AUTOMATYKA • 2007 • Tom 11 • Zeszyt 1–2

Jan Tadeusz Duda*

Statystyczne i statystyczno-regułowe metody segmentacji szeregów czasowych

1. Wprowadzenie

Segmentacja szeregów czasowych, tj. ich podział na podszeregi (odcinki) o hipotetycznie jednorodnych właściwościach statystycznych, jest jednym z ważnych elementów przetwarzania danych w systemach komputerowego sterowania i nadzorowania [1, 2]. Odgrywa ona szczególnie istotną rolę w sterowaniu procesów ciągłych prowadzonych nominalnie w reżimach stacjonarnych, w których wejścia i wyjścia procesowe winny być utrzymywanie na stałym poziomie z dopuszczeniem losowych odchyłek. Wydzielone algorytmicznie okresy, w których szereg czasowy można uznać za stacjonarny, dają podstawę do miarodajnej estymacji parametrów rozkładu prawdopodobieństwa odchyłek losowych, głównie odchylenia standardowego, a także określenia typu rozkładu. Informacje te mogą zostać wykorzystane do diagnozowania stanu sprawności aparatury pomiarowej oraz wykonawczej, a ogólnie do monitorowania przebiegu procesu metodami statystycznymi (*Statistical Proces Control* – SPC [3]). Wartości oczekiwane tych rozkładów, które można estymować średnią arytmetyczną szeregów w poszczególnych okresach stacjonarności, są z kolei podstawą do oceny efektywności procesu oraz diagnozowania sprawności aparatury technologicznej [4]. Stwierdzenie stacjonarności wszystkich zmiennych procesowych oznacza stacjonarność całego procesu. Wartości średnie zmiennych w tym okresie wyznaczają punkt pracy, który może być zarejestrowany jako obserwacja do identyfikacji charakterystyk statycznych lub wykorzystany do diagnozowania procesu metodą ortogonalizacji przestrzeni jego cech i analizy składowych głównych (*Principal Component Analysis* – PCA [1]). Rejestracja obserwacji w okresach stacjonarności wejść (przy niestacjonarnych wyjściach) usprawnia identyfikację modeli Hammersteina [5]. Z kolei rejestracja szeregów czasowych w okresach niestacjonarności pojedynczych wejść pozwala na selektywną identyfikację modeli dynamiki odpowiadających im torów, a analiza rozstępów czasowych między zakończeniem okresu stacjonarności wejść i wyjść umożliwia szybką identyfikację opóźnień [6].

* Katedra Marketingu i Zarządzania Produkcją, Wydział Zarządzania, Akademia Górniczo-Hutnicza w Krakowie

Podstawą segmentacji jest detekcja zmian właściwości statystycznych szeregu (głównie składowych wolnozmiennych) w oparciu o analizę kolejnych próbek w niedługich oknach. Podstawowe metody takiej detekcji są ujęte w normach międzynarodowych jako zasady statystycznej kontroli procesów – SPC [3]. Wykorzystują one statystyczne testy istotności pojedynczych odchyłek zmiennych od wartości średniej (z założenia znanej), serii odchyłek oraz tzw. sum skumulowanych. Bardziej zaawansowane podejście, oparte na analizie stosunku funkcji wiarygodności alternatywnych hipotez jest stosowane do wykrywania skokowych zmian wartości średniej szeregu (algorytm Page'a–Hinkleya [7]). W niniejszym artykule przedstawiono metodę wyznaczania parametrów tego algorytmu, sygnalizując ograniczenia jego praktycznej stosowalności ze względu na dopuszczalne opóźnienie detekcji. Następnie pokazano, że można go uogólnić na potrzeby detekcji zmian nieskokowych, z uwzględnieniem autokorelacji składowej losowej. Pozwala to na segmentację szeregów także w okresach niestacjonarności, co zwiększa możliwości komputerowej diagnostyki i monitorowania sytuacji procesowych. Zaprezentowano również inny algorytm segmentacji, oparty na wieloaspektowych testach statystycznych zmian szeregu w ruchomym oknie, wsparty regułowo i uwzględniający właściwości widmowe składowych losowych [8, 6]. Na reprezentatywnych przykładach pokazano bardzo wysoką skuteczność tej metody w wykrywaniu słabych zmian wartości średniej.

2. Algorytm Page'a–Hinkleya i jego uogólnienia

Podstawy teoretyczne algorytmów detekcji nagłych zmian parametrów statystycznych sygnałów zostały przedstawione w monografii [9], a ich zastosowania w automatyce i przetwarzaniu sygnałów omawia artykuł przeglądowy [7]. Rozważa się zmienną y będącą sumą deterministycznego sygnału o wartości stałej w pewnych przedziałach czasu oraz szumu stacjonarnego z o zerowej wartości średniej. Dostępny jest ciąg próbek zmiennej ze stałym rozstępem czasowym. Jego wartość w chwili k-tej ma postać

$$y_k = x_k + z_k \tag{1}$$

gdzie:

x_k – chwilowa wartość oczekiwana,
z_k – nieskorelowana zmienna losowa o zerowej wartości oczekiwanej.

Zakłada się, że w sytuacjach normalnych wartości x_k są wyznaczone dokładnie, a zmienna z ma rozkład Gaussa o znanej wariancją σ^2. Zadanie detekcji sprowadza się do przyjęcia hipotezy H_0, że ciąg x_k jest stały dla próbek od 0 do n, albo hipotezy alternatywnej H_1, że od pewnej chwili r ma inną (znaną) wartość. Wybór może być dokonany na podstawie stosunku wiarygodności (*Likelihood Ratio* – LR), według którego hipotezę H_1 przyjmuje się, gdy spełniony jest warunek

$$\Lambda_{rn} \overset{def}{=} \prod_{k=r}^{n} \frac{p(y_k \mid H_1)}{p(y_k \mid H_0)} > \beta \tag{2}$$

gdzie:

β – arbitralnie dobrana stała ($\beta > 1$),
r – chwila wystąpienia zmiany.

Jeżeli ciąg $\{z_k: k = r, ..., n\}$ jest dyskretnym szumem białym o rozkładzie gaussowskim, jego wartość oczekiwana przed skokiem wynosi μ_0, a w chwili r wzrosła o wartość ν, to kryterium (2) można zapisać w postaci:

$$\log(\Lambda_{rn}) = \frac{\nu}{\sigma^2} \cdot \sum_{k=r}^{n} \left(y_k - \mu_0 - \frac{\nu}{2} \right) \overset{def}{=} \frac{S_r^n(\mu_0, \nu)}{\sigma^2} > h, \quad h \overset{def}{=} \ln(\beta) \tag{3}$$

gdzie $S_r^n(\mu_0, \nu)$ jest tzw. sumą skumulowaną sygnału od r-tej do n-tej próbki.

Największą wiarygodność detekcji daje algorytm Page'a–Hinkleya [7]. Przyjmuje się w nim pewną maksymalną długość okna L, wyrażającą dopuszczalne opóźnienie detekcji i po uzyskaniu kolejnej próbki sygnału y_n wyznacza się $S_r^n(\mu_0, \nu)$ dla $r = n - L + 1, ..., n$. Hipotezę H_1 przyjmuje się jeśli

$$\max_{n-L<r\le n} \sup_{\nu} S_r^n(\mu_0, \nu) > \sigma^2 h \tag{4}$$

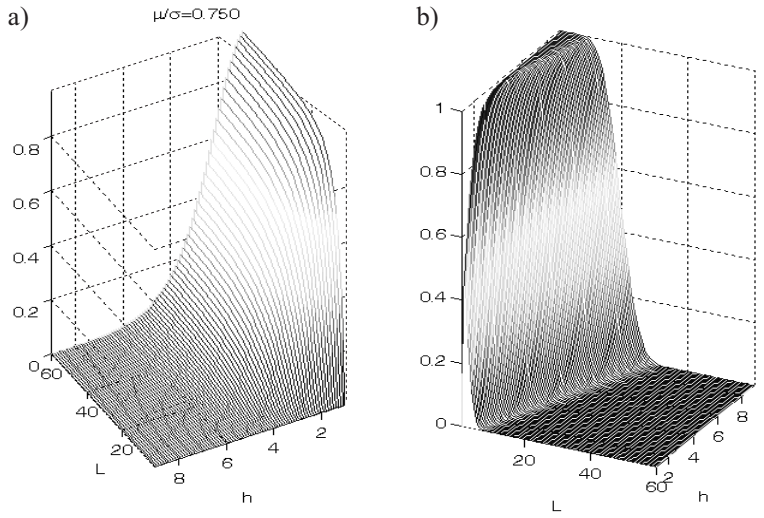
Jak widać, detekcja wymaga znajomości wartości oczekiwanej μ_0 sygnału przed zmianą, jego dyspersji σ i wielkości skoku ν. Testy mogą być prowadzone równolegle dla zestawu ustalonych względnych wartości skoku $\nu_w = \nu/\sigma$ (dodatnich i ujemnych), a kryterium decyzyjne dla każdego ν_w ma postać:

$$\begin{aligned} &T_j = \sum_{k=n-L+1}^{j} \left(\frac{y_k - \mu_0}{\sigma} - \frac{\nu_w}{2} \right), \; j = \tilde{n}_{L+1}, ..., n, \\ &\mathrm{H}_1 : \text{gdy} \max_{n-L<j<n} \left(\mathrm{sgn}(\nu_w)\left(T_n - T_j\right) \right) > h / |\nu_w| \end{aligned} \tag{5}$$

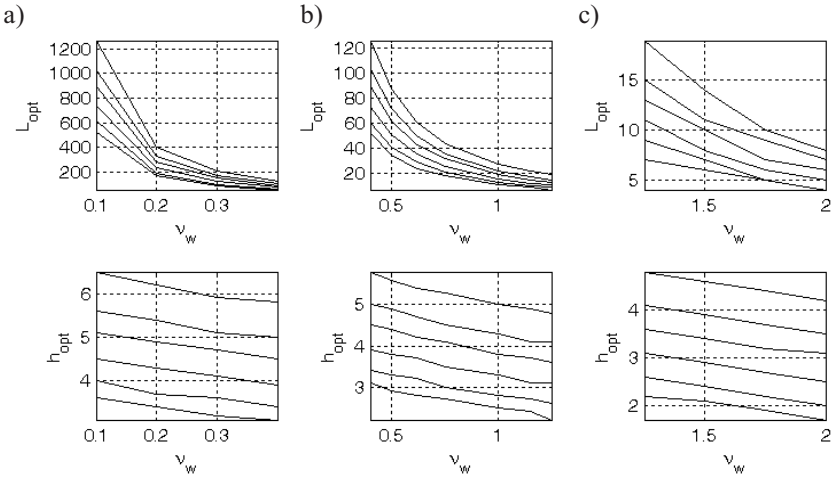
W alternatywnym podejściu wartość przyrostu v oblicza się na podstawie średniej arytmetycznej w oknie L_r, tj. dla próbek od r do n. Wykorzystanie uzyskanej w ten sposób estymaty ν_{rn} we wzorze (3) daje największą wartość $S_r^n(\mu_0, \nu_{rn})$, a warunek przyjęcia hipotezy H_1 ma wówczas postać nierówności:

$$\max_{n-L<r\le n} L_r \left(\nu_{rn} / \sigma \right)^2 > 2h \tag{6}$$

Parametry algorytmu, tj. maksymalną długość okna L oraz wartość parametru h można dobrać tak, aby uzyskać założone prawdopodobieństwo φ_{nvL} pominięcia skoku o wartości względnej $v_w = v/\sigma$ oraz prawdopodobieństwo fałszywego alarmu φ_{fL} ($\varphi_{fL} = 1 - \varphi_{n0L}$). Wymaga to scałkowania odpowiednich rozkładów warunkowych dla różnych założonych wartości względnych przyrostów v_w. Przykładowe wyniki takich obliczeń dla testu (5) pokazano na rysunku 1. Pozwalają one określić szerokość okna L_{opt} i odpowiadającą mu wartość h_{opt} w zależności od v_w, dla których uzyskuje się założone wartości φ_{fL} i φ_{nvL}. Wartości uzyskane dla różnych skoków względnych ν_w przyjętych w (5) przedstawiono na rysunku 2.



Rys. 1. Prawdopodobieństwo fałszywego alarmu φ_{fL} (a) oraz prawdopodobieństwo φ_{nvL} niewykrycia skoku wartości oczekiwanej o 0,75σ (b) dla testu (5) jako funkcja szerokości okna L i wartości arbitralnego parametru h



Rys. 2. Parametry L_{opt} i h_{opt} algorytmu Page'a–Hinkleya w wersji (5) odpowiadające różnym wartościom v_w. Kolejne krzywe od góry uzyskano dla $\varphi_{fL} = \varphi_{nvL}$ = {1%, 2%, 3%, 5%, 7,5% i 10%}. W celu poprawienia czytelności rysunków pokazano je w trzech zakresach v_w: od 0,1 do 0,4 (a), od 0,4 do 1,25 (b) i od 1,25 do 2 (c)

Z powyższych rysunków wynika, że uzyskanie odpowiedniej wiarygodności detekcji, tj. wartości φ_{fL} i φ_{nvL} rzędu 1%, dla $v_w < 0{,}5$ wymaga długich okien ($L_{\text{opt}} > 90$), co w praktyce oznacza niedopuszczalnie duże opóźnienie. Niemniej, okno zawierające 30 próbek po-

zwala już wykrywać wiarygodnie skoki nie mniejsze niż σ, a 13 próbek wystarcza do detekcji skoków od wartości 2σ.

Dla sygnału typu (1) algorytm Page'a–Hinkleya można łatwo dostosować do wykrywania nieskokowych zmian wartości x_n. W pracy [6] zbadano właściwości testu LR dla ciągu $\{x_i: i = r, \ldots, n\}$ narastającego liniowo od wartości μ_0 z szybkością a. Niech a_{rn} oznacza minimalno-kwadratową estymatę współczynnika a obliczoną dla próbek sygnału y_i od $i = r$ do n. Można wykazać, że do weryfikacji hipotezy H_1 według reguły typu (6) należy wówczas należy zastosować sumę skumulowaną $S_r^n(\mu_0, a)$, którą wyraża wzór

$$S_r^n(\mu_0, a_{rn}) = \frac{(a_{rn})^2}{2\sigma^2}\left(\frac{L_r(L_r+1)(2L_r+1)}{6}\right) \tag{7}$$

W praktyce, w czasie rzeczywistym winny być równocześnie wykonywane testy dla zmian skokowych i liniowych. W pracy [6] wykazano, że w przypadku zmian liniowych niezawodność samej detekcji zmiany jest dla testów (6) i (7) zbliżona, przy czym test zmodyfikowany (7) daje większe prawdopodobieństwo uzyskania poprawnej estymaty chwili rozpoczęcia zmiany, niż test klasyczny, pod warunkiem, że stosuje się go począwszy od $L_r > 3$. Z kolei w sytuacji, gdy zmiana sygnału użytecznego jest skokowa, test klasyczny (6) wykrywa ją z większym prawdopodobieństwem niż zmodyfikowany. Oznacza to, że równoczesne stosowanie obydwu testów zwiększa prawdopodobieństwo prawidłowego rozpoznania typu zmiany.

Algorytm Page'a–Hinkleya jest optymalną metodą wykrywania istotnych błędów optymalnej estymacji wartości oczekiwanej sygnału x_n za pomocą obserwatorów stanu [6, 10]. Może być także bezpośrednio wykorzystany do wiarygodnej detekcji istotnych niesprawności aparatury pomiarowej (uszkodzeń), np. na etapie analizy diagnostycznej torów pomiarowych w ramach cyfrowej filtracji antyaliasingowej [11]. Warto podkreślić, że zarówno test klasyczny (6), jak i zmodyfikowany (7) służą tylko do detekcji zmian sygnału użytecznego po dłuższym okresie jego stałości. Bezpośrednio po wykryciu zmiany analiza sygnału musi być zawieszona, aż do uzyskania miarodajnych estymat wartości μ_0 i σ. Ogranicza to możliwości jego wykorzystania jako narzędzia segmentacji szeregów, a także w diagnozowaniu przebiegów zmiennych procesowych (np. przewidzianych do stabilizacji), gdy badany sygnał jest niestacjonarny lub wykazuje istotną autokorelację, a estymaty wartości oczekiwanej zawierają składową losową. Sama idea testów LR może być jednak zastosowana jako narzędzie segmentacji takich szeregów, po odpowiednim uogólnieniu kryteriów typu (6) i (7).

Rozważmy ciąg v_n będący wyjściem pewnego procesu dynamicznego pobudzonego szumem białym z_n oraz sygnałem f_n, którego przebieg jest estymowany statycznym modelem regresyjnym. Przyjmijmy dla uproszczenia, że właściwości dynamiczne tego procesu opisuje adekwatnie model autoregresyjny pierwszego rzędu, którego parametr α i odchylenie standardowe σ szumu z są wyznaczone z pomijalnym błędem i nie ulegają zmianom. Proces może być opisany modelem Hammersteina [12]

$$v_n = \alpha v_{n-1} + f_{n-1} + z_n \tag{8}$$

lub złożeniem formuły trendu reprezentowanego ciągiem f_n i skorelowanych odchyłek losowych (reziduów)

$$v_n = \alpha(v_{n-1} - f_{n-1}) + f_n + z_n \tag{9}$$

Uogólniając ideę algorytmu Page'a–Hinkleya, można przyjąć, że celem analizy diagnostycznej ciągu (8) jest stwierdzenie, czy formuła wykorzystywana do estymacji ciągu f_i dla $i < n - L_r$ pozostaje adekwatna dla ostatnich L_r próbek, tj dla $i = r–1, r, ..., n–1$. Kwestię tę można rozstrzygać, wykorzystując testy stosunku wiarygodności dla „wybielonej" reprezentacji diagnostycznej y_n szeregu v_n. Ciąg diagnostyczny y_i dla $i = r, r + 1, ..., n$ oblicza się według wzoru

$$y_i \overset{def}{=} v_i - \alpha v_{i-1} - f_{i-1} \tag{10}$$

Właściwości uzyskanego ciągu y_i opisuje wzór (1), w którym x_i wyraża błąd estymacji składowej f_{i-1}. Modele typu (8) identyfikuje się na podstawie długich ciągów obserwacji [13], co pozwala przyjąć, że jeśli model jest adekwatny, to jego błąd losowy jest pomijalny, a wartość oczekiwana jest zerowa. Zatem, analogicznie jak w przypadku testu (6), analiza może być prowadzona z wykorzystaniem testu (3) dla sum skumulowanych obliczanych według wzoru

$$S_r^n(0, \hat{x}) = \frac{1}{\sigma} \sum_{i=r}^{n} \hat{x}_i \left(y_i - \frac{\hat{x}_i}{2} \right) \tag{11}$$

gdzie $\hat{x}_i$ oznacza i-tą wartość minimalno-kwadratowej aproksymaty błędu x_i, wyznaczonej na podstawie ciągu y_i dla $i = r, \ldots, n$.

Model (9) wykorzystuje się w analizach niestacjonarnych procesów stochastycznych o nieznanych wejściach. Jest on przedmiotem badań zespołu autora ukierunkowanych na opracowanie metod detekcji wczesnych symptomów dużych zmian trendów w szeregach finansowych ([14, 15, 16]). W metodzie segmentacji opracowanej w zespole autora [16] formułę f^0 trendu f wyznacza się dla bieżącego segmentu jako jawną funkcje czasu, stosując uogólnioną metodą najmniejszych kwadratów z macierzą autokorelacji zakłóceń odpowiadającą modelowi (9) [13]. W kolejnej chwili r należy rozstrzygnąć, czy formuła f^0 akceptowana w segmencie obejmującym pewną liczbę próbek wcześniejszych ($r - 1, r - 2, \ldots$) może być zastosowana dla próbki r-tej i dalszych (hipoteza H_0), czy też próbka r-ta winna rozpocząć nowy segment (hipoteza H_1). Wybielony sygnał umożliwiający testowanie tych hipotez wyraża się wzorem

$$y_i \overset{def}{=} v_i - f_i - \alpha(v_{i-1} - f_{i-1}) \tag{12}$$

gdzie f_i oznacza nieznaną, faktyczną wartość trendu (funkcji regresji 1. rodzaju) w chwili i-tej.

Jeśli do chwili r–1 hipoteza H_0 miała wiarygodne uzasadnienie, to od chwili r diagnozowanie ciągu (9) prowadzi się analizując ciąg wartości y_i^0 obliczonych ze wzoru (12), w którym jako f_i podstawia się predykcję $f_{r-1,i}^0$ i-tej wartości trendu, uzyskaną dla chwili i-tej na podstawie minimalno-kwadratowej aproksymaty f_{r-1}^0 ciągu w bieżącym segmencie, obejmującym próbki do $(r-1)$ włącznie. Faktyczne właściwości ciągu y_i^0 opisuje równanie

$$y_i^0 = \delta f_i - \alpha \delta f_{i-1} + z_i \tag{13}$$

gdzie δf_i jest błędem predykcji $f_{r-1,i}^0$ dla i-tej próbki.

Jeśli pierwsza odchyłka y_r^0 od zera jest statystycznie nieistotna (wg testu Studenta), to hipotezę H_0 uznaje się za potwierdzoną, co oznacza dołączenie punktu r do bieżącego segmentu i obliczenie zaktualizowanej aproksymaty f_r^0 trendu w segmencie. W przeciwnym wypadku, oblicza się predykcje $f_{r-1,n}^0$ kolejnych wartości trendu f_n i metodami SPC [3] sprawdza się istotność serii odchyłek $\{y_r^0, y_{r+1}^0, ..., y_n^0\}$ dla kolejnych n, do chwili $n = r + L_0 - 1$, gdzie L_0 jest minimalną długością ciągu, dla którego można wyznaczyć miarodajną aproksymatę wymaganego rzędu (np. dla trendu liniowego L_0 winno wynosić conajmniej 6). Warto zwrócić uwagę, że dla $i > r$ wzór (13) wykazuje mniejsze wartości odchyłek, niż wynikałoby to z faktycznych błędów δf_i predykatora f_{r-1}^0. W związku z tym należy zastosować testy dopuszczające względnie duże prawdopodobieństwo błędnego uznania istotności odchyłek. Nieistotność serii dla $n < r + L_0$ jest traktowana jako potwierdzenie hipotezy H_0 i skutkuje rozszerzeniem segmentu do próbki n-tej oraz aktualizacją modelu f_n^0. Stwierdzenie serii L_0 istotnych odchyłek powoduje uruchomienie dla kolejnych danych $n = r + L_0$, $r + L_0 + 1$, $r + L_0$+2, ..., testu stosunku wiarygodności opisanego niżej. Jeśli test potwierdzi hipotezę H_1 dla $n < r + L$, gdzie L jest przyjętym maksymalnym opóźnieniem decyzji, to szereg v_i dla $i = r, r + 1, ..., r + L_{smin} - 1$ jest traktowany jako nowy segment, gdzie L_{smin} oznacza minimalną założoną długość segmentu, pozwalającą na miarodajne testowanie kolejnych dalszych próbek $n = r + L_{smin}, r + L_{smin} + 1, ...$ Wartość L_{smin} winna też odzwierciedlać cel segmentacji, precyzując np. pojęcie trendów długoterminowych, jeśli takie mają być wyznaczane. W okresie od chwili $i = r + L$ do $r + L_{smin} - 1$ test jest zawieszony, ale dla kolejnych próbek i następuje aktualizacja modelu f_i^1 trendu f_i w nowym, kolejno rozszerzanym segmencie. Oczywiście, nie potwierdzenie hipotezy H_1 do chwili $n = r + L - 1$ skutkuje przyjęciem hipotezy H_0, a więc rozszerzeniem poprzedniego segmentu o badaną serię L próbek i aktualizacją modelu f_{r+L-1}^0.

Jeśli przyjąć, że błędy δf_i formuły f_i^0 w segmencie 0 są pomijalne, to test stosunku wiarygodności można przeprowadzić w taki sam sposób jak dla modelu (8), według sum skumulowanych obliczanych ze wzoru (11), w którym $\hat{x}_i$ oznacza i-tą wartość minimalno-kwadratowej aproksymaty ciągu y_i^0 ujawniającej rozbieżności Δf_i między faktycznymi wartościami trendu f_i, a ich predykcją $f_{r-1,i}^0$:

$$\hat{x}_i \cong \Delta f_i - \alpha \Delta f_{i-1} \quad \text{dla} \quad i = r, r+1, \ r+2, ..., n \tag{14}$$

Jak wspomniano wcześniej, jeśli w chwili r nastąpiła zmiana parametrów trendu f_i, to zgodnie ze wzorem (14) tylko pierwszy element e_r^0 testowanego ciągu odzwierciedla prawidłowo błąd modelu f_i^0, gdyż $\delta f_{r-1} \cong 0$. Dalsze wartości $\hat{x}_i$ są mniejsze niż faktyczny błąd formuły f_i^0 w hipotetycznie nowym segmencie, co obniża efektywność jego estymacji, a następnie wykrywania wg kryterium (4). Oznacza to także, że formułę $\hat{x}$ w wzorze (14) należy wyznaczać z pominięciem próbki r-tej. Przy niedużej długości L_r testowanego ciągu ($L_r = n - r + 1$) obniża to dodatkowo efektywność estymatorów błędu modelu. Wynika stąd, że niezawodność detekcji zmian trendu tą metodą może być znacznie niższa niż dla modelu (8).

Mając na względzie powyższe niekorzystne właściwości takiego podejścia, w pracy [16] zaproponowano obliczanie testu LR, z wykorzystaniem oddzielnego ciągu testowego y_i^1 dla hipotezy H_1. Ciąg y_i^1 oblicza się wg wzoru (12), podstawiając jako wartości trendu f_i dla $i \geq r$ ich estymaty $f_{n,i}^1$ obliczone według minimalno-kwadratowego modelu regresyjnego uzyskanego metodą uogólnionych najmniejszych kwadratów dla oryginalnego ciągu v_i w oknie obejmującym nowy hipotetyczny segment, tj. dla $i = r, ..., n$. Pomijając błędy predykatora f_{r-1}^0, można wówczas obliczyć ciąg $\hat{x}_i$ wg wzoru (14), podstawiając $\Delta f_{r-1} = 0$ oraz $\Delta f_i = f_{r-1,i}^0 - f_{n,i}^1$ dla $i \geq r$, a następnie zastosować kryterium (4) dla sumy (11).

Podejście to pozwala uwzględnić również fakt, że w praktycznych zadaniach analizowane segmenty nie są na tyle długie, aby błąd predyktora $f_{r-1,i}^0$ można było uznać za pomijalny. Niech $s_{f0,i}$ oznacza dyspersję błędu predyktora 2 dla chwili i-tej (tj. dla wyprzedzenia $L_r = i - r + 1$), $k_{f0,i\,i-1}$ – współczynnik kowariancji błędów tego predyktora dla chwil (i, i–1). Stosownie do wzoru (13) dyspersję i-tego elementu ciągu y_i^0 wyraża formuła

$$s_{y0,i} = \sqrt{s_{f0,i}^2 - 2\alpha\, k_{f0,i\,i-1} + \alpha^2 s_{f0,i-1}^2 + \sigma^2} \tag{15}$$

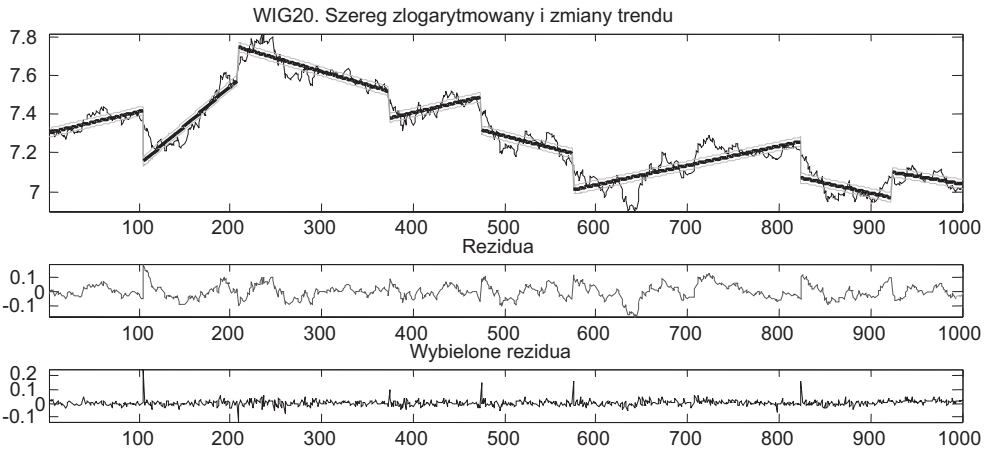
Niech $s_{y1,i}$ oznacza dyspersję wybielonych reszt modelu $f_{n,i}^1$ obliczoną jak we wzorze (15) dla błędów formuły $f_{n,i}^1$. Jeśli elementy ciągów y_i^0 oraz y_i^1 mają rozkłady N(0, $s_{y1,i}$) i N(0, $s_{y0,i}$) dla odpowiadających im hipotez, to iloraz iloczynów prawdopodobieństw y_i^1 i y_i^0 (patrz wzór (3)) dla hipotez H_1 i H_0 (w przedziale <r, n> o długości $L_r = n - r + 1$) ma postać

$$\log(\Lambda_{rn}) = \ln\left(\prod_{i=r}^{n} \frac{s_{y0,i}}{s_{y1,i}}\right) + \frac{1}{2}\sum_{i=r}^{n}\left(\left(\frac{y_i^0}{s_{y0,i}}\right)^2 - \left(\frac{y_i^1}{s_{y1,i}}\right)^2\right) \tag{16}$$

Z formalnego punktu widzenia wyrażenie (16) nie jest stosunkiem funkcji wiarygodności, gdyż błędy δf_i^0 predyktora $f_{r-1,i}^0$ są skorelowane dla kolejnych i ($k_{f0,i\,i-1} \neq 0$), a dodatkowa korelacji wynika z zależności (13). Niemniej może być ono widziane jako przybliżenie LR (tzw. czynnik Bayesa [17]), dające miarodajne podstawy do segmentacji

szeregów niestacjonarnych typu (9). Efekty pominiętej autokorelacji kompensuje częściowo obliczanie funkcji (16) według prawdopodobieństw całkowitych z dyspersjami $s_{y0,\,i}$, $s_{y1,\,i}$, zamiast warunkowych (o stałej dyspersji σ, co prowadziłoby do wzoru (11)).

Badania wykazały [16], że algorytm zmodyfikowany w powyższy sposób pozwala uzyskać bardzo dobre wyniki algorytmicznej detekcji względnie dużych zmian trendów długookresowych ($L_{s\min} > 30$), które są przedmiotem szczególnej uwagi w analizach szeregów finansowych. Ilustruje to rysunek 3.



Rys. 3. Segmentacja szeregu czasowego notowań dziennych indeksu WIG20 w okresie około 4 lat, z wykorzystaniem testu LR zmodyfikowanego wg wzoru (16). Parametry modelu dynamiki (9): $\sigma = 0{,}016$, $\alpha = 0{,}83$, trend f – liniowy. Parametry algorytmu LR: $h = 46$, $L_0 = 6$, $L = \max L_r = 15$, $L_{s\min} = 30$

3. Segmentacja szeregów czasowych metodą badania istotności trendu

W zadaniach segmentacji ukierunkowanych na pozyskiwanie danych do identyfikacji modeli regresyjnych procesów celowe jest wydzielanie sekwencji stosunkowo krótkich, ale licznych segmentów, często o niewiele różniących się wartościach średnich. Analizy przeprowadzone w rozdziale 2 pokazują, że testy typu LR mają wówczas słabą efektywność (wykrywanie słabych zmian wymaga długich ciągów testowych – patrz rys. 2), i są dedykowane raczej do wykrywania silnych zdarzeń rzadkich.

W pracy [8] zaproponowano algorytm lepiej dostosowany do specyfiki zadań segmentacji. Opiera się on na badaniu istotności statystycznej wygładzonej szybkości zmian przebiegu (zwanej trendem), obliczonej w oknie o założonej długości N, a także na analizie istotności różnic wartości średnich sygnału w różnych przedziałach czasu. Zwrócono uwagę na fakt, że jeśli segmentacja ma na celu wydzielenie stosunkowo krótkich segmentów, to odchyłki od wartości średniej (trendu) w segmencie nie zawierają niskoczęstotliwościowej części widma (usuniętej w wyniku uśredniania). Uwzględnienie tej właściwości szumów znacząco zwiększa niezawodność detekcji słabych zmian wartości średnich.

Przedział czasowy (t_p, t_k), w którym zachodzą istotne zmiany sygnału x nazywa się okresem niestacjonarności. Przyjmuje się, że w chwili t_n proces może znajdować się w jednym z trzech stanów:

1) stacjonarnym (H_0),
2) niestacjonarnym (H_1),
3) słabo stacjonarnym (H_2).

Hipoteza H_2 ma charakter roboczy w okresie, gdy nie ma podstaw ani do przyjęcia hipotezy H_1, ani H_0, a po okresie przejściowym musi nastąpić kwalifikacja do jednego z powyższych stanów podstawowych H_0 lub H_1. Stwierdzenie stacjonarności skutkuje obliczeniem wartości średniej $\bar{y}_m$ szeregu od chwili t_{k+1} do $t_m = t_{n-N}$, wykorzystywanej w dalszych testach. Segment stacjonarny jest zamykany dla ostatniej t_m, jeśli w chwili t_n nastąpiło potwierdzenie hipotezy H_1.

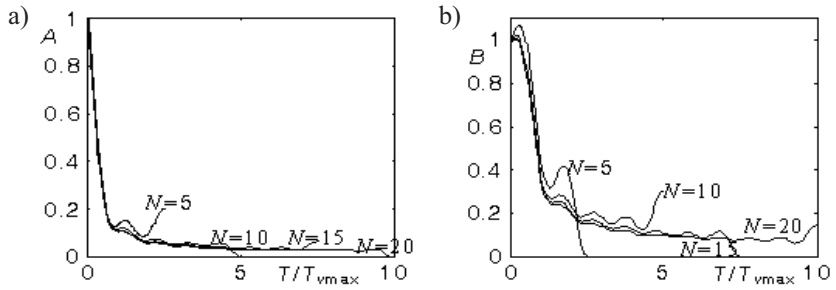
Algorytm opiera się na analizie właściwości statystycznych parametrów liniowej ortogonalnej aproksymacji minimalnokwadratowej ciągu $y_i = y(t_i)$ w oknie obserwacji zawierającym N próbek i kończącym się w chwili bieżącej n:

$$\hat{y}_{n-N+i} = a_y(t_n) + \Delta t[i - (N+1)/2]b_y(t_n), \quad i = 1, ..., N \tag{17}$$

gdzie:

$a_y(t)$ – wartość średnia ciągu $\{y_{n-N+i}, i=1, ..., N\}$,
$b_y(t)$ – średnia prędkość zmian tego ciągu (zwaną dalej trendem), obliczana metodą najmniejszych kwadratów.

Podstawą weryfikacji hipotezy o stacjonarności sygnału w oknie N są testy istotności następujących statystyk: trendu b_y, różnicy średniej a_y i średniej $\bar{y}_m$ w ostatnim okresie stacjonarności oraz wielkości pojedynczych odchyłek $|\bar{y}_m - y_n|$. Zakłada się, że szumy z_n we wzorze (1) reprezentują stacjonarne przebiegi czasowe o rozkładzie Gaussa, których widmo leży poza zasadniczym pasmem przenoszenia badanego obiektu, ograniczonym częstotliwością $\omega_d = 2\pi/T_{v\max}$, gdzie $T_{v\max}$ oznacza okres najniższej harmonicznej obecnej w sygnale. Przyjmuje się, że ich widmo jest w przybliżeniu równomierne w zakresie od ω_d do częstotliwości Nyquista π.



Rys. 4. Wariancja średniej arytmetycznej A (a) i trendu B szumu wysokoczęstotliwościowego (b) względem tych samych estymatorów dla szumu białego w funkcji stosunku szerokości okna $T=N\Delta t$ do okresu $T_{v\max}$ najniższej harmonicznej szumu, dla różnej liczby danych w oknie

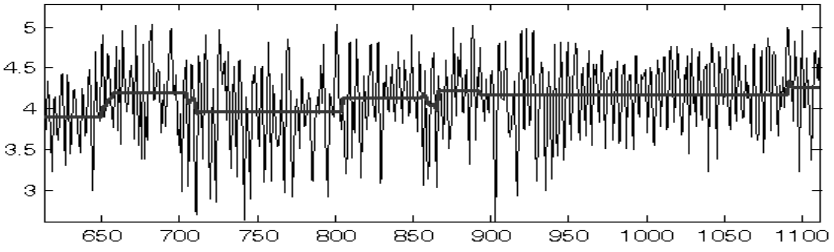
Zakłada się rozkład normalny badanych statystyk, a ich odchylenia standardowe oblicza się z uwzględnieniem wysokoczęstotliwościowego charakteru widma składowej losowej, co daje wartości znacznie mniejsze niż dla szumu białego (patrz rys. 4).

Wykorzystanie tej oczywistej właściwości odchyłek sygnału od hipotetycznej średniej $\bar{y}_m$ szeregu w niezbyt długich segmentach stacjonarnych daje radykalną poprawę selektywność detekcji zmian wartości średniej w typowych zadaniach segmentacji. Testy istotności uzupełnia analiza logiczna estymowanych sygnałów użytecznych, a także wprowadzenie mechanizmów regułowej detekcji początku okresu stacjonarności. Warunkiem przyjęcia hipotez H_1 i H_0 jest również przekroczenie minimalnej długość okresu, odpowiednio, niestacjonarności $T_{d\min}$ i stacjonarności $T_{s\min}$.

Wymagane parametry algorytmu to dyspersja szumu σ_v oraz wartość $T_{v\max}$. Z rysunku 4 wynika, że szerokość okna analizy $T=N\Delta t$ winna zawierać się w przedziale ($T_{v\max}$, $1{,}3T_{v\max}$). Dla mniejszych T selektywność testów szybko maleje, dla większych jej wzrost jest nieistotny. Parametr $T_{s\min}$ przyjmuje się arbitralnie ($T_{s\min} > T$), natomiast $T_{d\min}$ jest wyznaczany na podstawie teoretycznej funkcji autokorelacji ciągu $b_v(t_i)$ dla szumu z, jako dopuszczalna długość ciągu istotnych odchyłek losowych wynikająca z widma sygnału b_v.

Powyższy algorytm był wszechstronnie badany na symulowanych przebiegach czasowych, a także został sprawdzony na danych zarejestrowanych w kilku procesach przemysłowych. Wyniki tych badań przedstawiono w publikacjach [8, 18, 19, 20]. Symulacje wykazały [8], że algorytm ten pozwala uzyskać bardzo dobrą jakość klasyfikacji przebiegów nawet w przypadku, gdy stosunek dyspersji szumu σ_V do średniokwadratowej amplitudy s_x zmian sygnału $x(t)$ przekracza 3. Błędy estymacji wartości sygnału $x(t)$ utrzymują się wówczas na poziomie $0{,}01\cdot\sigma_V$. Przy większych wartościach (σ_V/s_x) obserwuje się stosunkowo często błędną klasyfikację przebiegu przy liniowych zmianach składowej $x(t)$. Zastosowane algorytmu Page'a–Hinkleya przy takim poziomie szumów daje albo dużą częstość fałszywych alarmów, albo ignorowanie zmian.

Celowość stosowania tego algorytmu do selekcji danych do identyfikacji charakterystyk statycznych procesu chemicznego pokazano w pracy [18]. Z ciągów zmiennych procesowych zawierających około 30 tys. próbek wybrano algorytmicznie około 100 obserwacji w stanach stacjonarnych i około 300 – odpowiadających stacjonarnym przebiegom wejść. Przykładowe wyniki segmentacji jednego z szeregów wykorzystanych w tym modelu przedstawia rysunek 5.

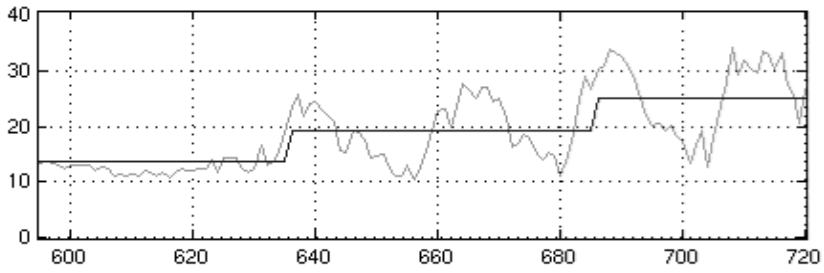


Rys. 5. Wyniki segmentacja metodą badania istotności trendu sygnału ciśnienia (próbkowanego co 1min.) w reaktorze syntezy kwasu azotowego [18]. Dla szumu przyjeto $\sigma_v = 0{,}65$, ograniczenie dolne widma $T_{v\max} = 15$. Parametry algorytmu segmentacji: $T = N = 18$, $T_{d\min}=10$, $T_{s\min}=25{,}0$

Zwraca uwagę bardzo wysoka skuteczność detekcji małych zmian wartości średniej na tle szumu wysokoczęstotliwościowego (stosunek odchylenia standardowego szumu do odchylenia standardowego zmian średniej pokazanych na rys. 5 wynosi 4,44). Modele charakterystyk statycznych i modele Hammersteina uzyskane na podstawie tak dobranych obserwacji wykazywały znacząco lepsze właściwości predykcyjne niż wyniki identyfikacji opartej na danych uśrednianych w stałych przedziałach czasu.

Przedstawiony algorytm charakteryzuje się również dobrą odpornością na błędy oceny parametrów charakterystyki widmowej szumu z. Przyjęcie zbyt dużej wartości $T_{v\max}$ lub σ_v nie wpływa na jakość klasyfikacji przebiegu. Nie ma również większego znaczenia kształt widma szumu w zakresie od $T_{v\max}$ do $2\Delta t$ (patrz rys. 6). Jednakże założenie wartości σ_v niższej niż rzeczywista dyspersja szumu lub znacznie niższej wartości $T_{v\max}$ powoduje istotne zwiększenie błędów segmentacji. Wskazane jest zatem stosowanie nadmiarowych oszacowań powyższych parametrów charakterystyki widmowej szumu z.

Skuteczność segmentacji szeregu niespełniającego ściśle założeń algorytmu ilustruje rysunek 6. Celem segmentacji było tu zebranie danych do identyfikacji charakterystyk statycznych około 1 tys. instalacji ogrzewania budynków, dla potrzeb ich scentralizowanego nadzorowania. Aby wyeliminować wpływ dynamiki procesu ogrzewania, należało odpowiednio uśrednić składowe losowe i przejściowe jednostkowego zużycia energii [19]. Charakterystyki te były następnie grupowane w celu wyodrębnienia klas sprawności instalacji [19, 20].



Rys. 6. Segmentacja przykładowego szeregu czasowego jednostkowego zużycia mocy w domowej instalacji grzewczej (okres próbkowania – 1h) [19]. Założone parametry szumu $\sigma_v = 3{,}7$, $T_{v\max} = 24$. Parametry algorytmu segmentacji: $T = N = 27$, $T_{d\min} = 15$, $T_{s\min} = 48$

4. Podsumowanie

Klasyczny algorytm Page'a–Hinkleya oparty na testach LR nie jest odpowiednim narzędziem segmentacji szeregów czasowych, ukierunkowanych na pozyskiwanie miarodajnie uśrednionych informacji o właściwościach szeregu stosunkowo krótkich segmentach. Wynika to z jego dostosowania do wykrywania rzadko występujących zmian wartości oczekiwanej (a nie średniej) w obecności białych szumów. Daje to podstawy do pominięcia błędów estymacji wartości oczekiwanej przed jej zmianą.

Zaproponowana w tym artykule, zmodyfikowana wersja algorytmu LR pozwala uwzględnić błędy estymacji parametrów szeregu w segmentach, ale jej efektywność dla

małych zmian tych parametrów jest niska. Metoda to może być rekomendowana do segmentacji szeregów niestacjonarnych, w celu ich dekompozycji na składową niestacjonarną aproksymowaną przyjętą formułą trendu oraz na stacjonarny sygnał resztowy opisany modelami autoregresyjnymi.

W artykule pokazano, że niezawodność detekcji zdarzeń procesowych rozpoczynających i kończących okresy stacjonarności typowych zmiennych procesowych można znacznie zwiększyć, wykorzystując wieloaspektowe testy statystyczne wsparte regułami, z uwzględnieniem specyficznych właściwości widmowych sygnałów w typowych (stosunkowo krótkich) okresach stacjonarności. Opracowane przez autora testy uwzględniają fakt, że odchyłki od wartości średniej w krótkich segmentach nie zawierają niskoczęstotliwościowej części widma sygnału (usuniętej w wyniku uśredniania), a więc mają charakter wysokoczęstotliwościowy. Zwiększa to radykalnie selektywność typowych statystyk, w podejściu klasycznym obliczanych przy założeniu braku autokorelacji zakłóceń. Algorytm segmentacji oparty na takich testach umożliwia wydzielanie segmentów o małych różnicach wartości średnich, kilkakrotnie mniejszych niż dyspersja sygnału resztowego (zakłóceń).

Literatura

[1] Chen Z.: *Data Mining and Uncertain Reasoning.* N. York, J. Wiley&Sons 2001

[2] Wang X.Z.: *Data Mining and Knowledge Discovery for Process Monitoring and Control.* Springer-Verlag London, 1999

[3] Wetherill G.B., Brown D.W.: *Statistical Process Control. Theory and practice.* Chapman and Hall, 1991

[4] Korbicz J., Kościelny J.M., Kowalczuk Z., Cholewa W. (red.): *Diagnostyka procesów – modele, metody sztucznej inteligencji, zastosowania.* Warszawa, WNT 2002

[5] Duda J.T., Stanek P., Janik K.: *Algorytmiczna selekcja danych do identyfikacji on-line statycznych modeli procesów wolnozmiennych*. Elektrotechnika, t. 14, z. 3, 1995, 193–203

[6] Duda J.T.: *Modele matematyczne, struktury i algorytmy nadrzędnego sterowania komputerowego*. Kraków, UWND AGH 2003

[7] Basseville M.: *Detecting Changes in Signals and Systems – Survey.* Automatica, vol. 24, No. 3, 1988, 309–326

[8] Duda J.: *Wybrane zagadnienia syntezy algorytmów sterowania nadrzędnego wolnozmiennymi, ciągłymi procesami przemysłowymi*. Zeszyty Naukowe AGH, nr 1420, Automatyka, z. 56, 1991

[9] Basseville M., Benveniste A. (Eds): *Detection of Abrupt Changes in Signals and Dynamical Systems.* LNCIS No. 77, Berlin, Springer 1986

[10] Byrski W.: *Obserwatory i ich zastosowanie w systemach sterowania adaptacyjnego*. Kraków, ZN AGH, Nr 1551, Automatyka z.65, 1993

[11] Augustyn J., Duda J.T.: *Wykorzystanie procesorów sygnałowych DSP do analizy sprawności układów kontrolno-pomiarowych.* Materiały II Krajowej Konferencji Naukowo-Technicznej "Diagnostyka Procesów Przemysłowych", Łagów k. Zielonej Góry, 8–11 września 1997, 150–155

[12] Niederliński A.: *Systemy komputerowe automatyki przemysłowej. t.2 Zastosowania.* Warszawa, WNT 1985

[13] Mańczak K., Nachorski Z.: *Komputerowa identyfikacja obiektów dynamicznych.* Warszawa, PWN 1983

[14] Pełech T., Duda J: *Event Detection in Financial Time Series By Immune-Based Approach. New Trends in Intelligent Information Processing.* Lecture Notes (Springer), 2006

[15] Duda J., Pełech T.: *Wykrywanie zdarzeń w szeregach finansowych z wykorzystaniem metod statystycznych.* Rozdział w monografii: „Systemy Ekspertowe", Tom 2. Monografia zbiorowa – redaktor Adam Grzech, Oficyna Wydawnicza Politechniki Wrocławskiej, Wrocław, czerwiec 2006, 317–327
[16] Duda J.T., Pełech-Pilichowski T.: *Detekcja wczesnych symptomów zmian trendu szeregów finansowych*. WZ AGH 2007 (materiał przygotowywany do publikacji)
[17] Jeffreys H.: *The Theory of Probability (3e).* Oxford, 1961
[18] Duda J.T., Stanek P., Janik K.: *Algorytmiczna selekcja danych do identyfikacji on-line statycznych modeli procesów wolnozmiennych.* Elektrotechnika, Tom 14. Zeszyt 3, 1995, 193–203
[19] Kiluк S., Duda J.T.: *Diagnozowanie sprawności urządzeń grzewczych w systemach masowego nadzorowania.* Pomiary Automatyka Kontrola, 2005, nr 9, 188–190
[20] Kiluk S., Duda J.T.: *Remote Diagnosis of Heating Systems.* 4th International Congress on Intelligent Building Systems InBuS 2006, Cracow 2006, 53–58