekcji wybrała wszystkie ważne predyktory, a nawet kilka takich, które nic nie wnoszą. Zauważmy jednak, że powinniśmy wybrać raczej „za dużo” niż „za mało” zmiennych. Zauważmy, że moduł automatycznie usuwa z dalszej analizy również „puste” zmienne: przyjmujące jedną wartość dla wszystkich obserwacji lub wypełnione brakami danych.

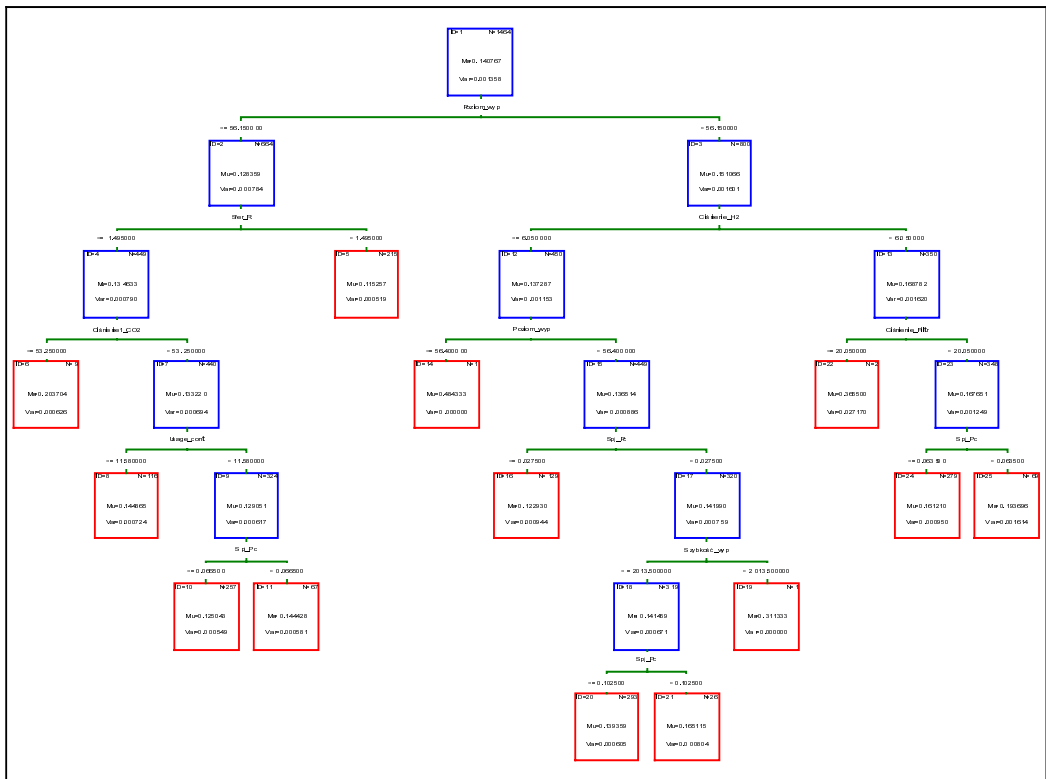




Moduł *Feature selection* automatycznie tworzy nowe źródło danych z wyselekcjonowanymi predyktorami – właśnie na tym źródle danych będziemy budować model (rys. powyżej).

Standardową procedurą *data mining* jest podział zbioru danych na próbę uczącą i testową. Dane z próby uczącej służą do budowy modelu (doboru jego parametrów). Natomiast dla danych testowych stosujemy gotowy model i sprawdzamy, na ile trafne są przewidywania modelu. W naszej analizie próbę uczącą będzie stanowiło 70% losowo wybranych przypadków, a pozostałe 30% będzie stanowiło próbę testową.

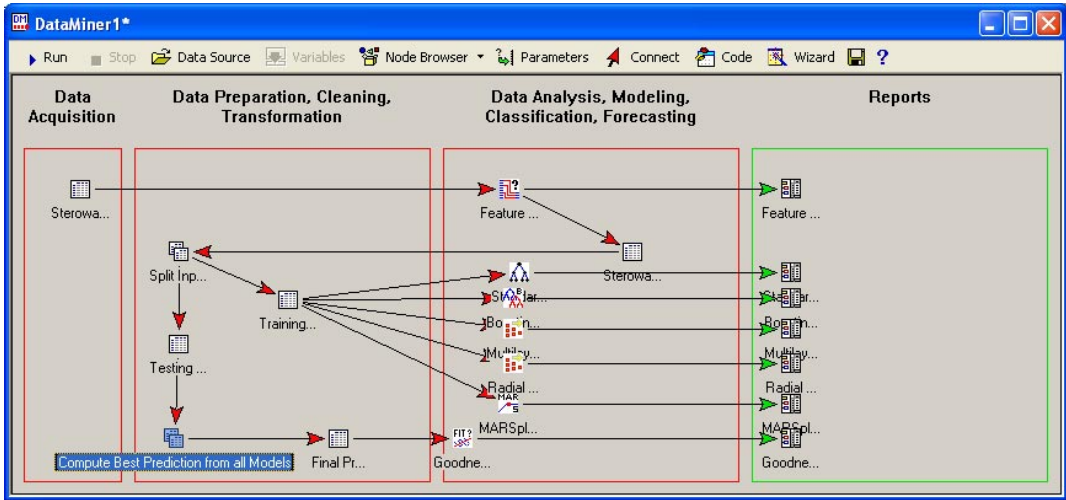
Jako metod modelowania użyjemy drzew regresyjnych (algorytm typu C&RT), sieci neuronowych (perceptronu wielowarstwowego i RBF), metody MARSplines i drzew regresyjnych ze wzmacnianiem. Drzewa regresyjne są najprostszą, najszybszą i najłatwiejszą w interpretacji z tych metod. Sieci neuronowe i drzewa ze wzmacnianiem są w stanie opisać nawet bardzo złożone zależności, jednak dopasowanie modelu jest czasochłonne, a możliwości jego interpretacji ograniczone. Metoda MARSplines jest pewnym kompromisem między mocą predykcyjną a interpretowalnością modelu i szybkością przetwarzania.



Na powyższym rysunku widzimy drzewo regresyjne uzyskane przy minimalnej liczbie obiektów w dzielonym węzle równej 300.



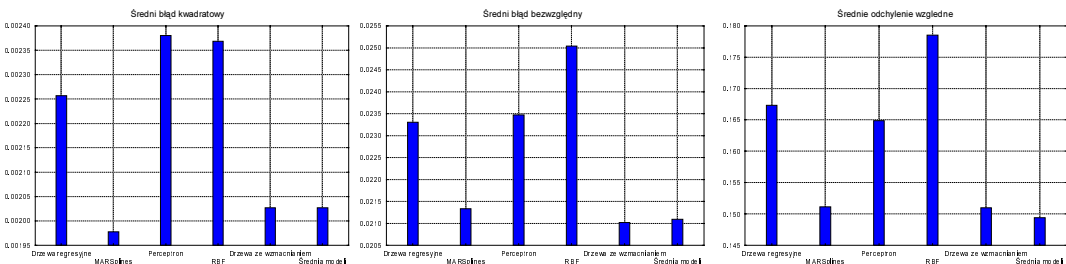
Kompletny projekt *STATISTICA Data Miner* widzimy na rysunku poniżej. Zwróćmy uwagę, że węzeł *Compute Best Prediction from all Models* stosuje model dla danych testowych i buduje model zagregowany: średnią przewidywań wszystkich modeli.



Do oceny jakości modeli i wyboru najlepszego z nich użyjemy modułu *Goodness of fit*. Jako wskaźników jakości modelu użyjemy trzech wielkości:

1. Średniego błędu kwadratowego (dla każdej obserwacji z próby testowej obliczamy różnicę między wartością przewidywaną i obserwowaną, podnosimy ją do kwadratu i obliczamy średnią dla całej próby testowej),
2. Średniego błędu bezwzględnego (dla każdej obserwacji z próby testowej obliczamy wartość bezwzględną różnicy między wartością przewidywaną a obserwowaną i obliczamy średnią dla całej próby testowej),
3. Średniego odchylenia względnego (dla każdej obserwacji z próby testowej obliczamy wartość bezwzględną różnicy między wartością przewidywaną a obserwowaną, dzielimy ją przez wartość obserwowaną i obliczamy średnią dla całej próby testowej).

Na poniższym rysunku widzimy, jak zmieniają się te wskaźniki dla zastosowanych metod.





Najlepszy model danych metoda MARSplines, która ma najmniejszy średni błąd kwadratowy, a inne wskaźniki niewiele gorsze od dużo bardziej skomplikowanych metod: drzew ze wzmacnianiem i modelu zagregowanego (średniej wszystkich modeli).

Model procesu uzyskany metodą MARSplines możemy zapisać w postaci kodu PMML (jest to dialekt XML opracowany specjalnie dla stosowania i przenoszenia modeli *data mining* między różnymi aplikacjami) lub kodu języka C w celu stosowania modelu dla nowych danych (także poza środowiskiem *STATISTICA*).

## Literatura

- [1]. Berry M.J.A., Linoff G., *Data mining techniques: for marketing, sales, and customer support*, John Willey & Sons 1997.
- [2]. Braha D. (ed), *Data Mining for Design and Manufacturing. Methods and Applications*, Kluwer Academic Publishers 2001.
- [3]. Giudici P., *Applied Data Mining. Statistical Methods for Business and Industry*, John Wiley & Sons Ltd, 2003.
- [4]. Weiss S.M, Indurkha N., *Predictive data mining. A practical guide*, Morgan Kaufman Publishers 1998.
- [5]. Berry M.J.A., Linoff G., *Mastering data mining*, John Willey & Sons 2000.
- [6]. Montgomery D. C., "Introduction to Statistical Quality Control", wyd. III, John Wiley & Sons, Inc.