



MODELOWANIE PROCESÓW PRODUKCYJNYCH

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

Tematem artykułu jest tworzenie modeli procesów produkcyjnych za pomocą technik analizy danych: statystyki i *data mining*. W „Słowniku języka polskiego” (PWN 1979) model definiowany jest jako „Układ względnie odosobniony, możliwie mało skomplikowany, działający analogicznie do oryginału, którym może być istota żywa, maszyna, zakład przemysłowy, organizacja społeczna itd.”. Ta poważna definicja nie powinna przesłonić tego, że w codziennym życiu bardzo często wykorzystujemy modele nie tylko do pracy, ale również do zabawy.

W zależności od przeznaczenia model dokładniej przedstawia pewne cechy rzeczywistego obiektu, pomijając inne, które mogą być kluczowe dla innego zastosowania. Przykładowo jeśli interesuje nas opór powietrza dla nowo projektowanego samochodu, to nasz model musi dokładnie odzwierciedlać kształt nadwozia, a silnik i cały układ napędowy nie jest istotny. Natomiast jeśli chcemy ocenić układ napędowy (np. zbadać przebieg momentu obrotowego), to kształt nadwozia jest zbędnym szczegółem.

Zajmiemy się statystycznymi modelami procesów produkcyjnych. Są to modele mające postać zbioru reguł logicznych lub równań, uzyskane na podstawie danych z przeszłości uwzględniające losowość właściwości procesu. Przykładem modelu statystycznego jest np. równanie (więcej informacji o tym modelu znajduje się w podręczniku [1]):

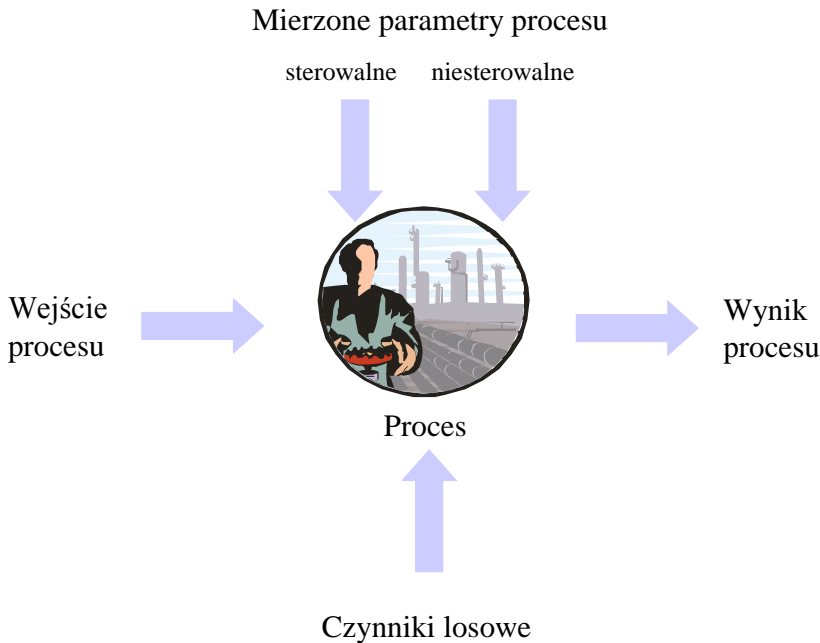
$$\text{Wytrzymałość} = 16,3 + 1,57 \cdot \text{ciśnienie formowania} + 4,16 \cdot \text{stężenie kwasu} + \varepsilon$$

W modelu *Wytrzymałość* jest zmienną zależną (lub objaśnianą), a *ciśnienie formowania* oraz *stężenie kwasu* predyktorami (używane są również nazwy zmienne niezależne, objaśniające lub predyktory). Losowość uwzględniamy poprzez składnik ε , który oznacza błąd losowy i zawiera w sobie m.in. wpływ niemierzonych i niemierzalnych czynników. Z praktycznego punktu widzenia wartości ε powinny być niewielkie, a w przypadku stosowania tradycyjnych metod statystycznych często zakłada się, że ma on rozkład normalny o wartości oczekiwanej 0.

Szersze i bardzo przystępne wprowadzenie do modelowania statystycznego znajduje się w podręczniku [2].

Na schemacie poniżej widzimy zmienne (właściwości, czynniki) dotyczące procesu. Najczęściej model będzie miał za zadanie odtworzyć zależności między parametrami na

wejściu procesu (np. cechami surowców wykorzystywanych w procesie) i mierzonymi właściwościami procesu (sterowalnymi, takimi jak np. ciśnienie w reaktorze i niesterowalnymi takimi jak np. ciśnienie atmosferyczne) a wynikami procesu (np. parametrami finalnego produktu, wydajnością procesu itd.). Inne przykłady zmiennych dotyczących procesu znajdują się w artykule [4].



W praktyce zazwyczaj mamy do dyspozycji bardzo dużo zmiennych, jednak najczęściej tylko niewielka część z nich jest ważna. Jednym z celów budowy modelu jest właśnie znalezienie tych ważnych zmiennych.

Przeznaczenie modeli możemy podzielić na dwie zasadnicze grupy:

1. Opisanie danych i odkrycie ważnych zależności, prawidłowości i wzorców.
2. Przewidywanie wartości zmiennych wyjściowych.

Modele opisowe powinny być łatwe do zrozumienia przez człowieka (wyklucza to podejścia typu „czarna skrzynka”, gdzie zależności między zmiennymi są niejawne). Model opisowy nie musi z dużą trafnością przewidywać wartości zmiennej zależnej, a w niektórych metodach (tzw. nieukierunkowanych lub bez nauczyciela) nawet nie wyróżniamy zmiennej zależnej.

Ciekawym przypadkiem zastosowań typu 1 są zadania, w których naszym celem jest stwierdzenie, które zmienne istotnie wpływają na zmienną zależną – przykład takiego właśnie modelu przedstawimy w dalszej części artykułu.



Wiedzę o wpływie poszczególnych zmiennych na właściwości finalnego produktu możemy wykorzystać do ustalenia dopuszczalnego zakresu zmienności i dokładności sprawdzania zmiennych. Jeśli jakiś czynnik nie wpływa na jakość produktu, możemy ustalić dla niego szerokie granice specyfikacji i granice kontrolne oraz rzadziej dokonywać jego pomiarów, natomiast większy nacisk położyć na naprawę ważne zmienne.

Modele powinny dobrze opisywać typowy przebieg procesu. Jeśli więc pojawi się jednostka lub partia, która jest źle opisywana przez sprawdzony, dobrze do tej pory spisujący się model, jest to sygnał o możliwym rozregulowaniu procesu. Na tej zasadzie działają karty kontrolne Shewharta (por. [3]). Przyjmują one bardzo prosty model: właściwość procesu ma stałą średnią i zmienność, które wyznaczamy na podstawie zebranych wcześniej pomiarów. Aktualna wartość właściwości (dla partii lub jednostki) może wahać się losowo zgodnie z założonym rozkładem (zazwyczaj przyjmuje się rozkład normalny dla właściwości i dopuszczalne odchylenie ± 3 sigma). Jeśli pojawi się wartość spoza dopuszczalnego zakresu, to mamy sygnał o rozregulowaniu. Takie podejście można uogólnić na bardziej złożone, wielowymiarowe zależności – przykładem takiego podejścia jest *MSPC (Multivariate Statistical Process Control)* stosowane do wykrywania rozregulowań dla procesów o setkach, a nawet tysiącach właściwości (więcej informacji można znaleźć w artykule „Monitorowanie i sterowanie jakością procesów wsadowych” w [5]).

W przypadku gdy naszym celem jest przewidywanie wartości zmiennej, często możliwości interpretacji uzyskanego modelu schodzą na dalszy plan i stosujemy złożone obliczeniowo i trudne w interpretacji metody, takie jak sieci neuronowe.

Model przewidujący wartości zmiennych często wykorzystuje się do sterowania procesem. Przykładowo jeśli prognozowana wartość właściwości procesu jest nieodpowiednia, to możemy skorygować ustawienia parametrów procesu, tak aby usunąć problem. Inny model może nam podpowiedzieć, który parametr procesu należy zmienić i o jaką wartość.

W przypadku procesów wieloetapowych model może przed zakończeniem procesu przewidywać, czy finalny produkt będzie spełniał wymagania. Jeśli nie, to pomijamy kolejne etapy procesu, ponieważ wiemy, że i tak nie uzyskamy użytecznego produktu. Takie zastosowania występują np. w przemyśle półprzewodnikowym.

Sposób tworzenia modeli

Autor artykułu [4] zaproponował dwa sposoby tworzenia modeli

1. Odkrywanie zależności.
2. Testowanie.

W pierwszym z nich badamy wyjścia procesu, aby stwierdzić czy ich zmiany są czysto losowe i wykryć ewentualne systematyczne przyczyny zmian. Innymi słowy naszym celem jest **odkrycie** czynników najsilniej wpływających na badaną zmienną.

Do odkrywania zależności stosujemy rozmaite wykresy (zwłaszcza wykres sekwencji), karty kontrolne i interakcyjne narzędzia eksploracji danych.



Wynikiem takiego badania może być wniosek o konieczności mierzenia i zbierania wartości jakiejś zmiennej w celu wykorzystania ich w modelu.

W podejściu testowym zaczynamy od określenia zbioru zmiennych, które są dostępne i potencjalnie wpływają na zmienną wyjściową. Następnie, stosując odpowiednią technikę analizy danych (statystyki lub *data mining*), stwierdzamy, czy zmienne te istotnie wpływają na zmienną zależną. W podejściu tym najczęściej stosowane metody to: regresja liniowa, analiza korelacji, regresja logistyczna i dyskryminacyjna.

Innym podziałem modeli jest rozróżnienie na klasyczną statystykę i *data mining*. Stosowanie metod statystycznych, także do modelowania, reguluje metodyka Six Sigma. Przegląd przewidzianych przez nią metod można znaleźć w podręczniku [6], a informacje o zastosowaniach *data mining* w przemyśle w [5]. Zasadnicze różnice między *data mining* a tradycyjnymi metodami to mniej restrykcyjne podejście do założeń modeli w przypadku *data mining*, częstsze stosowanie podejścia typu czarna skrzynka i używanie algorytmów uczących się. W *data mining* zazwyczaj człowiek podejmuje mniej decyzji, a więcej działań odbywa się automatycznie. Ponadto w modelach *data mining* dysponujemy zwykle większymi zbiorami danych.

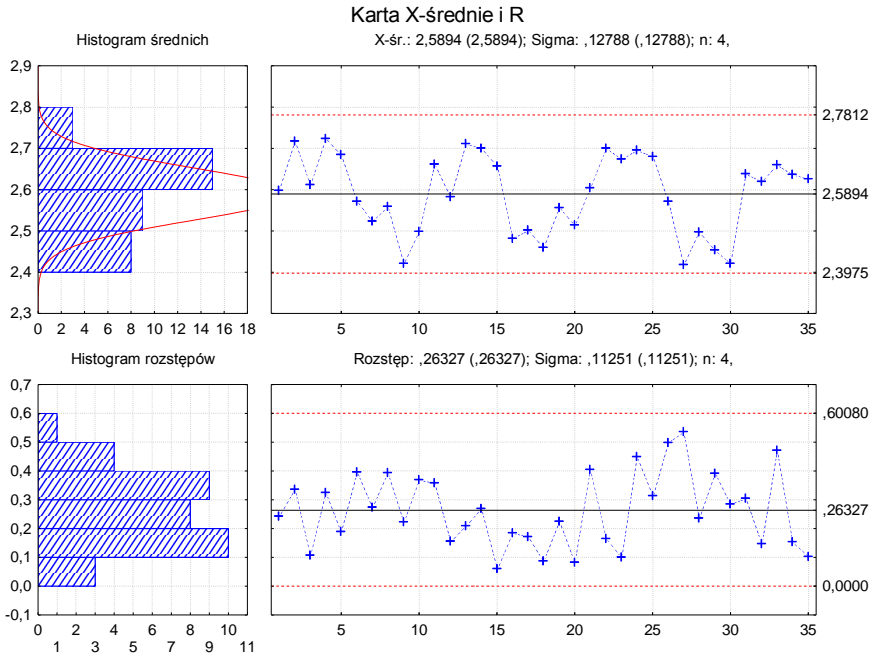
Specjalną, całościową strategią budowy, oceny i stosowania modeli jest opracowany przez Caterpillar PROCEED™ (opisany na stronach: www.statsoft.pl/press/caterpillar.html oraz proceed.statsoft.com). System ten jest tematem artykułu „PROCEED - modelowanie, optymalizacja i symulacja złożonych procesów produkcyjnych” w dalszej części niniejszej publikacji.

Przykłady

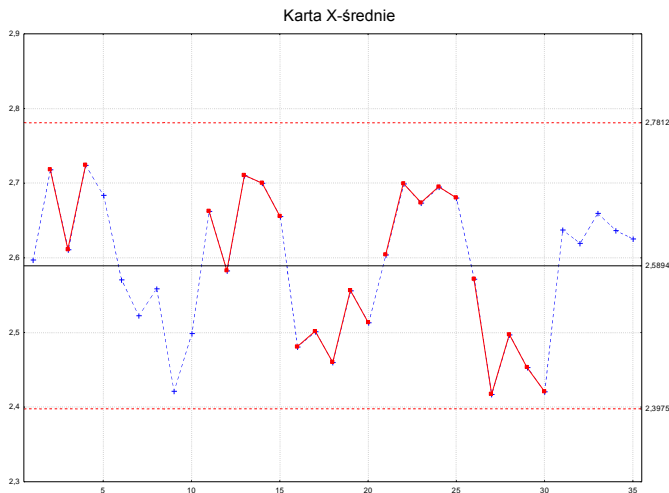
Różnicę między dwoma sposobami tworzenia modelu zobaczymy na dwóch prostych przykładach.

Najpierw zastosujemy „odkrywanie zależności”. Dla pewnej właściwości procesu chcemy stwierdzić, czy wykazuje ona wyłącznie losowe wahania wokół średniej, a jeśli nie, to jakie czynniki wpływają na jej wartości.

Zacniemy od zwykłej karty kontrolnej \bar{X} -średnie i R (znajduje się ona na rysunku poniżej). Na pierwszy rzut oka wydaje się, że nie ma przyczynowej zmienności, bo na karcie brak sygnałów o rozregulowaniu. Jednak po bliższym przyjrzeniu się karcie wartości średniej zauważymy podejrzane „falowanie” średniej z próbek. Również histogram średnich w próbkach wydaje się odbiegać od normalnego.



Losowość sekwencji średnich sprawdzimy za pomocą testów konfiguracji. Na rysunku poniżej widzimy, że po włączeniu wyświetlania wyników tych testów na kracie pojawiło się wiele sygnałów o rozregulowaniu procesu.



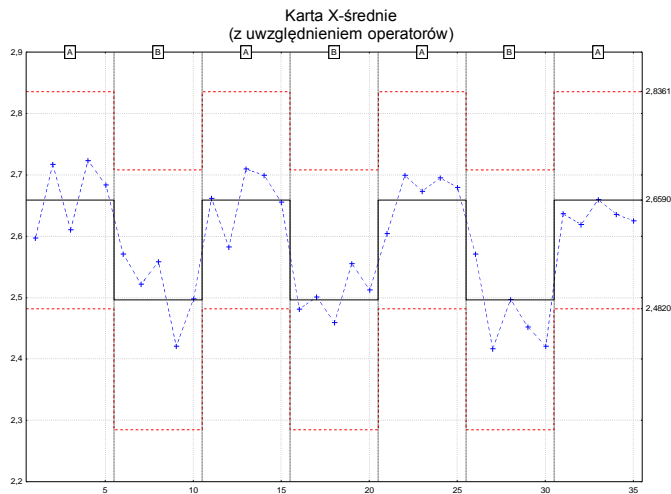
W poniższej tabeli znajduje się lista testów konfiguracji. Mamy cztery serie próbek, które są poza strefą C (najbliższą linii centralnej). Takie zachowanie jest typowe dla procesów, które mają wiele źródeł (np. mierzone produkty pochodzą z dwóch maszyn) lub pewien



czynnik włącza się i wyłącza regularnie, w pewnych odstępach czasu. Z karty kontrolnej możemy odczytać, kiedy takie włączanie i wyłączenie następowało, co ułatwia wytropienie czynnika wpływającego na właściwość produktu.

Test konfiguracji	Od próbki nr	Do próbki nr
9 po tej samej stronie l. centralnej	Brak	Brak
6 w trendzie rosnącym/malejącym	Brak	Brak
14 naprzemiennie w górę i w dół	Brak	Brak
2 z 3 w strefie A lub dalej	2	4
	27	29
4 z 5 w strefie B lub dalej	11	15
	16	20
	21	25
	26	30
15 w strefie C	Brak	Brak
8 poza strefą C	Brak	Brak

W naszym przypadku okazało się, że poszukiwanym czynnikiem jest operator. Na poniższej karcie, uwzględniającej wpływ operatora, widzimy, że średnia dla operatora A jest zauważalnie wyższa niż dla operatora B. Oczywiście wpływ operatora to jedna z tych rzeczy, która jest często spotykana i którą zawsze powinniśmy sprawdzać, ale identyczne podejście może wykryć czynnik, którego nie podejrzewalibyśmy o wywieranie wpływu na właściwość finalnego produktu.



Uzyskany przez nas model jest prosty: wiemy, że dla operatora A wartość właściwości jest średnio wyższa niż dla operatora B. W naszym przypadku moglibyśmy uściślić model, ale zdarza się, że wszystko, co możemy wydobyć z danych, to właśnie informacja tego typu.

Zobaczymy teraz, jak wygląda drugie podejście. Użyjemy nieco zmodyfikowanych danych z przykładów regresji omówionych w podręcznikach [1] i [6]. Interesuje nas wytrzymałość pewnego produktu. Zbiór zmiennych niezależnych tworzą *Ciśnienie*, *Stężenie*, *Przepływ* oraz *Temperatura*.

Zacniemy od analizy korelacji liniowej. W tabeli poniżej znajdują się współczynniki korelacji liniowej dla naszych zmiennych. Dostyc częstym błędem w interpretacji macierzy korelacji jest wyciąganie wniosków wyłącznie z wartości współczynników korelacji. Powinniśmy zawsze sprawdzić, czy współczynnik korelacji jest istotny statystycznie. Jeżeli nie jest, to jego wartość może być przypadkowa i tak naprawdę niewiele nam mówi. W naszej tabeli istotne współczynniki zostały oznaczone pogrubieniem. Ze zmienną zależną istotnie skorelowane są *Ciśnienie* i *Stężenie*. Zauważmy, że mamy również istotne związki między zmiennymi niezależnymi.

Zmienna	Korelacje (regresja.sta)				
	Oznaczone wsp. korelacji są istotne z $p < ,05000$				
	Wytrzymałość	Ciśnienie	Stężenie	Przepływ	Temperatura
Wytrzymałość	1,00	0,52	0,88	-0,27	0,17
Ciśnienie	0,52	1,00	0,15	0,46	0,44
Stężenie	0,88	0,15	1,00	-0,51	0,06
Przepływ	-0,27	0,46	-0,51	1,00	0,34
Temperatura	0,17	0,44	0,06	0,34	1,00

Teraz zbudujemy modele regresji wielorakiej (używany jest również termin regresja wielokrotna). Poniżej widzimy wyniki regresji uwzględniającej wszystkie zmienne. Podobnie jak w przypadku korelacji, powinniśmy zwrócić uwagę na istotność współczynników regresji. W modelu występują dwa nieistotne statystycznie współczynniki dla zmiennych i, aby uzyskać poprawny model, powinniśmy się ich pozbyć (uwaga: czasami w modelach regresji uwzględnia się współczynniki nieistotne statystycznie, jednak wymaga to „zewnątrznego” uzasadnienia, np. wynikającego z teorii fizycznej danego zjawiska lub doświadczenia).

N=20	Podsumowanie regresji zmiennej zależnej: Wytrzymałość (regresja.sta)					
	R= ,96647743 R2= ,93407863 Skoryg. R2= ,91649959 F(4,15)=53,136 $p < ,00000$ Błąd std. estymacji: 15,402					
	BETA	Błąd st. BETA	B	Błąd st. B	t(15)	poziom p
W. wolny			72,87590	67,67325	1,076879	0,298550
Ciśnienie	0,458030	0,088789	1,81518	0,35187	5,158663	0,000117
Stężenie	0,782917	0,089988	3,97608	0,45701	8,700260	0,000000
Przepływ	-0,061790	0,101435	-0,20042	0,32901	-0,609156	0,551537
Temperatura	-0,062573	0,075686	-0,15864	0,19188	-0,826747	0,421338

Stosuje się różne strategie eliminacji lub wstawiania zmiennych do modeli regresyjnych. My zastosujemy regresję krokowa wsteczną, bazująca na wartości statystki F. Poniżej widzimy wyniki tej procedury.



N=20	Podsumowanie regresji zmiennej zależnej: Wytrzymałość					
	BETA	Błąd st. BETA	B	Błąd st. B	t(17)	poziom p
W. wolny			16,27659	44,30005	0,36742	0,717842
Ciśnienie	0,396611	0,065759	1,57178	0,26060	6,03132	0,000013
Stężenie	0,819710	0,065759	4,16294	0,33396	12,46544	0,000000

Bardzo często do oceny modeli stosuje się współczynnik R^2 . Informuje on nas, jaki procent zmienności zmiennej zależnej wyjaśnia model. Jak zauważa R. D. Snee (zob. [4]), dążenie do uzyskania jak największego R^2 prowadzi do błędnego wstawiania do modelu dużej liczby zmiennych, a przecież chodzi nam o informację, co jest ważne, i uzyskanie stabilnego i odpornego modelu. W naszym przypadku model z eliminacją zmiennych ma nieco gorsze R^2 , ale jest lepszy, bo mówi nam, które zmienne są naprawdę ważne i będzie lepiej działał dla nowych danych, bo nie ma w nim „przypadkowych” współczynników.

Zauważmy, że model z nieistotnymi zmiennymi jest przykładem tzw. *przeuczenia*, o którym najczęściej mówi się w kontekście metod *data mining*. Przeuczenie polega na tym, że mamy model dobrze spisujący się dla danych, na których go zbudowano, ale dla nowych danych będzie najprawdopodobniej spisywał się dużo gorzej.

Ocena modeli

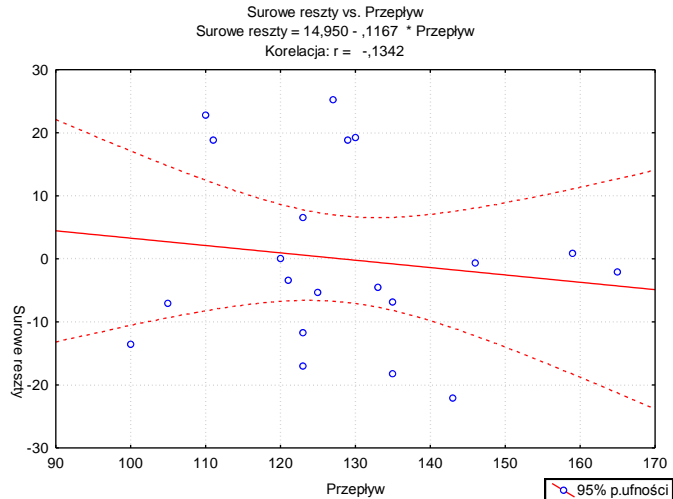
Do oceny modelu wykorzystuje się specjalnie zaprojektowany eksperyment, jednak nie zawsze jest to możliwe. W praktyce bardzo dobrym sprawdzianem jest zastosowanie modelu dla danych, których nie stosowano przy tworzeniu modelu. Jest to typowe podejście dla *data mining*, a jego wielką zaletą jest łatwa i intuicyjna interpretacja wyników. W szczególności nie musimy rozumieć metody modelowania, znać się na użytych metodach – po prostu jako wskaźnik jakości modelu dostajemy np. średnią wartość błędu dla nowych danych.

W przypadku modeli statystycznych zazwyczaj określone są wskaźniki dobroci dopasowania i można przeprowadzić rozmaite testy założeń modelu. Przykładem takich wskaźników jest współczynnik R^2 , o którym wspomnieliśmy, omawiając przykład modelu regresyjnego.

Jeśli mamy wgląd w postać modelu, to powinniśmy sprawdzić, czy wyniki są zgodne z ogólną wiedzą i doświadczeniem. Jeśli np. z modelu wynika, iż większej mocy silnika towarzyszy mniejsze zużycie paliwa, to jest to bardzo podejrzane i zapewne przy tworzeniu modelu popełniliśmy jakiś błąd.

Innym sposobem oceny modelu jest analiza reszt (tzn. różnic między wartościami przewidywanymi a obserwowanymi). Jeśli model wyjaśnia całą „przyczynową” zmienność, to reszty powinny rozkładać się czysto losowo; w przypadku wielu modeli (np. regresji liniowej) rozkład reszt powinien być normalny. Reszty nie powinny być skorelowane z żadną zmienną braną pod uwagę w analizie. Poniżej widzimy wykres

pokazujący zależności reszt w modelu regresji utworzonym w przykładzie powyżej od zmiennej *Przepływ* (zawierającym zmienne *Ciśnienie* i *Stężenie*). Jak widać reszty są rozłożone losowo, nie ma żadnego wyraźnego wzorca lub tendencji. Dopasowana prosta co prawda nie ma nachylenia 0, ale jest ono niewielkie, a przedział ufności dla położenia prostej (zaznaczony przerywaną linią) jest stosunkowo szeroki i obejmuje prostą o zerowym nachyleniu.



Jeżeli z modelu korzystamy rutynowo, przez dłuższy czas, to powinniśmy w sposób ciągły sprawdzać, czy jest on odpowiedni. Możemy do tego celu zastosować np. kartę kontrolną dla reszt modelu.

Przykład budowy modelu *data mining*

Mamy dane o pewnym ciągłym procesie. Naszym celem jest zbudowanie modelu przewidującego wartość parametru wyjściowego uzyskiwanej substancji. Proces przebiega w ten sposób, że zmiana ustawień wpływa na badaną cechę produktu z pewnym opóźnieniem. Aby wprowadzić odpowiednie korekty w procesie, musimy przewidzieć, jaka będzie przyszła wartość parametru.

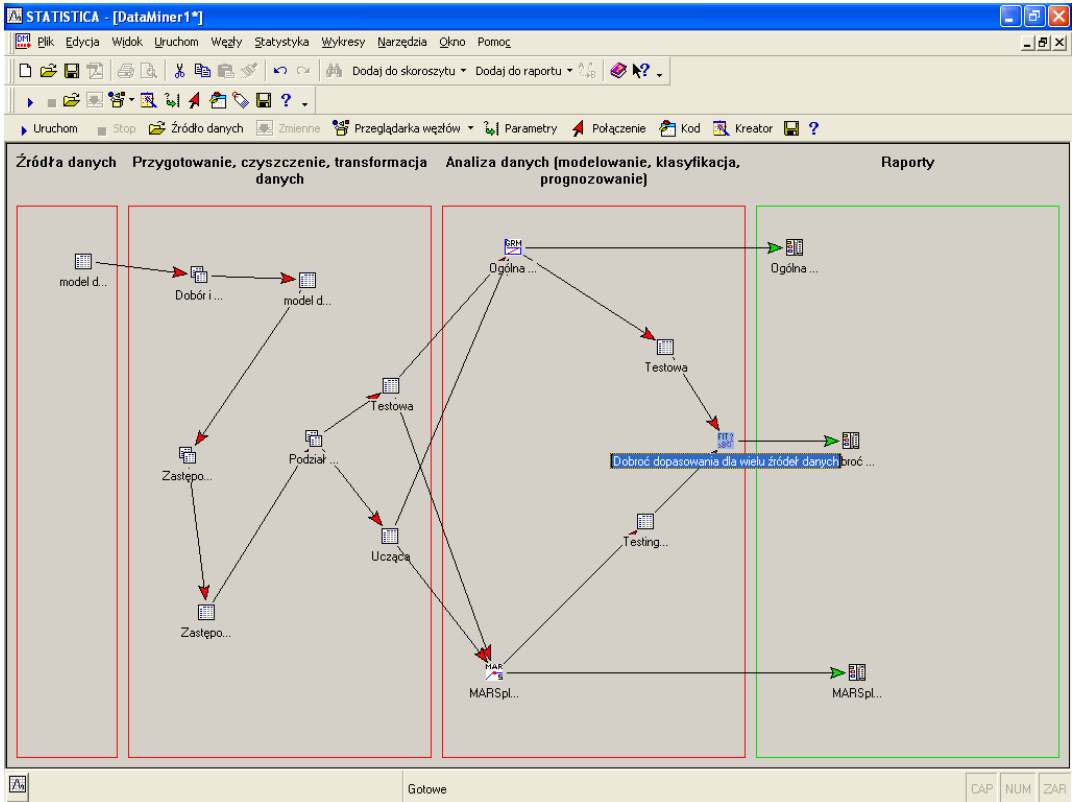
Dane mają już odpowiednią postać do prognozowania: zawierają bieżące wartości właściwości procesu oraz niektóre wartości z poprzedniego okresu, które podejrzewamy o wpływ na badaną zmienną. Łącznie dla każdej obserwacji dysponujemy 82 predyktorami (zmiennymi niezależnymi), a wszystkich przypadków jest 1298.

W przypadku tradycyjnego podejścia powinniśmy zbadać rozkład zmiennych, ich wzajemne powiązania, sprawdzić założenia metod modelowania planowanych do zastosowania w modelowaniu. Jest to dosyć pracochłonne i dlatego zastosujemy podejście *data miningowe* i metody, które nie wymagają jawnego określenia postaci modelu, spełnienia założeń itp.



Następny krok projektu to podział danych na próbę uczącą i testową. Próbę uczącą wykorzystamy do dopasowania parametrów modelu, a testową do jego oceny. Utworzymy losową próbę testową zawierającą 20% wszystkich przypadków.

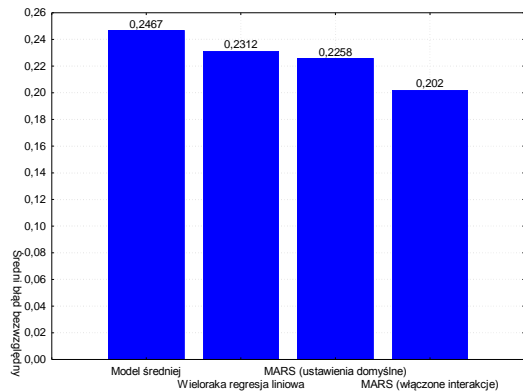
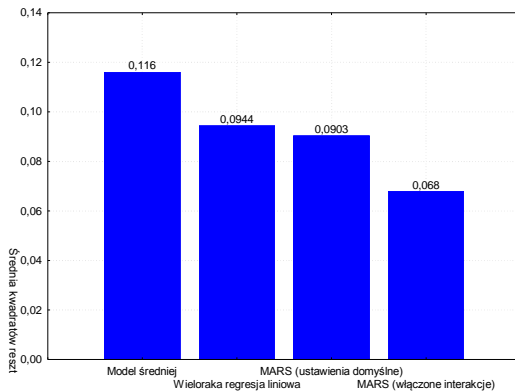
Kolejny krok to zastosowanie odpowiedniej metody modelowania. Na początek użyjemy dwóch metod: wielorakiej regresji liniowej z krokowym doбором zmiennych oraz metody MAR Splines. Do oceny jakości modeli użyjemy węzła *Dobroć dopasowania dla wielu źródeł danych*. Poniżej widzimy pełny projekt STATISTICA Data Miner.



Przy ocenie jakości modeli powinniśmy jako punkt odniesienia wziąć najprostszy model, czyli średnią w próbie testowej.

Model liniowy jest tylko nieco lepszy od naiwnego modelu średniej. Średnia kwadratów reszt jest w jego przypadku mniejsza o około 20%, a średni błąd bezwzględny o około 6% (por. rysunek poniżej).

Przy domyślnych ustawieniach dla metody MAR Splines uzyskiwany model jest niewiele lepszy niż regresja liniowa: wskaźniki błędów są lepsze tylko o kilka procent. Sytuacja zmienia się, gdy w modelu uwzględnimy interakcje 2 rzędu (pozostawiając domyślne wartości innych parametrów). Taki model MAR Splines ma wskaźniki błędów wyraźnie mniejsze, zarówno od modelu średniej, jak i modelu liniowego.



Zauważmy, że wykrycie tak silnego wpływu interakcji jest bardzo cenne. Wiemy, że na wartość parametru procesu pewne zmienne wpływają „wspólnie”, a wpływu tego nie da się opisać, patrząc oddzielnie na poszczególne zmienne. W naszym przypadku najsilniejszy wpływ na zmienną zależną wywierają interakcje opóźnionej temperatury wylotowej z temperaturą wlotową.

Model MAR Splines z interakcjami wydaje się być do zaakceptowania: jest zdecydowanie lepszy od naiwnego modelu średniej. Możemy go wykorzystać jako źródło wskazówek przy sterowaniu procesem, tym bardziej że nie chodzi nam o bardzo dokładne przewidzenie przyszłej wartości, a raczej o wychwycenie tendencji i zapobieganie niekorzystnym zmianom.

Literatura

1. J.M. Juran, *Juran's Quality Control Handbook*, wyd. IV, McGraw-Hill.
2. W.J. Krzanowski, *Statistical Modelling*, Arnold, 1998.
3. T. Greber, *Statystyczne sterowanie procesami*, StatSoft 2000.
4. R.D. Snee, *Develop Useful Models*, Quality Progress vol. 35/nr 12 (grudzień 2002).
5. *Statystyka i data mining w praktyce*, StatSoft Polska 2004.
6. F.W. Breyfogle. *Implementing Six Sigma*, Wiley 1999.
7. Podręcznik elektroniczny *STATISTICA*. StatSoft, Inc. (2005). *STATISTICA* (data analysis software system), version 7.1. www.statsoft.com.