STUDIA INFORMATICA 2008
Volume 29 Number 3A (78)

Marcin MICHALAK, Katarzyna STĄPOR
Politechnika Śląska, Instytut Informatyki

ESTYMACJA JĄDROWA W PREDYKCJI SZEREGÓW CZASOWYCH

Streszczenie. Artykuł opisuje problem predykcji w szeregach czasowych. Autorzy opisują modyfikację algorytmu jądrowego, którą porównują z metodą dekompozycji. W wyniku zastosowania zmodyfikowanego algorytmu zmniejszono błąd predykcji o jedną trzecią w porównaniu z najlepszym wynikiem uzyskanym metodą dekompozycji. Do eksperymentów użyto zarówno danych syntetycznych, jak i rzeczywistych.

Słowa kluczowe: analiza danych, funkcja regresji, estymacja nieparametryczna, estymatory jądrowe, szeregi czasowe, predykcja w szeregach czasowych, predykcja jądrowa

KERNEL ESTIMATION IN THE TIME SERIES PREDICTION

Summary. This paper raises a problem of time series prediction. Authors describe a modification of kernel prediction and compare it with the time series decomposition. The final prediction error was decreased by one third in comparison with the best result of time series decomposition. Experiments were conducted on the real and synthetic data.

Keywords: data analysis, regression function, nonparametric estimation, kernel estimators, time series, time series prediction, kernel prediction

1. Wstęp

Predykcja w szeregach czasowych wiąże się ściśle z zagadnieniem estymacji funkcji regresji. W przypadku ogólnym analizie regresji poddawane są dane (np. zbiór wartości pewnej cechy obserwowanej wśród obiektów), dla których nie uwzględnia się zależności czasowej pomiędzy poszczególnymi pomiarami. Można zatem przyjąć, że wszystkie pomiary

miały miejsce jednocześnie bądź traktuje się je jako niezależne od czasu pomiaru. W przypadku szeregów czasowych mamy natomiast do czynienia z danymi, pomiędzy którymi istnieje jednoznaczna zależność czasowa (chronologiczna). Doskonałym przykładem są dane związane ze zjawiskami ekonomicznymi, geograficznymi czy socjologicznymi.

Do standardowych metod analizy szeregów czasowych można zaliczyć metodę dekompozycji [1], metody stochastyczne [3] czy też metody jądrowe (w postaci zwykłych estymatorów jądrowych [8, 15, 16] (np. Nadarayi-Watsona) bądź to bardziej zaawansowanych, jakimi są maszyny wektorów podpierających (ang. *Support Vector Machines*) [2, 5]). W przypadku tej ostatniej grupy opisana wcześniej zależność czasowa w danych powoduje, że metody te wymagają pewnych modyfikacji. Głównie polega ona na zmianie dziedziny w której dokonywana jest estymacja (przekształcenie zostało szerzej opisane w punkcie 2.4).

W pierwszej części artykułu opisano dokładnie predykcję w szeregach za pomocą metody dekompozycji oraz predykcję z użyciem prostych estymatorów jądrowych, wraz ze wskazaniem jej słabych stron. W kolejnej części zaprezentowano szczegółowo kolejne etapy modyfikacji predykcji jądrowej, ilustrując wyniki poszczególnych kroków algorytmu wykresami przedstawiającymi rezultaty predykcji. Sam algorytm jest dokładnie opisany w ostatniej części, która zawiera także zestawienie wyników predykcji metodą dekompozycji i algorytmem jądrowym oraz podsumowuje uzyskane rezultaty.

2. Analizowane dane i zastosowane narzędzia

2.1. Dane poddawane analizie

Analizie poddano dwa szeregi czasowe. Pierwszy z nich zaczerpnięto z literatury [3] (oznaczony jako „szereg G") i jest on szeregiem rzeczywistym, drugi z nich („szereg M") stworzono syntetycznie. Oba szeregi przedstawia rysunek 1.

W przypadku szeregu G przewidywaną wartością był miesięczny przewóz pasażerów amerykańskimi liniami lotniczymi. Szereg G składa się ze 144 wartości. Pierwszej odpowiada wartość szeregu w miesiącu styczniu roku 1949, natomiast ostatniej (miesiąc 144) wartość szeregu w miesiącu grudniu roku 1960.

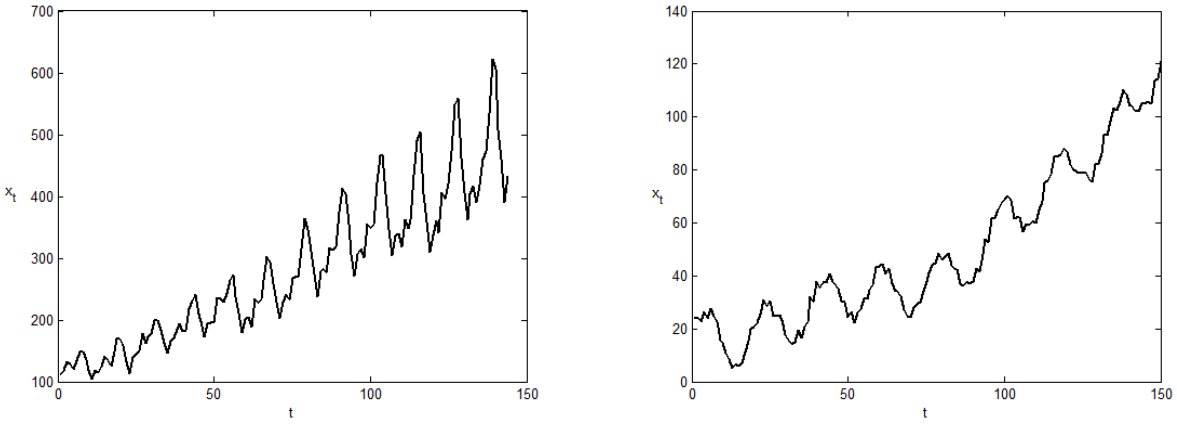
W przypadku szeregu M obserwowana wartość nie miała żadnego odniesienia do sytuacji rzeczywistej.

2.2. Cel predykcji

Głównym zadaniem stawianym przed tworzonymi przez autorów predyktorami było wyznaczanie wartości szeregu w czasie równym jednemu okresowi od momentu ostatniego

pomiaru wartości w szeregu. Upraszczając – efektem zastosowania predyktora był przewidywany zestaw wartości szeregu na jeden pełny okres „w przód".

a) szereg G b) szereg M



Rys. 1. Przykład dwóch szeregów czasowych: a) szereg rzeczywisty, b) szereg wygenerowany syntetycznie
Fig. 1. The example of two time series: a) the real time series, b) the syntetic time series

2.3. Metoda dekompozycji

Metoda dekompozycji [1, 14] ma na celu rozbicie oryginalnego szeregu czasowego na kilka odrębnych składowych:

T – trend szeregu; wolnozmienna składowa szeregu; rosnąca, malejąca bądź stała,

I – okresowość szeregu; periodyczne fluktuacje, charakteryzujące się stałością okresu,

C – cykl szeregu; charakterystyczne zmiany, posiadające naprzemienne maksimum i minimum,

e – błąd (składowa losowa).

Należy zaznaczyć, że nie wszystkie pozycje literaturowe uwzględniają cykl. W ogólności możemy jednak zapisać równanie predykcji w następującej postaci:

$$x_i = f(trend, okresowość, cykl, e) \tag{1}$$

przy czym funkcja f może być wyrażona jako suma bądź iloczyn poszczególnych składowych. Mówi się wtedy odpowiednio o modelu addytywnym bądź modelu multiplikatywnym. Równania obu szeregów wyglądają wtedy następująco:

$$x_i = T_i + I_i + C_i + e_i \tag{2}$$

$$x_i = T_i \cdot I_i \cdot C_i \cdot e_i \tag{3}$$

Pierwszym etapem analizy jest wyodrębnienie okresu (N) szeregu czasowego. W kolejnym kroku wyznaczamy szereg będący średnią ruchomą oryginalnego szeregu, przez

co uzyskuje się szereg posiadający jedynie trend oraz cykl. Wzór na średnią ruchomą $\{\overline{x}_w\}$, wyznaczoną w oknie o szerokości *w*, przedstawia się następująco :

$$\begin{aligned} \overline{x}_{2n+1} &= \frac{1}{2n+1}\sum_{i=-n}^{n} x_i \\ \overline{x}_{2n} &= \frac{1}{2n}\left(\frac{1}{2}x_{i-n} + \sum_{i=-n+1}^{n-1} x_i + \frac{1}{2}x_{i+n}\right) \qquad n \in \mathbb{N} \end{aligned} \tag{4}$$

widać zatem różnicę przy wyznaczaniu średniej ruchomej pomiędzy średnią wyznaczaną w oknie o szerokości będącej liczbą parzystą a średnią wyznaczaną w oknie o szerokości będącej liczbą nieparzystą. W przypadku metody dekompozycji wyznacza się szereg średnią ruchomą w oknie o długości równej okresowi analizowanego szeregu czasowego.

Następnie od oryginalnego szeregu $\{x_i\}$ odejmuje się wyznaczoną średnią ruchomą $\{\overline{x}_N\}$, co powoduje otrzymanie szeregu $\{y_i\}$, składającego się jedynie ze składowej okresowej $\{I_i\}$ oraz błędu $\{e_i\}$. Dla każdej z faz okresu tak otrzymanego szeregu $\{y_i\}$ wyznaczamy średnią wartość, która powoduje wyeliminowanie składowej losowej (błędu). Odejmując (przy założeniu że posługujemy się modelem addytywnym) tę wartość od szeregu, otrzymujemy rozbicie na obie składowe: okres i błąd. Jeśli zakłada się, że model dekompozycji jest modelem multiplikatywnym, należy w opisanym przed chwilą kroku zastąpić działanie odejmowania dzieleniem. W ostatnim etapie wyznaczamy równanie krzywej trendu w szeregu składającym się z trendu i cyklu, co pozwala nam ostatecznie wyznaczyć trend i cykl.

Właściwy model predykcji, przy założeniu że wyznaczone są trzy szeregi składowe, wygląda następująco:

$$x_{i+p} = T_{i+p} + I_{i+p} + C_{i+p} \tag{5}$$

$$x_{i+p} = T_{i+p} \cdot I_{i+p} \cdot C_{i+p} \tag{6}$$

gdzie *p* oznacza horyzont predykcji (oznaczany standardowo jako *h*, tu jednak autorzy wprowadzają oznaczenie *p*, by uniknąć konfliktu oznaczeń z oznaczeniem parametru wygładzającego, związanego z pojęciem estymatora jądrowego).

2.4. Estymacja jądrowa

Drugim wykorzystywanym przez autorów w charakterze predyktora narzędziem jest estymator jądrowy. Estymator jądrowy [16] w ogólności to funkcja $\widehat{f}: \mathbb{R}^m \to \mathbb{R}$ opisana wzorem:

$$\widehat{f}(x) = \frac{1}{nh^m}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{7}$$

gdzie n oznacza liczność próby losowej (albo liczbę próbek szeregu czasowego w zbiorze uczącym), m jest liczbą zmiennych objaśniających, a funkcja K jest funkcją borelowską, spełniającą warunek:

$$\int_{\mathbb{R}^m} K(x)dx = 1 \tag{8}$$

Dodatkowo zakłada się, że funkcja $K(x)$ (określana również jądrem bądź funkcją jądra) jest symetryczna względem zera i posiada w tym punkcie słabe maksimum globalne:

$$\begin{aligned} &\forall x \in \mathbb{R}^m \; K(x) = K(-x) \\ &\forall x \in \mathbb{R}^m \; K(0) \geq K(x) \end{aligned} \tag{9}$$

Dodatnia liczba h jest parametrem wygładzającym, będącym, obok funkcji jądra, głównym czynnikiem decydującym o jakości estymatora jądrowego. Dla jakości estymacji dobór właściwej wartości parametru h jest o wiele bardziej istotny niż dobór kształtu funkcji jądra [24]. Małe wartości parametru wygładzającego powodują zbytnie dopasowanie estymatora do danych i powodują ukazanie nieprawdziwych cech badanej populacji, podczas gdy duże wartości h prowadzą do uzyskania estymatora zbytnio upraszczającego własności populacji. W szczególności efekt ten jest widoczny dla cech, których wartości opisane są rozkładami wielomodalnymi [23].

Jedną z najprostszych metod oceniających jakość estymatora jądrowego jest średni scałkowany błąd kwadratowy (ang. *Mean Integrated Squared Error, MISE*), który z kolei można sprowadzić do postaci sumy scałkowanego kwadratu obciążenia estymatora (ang. *Integrated Squared Bias, ISB*) i scałkowanej wariancji estymatora (ang. *Integrated Variance, IV*). Minimalizowanie tak wyrażonego błędu prowadzi do następującej, optymalnej wartości parametru wygładzającego:

$$h^* = \left[\frac{R(K)}{\sigma_K^4 R(f'')} \right]^{\frac{1}{5}} n^{-\frac{1}{5}} \tag{10}$$

gdzie $R(\bullet)$ oznacza, dla każdej funkcji całkowalnej w kwadracie, wyrażenie $R(L) = \int L^2(x)dx$. Szczegóły wyprowadzenia można znaleźć w [21].

Zakładając, że rozkład danych uczących w przestrzeni zmiennych objaśniających jest rozkładem normalnym [22, 13], można wzór (10) uprościć do postaci:

$$h^* = 1,06\hat{\sigma}n^{1/5} \tag{11}$$

gdzie $\hat{\sigma}$ jest odchyleniem standardowym z próby, bądź korzystając z rozstępu międzykwartylowego $\hat{R}$, do postaci:

$$h^* = 1,06 \min\left(\hat{\sigma}, \frac{\hat{R}}{1,34} \right) n^{1/5} \tag{12}$$

Z zasady maksymalnego wygładzania (ang. *Maximal Smoothing Principle*) sformułowanej przez Turlacha [24], stanowiącej, że z założenia o dolnym ograniczeniu wyrażenia $R(f'')$ wynika górne ograniczenie na wartość h, optymalna wartość parametru wygładzania określona jest poniższym wzorem:

$$h^* = 3 \cdot 35^{-1/5} \hat{\sigma} \left[R(K) \right]^{1/5} n^{-1/5} \tag{13}$$

Ponieważ przedstawione metody mogą powodować utratę danych w przypadku funkcji wielomodalnych [23], rozwinięto także inne grupy metod wyznaczania parametru wygładzającego. Można wśród nich wskazać metody walidacji krzyżowej (ang. *cross-validation methods*), a także metody zagnieżdżone (ang. *plug-in methods*).

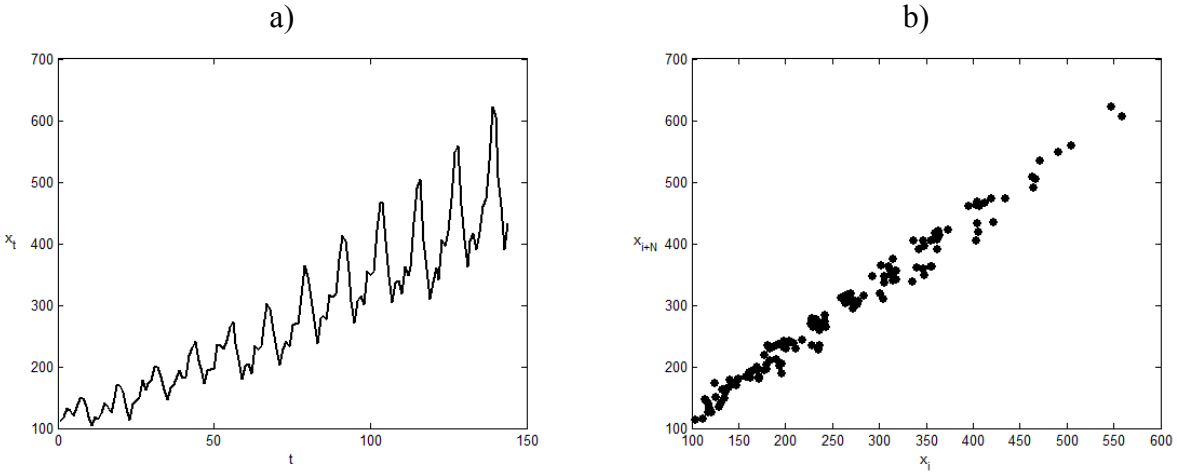
W ogólności, metody walidacji krzyżowej polegają na wielokrotnym sprawdzeniu błędu regresji, przy wykorzystaniu za każdym razem innego zbioru treningowego i testowego. W zależności od przyjętego schematu pojedynczego eksperymentu i estymatorów nieznanych wartości (takich jak np. pochodna funkcji estymowanej) uzyskano wiele różnych wariantów tej metody, jak na przykład walidacja krzyżowa pseudowiarygodnościowa [6, 11], walidacja krzyżowa najmniejszych kwadratów [4, 18] (zwana także nieobciążoną walidacją krzyżową), obciążona walidacja krzyżowa [20] czy też wygładzana walidacja krzyżowa [12].

Podstawą zagnieżdżonych metod wyznaczania parametru h jest wzór (10), w którym jako jedna z niewiadomych występuje pochodna funkcji estymowanej. Woodroofe [25] proponuje, by użyć parametru wygładzającego h_1 do wyznaczenia $\hat{f}_{h_1}(x)$, następnie uzyskany estymator użyć do wyznaczenia $\hat{R}(f'') = R(\hat{f}_{h_1}(x))$ i tak uzyskaną wartość $\hat{R}(f'') = R(\hat{f}_{h_1}(x))$ podstawić do wyrażenia (10) celem wyznaczenia optymalnej wartości h_2. Znane są również iteracyjne metody, polegające na cyklicznym wykonywaniu kroków opisanych powyżej tak długo, aż zostanie spełnione kryterium zbieżności [19]. W pracy [9] natomiast przedstawiono iteracyjną metodę wyznaczania wartości parametru wygładzającego, bazującą na estymatorze Gassera-Mullera [10], przy czym liczba iteracji jest z góry ustalona (za optymalne przyjmuje się h wyznaczone po 11 iteracjach).

Zastosowanie estymatora jądrowego do predykcji w szeregach czasowych wymaga zmodyfikowania przestrzeni danych pomiarowych. Zamiast zbioru punktów $(t, x(t))$ tworzymy przestrzeń punktów (x_i, x_{i+pmax}), gdzie p_{max} jest maksymalnym interesującym horyzontem predykcji. W artykule przyjęto $p_{max} = N$, gdzie N jest wyznaczoną (bądź oszacowaną) wcześniej długością okresu. Na rysunku 2 pokazano, jak wygląda ten sam szereg czasowy przedstawiony w obu przestrzeniach.

Począwszy od tego momentu, o ile nie zostanie to odpowiednio zasygnalizowane w treści artykułu, estymacja jądrowa przebiega w przestrzeni, w której zmienną opisującą jest wartość

szeregu w chwili t, a zmienną objaśnianą wartość szeregu w chwili odległej o maksymalny horyzont predykcji późniejszej.



Rys. 2. Ten sam szereg czasowy (szereg G) przedstawiony w dwóch różnych dziedzinach: a) w dziedzinie (t, $x(t)$), b) w dziedzinie (x_i, x_i+N)
Fig. 2. The same time series (G series) shown in two different domains: a) the (t, $x(t)$) domain, b) the (x_i, x_i+N) domain

W niniejszym artykule do predykcji szeregu czasowego $\{x_i\}$, poprzez zmienną objaśnianą $\hat{f}(x)$, użyto jednego z najprostszych estymatorów jądrowych, jakim jest estymator Nadarayi-Watsona [15]:

$$\hat{f}(x) = \frac{\sum_{i=1}^{n} y_i K\left(\frac{x - x_i}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)} \tag{14}$$

z jądrem Epanecznikowa jako funkcją jądra:

$$K(x) = \frac{3}{4}(1 - x^2) \qquad dla\ |x| \le 1 \tag{15}$$

Za takim doborem estymatora przemawia fakt, że cechuje się on niskim błędem estymacji oraz krótkim czasem jej trwania w porównaniu z estymatorami działającymi na podstawie funkcji jądra zdefiniowanej na ograniczonym nośniku [17]. Jako estymatora optymalnej wartości parametru wygładzającego użyto wartości wynikającej ze wzoru (12).

Błąd regresji w każdym pojedynczym eksperymencie był wyznaczany jako pierwiastek ze średniej wartości kwadratu różnicy wartości estymowanej i wyznaczonej przez estymator (ang. *Root Mean Squared Error, RMSE*).

$$err = \sqrt{\frac{1}{k}\sum_{i=1}^{k}(x_i - \hat{x}_i)^2} \tag{16}$$

gdzie k oznacza liczbę obiektów w zbiorze testowym bądź też liczbę przewidywanych wartości szeregu czasowego w pojedynczym eksperymencie predykcji. Jak już wspomniano,

pojedynczy eksperyment predykcji obejmował przewidywanie N kolejnych wartości szeregu czasowego. Tak zdefiniowane zadanie predykcji będzie w dalszej części artykułu zwane predykcją całego kolejnego okresu szeregu czasowego lub w skrócie predykcją całego okresu.

2.5. Zbiory uczące i zbiory testowe

Rozważając analizę szeregu czasowego metodą dekompozycji, przez zbiór uczący autorzy rozumieją zbiór obserwacji począwszy od jakiegoś ustalonego momentu w przeszłości po chwilę obecną. Zbiór ten służy wyznaczeniu wspomnianych już wcześniej czterech składowych szeregu czasowego. Tak wyznaczone składowe są z kolei stosowane przy wyznaczaniu wartości szeregu czasowego dla kolejnych, różnych horyzontów predykcji, przy czym rzeczywiste wartości szeregu czasowego w tych chwilach są znane i stanowią zbiór testowy dla metody dekompozycji.

Z kolei w przypadku jądrowej predykcji szeregu czasowego w zmodyfikowanej przestrzeni, przez zbiór uczący rozumie się zbiór par wartości przyjętych przez szereg czasowy, w odstępie równym horyzontowi predykcji, na podstawie to którego zbioru wyznaczana jest funkcja regresji. Zbiór testowy dla predykcji jądrowej stanowi również zbiór par obserwacji wartości szeregu czasowego, odległych o horyzont predykcji, przy czym ostatni element tak zdefiniowanej (pojedynczej) pary uporządkowanej stanowi punkt odniesienia dla wartości wynikającej z wyznaczenia wartości funkcji regresji dla argumentu będącego pierwszym elementem pary uporządkowanej (należącej do rozważanego obiektu ze zbioru testowego).

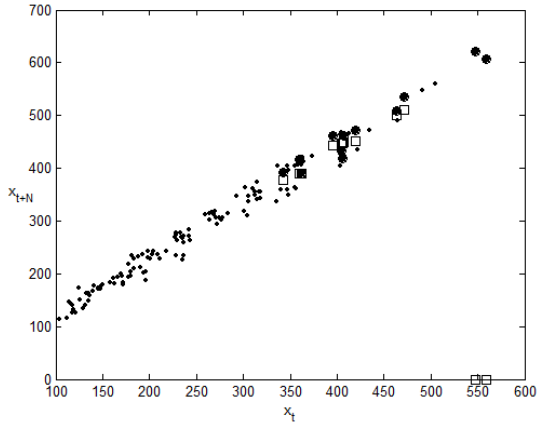
Pojęcie zbioru strojącego (tuningowego) odnosi się do pewnego podzbioru zbioru uczącego. Otóż rozważając wybrany wcześniej zbiór uczący, można go potraktować jako kompletny zbiór obiektów i podzielić go na zbiór uczący (drugiego poziomu) oraz testowy. Otrzymany w ten sposób zbiór uczący nazywa się w dalszej części zbiorem uczącym, natomiast zbiór testowy jest rozróżniany jako zbiór strojący.

3. Modyfikacja estymatora jądrowego

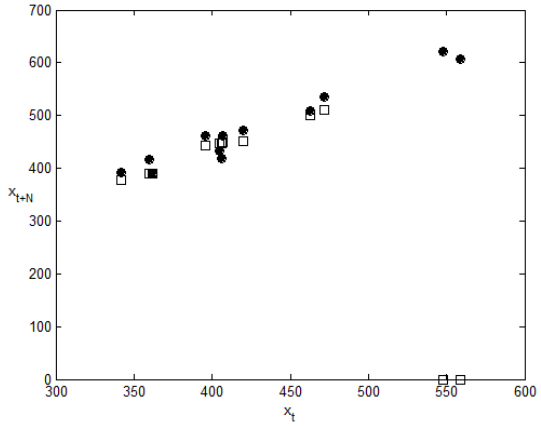
Na rysunku 2 przedstawiono przebieg szeregu G w zmienionej przestrzeni. Można zauważyć, że wraz ze wzrostem argumentu x spada gęstość punktów na wykresie i to zarówno na płaszczyźnie, jak i na osi x. Ponieważ estymacja jądrowa wymaga podania wartości parametru wygładzającego h, wartość wyznaczona globalnie na zbiorze uczącym nie znajdzie efektywnego zastosowania w górnej części dziedziny szeregu, gdzie należy się spodziewać

obiektów ze zbioru testowego. Można zatem spodziewać się, że tak przeprowadzona predykcja zostanie obarczona wysokim błędem. Wynik predykcji przedstawia rysunek 3.

Na wykresie tym małe czarne punkty reprezentują zbiór uczący, duże czarne kółka to zbiór testowy, natomiast kwadraty to wynik predykcji. Uwagę zwracają dwa punkty leżące na osi x w okolicy wartości $x = 550$. Dla zwiększenia przejrzystości rysunek 4 przedstawia wynik tej samej predykcji co rysunek 3, jednak z rysunku 4 usunięto punkty uczące.



Rys. 3. Wynik standardowej predykcji z użyciem estymatora jądrowego: całkowicie błędne rezultaty dla skrajnych obiektów
Fig. 3. The result of the standard kernel prediction: completely wrong results for extreme objects



Rys. 4. Wynik standardowej predykcji z użyciem estymatora jądrowego: całkowicie błędne rezultaty dla skrajnych obiektów (wykres z pominięciem obiektów uczących)
Fig. 4. The result of the standard kernel prediction: completely wrong results for extreme objects (excluding training objects)

Najbardziej skrajne dwa punkty testowe znalazły się tak daleko od punktów uczących, że ich otoczenie (wyznaczone przez niezerową wartość funkcji jądra) nie zawierało żadnego punktu uczącego, na podstawie którego mogłaby zostać wyznaczona estymowana wartość. Błąd predykcji całego okresu wyniósł $RMSE = 275{,}25$, podczas gdy błąd predykcji wartości, które okazały się niezerowe (błąd predykcji w punktach, które okazały się nie leżeć na osi X),

wyniósłby jedynie *RMSE* = 18,80. Przyjrzyjmy się zatem uważniej, jak zmienia się wartość parametru *h*, wyrażonego wzorem (12), w zależności od danych:

Tabela 1

Zależność wartości parametru *h* od zbioru danych

Zbiór	Liczba elementów w zbiorze	*h*
Zbiór uczący	120	32,7993
Zbiór testowy	12	36,1614
Zbiór uczący i testowy	132	36,1402

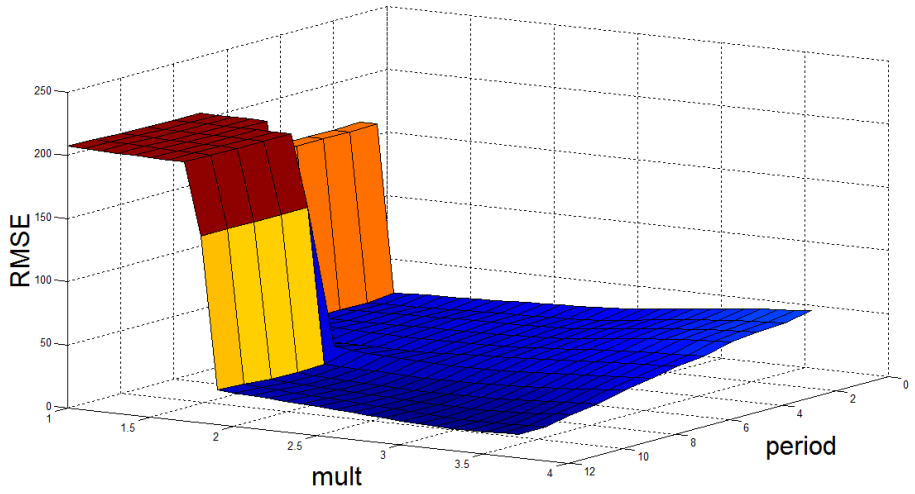
Można zatem zauważyć, że stosunkowo niewielki zbiór testowy znacząco wpływa na zmianę wielkości parametru *h*. Ma to swoje uzasadnienie w fakcie, że rozkład obiektów w przestrzeni zmiennej objaśniającej zbioru testowego jest znacząco różny od rozkładu obiektów należących do zbioru uczącego. Stąd też postanowiono adaptacyjnie wyznaczać zwiększanie wielkości parametru *h* przed samą predykcją.

W tym celu zbiór uczący podzielono na mniejszy zbiór uczący oraz zbiór strojący. Następnie zbadano, jak wpływa zwiększanie parametru *h* wyznaczonego na podstawie zbioru uczącego na błąd predykcji na zbiorze strojącym. Począwszy od *mult* = 1 zwiększano stopniowo mnożnik parametru *h* i obserwowano błąd predykcji na zbiorze strojącym. Następnie wyznaczano przybliżone minimum tak uzyskanej zależności *RMSE*(*mult*). By tak uzyskany wynik (wartość parametru *mult*) był bardziej niezależny od fazy okresu szeregu, ten sam schemat powtórzono dla każdej z faz okresu szeregu. Wynik przedstawia wykres na rysunku 5.

Jak widać, stopniowe zwiększanie wartości parametru *mult* powoduje najpierw gwałtowny spadek błędu predykcji, po czym następuje powolny jego wzrost. Pierwsze zjawisko jest związane z sytuacją, kiedy sąsiedztwo punktów testowych przestaje być zbiorem pustym, drugie natomiast to wspomniany już w części 2.4 problem zbyt dużej wartości parametru *h*, powodującej zbytnie wygładzenie zależności w danych.

Efektem wykonania powyższych obliczeń jest zestaw wartości *mult* dla każdej z faz okresu. Kolejnym krokiem jest wyznaczenie ostatecznej wartości *mult*, która ostatecznie zostanie wykorzystana we właściwej predykcji. Spośród intuicyjnie najprostszych funkcji, takich jak maksimum, średnia czy też różnego rzędu kwantyle, postanowiono zastosować medianę. Warte rozważenia jest również zastosowanie funkcji maksimum, jako że efekt zbytniego wygładzenia danych będzie usuwany w kolejnym kroku analizy.

Zakładając, że *rmse*(*t*, *p*, *h*) to błąd predykcji jądrowej *p* kolejnych wartości szeregu czasowego, dla chwil czasu z przedziału [*t*+1, *t*+*p*], przy zastosowaniu parametru wygładzającego *h*, wzór na wartość parametru *mult* przedstawia się w postaci wyrażonej wzorem (17).



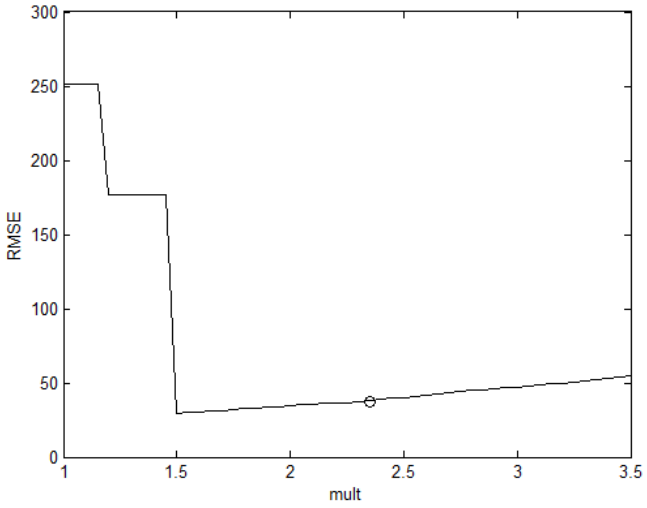
Rys. 5. Zależność błędu regresji (RMSE) od wielkości współczynnika mult oraz fazy okresu (period)
Fig. 5. The dependency of the regression error (RMSE) from the mult coefficient and the period phase (period)

$$mult = \text{med}\{mult_i,\ i = 0, 1, ..., p_{\max} - 1\}$$
$$mult_i = \arg\min_m rmse(t - i, p_{\max}, h \cdot m) \tag{17}$$

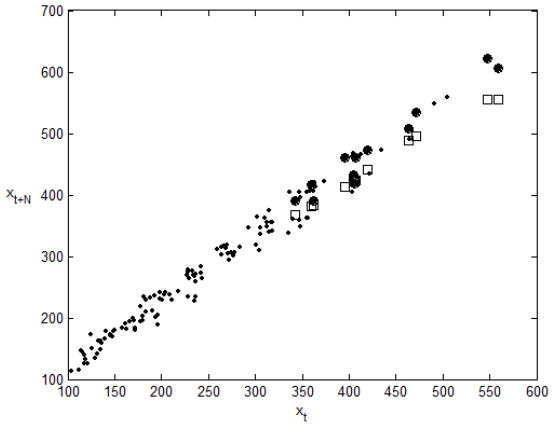
Wartość parametru *mult* wyznaczona w przedstawiony powyżej sposób wyniosła *mult* = 2,35. Dla porównania na wykresie 6 przedstawiono, jak zmienia się błąd predykcji na zbiorze testowym w zależności od parametru *mult*. Czarnym kółkiem oznaczono wartość parametru *mult* wyznaczoną za pomocą wzoru (17). Dla tak wyznaczonego $h' = h \cdot mult$ błąd predykcji wyniósł 37,39, podczas gdy dla $h' = h \cdot 1{,}55$ błąd predykcji wynosił 29,89. Na rysunku 7 przedstawiono predykcję z użyciem wyznaczonej wartości parametru *mult*=2,35, rysunek 8 przedstawia ten sam wynik, z pominięciem zbioru uczącego.

Jak już zostało wspomniane wcześniej, jednym z niepożądanych efektów związanych ze zwiększaniem wielkości parametru *h* jest zbytnie wygładzenie zależności w danych. W konkretnych przypadkach oznacza to, że przewidywane wartości leżą znacząco pod lub znacząco nad rzeczywistymi wartościami. Można zatem wprowadzić parametr nazwany przez autorów niedoszacowaniem, który oznacza w jakiej części dane przewidywane odzwierciedlają żądaną tendencję. Niedoszacowanie wyraża się zatem ilorazem wartości przewidywanej i wartości rzeczywistej:

$$\alpha = \frac{1}{k}\sum_{i=1}^{k}\frac{\hat{y}_i}{y_i} \tag{18}$$

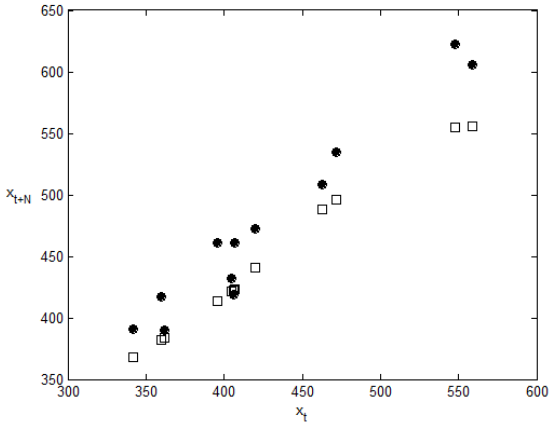


Rys. 6. Minimum błędu predykcji wyznaczone empirycznie (linia ciągła) w odniesieniu do minimum wyznaczonego algorytmem (kółko)
Fig. 6. The prediction error minimum, computed empirically (the solid line), related to the minimum obtained with the algorithm



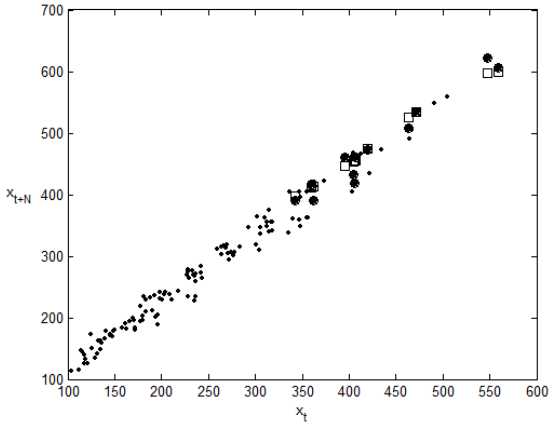
Rys. 7. Wynik predykcji z użyciem estymatora jądrowego ze zmodyfikowanym parametrem *h*: poprawienie efektów dla skrajnych obiektów
Fig. 7. The result of the kernel prediction with the modified parameter *h*: the improvement of results for extreme objects

Niedoszacowanie najprościej można wyznaczyć jako średnią wartość niedoszacowania na *k*-elementowym zbiorze danych strojących. W analizowanym przypadku średnie niedoszacowanie α wyniosło 0,9287.



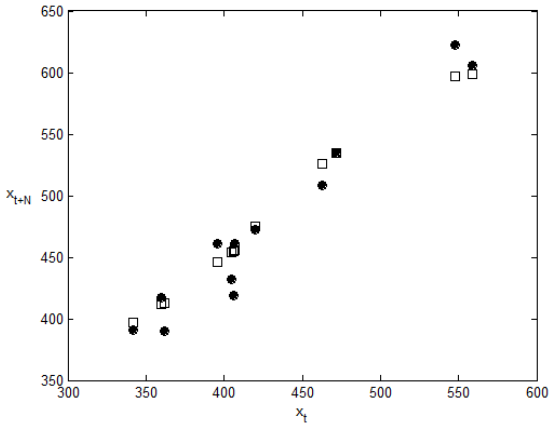
Rys. 8. Wynik predykcji z użyciem estymatora jądrowego ze zmodyfikowanym parametrem h: poprawienie efektów dla skrajnych obiektów (wykres z pominięciem obiektów uczących)
Fig. 8. The result of the kernel prediction with the modified parameter h: the improvement of results for extreme objects (excluding training objects)

Ostatecznie zatem dane uzyskane z predykcji metodą jądrową ze zmodyfikowanym parametrem h należy podzielić przez średnie niedoszacowanie α, wyznaczone na zbiorze strojącym. Wynik ostatniego kroku algorytmu, a jednocześnie końcową predykcję przedstawia wykres 9 oraz wykres 10 (z pominięciem obiektów uczących).



Rys. 9. Końcowy wynik predykcji zmodyfikowanym estymatorem jądrowym
Fig. 9. The final prediction result, with the usage of the modified kernel estimator

Wykres 10 pokazuje końcowy wynik predykcji. Można zauważyć znaczące polepszenie w porównaniu do rezultatów uzyskanych w poprzednich dwóch etapach. Kompleksowy opis kolejnych kroków algorytmu, a także szczegółowe porównanie błędu predykcji na każdym z poszczególnych etapów przedstawiono w kolejnej części artykułu.



Rys. 10. Końcowy wynik predykcji zmodyfikowanym estymatorem jądrowym (wykres z pominięciem obiektów uczących)
Fig. 10. The final prediction result, with the usage of the modified kernel estimator (excluding training objects)

4. Porównanie predyktorów

4.1. Zmodyfikowany algorytm predykcji jądrowej

W poprzedniej części przedstawiony został zmodyfikowany algorytm predykcji szeregów czasowych. Każdy kolejny krok tego algorytmu był formułowany na podstawie obserwacji wyników otrzymanych w wyniku wykonania poprzedniego kroku algorytmu. Poniżej zaprezetowano zebrane razem poszczególne, przedstawione wcześniej, kolejne kroki algorytmu:

1. Zdefiniować maksymalny interesujący horyzont predykcji p_{max}.
2. Przenieść dane pomiarowe szeregu czasowego z przestrzeni $(t, x(t))$ do przestrzeni (x_i, x_{i+pmax}). Jeśli zbiór danych w dziedzinie czasu składał się z m próbek, to w zmodyfikowanej przestrzeni całkowity zbiór danych będzie składał się z $m - p_{max}$ elementów.
3. Ze zbioru wszystkich obiektów wyodrębnić p_{max} obiektów, będących zbiorem testowym. Zbiór ten stanowią pary uporządkowane, w których drugimi elementami są ostatnie p_{max} zaobserwowane wartości szeregu czasowego.
4. Z pozostałych obiektów wyodrębniamy na podobnej zasadzie p_{max} obiektów, które staną się zbiorem strojącym.
5. Dla horyzontów predykcji od $p = 1$ do $p = p_{max}$ znaleźć wartości $mult_i$, które minimalizują błąd *RMSE* na zbiorze strojącym, a następnie wyznaczyć medianę z tych wartości, która stanie się ostateczną wartością parametru *mult* (17).
6. Wyznaczyć średnią wartość niedoszacowania α na zbiorze strojącym, według wzoru (18).
7. Opierając się na zmodyfikowanej parametrem *mult* wartości parametru wygładzającego h, a także na średnim niedoszacowaniu α, wyznaczyć metodą jądrowej estymacji funkcji

regresji wartość szeregu czasowego w chwili $t + p_{max}$, jako predykcji o horyzoncie p_{max} dla chwili bieżącej t, na podstawie następującego wzoru (m jest liczebnością próbek szeregu czasowego w dziedzinie czasu):

$$x_{t+p_{\max}} = \hat{f}(x_t) = \frac{1}{\alpha} \frac{\sum_{i=1}^{m-p_{\max}} y_i K\left(\frac{x_i - x_t}{h \cdot mult}\right)}{\sum_{i=1}^{m-p_{\max}} K\left(\frac{x_i - x_t}{h \cdot mult}\right)} \tag{19}$$

Należy zwrócić uwagę na fakt, że w sytuacji rzeczywistej, kiedy nie będziemy dysponowali zbiorem testowym, należy pominąć krok 3. algorytmu i podzielić zbiór par na zbiór uczący i strojący. W tej sytuacji m we wzorze (19) będzie oznaczało liczebność pełnego zbioru par (liczba obiektów w zbiorze uczącym powiększona o liczbę obiektów w zbiorze strojącym).

Natomiast w sytuacji, gdy stworzony już został predyktor jądrowy dla horyzontu predykcji p_{max}, a interesuje nas predykcja o horyzoncie $p < p_{max}$, wystarczy tak dobrać chwilę czasu w przeszłości t_{past}, by był spełniony warunek $t_{past} + p_{max} = t + p$, a następnie zastosować wzór (19).

4.2. Przebieg badań

W tabelach 2 – 5 przedstawiono wyniki predykcji wartości szeregów G i H metodą dekompozycji, estymatorem jądrowym oraz estymatorem jądrowym zmodyfikowanym w sposób zaproponowany przez autorów. Przyjęty horyzont predykcji zmieniał się od $p = 1$ aż do p_{max}, wynoszącego dla szeregów G i M odpowiednio 12 i 17. W przypadku metody dekompozycji jako zbiór uczący użyto obiektów z przedziału [1, 132] (w przypadku szeregu M – [1, 133]). Prognozy były wyliczane na zbiorze testującym, składającym się z obiektów z przedziału [133, 144] (w przypadku szeregu M – [134, 150]). Dla zadania predykcji jądrowej zbiór 132 par obserwacji szeregu G (133 par dla szeregu M) podzielono na zbiór 108 (116) obiektów w zbiorze uczącym, 12 (17) obiektów w zbiorze strojącym oraz 12 (17) obiektów w zbiorze testowym. Jako kryterium porównawcze wykorzystano *RMSE* na zbiorze testującym.

4.3. Analiza szeregu G

Tabela 2 przedstawia wyniki predykcji metodą dekompozycji dla szeregu G. Rozważono cztery przypadki: trend liniowy i wykładniczy, oraz model addytywny i multiplikatywny.

Tabela 2

Zestawienie błędów predykcji metodą dekompozycji dla szeregu G

	Trend	Wykładniczy		Liniowy	
	Model	addytywny	multiplikatywny	addytywny	multiplikatywny
Miesiąc	Przewóz				
I 60	417	422,30	386,44	382,60	357,09
II 60	391	416,94	379,97	374,76	349,83
III 60	419	460,43	445,96	415,69	409,00
IV 60	461	458,52	436,85	411,14	399,04
V 60	472	469,88	448,54	419,78	408,01
VI 60	535	510,93	510,57	458,03	462,46
VII 60	622	542,38	566,27	487,70	506,05
VIII 60	606	546,92	570,03	489,61	507,86
IX 60	508	508,45	498,40	448,42	442,59
X 60	461	479,80	439,47	416,97	388,89
XI 60	390	453,74	385,60	388,03	339,97
XII 60	432	479,49	432,65	410,82	380,00
	RMSE	**40,32**	**26,60**	**64,63**	**68,52**

W przypadku szeregu G najmniejszy błąd predykcji uzyskano stosując model multiplikatywny z wykorzystaniem trendu wykładniczego.

Tabela 3 przedstawia wyniki predykcji estymatorem jądrowym kolejno: standardowym estymatorem (1), estymatorem z adaptacyjnie modyfikowanym h (2), estymatorem z modyfikowanym h oraz doszacowaniem wartości końcowych (3).

Tabela 3

Zestawienie błędów predykcji metodą jądrową dla szeregu G

Miesiąc	Przewóz	(1)	(2)	(3)
I 60	417	389	382	411
II 60	391	377	367	395
III 60	419	448	421	453
IV 60	461	442	412	444
V 60	472	451	437	471
VI 60	535	514	493	531
VII 60	622	0	553	595
VIII 60	606	0	555	598
IX 60	508	501	487	524
X 60	461	448	421	454
XI 60	390	390	383	412
XII 60	432	447	420	452
	RMSE	**275,26**	**37,39**	**17,18**

4.4. Analiza szeregu M

W tabeli 4 przedstawiono wyniki predykcji szeregu M metodą dekompozycji.

Tabela 4

Zestawienie błędów predykcji metodą dekompozycji dla szeregu M

	Trend	Wykładniczy		Liniowy	
	Model	addytywny	multiplikatywny	addytywny	multiplikatywny
t	x_t				
134	98,59	84,63	70,25	74,09	63,61
135	103,49	84,51	68,78	73,21	61,98
136	102,38	86,19	72,99	74,12	65,43
137	105,71	84,22	69,33	71,35	61,84
138	110,27	85,64	73,74	71,95	65,43
139	108,38	87,53	76,98	72,98	67,92
140	104,14	89,43	82,13	74,00	72,02
141	104,04	90,65	83,71	74,31	72,92
142	102,19	96,53	95,44	80,74	84,49
143	102,11	97,39	94,33	80,72	83,01
144	104,80	101,01	101,45	83,44	88,71
145	104,92	103,30	100,21	83,62	82,54
146	105,82	104,90	101,88	84,31	83,48
147	105,03	107,25	102,10	85,72	83,22
148	113,47	107,50	98,63	85,00	80,00
149	114,25	107,26	92,23	83,77	74,44
150	120,94	108,44	91,11	83,92	73,21
	RMSE	**13,36**	**23,23**	**28,47**	**33,68**

W przypadku szeregu M najmniejszy błąd predykcji uzyskano stosując model addytywny z wykorzystaniem trendu wykładniczego.

Tabela 5

Zestawienie błędów predykcji metodą jądrową dla szeregu M

t	x_t	(1)	(2)	(3)
134	98,59	93,11	83,44	99,52
135	103,49	93,06	83,69	99,82
136	102,38	93,06	84,19	100,42
137	105,71	93,06	83,87	100,03
138	110,27	91,17	82,73	98,68
139	108,38	89,44	82,36	98,24
140	104,14	88,95	82,12	97,95
141	104,04	88,43	81,83	97,60
142	102,19	88,19	81,70	97,44
143	102,11	88,70	81,99	97,79
144	104,80	87,13	80,92	96,52
145	104,92	86,41	79,81	95,20
146	105,82	91,06	82,72	98,66
147	105,03	91,93	82,83	98,79
148	113,47	93,06	83,72	99,86
149	114,25	0,00	86,40	103,05
150	120,94	0,00	86,26	102,88
	RMSE	**42,77**	**23,96**	**8,74**

Tabela 5 przedstawia wyniki predykcji estymatorem jądrowym kolejno: standardowym estymatorem (1), estymatorem z adaptacyjnie modyfikowanym h ($mult = 4{,}6$) (2), estymatorem z modyfikowanym h oraz doszacowaniem wartości końcowych (3).

5. Podsumowanie

W artykule porównano dokładność predykcji dwóch szeregów czasowych metodą dekompozycji i metodą wykorzystującą estymatory jądrowe. Estymatory jądrowe z powodzeniem realizują zadanie estymacji funkcji regresji, jednak w przypadku predykcji szeregów czasowych pojawiają się pewne niepożądane efekty związane z rozmieszczeniem próbek uczących i testowych (puste sąsiedztwo dla niektórych obiektów testowych). Zaproponowany w artykule zmodyfikowany adaptacyjny algorytm jądrowy pozbawiony jest przedstawionej wady.

Wyniki uzyskane za pomocą algorytmu odniesione zostały do wyników otrzymanych klasyczną metodą dekompozycji. Porównanie miało miejsce zarówno na danych rzeczywistych (szereg G), jak i na sztucznie wygenerowanych (szereg M). W przypadku obu szeregów zastosowanie zmodyfikowanego algorytmu spowodowało znaczącą poprawę predykcji – błąd zmalał o 35%.

LITERATURA

1. Bielińska E.: Metody prognozowania. Wydawnictwo Śląsk, Katowice 2002.
2. Boser B. E., Guyon M. I., Vapnik V.: A training algorithm for optimal margin classifiers. Fifth Annual Workshop on Computational Learning Theory, Pittsburgh 1992.
3. Box J. E. P., Jenkins G. M.: Analiza szeregów czasowych. PWN, Warszawa 1992.
4. Bowman A.: An alternative method of cross-validation for the smoothing of density estimates. Biometrika, 71, 1984, s. 353÷360.
5. Cao L. J., Tay F.: Support Vector Machine With Adaptive Parameters in Financial Time Series Forecasting. IEEE Transactions on Neural Networks, Volume 14, Issue 6, Nov. 2003, s. 1506÷1518.
6. Duin R. P. W.: On the choice of smoothing parameters of Parzen estimators of probability density functions. IEEE Transactions on Computers, C-25, 1976, 1175÷1179.
7. Fan J., Gijbels I.: Variable bandwith and local linear regression smoothers. The Annals of Statistics, 1992, Vol. 20, No. 4, s. 2008÷2036.
8. Gajek L., Kałuszka M.: Wnioskowanie statystyczne. WNT, Warszawa 2000.

9. Gasser T., Kneip A., Kohler W.: A flexible and fast method for authomatic smoothing. Journal of the American Statistical Association, 1991, Vol. 86, No. 415, s. 643÷652.
10. Gasser T., Muller H. G.: Estimating Regression Function and Their Derivatives by the Kernel Method. Scandinavian Journal of Statistics, 11, 1984, s. 171÷185.
11. Habbema J. D. F., Hermans J., van den Broek K: A stepwise discrimination analysis program using density estimation. Compstat 1974: Proceedings in Computational Statistic, Physica Verlag, Vienna 1974.
12. Hall P., Marron J. S., Park B. U.: Smoothed Cross-Validation. Probability Theory and Related Fields, Vol. 92, No. 1, 1992, s. 1÷20.
13. Hardle W.: Smoothing Techniques With Implementation in S. Springer, New York 1991.
14. Jóźwiak J., Podgórski J.: Statystyka od podstaw. PWE, Warszawa 2006.
15. Koronacki J., Ćwik J.: Statystyczne systemy uczące się. WNT, Warszawa 2005.
16. Kulczycki P.: Estymatory jądrowe w analizie systemowej. WNT, Warszawa 2005.
17. Michalak M., Stąpor K.: Ocena rzeczywistej wydajności wybranych regresorów. Studia Informatica, Vol. 29, No. 1 (75), s 23÷34, 2008.
18. Rudemo M.: Empirical choice of histograms and kernel density estimators. Scandinavian Journal of Statistics, 9, 1982, s. 65÷78.
19. Scott D. W., Tapia R. A., Thompson J. R.: Kernel density estimation revisited. Nonlinear Analysis, Theory, Methods and Applications, 1, 1977, s. 339÷372.
20. Scott D. W., Terrell G. R.: Biased and Unbiased Cross-Validation in Density Estimation. Journal of the American Statistical Association, 82, 1987, s. 1131÷1146.
21. Scott D. W.: Multivariate Density Estimation: Theory, Practice and Visualization. Wiley, 1992.
22. Silverman B,: Density estimation for statistics and data analysis. 1986, Monographs on Statistics and Applied Probability 26.
23. Terrell G. R.: The Maximal Smoothing Principle in Density Estimation. Journal of the American Statistical Association, 85, (1990), s. 470÷477.
24. Turlach B. A.: Bandwidth Selection in Kernel Density Estimation: A Review. CORE and Institut de Statistique.
25. Woodroofe M.: On Choosing a Delta-Sequence. The Annals of Mathematical Statistics, Vol. 41, No. 5, 1970, s. 1665÷1671.

Recenzent: Dr hab. inż. Ewa Bielińska

Wpłynęło do Redakcji 23 maja 2008 r.

Abstract

This short article raises a problem of time series prediction, that can be treated as a particular case of the estimation of the regression function. This relation motivated authors to develop the kernel method of time series prediction, that gives better results than time series decomposition.

Authors started from the time series decomposition, that is a simple and popular method of time series analysis. That method gave a reference point of possible prediction error for the second method of time series prediction: kernel prediction. The typical way of kernel prediction gave unsatisfactory results, burdened a with big prediction error.

Authors suggested some modifications of kernel prediction model. The most important part is the division of training set into two separate subsets, called training and tuning. The second element of algorithm are two parameters *mult* and α, that are evaluated adaptively, on the basis of the tuning set prediction error.

Presented modifications decreases the prediction error by one third in comparison with the prediction error obtained as the result of the time series decomposition prediction. The algorithm was performed on real and synthetic time series.

Adresy

Marcin MICHALAK: Politechnika Śląska, Instytut Informatyki, ul, Akademicka 16, 44-100 Gliwice, Polska, Marcin.Michalak@polsl.pl

Katarzyna STĄPOR: Politechnika Śląska, Instytut Informatyki, ul, Akademicka 16, 44-100 Gliwice, Polska, Katarzyna.Stapor@polsl.pl