



DATA MINING W PROGNOZOWANIU ZAPOTRZEBOWANIA NA NOŚNIKI ENERGII

Andrzej Sokółowski, Agnieszka Pasztyła

StatSoft Polska Sp. z o. o.; Akademia Ekonomiczna w Krakowie, Katedra Statystyki

Wprowadzenie

Metody prognozowania zjawisk klasyfikowane są z różnych punktów widzenia. Możemy prognozować wartość jednej zmiennej lub stan obiektu wielowymiarowego. W tym drugim przypadku prognoza może powstać z połączenia prognoz jednowymiarowych lub może być osiągnięta z modelu wielorównaniowego opisującego „jednocześnie” kształtowanie się wielu zmiennych i relacji między nimi. Powszechnie znane są tu wielorównaniowe modele makroekonomiczne.

Typ pojedynczej zmiennej prognozowanej określa w dużym stopniu zakres metod możliwych do zastosowania. Najbardziej popularne jest prognozowanie wartości zmiennej ciągłej i większość osób, myśląc o prognozowaniu, ma na myśli takie właśnie zagadnienie. Wszyscy jesteśmy odbiorcami prognozy pogody w części dotyczącej temperatury powietrza. Jeżeli zmienna prognozowana jest zmienną jakościową, to w zasadzie mamy do czynienia z zagadnieniem klasyfikacji (rozpoznania, jaki wariant prognozowanej cechy jakościowej wystąpi przy zadanej kombinacji zmiennych objaśniających). Posługując się dalej przykładem prognozowania pogody, zwróćmy uwagę na prognozę dotyczącą opadów. Zazwyczaj cecha ta ma trzy warianty: bez opadów, opady przelotne, opady ciągłe. Prognoza wskazuje nam wariant najbardziej prawdopodobny.

Pojęcie prognozowania ściśle łączymy z czasem. Mówimy przecież o prognozie pogody na okresy przyszłe, na najbliższy dzień, tydzień, ale nie na wczoraj. Takie rozumienie prognozowania jako przewidywania kształtowania się zjawisk w przyszłości jest, ogólnie rzecz biorąc, poprawne, choć zagadnienie rozpoznawania choroby na podstawie objawów jest też prognozowaniem czegoś, czego nie wiemy.

Przy prognozowaniu zjawisk rozpatrywanych w czasie mamy dwie zasadnicze grupy problemów: prognozowanie tylko na podstawie informacji o dotychczasowej historii zjawiska bądź prognozowanie oparte również na informacji dotyczącej czynników kształtujących poziom prognozowanego zjawiska, czynników towarzyszących i zakłócających.



Pierwsza grupa zagadnień jest rozwiązywana w ramach klasycznej analizy szeregów czasowych. Druga grupa – to szeroko rozumiane modele regresji.

Wiele prognozowanych zagadnień ma swoją ugruntowaną teorię i te informacje teoretyczne powinny być wykorzystywane w budowaniu modeli prognostycznych. Przykładem może być tu funkcja produkcji Cobb-Douglasa, która w najprostszej swej wersji opisuje związek pomiędzy pracą, kapitałem i wielkością produkcji. Informacja o charakterze powiązań między zmiennymi w modelach opisowych ma swoją analogię w analizie szeregów czasowych, gdzie wiele prognozowanych zjawisk ma strukturę harmoniczną ściśle powiązaną z kalendarzowym podziałem roku. Problemy prognostyczne, rozwiązywane poprzez budowę i weryfikację modelu wykorzystującego teoretyczny opis związków między zmiennymi, to zagadnienia tworzące tzw. *uczenie ukierunkowane*.

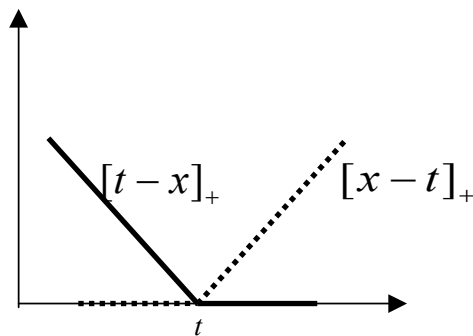
Inne podejście przewiduje poszukiwanie modelu najlepiej pasującego do danych, bez względu na to, jaki mechanizm te dane wygenerował. To jest *uczenie nieukierunkowane*. Tutaj uzyskany model nie ma służyć opisowi relacji, związków przyczynowych, a tylko prognozowaniu lub rozpoznawaniu. To są typowane zagadnienia *data mining*.

Celem tego opracowania jest przedstawienie wykorzystania metody do prognozowania zmiennych związanych z energetyką, niewymagającej wnikania w związki pomiędzy zmiennymi. Pokażemy tu wykorzystanie zarówno typowych zmiennych „czasowych”, jak i zmiennych o charakterze czynników kształtujących zmienną prognozowaną.

Koncepcja metody MARS

MARS (*Multivariate Adaptive Regression Splines*) to metoda nieparametryczna, należąca do szerszej grupy metod data mining, określanej jako *metody ukierunkowane* (ang. *Supervised learning*), i wykorzystywana do rozwiązywania problemów typu regresyjnego i klasyfikacji. Koncepcja metody rozszerza tradycyjne ujęcie zmiennych objaśniających w modelu regresyjnym. Poza całościowym uwzględnieniem wpływu predyktorów (tak jak w klasycznym modelu regresji) analizowane są wszystkie obserwacje danej zmiennej objaśniającej i obszar jej zmienności dzielony jest na przedziały, w których ma ona różny wpływ na badane zjawisko. Granice przedziałów oznaczane są za pomocą wartości progowych (tzw. węzłów: *knots*). Oznacza to, że w zależności od tego, czy wartość zmiennej objaśniającej znajduje się poniżej czy powyżej wartości progowej, to zmienna może być uwzględniana w modelu z różną wagą i różnym znakiem. Rozróżnienie wartości predyktora na mniejsze i większe od wartości progowej (t) jest dokonywane za pomocą funkcji: $\beta_i \max(0; X - t)$. Funkcja ta określana jest jako funkcja bazowa (*basis function*).

Działanie funkcji bazowej najlepiej można przedstawić za pomocą wykresu:



$$\text{gdzie: } (x-t)_+ = \begin{cases} x-t, & \text{dla } x > t \\ 0, & \text{dla } x \leq t \end{cases}.$$

Dodatkowym atutem metody MARS jest możliwość uwzględnienia interakcji pomiędzy zmiennymi objaśniającymi. Własność tę najlepiej można zilustrować przykładem. Wiadomo, że zapotrzebowanie na energię elektryczną np. w niedzielę o godz. siódmej rano jest inne niż zapotrzebowanie o tej samej godzinie w poniedziałek. Jednak pobór energii w poniedziałek wielkanocny będzie się różnił znacznie od poniedziałków „powszednich”. W klasycznym modelu regresji trudno jest odzwierciedlić zależności pomiędzy powyższymi zmiennymi: godziną, dniem tygodnia i świętem. Natomiast postać funkcyjna MARS pozwala na uwzględnienie interakcji między predyktorami. Powyższą sytuację możemy opisać za pomocą iloczynu funkcji bazowych

$$\max(0, \textit{godzina} - t_1) \cdot \max(0, \textit{dz_tygodnia} - t_2) \cdot \max(0, \textit{swieto} - t_3).$$

W ten sposób uzyskujemy lepsze dopasowanie prognozy do rzeczywistego układu czynników kształtujących prognozowane zjawisko.

Ogólnie postać funkcji MARS uzyskuje się poprzez sumowanie funkcji bazowych oraz iloczynów tych funkcji z odpowiednimi wagami, określonych wspólnie oznaczeniem $h_m(X)$:

$$f(X) = \beta_0 + \sum_{m=1}^M h_m(X).$$

Teoretycznie model ten można przedstawić jako ważoną sumę wybranych funkcji bazowych spośród wszystkich dostępnych funkcji bazowych uwzględniających wszystkie wartości zmiennych objaśniających. Innymi słowy, mamy do dyspozycji $2Nk$ możliwych funkcji bazowych, gdzie N oznacza liczbę obserwacji, a k liczbę predyktorów. Cyfra 2 wynika ze znaku funkcji bazowej (ujemny lub dodatni). Z tego zbioru, za pomocą specjalnie zaprojektowanego algorytmu, przeszukujemy przestrzeń obserwacji w celu wyznaczenia wartości progowych (węzłów) i interakcji między zmiennymi. W oparciu o wybrane węzły budowane są funkcje bazowe, które wraz z odpowiednimi wagami składają się na opis badanego zjawiska.

Algorytm

Pierwotny algorytm, zaproponowany w 1991 roku przez J. Friedmana [1], został zaprojektowany dla średniej wielkości zbiorów danych ($50 \leq N \leq 1000$), o maksymalnie 20 zmiennych objaśniających. Algorytm MARSplines dostępny w *STATISTICA Data Miner* umożliwia wykorzystanie dowolnej liczby obserwacji i zmiennych objaśniających oraz wielu zmiennych wyjściowych.

Algorytm jest oparty na metodzie rekurencyjnego podziału przestrzeni cech (*recursive partitioning*) [2] i składa się z dwóch etapów następujących na przemian, aż do osiągnięcia zadanego kryterium stopu. W pierwszym kroku zwiększana jest złożoność modelu przez dodawanie funkcji bazowych, dopóki nie zostanie wyczerpana maksymalna liczba funkcji zadana przez użytkownika. Następnie uruchamiana jest procedura usuwania najmniej istotnych funkcji bazowych z modelu, tzn. takich, których usunięcie spowoduje najmniej spadek miary dobroci dopasowania. Ten etap nosi nazwę *przycinanie* (ang. *pruning*) i jest opcjonalny w algorytmie MARSplines. Zatrzymanie algorytmu następuje po osiągnięciu minimum współczynnika określanego nazwą *uogólniony błąd w ocenie krzyżowej* (ang. *Generalized Cross Validation error - GCV*). Jest to miara dobroci dopasowania modelu do danych rzeczywistych, która uwzględnia nie tylko wielkość błędu resztowego, ale również stopień złożoności modelu:

$$GCV = \frac{\sum_{i=1}^N (y_i - f(x_i))^2}{\left(1 - \frac{C}{N}\right)^2} \quad \text{i } C = 1 + cd,$$

gdzie N oznacza liczbę przypadków w zbiorze danych, a d określa liczbę stopni swobody, która jest równa liczbie niezależnych funkcji bazowych. Wielkość c jest „karą” za dodanie kolejnej funkcji bazowej do modelu.

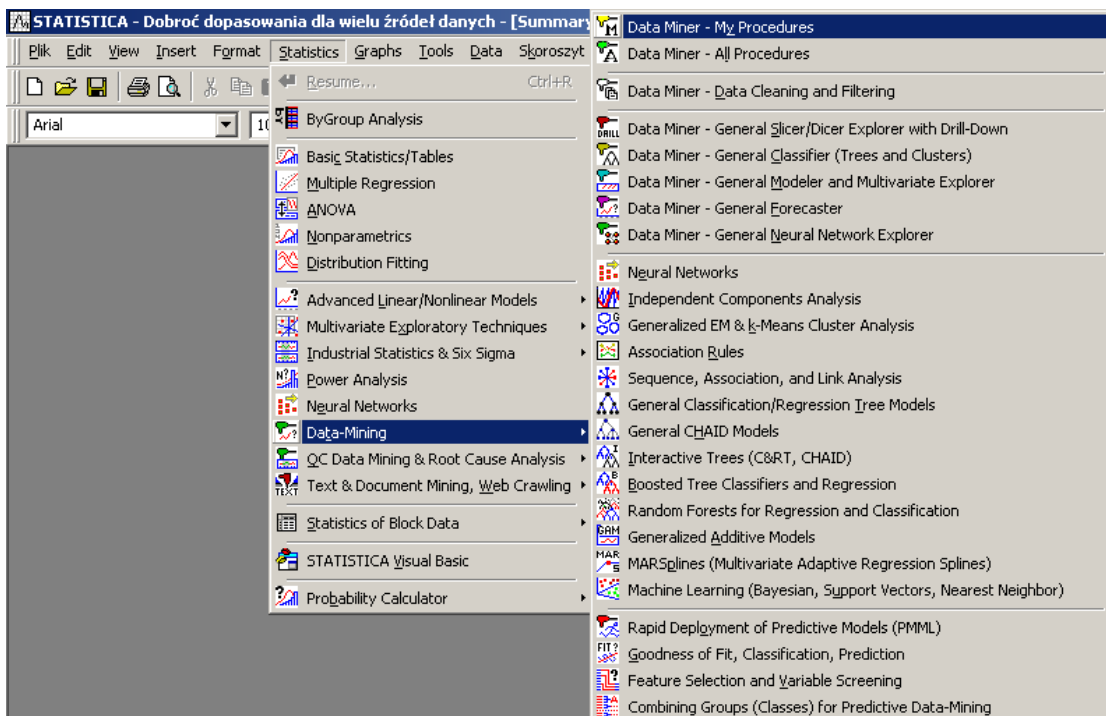
Przykład

Przykład dotyczy prognozowania zapotrzebowania na energię ciepłą. Dane obejmują okres czterech miesięcy (od stycznia do kwietnia) i zostały zebrane przez jedną z elektrociepłowni. Dla zachowania poufności, wartości niektórych zmiennych zostały zmienione. Zmiennymi prognozowanymi są przepływ rzeczywisty (zapotrzebowanie na energię ciepłą), moc rzeczywista i temperatura powrotu (wody) - zmienne ilościowe. Do zmiennych objaśniających należą:

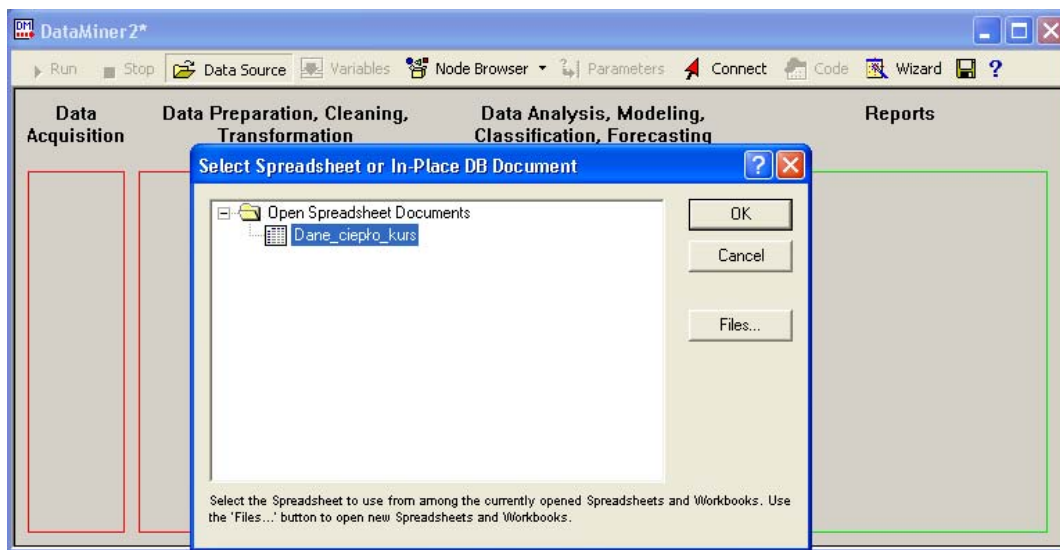
- ◆ zmienne ilościowe: temperatura żądana od EC, temperatura wymagana z tabeli regulacyjnej, temperatura zasilania, temperatura dobowa, temperatura godzinowa, wiatr (prędkość);
- ◆ zmienne jakościowe: godzina, miesiąc, dzień miesiąca, dzień tygodnia, dni robocze, zachmurzenie.



Łącznie zbiór danych zawiera 2880 przypadków. W celu zbudowania modelu predykcyjnego w menu *Statystyka* wybieramy moduł *Data-Mining*, a następnie *Data Miner - My Procedures*.

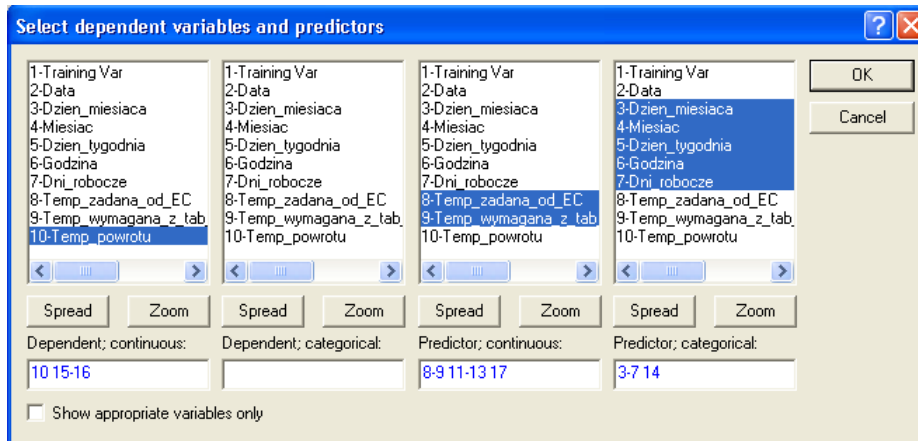


W nowym projekcie definiujemy źródło danych (*Data source*).

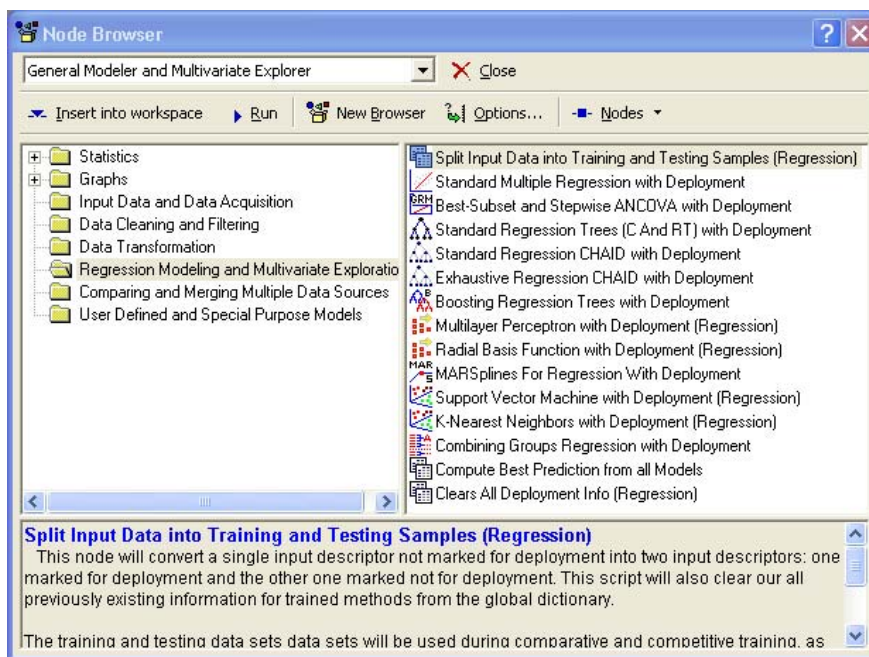





Gdy zbiór danych zostanie umieszczony w obszarze definiowania danych, wówczas możemy określić zmienne analizy. Klikamy prawym przyciskiem myszy na ikonę pliku danych i wybieramy opcję *Variable Selection* (*wybór zmiennych*). Zmienne definiujemy zgodnie z niżej podaną specyfikacją.

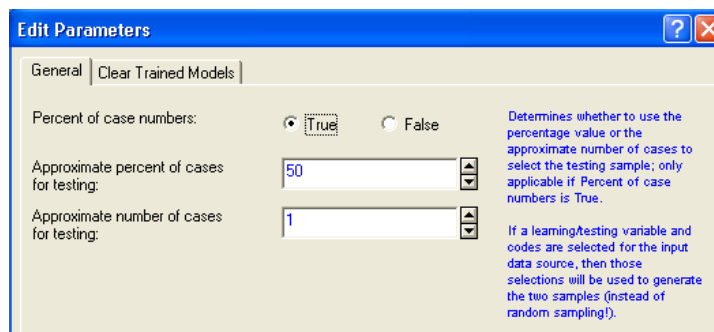



Następnie należy podzielić dane na zbiór uczący i testowy. Podziału dokonujemy za pomocą węzła *Split Input Data into Training and Testing Sets (Regression)* dostępnego w grupie *Regression Modelling and Multivariate Exploration*. W tym celu otwieramy przeglądarkę węzłów (*Node Browser*) i wybieramy opcję *General Modeler and Multivariate Explorer*. Następnie wstawiamy wybrany węzeł i umieszczamy go w obszarze roboczym za pomocą przycisku **Insert into workspace**.

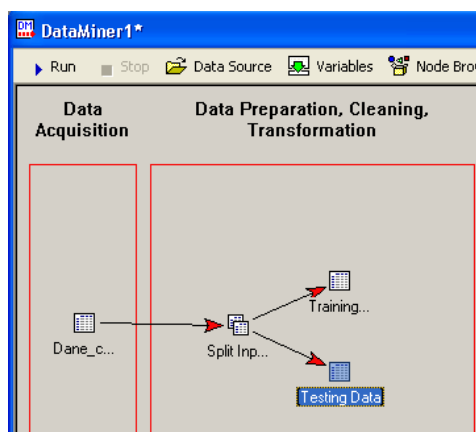






Gdy węzeł znajduje się w polu transformacji danych, należy określić liczbę przypadków, które będą należeć odpowiednio do zbioru testowego i uczącego. Aby podzielić przypadki na podzbiory, klikamy prawym przyciskiem myszy na ikonie węzła i wybieramy opcję  Edit Parameters... . Następnie w zakładce *General* określamy procentowy udział zbioru testowego (*Approximate percent of cases for testing*) lub jego liczebność (*Approximate number of cases for testing*).



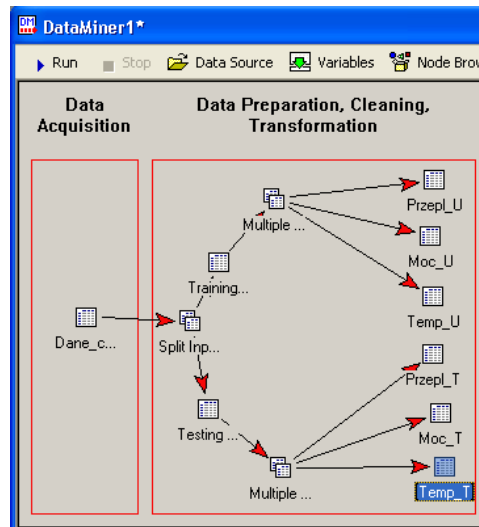
W trzecim kroku uruchamiamy działanie węzła za pomocą przycisku  Run umieszczonego w głównym menu i tworzymy dwa podzbiory danych: uczący i testowy.



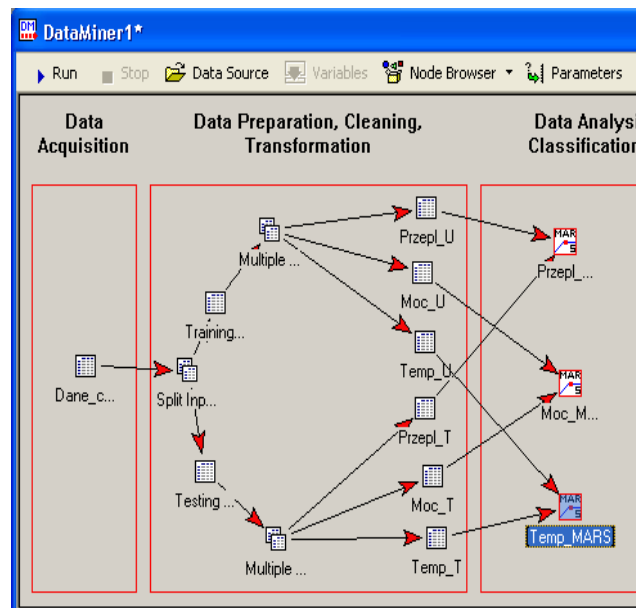
Nasz model będzie obejmował trzy zmienne prognozowane, dlatego dla każdej z nich należy przygotować osobną kopię zbioru uczącego i testowego. W tym celu wykorzystujemy węzeł *Multiple Copies of Data Source* dostępny w przeglądarce węzłów – opcja *General Modeler and Multivariate Explorer*, grupa *Input Data and Data Acquisition*. Po zaznaczeniu pliku (kliknięciem na ikonie), dla którego chcemy utworzyć kopie, otwieramy przeglądarkę węzłów i dodajemy wybrany węzeł. Czynność powtarzamy dla obydwóch podzbiorów. Następnie tworzymy trzy kopie, korzystając – jak poprzednio – z opcji  Edit Parameters... i określając ich liczbę dla każdego podzbioru osobno. Jako rodzaj kopiowania (*Type of copy operation*) wybieramy *tworzenie dodatkowej kopii źródła danych* (*Clone original data source*) i klikamy przycisk  Run . Otrzymujemy sześć plików



z danymi. Aby nie pomylić plików w dalszej analizie, warto na tym etapie nazwać je zgodnie z przeznaczeniem. Na potrzeby przykładu plikom zostały nadane nazwy zmiennych prognozowanych, z zaznaczeniem, czy jest to zbiór uczący (U) czy testowy (T). Nazwę zbioru danych możemy zmienić, klikając na ikonę prawym przyciskiem myszy i wybierając opcję **Rename...** lub za pomocą klawisza **F2**.



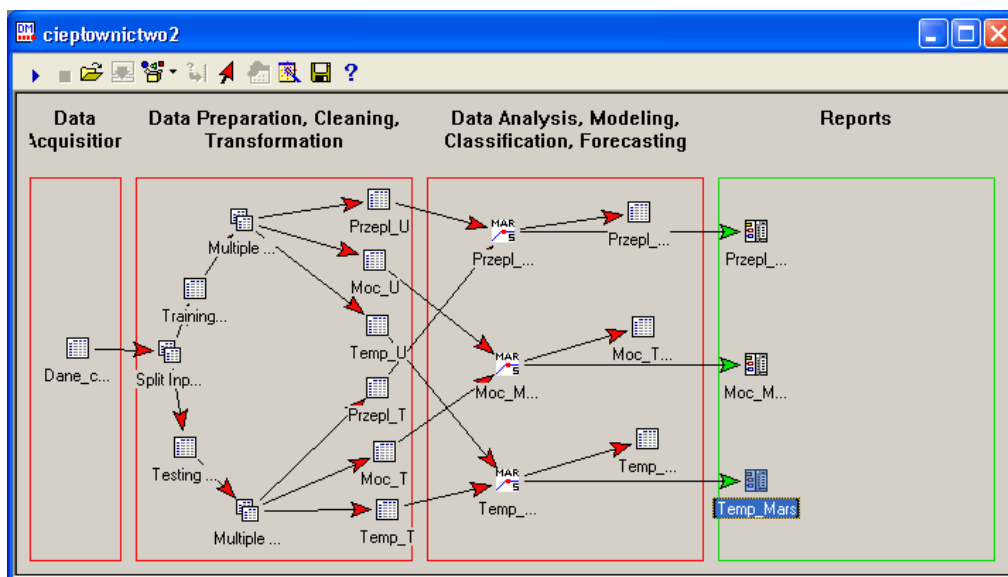
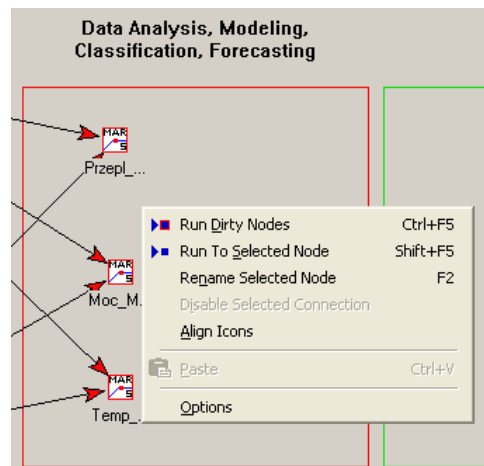
Teraz możemy przystąpić do zdefiniowania zmiennych dla każdego z utworzonych plików pod kątem tworzonych modeli predykcyjnych. Za pomocą opcji **Variable Selection...** dla każdej pary zbioru uczącego i testowego wybieramy inną zmienną objaśnianą, a następnie dodajemy osobny węzeł analizy *MARSplines*.





W celu dodania węzła analizy dla dwóch plików (o takich samych zmiennych zależnych i niezależnych) jednocześnie trzeba zaznaczyć oba za pomocą przycisku *Ctrl* i otworzyć przeglądarkę węzłów (*Node browser*), opcja *General Modeler and Multivariate Explorer*, grupa węzłów *Regression Modelling and Multivariate Exploration*. Przy przeprowadzaniu równoległe kilku analiz tego samego typu, również warto umieścić bardziej szczegółową informację o analizie w nazwie węzła. Robi się to tak samo, jak w przypadku plików danych.

W celu oszacowania parametrów modeli najpierw definiujemy parametry analizy, korzystając z opcji *Edit Parameters...*, a następnie uruchamiamy estymację za pomocą przycisku *Run Dirty Nodes* (*uruchom niezaktualizowane węzły*), który pojawia się, gdy klikniemy prawym przyciskiem myszy w puste miejsce w polu analizy.

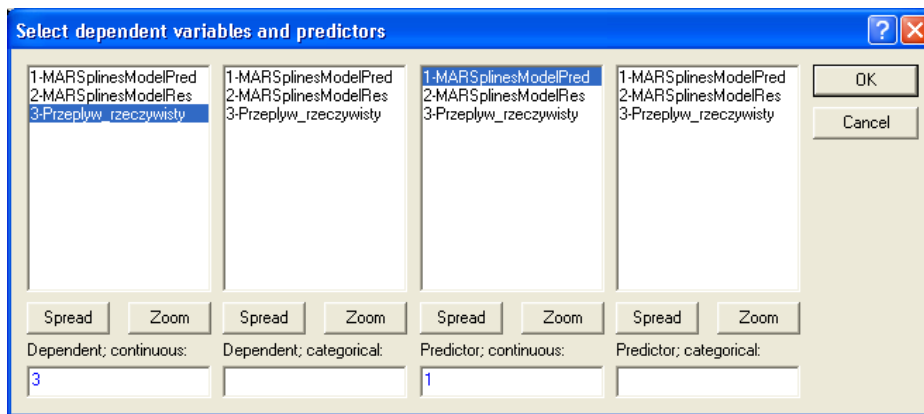




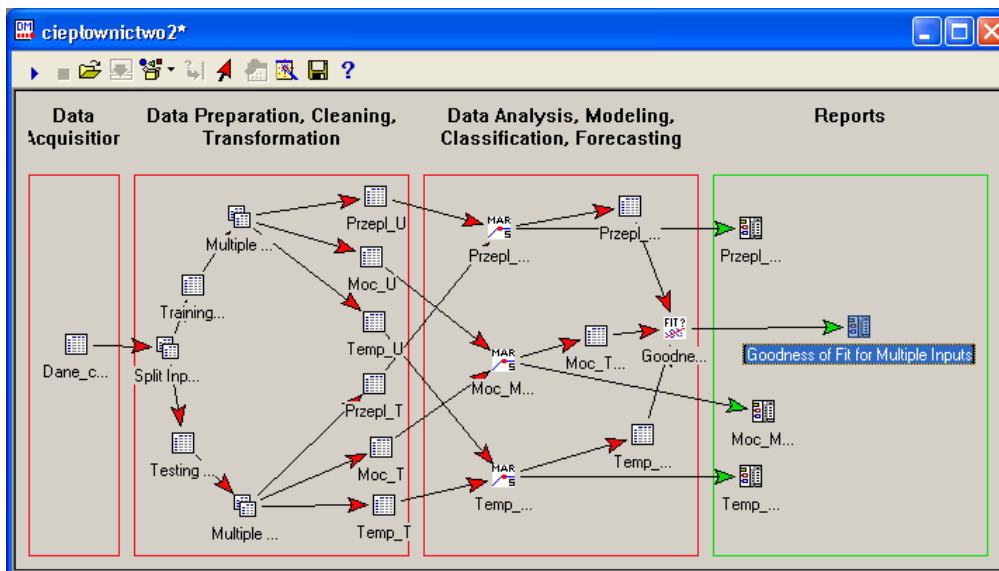
Wyniki dla każdego modelu możemy przeglądać po otwarciu raportów analiz. Przykładowe okno z podsumowaniem modelu znajduje się poniżej.

Regression statistics		Regression statistics (Sp	
Mean (observed)		Przepliw_rzeczywisty	18065.71
Standard deviation (observed)			3704.69
Mean (predicted)			18065.71
Standard deviation (predicted)			3672.92
Mean (residual)			-0.00
Standard deviation (residual)			484.15
R-square			0.98
R-square adjusted			0.98

Po oszacowaniu modeli dla trzech prognozowanych zmiennych możemy ocenić, jak sprawdzają się w zbiorze testowym. Miary dobroci dopasowania modelu oraz odpowiednie wykresy dla zbioru testowego łatwo uzyskać za pomocą węzła *Goodness of Fit for Multiple Inputs*, który znajduje się w zakładce *Statistics* → *Data Mining* → *Goodness of Fit*. Jednakże, aby uzyskać poprawne wyniki, należy dla każdego pliku prawidłowo określić zmienne. Rysunek poniżej ilustruje sposób definiowania zmiennych.



Po połączeniu węzła z trzema plikami testowymi, otrzymanymi po estymacji modelu, zdefiniowaniu zmiennych i uruchomieniu go, uzyskujemy miary dobroci dopasowania modeli do danych testowych.



Aby ułatwić oszacowanie jakości modeli prognostycznych, podajemy zakresy zmienności analizowanych zmiennych:

- ◆ temperatura powrotu: 20-62,
- ◆ moc rzeczywista: 190-1 500,
- ◆ przepływ rzeczywisty: 5 000-22 100.

Należy podkreślić, że powyższe obszary zmienności mogą nie odpowiadać wartościom zmiennych obserwowanym w warunkach rzeczywistych ze względu na dokonane modyfikacje danych. Z tego również powodu nie podajemy jednostek fizycznych zmiennych.

	1 Przeciętny kwadratowy błąd bezwzględny	2 Przeciętny błąd bezwzględny	3 Przeciętny kwadratowy błąd względny	4 Przeciętny błąd względny	5 Współczynnik korelacji
Moc rzeczywista	994.238475	24.2198924	0.00310880304	0.0410916045	0.988867465
Temperatura powrotu	0.876077659	0.582630639	0.000505393047	0.0135416825	0.971955068
Przepływ rzeczywisty	277336.457	416.202178	0.00117323254	0.0249308914	0.989159636

Literatura

1. Friedman J., 1991, Multivariate Adaptive Regression Splines, Annals of Statistics, 19, 1-141.
2. Gatnar H., 2001, Nieparametryczna metoda dyskryminacji i regresji, Wydawnictwo Naukowe PWN.



3. Hastie T., Tibshirani R., Friedman J., 2001, The Elements of Statistical Learning. Data mining, Inference and Prediction, Springer Verlag.
4. Zeliaś A., Pawełek B., Wanat S., 2004, Prognozowanie ekonomiczne. Teoria, przykłady, zadania. Wydawnictwo Naukowe PWN.