PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 107
Taksonomia 17 2010

Joanna Trzęsiok
Akademia Ekonomiczna w Katowicach

DOBÓR ZMIENNYCH DO MODELU REGRESYJNEGO ZBUDOWANEGO ZA POMOCĄ WYBRANYCH METOD NIEPARAMETRYCZNYCH

Streszczenie: W artykule poruszony został problem doboru zmiennych objaśniających do modelu regresyjnego, zbudowanego za pomocą wybranych nieparametrycznych metod regresji: POLYMARS oraz PPR. Przedstawione i porównane zostały dwie procedury selekcji zmiennych: eliminacja pojedynczych zmiennych oraz eliminacja blokiem. Wyniki przeprowadzonych analiz pokazują, że zastosowanie redukcji liczby zmiennych prowadzi do uzyskania modeli mniej złożonych i charakteryzujących się mniejszymi wartościami błędu średniokwadratowego niż modele zbudowane na komplecie zmiennych.

1. Wstęp

Dobór zmiennych do modelu jest klasycznym problemem podejmowanym w analizie danych. W wielu sytuacjach badacz musi zdecydować, czy w pierwszym etapie badania dokonać selekcji zmiennych i wprowadzić do modelu tylko część z nich, czy zbudować model na oryginalnym zestawie cech. Usunięcie zmiennych nieistotnych prowadzi do ograniczenia złożoności modelu. Jednak może się wiązać z utratą części informacji, co w konsekwencji doprowadzi do obniżenia dokładności predykcji modelu.

Zgodnie z klasyfikacją przedstawioną w pracy [Guyon i in. 2006] wyróżnić można trzy podejścia do problemu doboru zmiennych do modelu:

- filtrowanie zmiennych (*filters*), odbywające się na etapie przygotowania danych, niezależnie od zastosowanej metody regresji;
- symulacyjne przeszukiwanie podzbiorów zmiennych (*wrappers*), które polega na wielokrotnym budowaniu i porównywaniu modeli dla różnych zestawów zmiennych. Do metod tych zaliczamy między innymi eliminację zmiennych;
- metody zagnieżdżone (*embedded methods*), w których kryterium doboru zmiennych jest integralną częścią algorytmu metody.

Nieparametryczne metody regresji są podejściem alternatywnym w stosunku do metod klasycznych. Mają większy obszar zastosowań, ponieważ nakładają mniej restrykcyjnych założeń na badane zmienne. Ponadto wyniki przeprowadzonych badań empirycznych pokazują, że błędy prognoz modeli, zbudowanych za

pomocą tych metod, są często mniejsze niż w przypadku modeli klasycznych (zob. [Meyer, Leisch, Hornik 2002]). Wiele z metod nieparametrycznych ma wbudowane w postać algorytmu kryterium doboru zmiennych. Okazuje się jednak, że zastosowanie dodatkowo prostej procedury eliminacji zmiennych może znacznie zwiększyć dokładność predykcji tych modeli (zob. [Trzęsiok 2009]).

Nieparametryczne modele regresji działają często na zasadzie „czarnej skrzynki", co powoduje, że trudne lub wręcz niemożliwe staje się interpretowanie uzyskiwanych wyników. Jest to często największa wada tych metod. Okazuje się, że procedury doboru zmiennych objaśniających, oprócz redukcji złożoności modelu i poprawy jego dokładności predykcji, dają również możliwość pozyskiwania dodatkowej wiedzy o badanym zjawisku, a przez to otwierają dla badacza „czarną skrzynkę", jaką jest ten model.

Celem artykułu było przedstawienie i porównanie dwóch symulacyjnych metod przeszukiwania zbioru cech: eliminacji pojedynczych zmiennych oraz eliminacji blokiem. Zbadano również, jak zmienia się dokładność predykcji modelu, do którego wprowadzono tylko część zmiennych w stosunku do modelu uzyskanego na pełnym zestawie cech. Do analizy wykorzystano modele zbudowane na podstawie dwóch wybranych nieparametrycznych metod regresji: POLYMARS oraz PPR. Wszystkie obliczenia wykonano za pomocą programu statystycznego **R**.

2. Modele regresji wykorzystane w analizie

Metody doboru zmiennych, jakimi są eliminacja pojedynczych zmiennych oraz eliminacja blokiem, to metody uniwersalne, które mogą zostać wykorzystane do zagadnień zarówno regresji, jak i klasyfikacji. W artykule przedstawiono zastosowanie tych metod tylko w analizie regresji. Wykorzystano je w doborze cech do modeli zbudowanych za pomocą dwóch wybranych nieparametrycznych metod regresji:

- wielowymiarowej metody krzywych sklejanych POLYMARS (*Multivariate Adaptive Polynomial Spline Regression*),
- metody rzutowania PPR (*Projection Pursuit Regression*).

Są to jedne z najbardziej popularnych metod nieparametrycznych, które w wielu przypadkach generują modele charakteryzujące się najniższymi wartościami błędów predykcji (zob. [Meyer, Leisch, Hornik 2002]). Modele, które uzyskujemy za pomocą tych metod, są stabilne i odporne na występowanie w zbiorze uczącym szumu czy wartości oddalonych. Metody POLYMARS oraz PPR zostały szczegółowo omówiono m.in. w pracach: [Kooperberg, Bose, Stone 1997; Friedman, Stuetzle 1981; Trzęsiok 2004a; 2004b].

3. Dobór zmiennych do modelu regresyjnego

Do rozwiązania problemu wyznaczenia najlepszego zestawu cech dla wybranego modelu regresyjnego zastosować można wyczerpujące, symulacyjne przeszukiwania wszystkich możliwych podzbiorów zmiennych. Otrzymane w ten sposób roz-

wiązanie jest optymalne w sensie globalnym. Jest to jednak strategia zachłanna, która wymaga dużej mocy obliczeniowej komputera oraz czasu. Okazuje się, że wyznaczenie optymalnego, w sensie globalnym, zestawu zmiennych w przypadku zbiorów danych charakteryzowanych przez dużą liczbę zmiennych nie może być w praktyce efektywnie stosowane.

Alternatywą dla opisanego podejścia są m.in. metody eliminacji zmiennych, które polegają na przeszukiwaniu podzbiorów cech na podstawie strategii wspinaczki. Otrzymujemy w ten sposób rozwiązanie optymalne jedynie w sensie lokalnym, jednak niewątpliwą zaletą tym metod jest zarówno nieporównywalnie mniejsza złożoność algorytmów, jak i o wiele krótszy czas obliczeń.

3.1. Procedura eliminacji pojedynczych zmiennych

Procedura eliminacji pojedynczych zmiennych zaczerpnięta została z pracy [Guyon i in. 2006]. W pierwszym etapie zbudowany zostaje model na oryginalnym zbiorze wszystkich zmiennych. W każdym kolejnym kroku usuwamy jedną zmienną według ustalonego *a priori* kryterium i budowany jest model na pomniejszonym zbiorze cech. Kolejno eliminowane zostają zmienne w najmniejszym stopniu wpływające na zmienną zależną. Procedura jest powtarzana tak długo, aż w zbiorze zostanie tylko jedna cecha, ta która ma najsilniejszy wpływ na zmienną Y.

Wykorzystywanym kryterium jest tutaj minimalny błąd średniokwadratowy MSE:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}(\mathbf{x}_i)\right)^2, \tag{1}$$

gdzie y_i to obserwowalna wartość zmiennej zależnej, $\hat{f}(\mathbf{x}_i)$ – wartość teoretyczna, a n – liczba zmiennych. Błąd MSE za każdym razem obliczony jest metodą sprawdzania krzyżowego, która jest metodą estymacji. Polega ona na losowym podziale zbioru danych na k części, gdzie $k-1$ podzbiorów wykorzystanych jest do budowy modelu, pozostała zaś jedna część stanowi zbiór testowy. Na tak przygotowanym zbiorze zbudowanych zostaje k różnych modeli, które testowane są za każdym razem na innym podzbiorze. Otrzymane w kolejnych krokach wartości błędu średniokwadratowego zostają na końcu uśrednione.

Algorytm metody eliminacji pojedynczych zmiennych przedstawić można w następujących krokach:

Krok 1. Zbuduj model regresyjny na zbiorze uczącym D, wykorzystując pełen zestaw zmiennych. Utwórz pomocniczy zbiór uczący S będący kopią zbioru D.

Krok 2. Poprzez wyłączenie tymczasowo ze zbioru S kolejno każdej ze zmiennych wygeneruj wiele zmodyfikowanych zbiorów uczących na bazie S. Zbuduj na tak zmodyfikowanych zbiorach modele regresyjne.

Krok 3. Dla każdego modelu z kroku 2 oblicz metodą sprawdzania krzyżowego błąd średniokwadratowy MSE.

Krok 4. Zidentyfikuj model z wyłączoną zmienną, dla którego wartość błędu MSE jest najmniejsza, a następnie usuń ze zbioru S tę zmienną.

Krok 5. Powróć do kroku 2 i powtarzaj procedurę dopóki w S pozostanie tylko jedna zmienna.

Krok 6. Z otrzymanego ciągu modeli (z malejącą liczbą zmiennych) wybierz ten, dla którego wartość MSE jest najmniejsza. Jest to model końcowy.

3.2. Procedura eliminacji zmiennych blokiem

Procedura eliminacji zmiennych blokiem (zob. [Nagatani, Abe 2007]) jest modyfikacją metody przedstawionej w punkcie 3.1. Pozwala ona na usunięcie w jednym etapie całego bloku zmiennych, a nie tylko pojedynczej cechy.

Podobnie jak poprzednio na zmodyfikowanych zbiorach danych, z których wyłączono kolejno każdą ze zmiennych, zbudowane, a następnie porównywane pod względem błędu średniokwadratowego (liczonego zawsze metodą sprawdzania krzyżowego) zostają modele regresyjne. Następnie identyfikowane są cechy odpowiadające modelom, dla których błąd średniokwadratowy był mniejszy od MSE modelu otrzymanego na oryginalnym komplecie zmiennych. Metoda pozwala na ostateczne usunięcie z bloku wszystkich takich zmiennych, jeśli MSE modelu, zbudowanego na tak zmienionym zbiorze, jest nie większy od MSE początkowego modelu, uzyskanego na komplecie zmiennych. Oczywiście w wielu przypadkach pojedyncze usuwanie zmiennych daje model o mniejszym błędzie średniokwadratowym, jednak wyeliminowanie wszystkich tych cech jednocześnie znacznie pogarsza dokładność predykcji. Jeśli usunięcie całego bloku zmiennych powoduje wzrost błędu, to liczebność zbioru zmiennych – kandydatów do usunięcia, zmniejszamy o połowę, wybierając tę grupę, której odpowiadają najmniejsze wartości MSE.

Algorytm metody eliminacji zmiennych blokiem składa się z następujących kroków:

Krok 1. Zbuduj model regresyjny na zbiorze uczącym D, wykorzystując pełen zestaw zmiennych. Oblicz błąd średniokwadratowy tego modelu MSE_D metodą sprawdzania krzyżowego. Utwórz pomocniczy zbiór S będący kopią zbioru D.

Krok 2. Przez wyłączenie tymczasowo ze zbioru S kolejno każdej ze zmiennych wygeneruj wiele zmodyfikowanych zbiorów uczących na bazie S. Zbuduj na tak zmodyfikowanych zbiorach modele regresyjne.

Krok 3. Dla każdego modelu z kroku 2 oblicz metodą sprawdzania krzyżowego błąd średniokwadratowy MSE.

Krok 4. Zidentyfikuj wszystkie modele z wyłączoną zmienną, dla których wartość błędu *MSE* jest mniejsza niż wartość MSE_D dla modelu zbudowanego na komplecie zmiennych. Jeśli warunek nie jest spełniony dla żadnego modelu, to wszystkie zmienne są istotne, więc zakończ procedurę.

Krok 5. Usuń tymczasowo ze zbioru S wszystkie zmienne zidentyfikowane w kroku 4. Zbuduj na tym zbiorze nowy model i oblicz jego błąd średniokwadratowy.
Jeśli obliczony błąd jest mniejszy od wartości MSE_D, to zapamiętaj tak zredukowany zbiór S i powróć do kroku 2.

Krok 6. W przeciwnym przypadku przywróć zbiór S z kroku 2 i zastosuj algorytm połowienia do zidentyfikowania mniej licznego bloku zmiennych do usunięcia:
a) uporządkuj modele z kroku 4 rosnąco według wartości *MSE*,
b) tymczasowo usuń ze zbioru S pierwszą połowę zmiennych, odpowiadającą uporządkowanym w kroku 6b modelom.

Zbuduj model regresyjny na zbiorze S, z którego tymczasowo usunięto blok zmiennych o połowę mniej liczny niż uprzednio, i oblicz błąd *MSE*. Jeśli obliczony błąd jest mniejszy od wartości MSE_D, to pozostaw tak zredukowany zbiór S i powróć do kroku 2. W przeciwnym przypadku przywróć zbiór S z kroku 2 i rekurencyjnie zastosuj metodę połowienia dla mniej licznego bloku zmiennych (przejdź do kroku 6b).

4. Badania empiryczne

Do analizy porównawczej opisanych procedur doboru zmiennych wybrano dwie nieparametryczne metody regresji: POLYMARS oraz PPR. Wszystkie badania empiryczne przeprowadzone zostały z wykorzystaniem programu statystycznego

Tabela 1. Charakterystyki zbiorów danych wykorzystanych w analizie

Zbiór danych	Liczba zmiennych objaśniających	Liczba obserwacji
Triazines [1]	58	186
Peak [2]	50	200
Bank [3]	32	1000
Boston [4]	13	506

Źródło: opracowanie własne.

[1] Zbiór *Triazines* pochodzi z repozytorium instytutu „US Environmental Protection Agency ".

[2] Zbiór *Peak* został wygenerowany komputerowo za pomocą funkcji mlbench.peak zaimplementowanej w programie **R**.

[3] Zbiór *Bank* pochodzi z repozytorium „DELVE – Data for Evaluating Learning In Valid Experiments".

[4] Dane w zbiorze *Boston* zostały zebrane w roku 1978 przez Harisona i Rubinfelda. Publikacja w artykule [Harrison, Rubinfeld 1978].

R. Badanie wykonano na zbiorach danych standardowo wykorzystywanych do badania własności różnych nieparametrycznych metod regresji. Wybrane charakterystyki tych zbiorów przedstawione zostały w tab. 1.

Tabela 2. Wynik działania procedur doboru zmiennych dla zbioru *Bank* dla metody PPR

Eliminacja pojedynczych zmiennych						Eliminacja zmiennych blokiem		
etap	usunięta zmienna	*MSE* modelu	etap	usunięta zmienna	*MSE* modelu	etap	usunięte zmienne	*MSE* modelu
0	nic	0,0106	17	27	0,0081	0	nic	0,0106
1	1	0,0091	18	14	0,0081	1	1, 4, 5, 7, 8, 15, 18, 21, 23, 24, 25, 27, 28, 29, 30, 31, 32	0,0103
2	17	0,0090	19	9	0,0080			
3	28	0,0093	20	11	0,0081			
4	32	0,0092	21	2	0,0080			
5	8	0,0092	22	30	0,0080			
6	22	0,0095	23	3	0,0081	2	2, 3, 9, 10, 11, 13, 14, 16, 17, 19, 20, 22, 26	0,0100
7	4	0,0088	24	7	0,0079			
8	5	0,0092	25	16	**0,0078**			
9	20	0,0089	26	19	0,0079			
10	26	0,0089	27	29	0,0079	3	6, 12	
11	15	0,0088	28	24	0,0081			
12	25	0,0087	29	23	0,0086			
13	31	0,0086	30	18	0,0100			
14	10	0,0081	31	6	0,0113			
15	13	0,0090	32	12				
16	21	0,0080						

Źródło: opracowanie własne.

Tabela 3. Wynik działania procedur doboru zmiennych dla zbioru *Boston* dla metody POLYMARS

Eliminacja pojedynczych zmiennych						Eliminacja zmiennych blokiem		
etap	usunięta zmienna	*MSE* modelu	etap	usunięte zmienne	*MSE* modelu	etap	usunięte zmienne	*MSE* modelu
0	nic	13,724	7	1	11,902	0	nic	13,724
1	2	12,877	8	7	12,336	1	2, 4, 3	13,339
2	4	12,862	9	11	12,414	2	1, 7, 9, 12	12,630
3	9	11,894	10	8	14,334	3	5, 10	13,479
4	5	11,578	11	10	19,549	4	6, 8, 11, 13	
5	3	**11,578**	12	6	26,843			
6	12	11,715	13	13				

Źródło: opracowanie własne.

Zgodnie z przedstawionymi algorytmami na każdym zbiorze danych zbudowany został model na podstawie wybranej nieparametrycznej metody regresji. Następnie w modelach tych systematycznie dokonywano eliminacji zmiennych, pojedynczo oraz blokiem. Uzyskane wyniki przedstawione zostały w tab. 2 i 3. Ze względu na ograniczenia objętości tej pracy szczegółowo przedstawiono etapy omawianych procedur tylko w dwóch przypadkach, natomiast podsumowanie wyników wszystkich przeprowadzonych analiz przedstawiono w tab. 4 i 5.

W pierwszej części każdej tabeli zaprezentowano kroki procedury eliminacji pojedynczych zmiennych. W kolumnach 2 i 5 umieszczony został numer kolejnej zmiennej usuniętej z modelu w etapie *j*, natomiast w kolumnach 3 i 6 przestawiono wartości błędu średniokwadratowego modelu otrzymanego w tym kroku. Druga część tabeli zawiera wyniki w przypadku eliminacji blokiem. W kolumnie 8 przedstawiono numery zmiennych, które w danym etapie zostały usunięte, kolumna 9 zaś prezentuje wartości MSE modelu na tak zmodyfikowanym zestawie zmiennych. Wszystkie błędy średniokwadratowe obliczone zostały metodą sprawdzania krzyżowego.

W przypadku eliminacji pojedynczych zmiennych za najlepszy, w sensie przyjętego kryterium, uważamy ten model, który charakteryzuje się najniższym błędem średniokwadratowym. Dla metody eliminacji zmiennych blokiem ostateczny model otrzymujemy zawsze w ostatnim kroku algorytmu.

Z tabeli 2 wynika, że dla modeli zbudowanych metodą PPR na zbiorze *Bank* najniższą wartość błędu średniokwadratowego, po zastosowaniu procedury eliminacji pojedynczych zmiennych, otrzymujemy w kroku 25 (po wyrzuceniu 25 zmiennych). Zatem ostateczny, końcowy model, charakteryzujący się możliwie najmniejszym MSE, zbudowany został na zbiorze opisywanym przez 7 cech. Tylko te zmienne okazują się istotnie wpływać na zmienną zależną. Procedura eliminacji blokiem dla zbioru *Bank* i metody PPR miała jedynie trzy etapy, a ostateczny model, którego błąd średniokwadratowy jest równy 0,01, uzyskano na zbiorze charakteryzowanym przez trzy zmienne.

Tabela 4. Wyniki zastosowania procedur doboru zmiennych do modeli zbudowanych za pomocą metody POLYMARS dla każdego z analizowanych zbiorów danych

			Eliminacja pojedynczych zmiennych			Eliminacja zmiennych blokiem		
Zbiór danych	Oryginalna liczba zmiennych	*MSE* początk. modelu	liczba zmiennych wyeliminowanych	*MSE* końcowego modelu	czas [s]	liczba zmiennych wyeliminowanych	*MSE* końcowego modelu	czas [s]
Bank	32	0,0078	12	0,0077	2501,2	25	0,0078	245,7
Boston	13	13,7237	5	11,5780	292,7	9	13,4785	93,8
Peak	50	32,9208	29	17,3331	2454,4	47	24,7606	136,2
Triazines	58	0,0240	55	0,0198	1328,4	56	0,0217	70,89

Źródło: opracowanie własne.

Analogicznie w przypadku modeli budowanych metodą POLYMARS na zbiorze *Boston* eliminacja pojedynczych zmiennych wskazuje jako najlepszy model po wyrzuceniu pięciu zmiennych. W ostatnim etapie eliminacji zmiennych blokiem uzyskujemy model zbudowany na zestawie tylko czterech zmiennych istotnych.

Zestawienie wyników wszystkich przeprowadzonych analiz znajduje się w tab. 4 i 5.

Tabela 5. Wyniki zastosowania procedur doboru zmiennych do modeli zbudowanych za pomocą metody PPR dla każdego z analizowanych zbiorów danych

			Eliminacja pojedynczych zmiennych			Eliminacja zmiennych blokiem		
Zbiór danych	Oryginalna liczba zmiennych	*MSE* początk. modelu	liczba zmiennych wyeliminowanych	*MSE* końcowego modelu	czas [s]	liczba zmiennych wyeliminowanych	*MSE* końcowego modelu	czas [s]
Bank	32	0,0106	25	0,0078	5083,3	30	0,0100	585,88
Boston	13	14,0885	3	11,3309	282,4	1	11,9638	96,2
Peak	50	56,3735	19	36,3101	5549,6	49	48,3365	2788,2
Triazines	58	0,0644	50	0,0184	6622,7	57	0,0239	603,6

Źródło: opracowanie własne.

5. Podsumowanie

W algorytmach obu wykorzystanych w analizach nieparametrycznych metod regresji wbudowany jest mechanizm doboru zmiennych do modelu. W metodzie POLYMARS jest to etap eliminacji cech, w PPR zaś – transformacja zmiennych przez rzutowanie. Jednak, jak pokazują wyniki przeprowadzonych analiz (zob. tab. 4, 5), zastosowanie dodatkowej procedury eliminacji czy to pojedynczych zmiennych, czy blokiem, w każdym z omawianych przypadków powoduje obniżenie wartości błędu średniokwadratowego modelu. Większy spadek wartości *MSE* zaobserwować możemy w sytuacjach, gdy dobór zmiennych do modelu odbywał się przez eliminację pojedynczych zmiennych.

Celem procedury eliminacji pojedynczych zmiennych jest uzyskanie modelu charakteryzującego się jak najwyższą dokładnością predykcji. Dlatego w większości przypadków metoda ta będzie generowała modele o niższych wartościach *MSE* niż procedura eliminacji zmiennych blokiem. Zaletą tego podejścia jest również to, że pozwala ona na uzyskanie modelu o mniejszej złożoności niż model zbudowany na oryginalnym zestawie zmiennych objaśniających. Jej wadą jest znaczne zwiększenie czasu potrzebnego na zbudowanie modelu regresyjnego.

W wielu analizach danych dążymy do uzyskania modelu będącego pewnego rodzaju kompromisem pomiędzy dokładnością a prostotą. Nieparametryczne metody regresji, takie jak POLYMARS czy PPR, pozwalają na budowanie modeli cha-

rakteryzujących się często wysoką dokładnością predykcji. Mamy więc dobre modele, które jednak często są zbyt złożone. Celem metody eliminacji zmiennych blokiem jest jak największe uproszczenie modelu, przy jednoczesnym zachowaniu jego zdolności predykcyjnych. Wyniki przedstawione w tab. 4 i 5 pokazują, że w przypadku tej procedury zazwyczaj usuwamy z modelu więcej cech niż w eliminacji pojedynczych zmiennych. Metoda eliminacji blokiem jest też znacznie mniej czasochłonna.

Literatura

Friedman J.H., Stuetzle W., *Projection pursuit regression*, „Journal of the American Statistical Association" 1981 nr 76, s. 817-823.

Guyon I., Gunn S., Nikravesh M., Zadeh L. (red.), *Feature Extraction, Foundations and Applications*, Springer, 2006.

Harrison D., Rubinfeld D. L., *Hedonic Prices and the Demand for Clean Air*, „Journal of Environmental Economics and Management" 1978 no 8.

Kooperberg C., Bose S., Stone C.J., *Polychotomous Regression*, „Journal of the American Statistical Association" 1997 nr 92, s. 117-127.

Meyer D., Leisch F., Hornik K., *Benchmarking support vector machines*, Report no. 78, Vienna University of Economics and Business Administration, 2002, http://www.wuwien.ac.at/am/Download/ report78.pdf.

Nagatani T., Abe S., *Backward variable selection of support vector regressors by block deletion*, „International Joint Conference on Neural Networks" (IJCNN), Orlando 2007, s. 1540-1545.

Trzęsiok J., *Metoda rzutowania w budowie modelu regresyjnego*, [w:] *Postępy ekonometrii*, red. A.S. Barczak, AE, Katowice 2004b.

Trzęsiok J., *Ocena wpływu wymiaru przestrzeni zmiennych na jakość predykcji wybranych nieparametrycznych modeli regresji*, [w:] *Klasyfikacja i analiza danych*, Taksonomia 16, K. Jajuga, M. Walesiak (red.), Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu nr 47, UE, Wrocław 2009, s. 141-148.

Trzęsiok J., *Wybrane nieparametryczne metody regresji i ich zastosowania*, [w:] *Klasyfikacja i analiza danych – teoria i zastosowania*, Taksonomia 11, Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 1022, AE, Wrocław 2004a.

VARIABLE SELECTION FOR REGRESSION MODEL BUILT WITH THE USE OF CHOSEN NONPARAMETRIC METHODS

Summary: The paper aims to discuss and compare two procedures for variable selection for regression models built with the use of two nonparametric methods: POLYMARS and projection pursuit regression. The results obtained on the benchmark data sets show that using the procedure for the reduction of the number of predictors yields simpler models with smaller mean squared errors than models built on the complete set of the input variables.