INFORMATYKA EKONOMICZNA BUSINESS INFORMATICS 22 • 2011

ISSN 1507-3858

Krzysztof Drelczuk, Jerzy Korczak
Uniwersytet Ekonomiczny we Wrocławiu

NIENADZOROWANA DETEKCJA ANOMALII KONTEKSTOWYCH W FINANSOWYCH SZEREGACH CZASOWYCH O DUŻEJ CZĘSTOŚCI Z WYKORZYSTANIEM FALEK DAUBECHIES

Streszczenie: W niniejszej pracy autorzy zaprezentują algorytm oparty na dyskretnych transformatach falkowych do detekcji anomalii kontekstowych w finansowych szeregach czasowych. W tym celu wykorzystane zostaną falki ortogonalne Daubechies (D1-D30). Badanie zostanie przeprowadzone na danych pochodzących z rynku FOREX dla pary walutowej EUR/USD dla okresu 10 lat.

Słowa kluczowe: dyskretne transformaty falkowe, anomalie, finansowe szeregi czasowe, rynek FOREX.

1. Wstęp

Termin „detekcja anomalii" w danych oznacza proces wykrywania nietypowych wzorców lub próbek w danych, które w jakiś sposób odbiegają od oczekiwanego zachowania (*outliers*). Rozwiązania tego problemu znajdują szerokie zastosowanie w całej gamie aplikacji w różnych dziedzinach gospodarki czy też nauki [Ji 2008]. Jako przykłady zastosowań można podać ostrzeganie przed nieoczekiwanymi spadkami lub wzrostami kursu akcji, wykrywanie kradzieży kart kredytowych, sygnalizację nadużyć ubezpieczeniowych czy podatkowych, identyfikację włamań do systemów informatycznych itp. [Chandola i in. 2009].

Waga informacji o wykrytych anomaliach jest duża ze względu na to, że może nieść często bardzo ważną, a nawet czasami krytyczną informację. Dla przykładu, wykryta anomalia w transakcjach płatniczych karty kredytowej może oznaczać jej kradzież. W przypadku systemów informatycznych nietypowe połączenie użytkownika może być oznaką próby włamania się do systemu. W przypadku finansowych szeregów czasowych może wskazywać na próbę manipulowania rynkiem, a w przypadku szeregów czasowych o dużej częstości może wskazywać na pojawienie się na

rynku bardzo ważnej informacji bądź zbiorowej histerii lub euforii. Informacje takie z punktu widzenia inwestora mają kluczowe znaczenie [Lazarevic 2003].

W literaturze tematu wskazuje się na trzy rodzaje anomalii: punktową, kontekstową oraz zespołową [Chandola i in. 2009]. W celu rozpoznania anomalii stosowanych jest wiele metod i algorytmów, które można zaliczyć do jednej z trzech klas, mianowicie metod nadzorowanych, pół-nadzorowanych bądź nienadzorowanych. W pierwszej z nich zakłada się istnienie zbioru uczącego zawierającego wzorce uznane za normalne i anormalne. W drugiej w zbiorze uczącym umieszcza się jedynie wzorce normalnych zachowań. Każda rozbieżność względem zbioru uczącego traktowana jest jako anomalia. W ostatniej metodzie nie zakłada się istnienia zbioru uczącego, a anomalie lokalizowane są w czasie rzeczywistym na podstawie ustalonych metryk lub algorytmów.

W niniejszej pracy autorzy skoncentrują się na detekcji anomalii kontekstowych. Anomalie kontekstowe (*contextual anomalies*) to obserwacje, które uznawane są za anormalne tylko i wyłącznie w określonym kontekście [Chandola i in. 2009]. W identyfikacji anomalii przyjęto metodę nienadzorowaną, która pozwala na analizę w czasie rzeczywistym strumieni danych o dużej częstości zmian. Założenie to pozwoli na łatwą integrację rozwiązania z już istniejącymi systemami analizy szeregów czasowych oraz rozszerzy spektrum zastosowań opracowanych metod. Do detekcji anomalii wykorzystana zostanie dyskretna transformata falkowa. Generalnie, algorytmy falkowe są skuteczne w wyszukiwaniu osobliwości w sygnałach, co zostało potwierdzone w wielu dziedzinach. Wybór taki podyktowany został też dużą skutecznością oraz szybkością przetwarzania strumieni danych w czasie rzeczywistym [Antoniadis i in. 2010].

Główna hipoteza postawiona w artykule jest następująca: „Z wykorzystaniem transformacji falkowej z dużym prawdopodobieństwem (powyżej 60%) można zlokalizować anomalie występujące w finansowych szeregach czasowych, których następstwem jest znaczący wzrost lub spadek wartości notowań".

Weryfikacja takiej hipotezy wymagać będzie realizacji trzech zadań. W pierwszej kolejności będzie należało określić, czy istnieje pewna wartość, dla badanej pary walutowej, nazywana dalej poziomem anomalii λ, przy której hipoteza jest spełniona. Drugim zadaniem będzie określenie optymalnej wielkości okna przesuwnego, dla którego prawdopodobieństwo detekcji anomalii jest relatywnie największe. Ostatnim zadaniem będzie wskazaniem falki z rodziny Daubechies, dla których prawdopodobieństwo detekcji anomalii jest relatywnie największe.

Artykuł został podzielony na trzy główne części. W pierwszej przedstawiono dyskretną transformatę falkową. W drugiej części podano opis metody badawczej ze szczególnym uwzględnieniem finansowych szeregów czasowych. W trzeciej części zaprezentowano wyniki eksperymentów. Materiałem badawczym dla weryfikacji hipotezy były notowania pary walutowej EUR/USD na rynku FOREX

w ciągu 10 lat (od 16 czerwca 2001 do 16 czerwca 2011). W zakończeniu podsumowano doświadczenia i zarysowano program przyszłych prac.

2. Dyskretna transformata falkowa

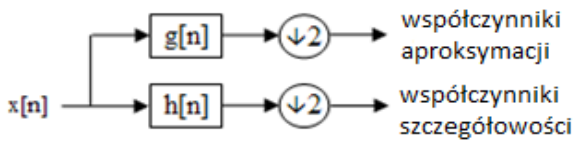
Analiza falkowa polega na dekompozycji sygnału i przedstawieniu go w postaci liniowej kombinacji funkcji bazowych, zwanych falkami. Cechami odróżniającymi tę metodę analizy sygnału od innych metod są wielostopniowa dekompozycja sygnału, zmienna rozdzielczość w dziedzinie czasu i częstotliwości oraz możliwość stosowania funkcji bazowych innych niż funkcje harmoniczne. Istnieją dwa podejścia do transformaty falkowej: dyskretnej DWT (*Discrete Wavelet Transform*) i ciągłej CWT (*Continuous Wavelet Transform*) [Daubechies 1992].

Dyskretna transformata falkowa sygnału x polega na przepuszczeniu go przez układ filtrów. Najpierw przez filtr dolnoprzepustowy g. Równocześnie sygnał jest przepuszczany przez filtr górnoprzepustowy h [Strang, Nguyen 1997]. Wyniki będące współczynnikami szczegółowości (z filtra górnoprzepustowego) oraz współczynnikami aproksymacji sygnału (z filtru dolnoprzepustowego) po połączeniu tworzą kwadratowy filtr lustrzany [Crochiere i in. 1976]. Sygnały wyjściowe opisują następujące formuły:

$$\begin{aligned} y_{\text{low}}[n] &= \sum_{k=-\infty}^{\infty} x[k]g[2n-k] \\ y_{\text{high}}[n] &= \sum_{k=-\infty} x[k]h[2n-k] \end{aligned} \tag{1}$$

gdzie: x – sygnał wejściowy,
h – filtr górnoprzepustowy,
g – filtr dolnoprzepustowy.

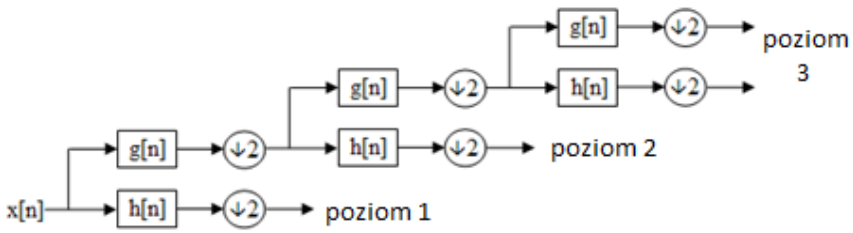
Zgodnie z teorią sygnałów dyskretną transformatę falkową możemy opisać za pomocą diagramu przedstawionego na rys. 1.



Rys. 1. Jednopoziomowa dyskretna transformata falkowa

Źródło: opracowanie własne.

Tak przeprowadzoną dekompozycję możemy powtarzać rekursywnie, otrzymując sygnał o coraz to mniejszej rozdzielczości. Przedstawia to rys. 2.



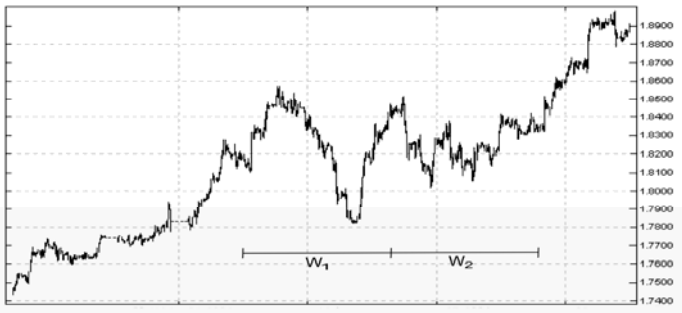
Rys. 2. Wielopoziomowa dyskretna transformata falkowa

Źródło: opracowanie własne.

Wynikiem każdego poziomu transformaty są dwa wektory o długości równej połowie wektora wejściowego (symbol ↓2). Z tego powodu długość sygnału wprowadzonego do systemu musi być wielokrotnością 2^n, gdzie n jest liczbą możliwych poziomów transformaty.

3. Metoda badania

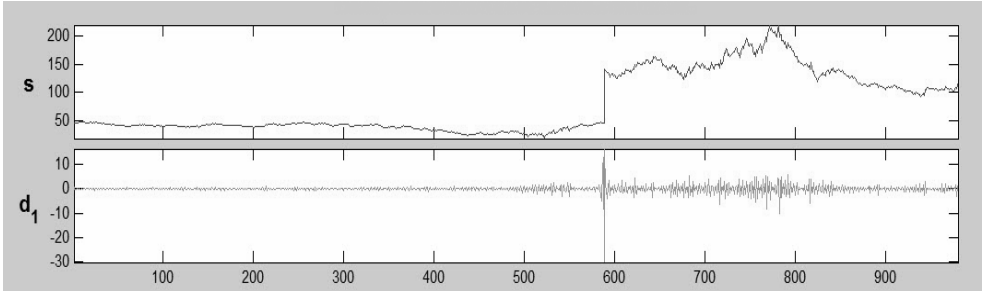
Badania zostały przeprowadzone z wykorzystaniem rodziny falek Daubechies o współczynnikach od 2 do 30. Analiza anomalii dotyczyła 10 lat notowań pary walut EUR/USD na międzynarodowym rynku FOREX (od 16 czerwca 2001 do 16 czerwca 2011). Ze względu na bardzo dużą ilość obserwacji notowania zostały zagregowane do 1 minuty. Do skonstruowania wektorów kodowych użytych do badań wykorzystano okna przesuwne o wielkości 32, 64, 128, 256 oraz 512. W celu weryfikacji hipotezy dla każdego wektora kodowego W_1 utworzono również wektor kontrolny W_2 o takiej samej wielkości. Ilustruje to rys. 3. Na wykresie na osi poziomej odłożono czas, na pionowej natomiast wartość szeregu. Skala czasu nie jest istotna, dlatego została na wykresie pominięta.



Rys. 3. Wektor kodowy i kontrolny

Źródło: opracowanie własne.

W badaniu autorzy przyjęli, że anomalia występuje wtedy i tylko wtedy, gdy różnica maksymalnej i minimalnej wartości szeregu współczynników szczegółowości h jest większa niż założony poziom λ. Wartość poziomu λ jest przedmiotem badania niniejszego projektu. Przykładową anomalię przedstawia rys. 4. Nagła zmiana sygnału s powoduje znaczące zwiększenie amplitudy detali pierwszego poziomu d_1.



Rys. 4. Sygnał wraz ze współczynnikami szczegółowości pierwszego poziomu po dekompozycji falką D4

Źródło: opracowanie własne.

Na tej podstawie zdefiniowano funkcję sukcesu/porażki:

$$f(W_1, W_2) \to \{0, 1\},$$

$$f(W_1, W_2) \to \begin{cases} 0 : [h < \lambda \wedge m > c] \vee [h > \lambda \wedge m < c] \\ 1 : [h < \lambda \wedge m < c] \vee [h > \lambda \wedge m > c] \end{cases},$$

$$m = \max\left(\frac{100 \cdot \max(W_2)}{W_1[s]}, \frac{100 \cdot \min(W_2)}{W_1[s]} \right), \tag{2}$$

$$h = \left| \max\left(h[W_1]\right) - \min\left(h[W_2]\right) \right|,$$

gdzie: s – wielkość okna ($W_1[s]$ oznaczać będzie ostatnią obserwację w oknie W_1),

$h[W_1]$ – szereg detali pierwszego poziomu po dekompozycji odpowiednią falką Daubechies,

c – założony poziom zmiany zdefiniowany w tab. 1.

Podsumowując, algorytm osiąga sukces, gdy wykryta anomalia ($h > \lambda$) skutkuje c procentowym wzrostem lub spadkiem (ta sama wartość dla obu przypadków) wartości szeregu czasowego w oknie kontrolnym w stosunku do ostatniej obserwacji okna badanego lub też jej brak ($h < \lambda$) nie powoduje takiej zmiany. Porażka al-

gorytmu zdefiniowana jest jako wzrost *c* procentowy wartości szeregu czasowego w oknie kontrolnym w stosunku do ostatniej obserwacji okna badanego bez wykrycia anomalii, jak i brak takowego przy anomalii wykrytej.

W zależności od wielkości okna parametr *c* został ustalony empirycznie, a jego wartości pokazuje tab. 1.

Tabela 1. Stosunek liczby zmian większych oraz mniejszych od 3% w zależności od wielkości okna

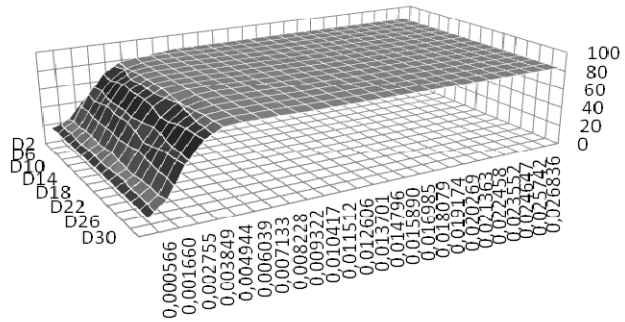
Wielkość okna	Liczba wektorów kodowych	Wartość parametru *c*	Liczba wektorów kontrolnych ze zmianą większą niż *c*	Liczba wektorów kontrolnych ze zmianą mniejszą niż *c*
32	120 240	0,306	60 000	60 240
64	60 060	0,488	29 820	30 240
128	30 000	0,671	15 000	15 000
256	14 940	1,025	7 440	7 500
512	7 440	1,290	3 720	3 720
1 024	3 660	2,504	1 800	1 860

Źródło: opracowanie własne.

Wartość *c* została ustalona inna dla każdej wielkości okna tak, aby ilość wektorów ze zmianą mniejszą oraz większą niż ten parametr była mniej więcej równa. Taki zabieg ma na celu wiarygodniejszą weryfikację hipotezy postawionej we wstępie niniejszego artykułu. Gdyby parametr *c* został ustalony arbitralnie, mogłoby dojść do takiej sytuacji, że liczba wektorów kontrolnych ze zmianą np. większa niż *c* wynosiłaby 100% wszystkich wektorów kontrolnych. Wiarygodna weryfikacja poprawności algorytmu byłaby wtedy niemożliwa do zrealizowania. Przy zbyt wysokim *c* funkcja zwracałaby zawsze ‘sukces’ i odwrotnie – przy zbyt niskim zawsze ‘porażkę’. Analogią może być tutaj analizowanie algorytmów prognozujących wzrost wartości finansowych szeregów czasowych w trendzie zwyżkowym i odwrotnie – algorytmów prognozujących spadek w trendzie spadkowym.

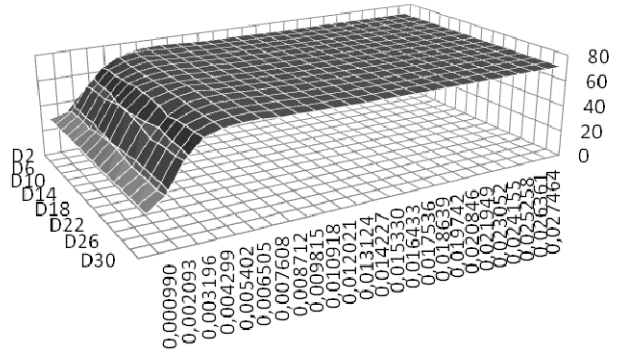
4. Wyniki eksperymentów

Istotą badania było wykazanie, że dzięki wykryciu anomalii można *a priori* przewidzieć, że wartość badanego szeregu czasowego zmieni się (to jest wzrośnie lub zmaleje) o więcej niż założona wartość *c*. Na rysunkach od 5 do 9 zaprezentowano uzyskane wyniki. Na osi poziomych odłożono rodzaj falki oraz współczynnik anomalii λ. Na osi pionowej natomiast skuteczność algorytmu zgodnie z funkcją sukcesu/porażki zdefiniowanej we wzorze 2 (po przeskalowaniu do skali procentowej – od 0 do 100).



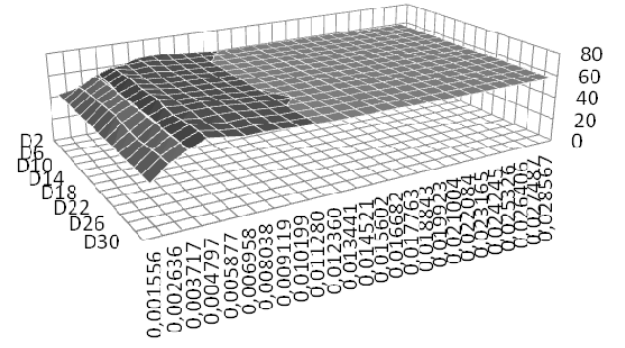
Rys. 5. Procentowa liczba poprawnych decyzji dla okna wielkości 32

Źródło: opracowanie własne.



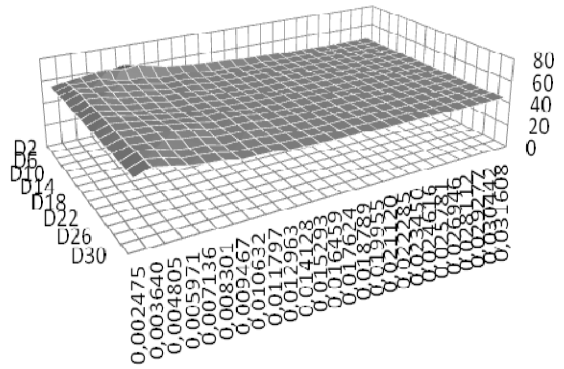
Rys. 6. Procentowa liczba poprawnych decyzji dla okna wielkości 64

Źródło: opracowanie własne.



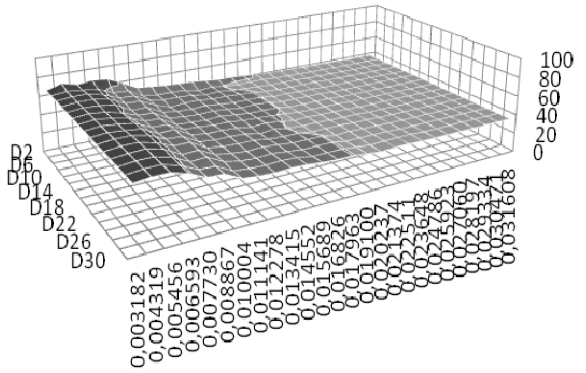
Rys. 7. Procentowa liczba poprawnych decyzji dla okna wielkości 128

Źródło: opracowanie własne.



Rys. 8. Procentowa liczba poprawnych decyzji dla okna wielkości 256

Źródło: opracowanie własne.

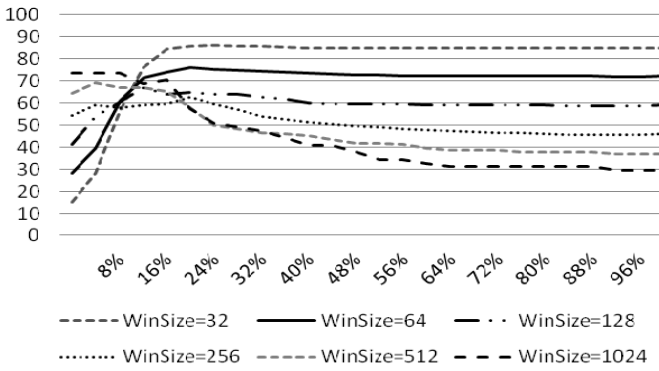


Rys. 9. Procentowa liczba poprawnych decyzji dla okna wielkości 512

Źródło: opracowanie własne.

Badania wykonano dla okien o wielkości od 32 do 512 (czyli od $2n$ do $2n+4$ dla n początkowego 5 przy kroku 1). Zagregowane (uśrednione w obrębie jednego okna) wyniki zaprezentowano na rys. 10. Na osi poziomej odłożono znormalizowany współczynnik anomalii λ. 100% oznacza największy występujący współczynnik w badaniu (największa zaobserwowana anomalia). 0% oznacza najmniejszy występujący współczynnik w badaniu (brak anomalii). Jak widać, powyżej pewnego poziomu współczynnika λ (około 20%) dla okien o wielkości 32 i 64 z dużym prawdopodobieństwem stwierdzamy, że wykryta anomalia poprzedza zmianę wartości szeregu o zadaną wartość. Należy zauważyć, że nie zachodzi relacja odwrotna. Przy niskim współczynniku anomalii dla tej wielkości okien nie jesteśmy w stanie stwierdzić, że wartość szeregu nie zmieni się o założoną wartość. Przy współczynniku λ poniżej 8% skuteczność takiej prognozy wynosiła mniej niż 50%.

Dla znacząco większych okien o wielkości 256 i 512 sytuacja wygląda dokładnie odwrotnie. Brak wykrycia anomalii determinuje brak zmiany wartości szeregu, natomiast jej obecność nie umożliwia stwierdzenia, czy wartość szeregu zmieni się czy nie. Skuteczność algorytmu spada nawet do 30%.



Rys. 10. Zestawienie średnich wyników dla wszystkich okien przesuwnych

Źródło: opracowanie własne.

W przypadku okien o średniej wielkości wykorzystanych w badaniu, czyli 128, średnia skuteczność oscylowała w granicach 50%; tak więc dla tych okien nie ma przesłanek do przyjęcia hipotezy postawionej w artykule.

Badanie nie wykazało znaczących różnic pomiędzy wyborem rodzaju falki z rodziny Daubechies a jakością lokalizacji anomalii. Wskazuje to na wybór, bez straty jakości dla algorytmu, falki najmniej złożonej obliczeniowo (D2 lub D4). Dodatkowo, ze względu na charakterystykę falek od D2 do D30, może to wskazywać na to, że na wykrycie anomalii ma tak samo mocny wpływ sekwencja ruchów notowań, jak i ich moc. Kwestią otwartą pozostaje, czy skuteczność algorytmu opartego na analizie sekwencji współczynników uzyskanych z transformacji falkowych byłaby równie wysoka.

5. Podsumowanie i dalsze prace

Autorzy niniejszej pracy przebadali notowania pary waluty EUR/USD na rynku FOREX w ostatnich dziesięciu latach. Dane były zagregowane do 1 minuty. Łącznie szereg czasowy poddany analizie miał długość 3,850,560 obserwacji. Szybkość obliczeń, z jakimi badanie zostało przeprowadzone, przyświadcza autorom pracy, że jest możliwa analiza anomalii w finansowych szeregach czasowych o dużej częstotliwości zmian w czasie rzeczywistym. Kwestią otwartą pozostaje optymalizacja algorytmu dla danych w postaci strumienia oraz zrównoleglenie algorytmu, co umożliwi zastosowanie algorytmu w chmurze obliczeniowej i analizowanie wielu rynków jednocześnie.

Jak pokazały przeprowadzone badania, wykorzystanie dyskretnej transformaty falkowej przyniosło bardzo dobre rezultaty w detekcji anomalii w finansowym szeregu czasowym dla pary EUR/USD. Materiał empiryczny wykorzystany do badań pozwala przypuszczać, że taka skuteczność nie jest dziełem przypadku, a powtarzalnym rezultatem. Należy mieć na uwadze jednak, że rezolucja czasowa badanych szeregów wynosiła 1 minutę. Tak więc wielkość okna wynosiła od 32 do 512 minut (około 8 godzin). W przypadku rynku FOREX duże wielkości okna nie są użyteczne, stąd aby potwierdzić skuteczność algorytmu, muszą być przeprowadzone dalsze badania przy mniejszych wielkościach oraz mniejszych rezolucjach czasowych. Dodatkowo należy pamiętać, że wykryta anomalia nie wskazuje na kierunek zmiany, a jedynie na fakt jej wystąpienia.

W najbliższym czasie będą przeprowadzone dalsze badania na innych parach walutowych o mniejszej agregacji czasowej. Podjęta zostanie także próba analizy anomalii w czasie rzeczywistym, po modyfikacjach algorytmu i przystosowaniu go do przetwarzania strumieni danych.

Literatura

Antoniadis A., Bigot J., Lambert-Lacroix S. [2010], *Peaks Detection and Alignment for Mass Spectrometry Data*, "Journal de la Société Française de Statistique" 151(1), s. 17–37.

Chandola V., Banerjee A., Kumar V. [2009], *Anomaly Detection: A Survey*, University of Minnesota, Res. Pap.

Crochiere R.E., Webber S.A., Flanagan J.L. [1976], *Digital Coding of Speech in Subbands*, "Bell System Technical Journal", no. 55, s. 1069–1085.

Daubechies I. [1992], *Ten Lectures on Wavelets*, Vermont: Capital City Press, Montpelier.

Ji Z. [2008], *Towards Outlier Detection for High-dimensional Data Streams using Projected Outlier Analysis Strategy*, Ph.D. thesis, Dalhousie University, Halifax.

Lazarevic A. [2003], *A Comparative Study of Anomaly Detection Schemes in Network Intrusion Detection*, SDM.

Strang G., Nguyen T. [1997], *Wavelets and Filter Banks*, Wesley-Cambridge Press.

UNSUPERVISED DETECTION OF CONTEXTUAL ANOMALIES IN HIGHLY FREQUENT TIME SERIES USING DAUBECHIES WAVELETS

Summary: This paper presents an algorithm to detect contextual anomalies of financial time series based on discrete wavelet transform. For this purpose, the orthogonal Daubechies wavelets (D1-D30) have been used. The study was conducted on the quotes of the currency pair EUR/USD extracted from the FOREX market covering a period of 10 years.

Keywords: discrete wavelet transform, anomalies, financial time series, FOREX market.