PRZEGLĄD STATYSTYCZNY
NUMER SPECJALNY 1 – 2012

DANIEL KOSIOROWSKI[1]

# GŁĘBIA POŁOŻENIA-ROZRZUTU W STRUMIENIOWEJ ANALIZIE DANYCH EKONOMICZNYCH

## 1. WPROWADZENIE

Współczesna gospodarka w sposób ciągły generuje gigantyczne zbiory danych. Analiza, monitorowanie, decydowanie w oparciu o wielkie zbiory danych stanową bez wątpienia wieloaspektowe wyzwanie dla współczesnej statystyki (por. Aggerwal, 2007).

Przykładem tego typu wyzwań jest tzw. analiza strumienia danych (strumieniowe przetwarzanie danych). Analiza taka przykładowo może obejmować monitorowanie setek tysięcy finansowych szeregów czasowych w celu znalezienia użytecznych inwestycyjnie zależności pomiędzy nimi, analizę danych generowanych przez stacje pogodowe w pewnym obszarze oceanu, monitorowanie centrum miasta za pomocą systemu kamer, decydowanie co do podjęcia interwencji na rynku zbóż w oparciu o dane dostarczane przez giełdy towarowe.

Ujmując zagadnienie nieprecyzyjnie możemy określić strumień danych jako „*nieokreślonej długości ciąg z reguły wielowymiarowych obserwacji*" (por. Szewczyk 2010). Należy zwrócić uwagę, że w przypadku tradycyjnie rozumianej analizy procesu stochastycznego, powiedzmy $\{X_t\}$, zakładamy ustalony przedział czasowy, powiedzmy $[0, T]$. Nasze obliczenia dotyczą tego przedziału a więc wnioskujemy na podstawie informacji uzyskanej do chwili $T$. W przypadku analizy strumienia danych nie ustalamy przedziału badania $[0, T]$. Każda kolejna chwila oznacza nową analizę procesu stochastycznego. Strumieniowe przetwarzanie danych, analizę strumienia danych można określić, jako sekwencję analiz procesu stochastycznego. Terminologia wywodzi się z informatyki, gdzie tego typu zagadnienia były rozważane po raz pierwszy. Oczywiście strumieniowe przetwarzanie danych można rozpatrywać na gruncie teorii procesów stochastycznych i w szczególności na gruncie teorii szeregów czasowych. W nawiązaniu do uwag jednego z Recenzentów dotyczących związków pomiędzy analizą procesów stochastycznych a analizą strumieni danych – autor uważa, że w obrębie znanych mu prac z zakresu zastosowań procesów stochastycznych w ekonomii najbliższe praktyki strumieniowego przetwarzania danych są prace duetu Aït-Sahalia i Jacod (por. Aït-Sahalia, Jacod, 2012 i odniesienia do literatury tamże) dotyczące badań procesów dyfuzji ze skokami. Prace te jednakże dotyczą jednej analizy procesu obserwowanego na pewnym przedziale

[1] Artykuł powstał w części dzięki wsparciu Narodowego Centrum Nauki w postaci grantu DEC-011/03/B/HS4/01138

czasu [0,T] – nie dotyczą rodziny takich analiz. Wspomniani autorzy skupiają swą uwagę na jednowymiarowych modelach parametrycznych o jednym reżimie. Przyjmują raczej mocne założenia odnośnie tychże modeli oraz odnośnie sposobu pobierania danych. Nie rozważają obserwacji odstających. Elegancja prezentacji zagadnień przez wspomniany duet uczonych stanowi punkt odniesienia i cel dla autora niniejszej pracy w przyszłości.

W literaturze dotyczącej analizy strumieni danych, strumieniowego przetwarzania danych w zasadzie nie podaje się wprost odwołania do probabilistycznego modelu danych. Jednakże wczytując się w tę literaturę można pokusić się o stwierdzenie, że analiza taka jest w istocie rodziną analiz procesu stochastycznego odznaczających się następującymi cechami:

1. Obserwacje generowane są przez proces, w którym ma miejsce nieliniowa zależność teraźniejszości od przeszłości.
2. Obserwacje modeluje się na ogół przez proces niestacjonarny, którego nie da się sprowadzić do procesu stacjonarnego za pomocą różnicowania, usunięcia deterministycznego trendu. Proces na ogół odznacza się występowaniem pewnej ilości reżimów. Typ niestacjonarności, liczba i charakterystyki reżimów mogą zmieniać się w czasie.
3. Analizę strumienia prowadzimy opierając się na stale aktualizowanej próbie – na podstawie ustalonej długości ruchomego okna (można rozważać okna różnej długości dla różnych skal czasu – sekund, minut, dni itd.). Na podstawie takiej stale aktualizowanej próby podejmujemy decyzje, na jej podstawie monitorujemy położenie, rozrzut strumienia.
4. Strumienie na ogół liczą setki tysięcy wielowymiarowych obserwacji. Z reguły dane z racji swej wielkości nie są magazynowane w pamięci komputera – muszą być przetwarzane na bieżąco (ang. *on–line processiong)*.
5. Dane napływają do obserwatora z reguły w nierównych odstępach czasu, w pakietach nierównej wielkości. Można założyć, że modelem strumienia jest proces stochastyczny z czasem ciągłym. Wówczas mamy na uwadze sytuację, gdy częstość próbkowania obserwacji ze strumienia jest zmienną losową. Można założyć stosownie skonstruowany proces dyskretny odwołując się np. do teorii procesów podporządkowanych, warunkowych procesów trwania, bądź tak jak w niniejszej pracy wyjść od takiego procesu, który losowo generuje sygnał (odpowiednio zdefiniowany) w chwilach równo od siebie odległych.
6. Do analizy strumieni stosuje się są na ogół procedury nieparametryczne, które muszą spełniać wysokie wymagania w zakresie złożoności obliczeniowej, które muszą radzić sobie z problemem „rzadkości danych" w wielu wymiarach (por. Hastie i in. 2009).

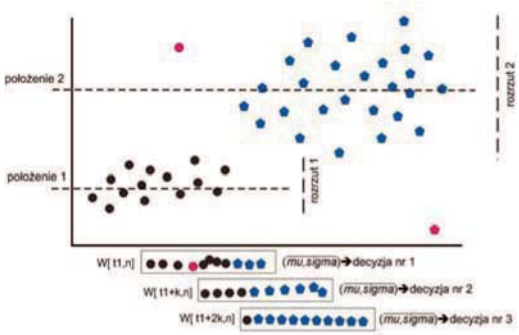
W niniejszej pracy zakładamy, że strumień generowany jest przez pewną konkretną postać ogólnego modelu określanego mianem CHARME (por. Stockis i in. 2010). Rozważamy tym samym proces stochastyczny z czasem dyskretnym o ustalonej liczbie reżimów. Zakładamy, że w obserwowanych przez nas danych występują obserwacje od-

stające. Mamy tutaj na uwadze sytuację, gdy na badany proces działa tzw. *addytywny proces odstawania* (AO) (ang. additive outliers process) – przyjmujemy ramy pojęciowe zaproponowane w klasycznym podręczniku Marona i in. (2006). Niech $x_t$ oznacza proces warunkowo stacjonarny[2], niech $v_t$ oznacza stacjonarny proces odstawania. Niech $P(v_t = 0) = 1 - \varepsilon$, co oznacza, że „niezerowa" część procesu $v_t$ pojawia się z prawdopodobieństwem $\varepsilon$. W modelu AO, zamiast $x_t$ obserwujemy $y_t = x_t + v_t$ przy czym zakłada się, że procesy $x_t$ i $v_t$ są wzajemnie niezależne. AO można określić, jako *proces błędów grubych*, obserwacje odstające na ogół są izolowane. W niniejszym artykule skupiamy naszą uwagę na procesie podejmowania decyzji na podstawie stale uaktualnianej niewielkiej próby ze strumienia. Decyzje dotyczą m.in. prognozowania kolejnych wartości strumienia, prognozowania i monitorowania charakterystyk rozrzutu, położenia i skośności, (bezwarunkowych i warunkowych względem obserwowanej próby w przeszłości), monitorowania zależności pomiędzy teraźniejszością i przeszłością strumienia. Naszym zadaniem jest stworzenie stosownych narzędzi umożliwiających nam odczytanie sygnału zawartego w strumieniu w sytuacji występowania obserwacji odstających. Należy jednakże podkreślić, że w przeciwieństwie do nauk inżynieryjnych (dane = deterministyczny sygnał+ losowy szum) przez sygnał rozumiemy relację pomiędzy charakterystykami liczbowymi probabilistycznego modelu danych[3]. W zasadzie w przyjętych przez nas dalej ramach pojęciowych odczytanie sygnału wiążemy ze wskazaniem reżimu procesu generującego strumień. Zagadnienie schematycznie przedstawiają rys. 1-2. W niniejszej pracy nie nawiązujemy bezpośrednio do teoretycznych trudności związanych z pomiarem odporności statystyki w przypadku analizy procesów stochastycznych. *Odporność naszych propozycji rozumiemy w duchu jednolitego i ogólnego podejścia Gentona i Lucasa (2003) jako odporność reguły decyzyjnej określonej na stale uaktualnianej próbie ze strumienia (za punkt odniesienie bierzemy np. medianę w przestrzeni decyzji, rozważamy różne funkcje straty np. LINEX).* Według Gentona i Lucasa (2003) krytyczna cecha estymatora sprowadza się do tego, że ten przyjmuje różne wartości dla różnych realizacji próby. Jeżeli możliwe jest kontinuum prób a estymator jest ciągły, to oczekujemy kontinuum jego wartości. Załamanie estymatora polega na tym, że ta jego własność zanika, estymator przyjmuje jedynie skończoną liczbę różnych wartości pomimo kontinuum możliwych prób. Jeżeli dla przykładu rozważany przez nas model strumienia dopuszcza powiedzmy 10 reżimów, a procedura mająca wskazać te reżimy wskazuje jedynie jeden z nich, to powiemy, że procedura łamie się.

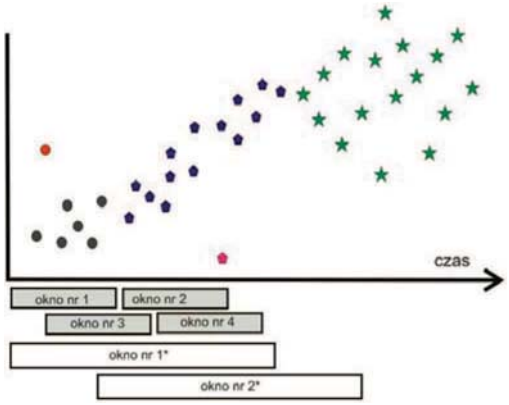
Można umownie wyróżnić dwa nurty podejść do analizy strumieni danych – nurt związany z metodami eksploracyjnej analizy bardzo wielkich zbiorów danych (ang. ve-

[2] Mówimy, że jednowymiarowy proces jest warunkowo stacjonarny, jeżeli jego rozkładu warunkowe są niezmiennicze względem przesunięć w czasie (por. Shalizi, Kantorovich, 2007 def. 51 str. 35)

[3] W niniejszym artykule zakładamy, że dane generuje pewien niestacjonarny proces stochastyczny. Sygnał utożsamiamy z charakterystykami liczbowymi jego modelu(li). Jednakże o ile zmienimy rozumienie sygnału – można rozważać strumienie generowane przez procesy stacjonarne bądź układy stricte deterministyczne. W kontekście zastosowań tematyki w ekonomii – przyjęte ramy wydają się być najwłaściwsze.




Rysunek 1. Ilustracja zagadnienia decydowania co do zmiany położenia – rozrzutu na podstawie ruchomego okna




Rysunek 2. Trzy reżimy strumienia danych. Dane zawierają obserwacje odstające, rozważamy ruchome okna różnej długości

ry big high-dimensional data mining) oraz nurt związany z klasyczną nieparametryczną analizą szeregów czasowych (por. Fan, Yao, 2005). W obrębie pierwszego nurtu (por. Aggerwal, 2007) wyróżnić można m.in.: dynamiczną redukcję wymiaru zagadnienia za pomocą tzw. mikro-skupisk, badanie dynamicznych klasyfikacji, stosowanie adaptacyjnej metody najbliższych sąsiadów, wykorzystywanie drzew regresyjnych i klasyfikacyjnych, wykorzystanie sieci neuronowych, sieci bayesowskich. Drugi nurt wiąże się z adaptacjami metod nieparametrycznej analizy szeregów czasowych. Mamy tutaj na uwadze adaptacje lokalnej liniowej, lokalnej wielomianowej regresji w tym szereg wariantów nieparametrycznej *regresji Nadaraya-Watsona* (patrz Hall i in., 1999), metody wykorzystujące wielomiany ortogonalne, regresję nieliniową z ograniczeniami (np. metody LOESS, LASSO por. Hastie i in., 2009), sklejki itd. Należy podkreślić, że w przypadku analizy strumieni danych na ogół wielowymiarowych niezmiernie istotne jest, aby procedura radziła sobie z tzw. *„przekleństwem wielowymiarowości" – rzadkość danych* (ang. sparse data) *w wielu wymiarach*. Owo przekleństwo sprawia m.in., że dla przykładu dobre statystyczne własności jednowymiarowej regresji Nadaraya-

Watsona zanikają w wielu wymiarach, istotność statystyczna oszacowań wielowymiarowych modeli stosowanych w empirycznych finansach budzi poważne wątpliwości (por. Kosiorowski, Snarska, 2012).

W literaturze jak dotychczas nie jest znanych wiele odpornych metod analizy strumieni danych. Wiąże się to między innymi z trudnościami z rozumieniem odstawania w przypadku strumieni generowanych przez model o wielu reżimach. Pojawia się dla przykładu pytanie *czy rozumienie odstawania powinno się w takim przypadku wiązać z konkretnym reżimem procesu*? Co ciekawe w przypadku analizy strumienia z jednostkami odstającymi stosowana procedura powinna być odporna, jednak nie bardzo odporna (tzn. jej punkt załamania nie powinien osiągać maksymalnej możliwej wartości) – tak, aby pomijała wpływ obserwacji odstających, lecz jednocześnie była wrażliwa na zmianę reżimu modelu.

W pracy proponujemy proste i przyjazne dla użytkownika metody analizy strumienia danych ekonomicznych odwołujące się do tzw. koncepcji głębi danych (por. Kosiorowski, 2012). W kolejnych częściach artykułu prezentujemy odpowiednio: w drugiej części wprowadzamy model strumienia danych, w trzeciej przedstawiamy wybrane zagadnienia związane z głębią położenia-rozrzutu Mizery i Müller, w czwartej przedstawiamy propozycje procedur wykorzystujących tę głębię, w piątej wyniki badań własności propozycji za pomocą symulacji, artykuł kończymy konkluzjami i literaturą.

## 2. MODEL STRUMIENIA DANYCH EKONOMICZNYCH

W literaturze nie jest znanych wiele modeli strumienia danych, do nielicznych należy zaliczyć propozycję Hahsler i Dunhamr (2010), w której rozważa się zmienny w czasie łańcuch Markowa dla mikro skupisk. Wydaje się jednak, że model strumienia danych można skonstruować na podstawie jednego z wykorzystywanych w ekonometrii modeli dla zjawisk o zmiennym reżimie np. model TAR (ang. treshold auroregressive model) bądź jego nieliniową wersję FAR (ang. functional auroregressive model) (por. Fan, Yao, 2005).

Wybór modelu strumienia danych wykorzystywanego w niniejszej pracy podyktowany jest względami wygody oraz jego elastycznością w zakresie opisu szerokiego spektrum możliwych zjawisk. Zdecydowano się budować model strumienia danych na bazie *warunkowego heteroskedastycznego nieparametrycznego modelu* CHARN postaci

$$X_t = m(X_{t-1}, ., X_{t-p}) + \sigma(X_{t-1}, ., X_{t-p})\epsilon_t, \tag{1}$$

o dowolnych lecz ustalonych funkcjach $m(\cdot)$ oraz $\sigma(\cdot)$ (np. $m(\mathbf{x}) = E(X_t|[X_{t-1}, ..., X_{t-p}] = \mathbf{x})$, $\sigma^2(\mathbf{x}) = Var(X_t|[X_{t-1}, ..., X_{t-p}] = \mathbf{x})$, gdzie $\mathbf{x} = (x_{t-1}, ..., x_{t-p})$ oraz o niezależnych o tym samym rozkładzie innowacjach $\varepsilon_t$ o wartości oczekiwanej zero (zobacz Fan, Yao, 2005).

Model (1) stanowi punkt wyjścia dla procesu budowy modelu strumienia danych ekonomicznych. Podkreślmy jednakże, że w kontekście analizy strumieni danych ekonomicznych z zasady nie zakładamy, że obserwowany proces ma tę samą funkcję trendu $m$ oraz tę samą zmienność $\sigma$ w każdej chwili. Nie zakładamy też, że funkcje

te zmieniają się powoli, stopniowo w czasie. W takim oto kontekście skupiamy naszą uwagę na modelu CHARME (ang. *conditional heteroscedastic autoregressive mixture of experts*) (zobacz Stockis i in., 2010). Model CHARME stanowi ogólne podejście do modelowania szeregów czasowych o zmiennym reżimie. Układ ekonomiczny oscyluje pomiędzy pojedynczymi stanami, których same dynamiką rządzi model CHARN (1). CHARME w przypadkach szczególnych obejmuje wiele nieliniowych szeregów czasowych jak np. modele dwuliniowe, modele progowe TAR. W modelu CHARME dynamiką strumienia $\{X_t\}$ rządzi ukryty łańcuch Markowa $\{Q_t\}$ na skończonym zbiorze stanów $\{1, 2, ., K\}$, sam model zdefiniowany jest w następujący sposób:

$$X_t = \sum_{k=1}^{K} S_{tk}(m_k(X_{t-1}, ., X_{t-p}) + \sigma_k(X_{t-1}, ., X_{t-p})\epsilon_t) + b_t\theta_t, \tag{2}$$

gdzie $S_{tk} = 1$ dla $Q_t = k$ oraz $S_{tk} = 0$ w przeciwnym przypadku, $m_k$, $\sigma_k$, $k = 1, ..., K$, są nieznanymi funkcjami, $\epsilon_t$ są niezależnymi zmiennym losowymi o średniej zero, człon $b_t\theta_t$ reprezentuje obserwacje odstające typu AO (por. Maronna i in., 2006), $b_t$ jest nieobserwowalną binarną zmienną losową wskazującą pojawienie się obserwacji odstającej w chwili $t$, natomiast $\theta_t$ oznacza losową wielkość obserwacji odstającej. Aby uniknąć „*przekleństwa wielowymiarowości*", postulujemy przyjąć $p = 1$ bądź $p = 2$.

Wprowadzając ukryty łańcuch Markowa rządzący zmianami reżimów dopuszczamy występowanie nagłych wartości strumienia. Dodatkowo przyjmujemy następujące założenia odnośnie strumienia i wykorzystywanego do jego opisu modelu CHARME:

1. Losowa liczba obserwacji odstających w strumieniu, pojawiających się do chwili $t$ dana jest za pomocą $N_t = \sum_{i=1}^{t} b_i$ oraz jest ograniczona według prawdopodobieństwa warunkowo względem $N_{t^*}$, $t^* < t$. Oznacza to, że zamiast ustalać z góry liczbę obserwacji odstających stosujemy ograniczenie na prawdopodobieństwo ich pojawienia się. Umożliwiamy tym samym rozróżnienie pomiędzy częstymi, zwykłymi szokami oraz rzadkimi zdarzeniami odstającymi.
2. Nie zakładamy jakiejkolwiek wiedzy, co do liczby i położenia obserwacji odstających, nie nakładamy też ograniczeń na strukturę zależności $\{b_t\}$.
3. Strumień, który jest modelowany za pomocą modelu CHARME jest *warunkowo stacjonarny* (por. Shalizi, Kantorovich, 2007)
4. Zakładamy, że ukryty łańcuch Markowa będzie zmieniał swą wartość rzadko, tzn. obserwowany proces będzie podlegał temu samemu reżimowi przez względnie długi okres czasu zanim nastąpi zmiana. Stawiamy tym samym ograniczenia, co do postaci macierzy przejścia $\mathbf{P} = [p_{rs}]$, $r, s = 1, ..., k$, dla łańcucha $Q_t$ postaci $p_{rr} >> p_{rs}$.

Niech $x_1, x_2, ...$ oznacza strumień danych generowany przez model (2). Przez okno $W_{i,n}$ rozumiemy ciąg punktów kończący się w punkcie $x_i$ i o długości $n$: $W_{i,n} = (x_{i-n+1}, ..., x_i)$. Czasem wygodnie jest rozważać okno $W_{[i,j]}$ – podciąg strumienia danych pomiędzy $i$-tą oraz $j$-tą obserwacją. Spora część technik analizy strumieni danych opiera się na monitorowaniu różnych odległości pomiędzy rozkładami empirycznymi

wyznaczanymi na podstawie dwóch bądź więcej okien $W_{i,n}$ i $W_{j,n}$. Można przy tym rozważać ustalone okna, ruchome okna itd. Rozważając wielowymiarowy strumień danych $\mathbf{x}_1, \mathbf{x}_2, ...$ podobnie badamy zachowanie się wielowymiarowych okien $\mathbf{W}_{i,n}$ (por. Kosiorowski, Snarska, 2012).

**PROBLEM 1**: Rozważmy sytuację, gdy w oparciu o stale uaktualniane okna $W_{i,n}, W_{i+1,n}, ...$ przewidujemy odpowiednio kolejne obserwacje $\hat{x}_{i+1}, \hat{x}_{i+2}, ...$ bądź kolejne okna $\hat{W}_{i+1,k}$,$\hat{W}_{i+2,k}$,. . . , $k \ll n$. Chcielibyśmy znaleźć optymalną procedurę prognostyczną, tzn. minimalizującą pewną funkcję straty w sytuacji, gdy strumień danych zawiera jednostki odstające.

**PROBLEM 2:** Na podstawie monitorowania ruchomego okna $W_{i,n}$, $i = 1, 2, ...$ zamierzamy wykryć bezwarunkowe zmiany w modelu generującym strumień danych. Jeżeli założymy pewien model postaci (2), to naszym celem jest wykrycie stanu $Q_k$ ukrytego łańcucha Markowa, a w konsekwencji funkcji $m_k$ oraz $\sigma_k$ występujących w (2).

**PROBLEM 3**: Monitorujemy strumień danych $x_1, x_2, ...$ oraz naszym celem jest wykrycie zmiany rozkładu warunkowego okna $W_{i+1,n}$, pod warunkiem zaobserwowanego okna $W_{i,n}$, $i = 1, 2, ...$, tzn. zmiany $P(W_{i+1,n} \in A | W_{i,n} = \mathbf{x})$, $A \subset \mathrm{R}$, dla $i = 1, 2, ....$ W ramach pojęciowych, wyznaczonych przez model (2) naszym celem jest wykrycie zmian macierzy przejścia ukrytego łańcucha $Q_k$, bądź zmiany funkcji $m_k$ oraz $\sigma_k$ oznaczających przykładowo warunkowe położenie i warunkowy rozrzut.

**PROBLEM 4**: Monitorujemy $d$-wielowymiarowy strumień danych $\mathbf{x}_1 = (x_{11}, ..., x_{1d})$, $\mathbf{x}_2 = (x_{21}, ..., x_{2d})$,. . . , a naszym celem jest wykrycie zmian łącznego (warunkowego) rozkładu $\mathbf{x}_i$ na podstawie $\mathbf{W}_{i-1,n}$, $i = 1, 2, ....$ W szczególności jesteśmy zainteresowani wykrywaniem zmian postaci związku liniowego pomiędzy współrzędnymi wektorów $\mathbf{x}_i$.

Stosowanie w zagadnieniu predykcji (problem 1) lokalnego liniowego, lokalnego wielomianowego modelowania wymaga, aby zmiany pomiędzy reżimami były gładkie (a w konsekwencji dawały się lokalnie aproksymować za pomocą funkcji liniowej bądź wielomianowej) oraz aby strumień nie zawierał obserwacji odstających. Stosowanie globalnych sklejek z jednej strony wymaga „zatrzymania" analizy, aby można było taki model oszacować, z drugiej strony napotykamy problemy związane ze złożonością obliczeniową oraz przekleństwem wielowymiarowości. Stosowanie wielomianów ortogonalnych z kolei zmusza nas od ograniczenia się do kowariancji jako miary zależności obserwacji w czasie – co nie jest właściwe w przypadku strumieni o na ogół nieliniowej strukturze zależności w czasie.

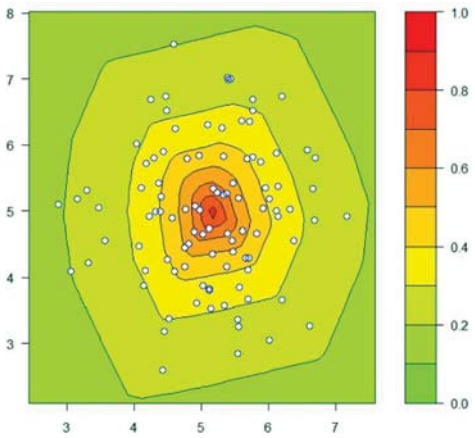
Jest powszechnie wiadomo, że oszacowania momentów procesów stochastycznych są użyteczne jedynie o ile poczynimy bardzo mocne założenia odnośnie tychże procesów. Mamy tutaj na uwadze m.in. istnienie momentów odpowiednich rzędów, jednoznaczność opisu rozkładu za pomocą momentów, postać funkcji autokowariancji itd. (por. Jacod, Shiryaev, 2003). Podobnie rzecz się przedstawia z prognozowalnością procesów. Strumienie na ogół nie spełniają takich założeń. Jednakże zamiast opisywać proces za pomocą miar konstruowanych na podstawie momentów możemy go opisywać

za pomocą miar wykorzystujących statystyki pozycyjne i porządkowe. Opis procesu w kategoriach indukowanych przez statystyki pozycyjne i porządkowe jest możliwy nawet w sytuacji, gdy nie da się opisać procesu za pomocą statystyk wykorzystujących momenty. W takim oto kontekście proponujemy wykorzystać narzędzia koncepcji głębi danych do odpornej analizy strumienia danych. W celu oszacowania niepewności związanej z analizą strumienia danych proponujemy wykorzystać metody Monte Carlo.

## 3. GŁĘBIA STUDENTA

Dla danego rozkładu prawdopodobieństwa $F$ na $\mathrm{R}^d$, $d \geq 2$ statystyczna funkcja głębi $D(\mathbf{x}, F)$ przyporządkowuje $\mathbf{x} \in \mathrm{R}^d$ liczbę z przedziału [0,1] będącą miarą centralności tej obserwacji względem rozkładu $F$. Statystyczne funkcje głębi kompensują brak naturalnego porządku w $\mathrm{R}^d$, $d \geq 2$, poprzez orientowanie punktów względem centrum – względem $d$-wymiarowej mediany indukowanej przez konkretną funkcje głębi. Wyższe wartości głębi reprezentują wyższy stopień centralności. Wprowadzenie do koncepcji głębi danych znajdziemy w pracach Serflinga (2006) oraz Kosiorowskiego (2012).

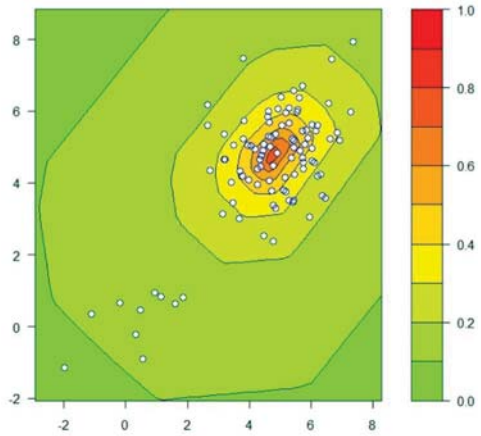
Na rysunku 3 przedstawiono empiryczną funkcję głębi projekcyjnej dla próby złożonej ze stu obserwacji wygenerowanych z dwuwymiarowego rozkładu normalnego. Rysunek 4 przedstawia empiryczną funkcję głębi projekcyjnej dla mieszaniny dwóch dwuwymiarowych rozkładów normalnych. Dwuwymiarowe mediany projekcyjne znajdują się wewnątrz najbardziej centralnych obszarów. W niniejszym artykule skupiamy naszą uwagę na szczególnym przypadku statystycznej funkcji głębi – na *głębi Studenta*. Głębi określonej dla pary: *miara położenia i miara rozrzutu,* odnoszącej się do jednowymiarowego zbioru danych.




Rysunek 3. Wykres konturowy głębi projekcyjnej z próby – 100 obserwacji z rozkładu normalnego 2d

Źródło: Obliczenia własne - pakiet środowiska R {depthproc 0.1}

Wychodząc od jednowymiarowego modelu położenia i rozrzutu Mizera i Müller 2004 wprowadzili pojęcie jednowymiarowej głębi położenia-rozrzutu oraz pokazali wybrane jej zastosowania. Ważny przypadek szczególny ich koncepcji to głębia Studenta oraz estymator maksymalnej głębi Studenta – mediana Studenta.




Rysunek 4. Wykres konturowy głębi projekcyjnej z próby – 100 obserwacji z mieszaniny rozkładów normalnych 2d

Źródło: Obliczenia własne – pakiet środowiska R {depthproc 0.1}

Mizera (2002) rozpoczyna swe rozważania od obserwacji $z_i$, $i = 1, ..., n$, następnie wprowadza funkcję kryterium $F_i = F_i(z_i)$ – dla danego dopasowania reprezentowanego przez $\theta$, funkcja kryterium $F_i$ wyraża brak dopasowania $\theta$ do konkretnego punktu $z_i$. Oznacza to, że $\theta^*$ odzwierciedla (pasuje do) $z_i$ lepiej niż $\theta$, jeżeli $F_i(\theta^*) < F_i(\theta)$.

Według propozycji Mizery ogólna głębia Tukey'a może zostać zdefiniowana jako miara dopuszczalności dopasowania zważywszy na zaobserwowane dane. *Możemy zdefiniować głębię dopasowania $\theta$ jako frakcję danych, której pominięcie sprawia, że $\theta$ staje się brakiem dopasowania, dopasowaniem, które może zostać zdominowane jednostajnie przez każde inne*. W oparciu o tę zasadę Mizera (2002) definiuje globalną głębię oraz bardziej operacyjną wersję tej głębi – głębię styczną – wynik przejścia od ogólnego kryterium optymalności do jego wersji różniczkowej. Biorąc pochodne w zagadnieniu optymalizacji z wykorzystaniem funkcji kryterium $F_i$ Mizera definiuje głębię styczną dopasowania $\theta$ jako:

$$d(\theta) = \inf_{\mathbf{u} \neq 0} \#\{i : \mathbf{u}^t \nabla_\theta F_i(\theta) \geq 0\}, \tag{3}$$

gdzie # oznacza względną proporcję zbioru indeksów – ich liczbę podzieloną przez $n$, $\nabla_\theta F_i(\theta)$ oznacza gradient funkcji kryterium $F$ w punkcie $\theta$ dla obserwacji $i$.

Teoretyczna innowacja Mizery i Müller (2004) polega na wykorzystaniu funkcji wiarygodności w charakterze funkcji kryterium. Niech $y_i$ oznaczają zmienne losowe o gęstości $f$.

DEFINICJA 1: Głębia położenia–rozrzutu Mizery i Müller $(\mu, \sigma) \in \mathrm{R} \times [0, \infty)$ względem próby $\mathrm{Y}^n = \{y_1, ..., y_n\}$ definiowana jest jako

$$D((\mu,\sigma), \mathrm{Y}^n) = \begin{cases} \inf\limits_{\mathbf{u}\neq 0} \# \left\{ i : (u_1, u_2) \begin{pmatrix} \psi(\tau_i) \\ \chi(\tau_i) - 1 \end{pmatrix} \geq 0 \right\} & dla, \quad \sigma > 0 \\ \#\{i : y_i = \mu\} & dla, \quad \sigma = 0 \end{cases} \tag{4}$$

gdzie znak mnożenia interpretujemy jako iloczyn skalarny, $\tau_i$ jest skrótem dla oraz funkcje $\psi, \chi$ zależą od ustalonej gęstości $f$, $\psi(\tau) = (-\log f(\tau))' = -f'(\tau)/f(\tau)$, oraz $\chi(\tau) = \tau\psi(\tau)$.

Definicja 1, wywodząc się z metody największej wiarogodności, wprowadza rodzinę głębi zależnych od przyjętej gęstości.

DEFINICJA 2: Głębia Studenta $(\mu, \sigma) \in \mathrm{R} \times [0, \infty)$ względem rozkładu prawdopodobieństwa $P$ na R dany jest jako

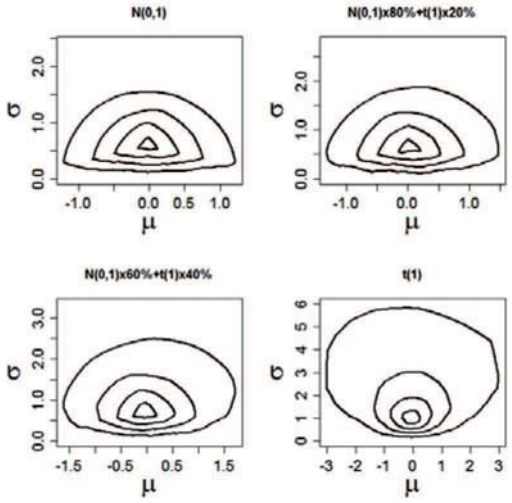
$$D(\mu, \sigma, P) = \inf_{(u_1,u_2)\neq 0} P\{y : u_1(y - \mu) + u_2((y - \mu)^2 - \sigma^2) \geq 0\} \tag{5}$$

Głębię Studenta z próby $\mathrm{Y}^n = \{y_1, ..., y_n\}$ otrzymujemy poprzez podstawienie w definicji 2 rozkładu empirycznego $P_n$ wyznaczonego na podstawie tej próby.

Głębia położenia-rozrzutu jest ekwiwariantna względem położenia i rozrzutu, mediana Studenta ma tę samą własność. Mediana Studenta jest bardzo dobrym estymatorem centrum symetrii dla małych zbiorów danych, rzędu 30-100 obserwacji. Rysunki 5-6 przedstawiają wykresy konturowe głębi Studenta dla mieszanin standardowego rozkładu normalnego, rozkładu Studenta o jednym stopniu swobody i dla skośnego rozkładu Studenta o jednym stopniu swobody i parametrze skośności -2 – t(1,-2) (por. pakiet {skewt}programu R i literatura tamże). Rysunki te sugerują możliwość dyskryminacji pomiędzy tymi rozkładami na podstawie wykresu konturowego sporządzonego na podstawie próby. Zaznaczmy, że wykresy konturowe głębi Studenta można potraktować jako uogólnienie jednowymiarowego wykresu kwantyl – kwantyl.

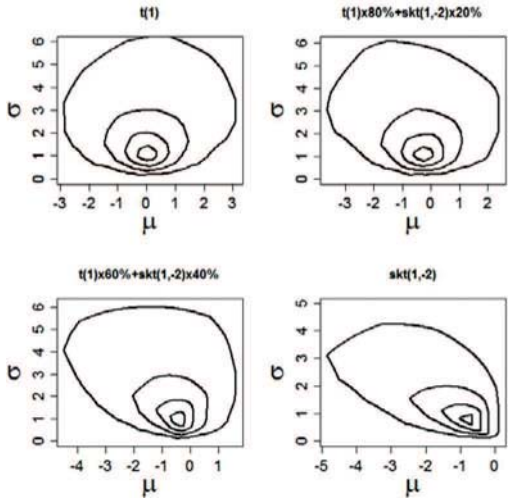
Mizera i Müller (2004), zakładając próbę losową prostą, pokazują jednostajną względem $(\mu, \sigma)$ zbieżność prawie na pewno estymatorów maksymalnej głębi Studenta oraz jej zadowalającą efektywność dla modelu normalnego. Pokazują też, że punkt załamania próby skończonej mediany Studenta jest w przybliżeniu równy 33% oraz mediana Studenta ma ograniczoną funkcję wpływu. Symulacje prowadzone przez autora niniejszego artykułu sugerują bardzo dobre własności mediany Studenta w sytuacji, gdy modele generujące dane odznaczały się heteroskedastycznością, nieliniową zależność pomiędzy obserwacjami oraz skośnością rozkładu. *Należy podkreślić, że w przeciwieństwie do estymatorów największej wiarygodności oraz ich uogólnień w postaci M-estymatorów Hubera – przy estymacji za pomocą maksymalnej głębi położenia – rozrzutu nie korzystamy wprost z założenia niezależności obserwacji w próbie. Korzystamy jedynie z „rankingu" dopasowania zważywszy na obserwowany zbiór danych. To*

*istotna cecha estymatora w kontekście jego zastosowania do analizy strumienia danych (gdzie mamy do czynienia z zależnością obserwacji w czasie).* Fakty te skłaniają autora do wykorzystania mediany Studenta w analizie strumienia danych ekonomicznych. W dalszej części wykorzystujemy algorytm {lsdepth}zaproponowany przez Ch. Müller dla jednoczesnego obliczania konturów głębi Studenta oraz mediany Studenta.




Rysunek 5. Wykresy konturowe głębi studenta dla mieszanin N(0,1) i t(1)

Źródło: Obliczenia własne, lsdepth.




Rysunek 6. Wykresy konturowe głębi studenta dla mieszanin N(0,1) i t(1,-2)

Źródło: Obliczenia własne, lsdepth.

## 4. PROPOZYCJE

Praktyka wymaga, aby odporna analiza strumieni danych (estymacja, predykcja i podejmowanie decyzji) prowadzona była w tempie odpowiadającym napływowi nowych danych, pojawianiu się istotnych merytorycznie zdarzeń. Taki postulat eliminuje wiele dobrych procedur rozważanych w statystyce odpornej. Napływ nowych informacji powinien poprawiać precyzję takiej analizy. Znaczna część prezentowanych w literaturze podejść do analizy strumieni danych da się zakwalifikować do jednej z dwóch kategorii: *analizy wsadowej* (ang. batch-incremental) oraz *analizy odtworzeniowej* (ang. regenerative approach). W przypadku analizy wsadowej analizujemy strumień partiami – wykorzystujemy uaktualniany model predykcyjny do momentu, gdy nie wykryjemy istotnej zmiany (trendu). W przypadku analizy odtworzeniowej tworzymy nowy model predykcyjny z każdego nowego okna (por. Aggerwal, 2007).

Bardzo popularną techniką analizy strumieni danych jest monitorowanie okna uczącego o ustalonej wielkości, zazwyczaj wyznaczonej wcześniej a priori przez użytkownika. Zwróćmy uwagę na dylemat przed jakim stoi użytkownik w takiej sytuacji: *wybrać krótkie okno tak aby to okno odpowiadało rozkładowi wyznaczonemu na podstawie obecnego stanu strumienia czy też wybrać większe okno, tak aby model był niejako bardziej reprezentatywny w okresach stabilności*. Warto zaznaczyć, że wykorzystując pewne klasyczne algorytmy zazwyczaj zakładamy, że okno uczące (przeszłość) i okno testowe (przyszłość) pochodzą z tego samego rozkładu. W kontekście analizy strumienia danych ekonomicznych rozkład z zasady zmienia się w czasie. Pojawia się pytanie, jeżeli użytkowany model wydaje się być niewłaściwy, to czy powinniśmy go modyfikować czy też odrzucić całkowicie. Zmiany modelu mogą pojawiać się stopniowo bądź nagle, dokładny punkt zmiany może być niewykrywalny.

W niniejszej pracy skupiamy naszą uwagę na statystycznych funkcjach głębi. Procedury indukowane przez głębie wykazują bardzo dobre własności w zakresie odporności, efektywności i łatwości interpretacyjnej. Stosowane procedury reprezentują tendencję wyrażoną przez większość obserwacji. Nasz sposób decydowania odznacza się konserwatyzmem jednakże zabezpieczamy się przed wpływem obserwacji odstających na nasze decyzje.

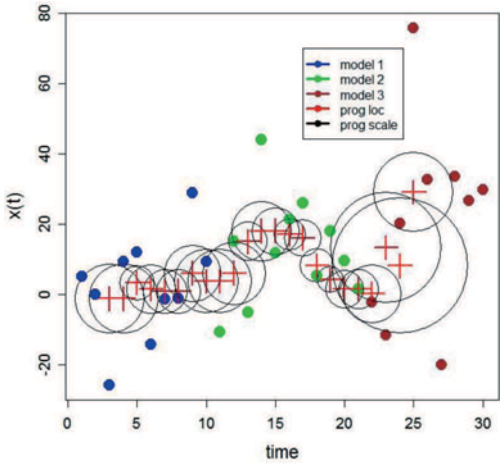
Oznaczmy przez $(\hat{\mu}_{i,n}; \hat{\sigma}_{i,n})$ medianę Studenta obliczoną na podstawie okna $W_{i,n}$. Niech $W_{Q(k),n}$ oznacza próbę wygenerowaną z $k-$tego reżimu $Q(k)$ modelu (2) o długości $n$.

PROPOZYCJA 1: W celu rozwiązania pierwszego problemu proponujemy przyjąć

$$\hat{x}_{t+1} = \hat{\mu}_{i,n}, i = 1, 2, ... \tag{6}$$

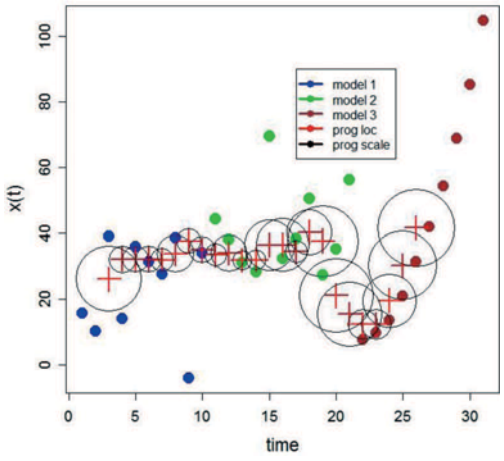
Jako przewidywanie następnej wartości strumienia danych bierzemy współrzędną położenia mediany Studena policzonej dla okna $W_{i,n}$. Jako „błąd standardowy" takiego przewidywania proponujemy wziąć $\hat{\sigma}_{i,n}$, tzn. współrzędną rozrzutu mediany Studenta policzonej dla okna $W_{i,n}$ (zobacz rys. 7-8).

PROPOZYCJA 2: W celu rozwiązania drugiego problem proponujemy ustalić próby referencyjne $W_{Q(1),n}, ..., W_{Q(k),n}$ wygenerowane z reżimów $Q(1), ..., Q(k)$ roz-




Rysunek 7. Ilustracja zastosowania pierwszej propozycji – heteroskedastyczność

Źródło: Obliczenia własne, lsdepth.




Rysunek 8. Ilustracja zastosowania pierwszej propozycji – lokalny liniowy trend

Źródło: Obliczenia własne, lsdepth.

patrywanego modelu (2), bądź wyznaczonych w związku z pewnym celem merytorycznym. Aby wykryć zmiany w bezwarunkowej wartości oczekiwanej procesu oraz w poziomie zmienności procesu, proponujemy monitorować zachowanie się następujących statystyk:

$$D_1 = \frac{\hat{\mu}_{i,n} - \hat{\mu}_{Q(j),n}}{\hat{\sigma}_{Q(j),n}}, \quad j = 1, ..., k, \quad i = 1, 2, ... \tag{7}$$

dla wykrycia zmian *poziomu charakterystyki położenia*

$$D_2 = \frac{min(\hat{\sigma}_{i,n}, \hat{\sigma}_{Q(j),n})}{max(\hat{\sigma}_{i,n}, \hat{\sigma}_{Q(j),n})}, \quad j = 1, ..., k, \quad i = 1, 2, ... \tag{8}$$

dla wykrycia *zmian w zmienności procesu*.

Dla wykrycia *zmian skośności procesu* proponujemy porównać wykres konturowy głębi Studenta z wybranymi wykresami referencyjnymi związanymi z interesującymi nas zagadnieniami.

PROPOZYCJA 3: W celu rozwiązania trzeciego problem proponujemy monitorować relacje pomiędzy współrzędnymi ruchomej mediany Studenta $(\hat{\mu}_{i,n}; \hat{\sigma}_{i,n})$ rozważanej dla okien różnej długości i porównywać te wykresy z wykresami referencyjnymi sporządzonymi dla danych wygenerowanych ze znanych modeli, bądź modeli ważnych ze względów merytorycznych

$$\hat{\sigma}_{i,n} vs. \hat{\sigma}_{i-l,n}, \quad l = 2, ..., k, \quad i = 1, 2, ... \tag{9}$$

$$\hat{\mu}_{i,n} vs. \hat{\sigma}_{i-l,n}, \quad l = 2, ..., k, \quad i = 1,2,... \tag{10}$$

Zauważmy, że stosując okna różnej długości możemy wykrywać zmiany następujące w różnych skalach czasowych (sekundy, godziny, dni), możemy jednocześnie wykorzystywać układ okien sporządzanych dla różnych skal czasowych. Istnieją, co najmniej dwa typy okien wykorzystywanych a analizie strumieni danych. W modelu *z dołączonymi oknami* (ang. adjacent windows) monitorujemy różnicę pomiędzy dwoma ruchomymi $W_{t,n}$ oraz $W_{t-n,n}$, gdzie $t$ oznacza aktualny czas. W modelu *z ustalonym oknem* mierzymy różnice pomiędzy ustalonym oknem $W_n$ i ruchomym oknem $W_{t,n}$. Pierwszy model lepiej wychwytuje „*intensywność zmian*" w danej chwili, podczas gdy drugi model jest lepszy do wykrycia stopniowych zmian, które mogą kumulować się w czasie. W praktyce zalecamy stosowanie co najmniej dwóch typów okien (dwóch częstości pobierania obserwacji).

## 5. WŁASNOŚCI PROPOZYCJI

W celu wykazania statystycznych własności propozycji w przypadku małej i umiarkowanej wielkości próby wykonano szereg symulacji w tym m. in. z następujących modeli CHARME o dwóch reżimach:

Posługując się symulatorem wchodzącym w skład pakietu {fGarch}środowiska R autorstwa Diethelma Wuertza i Rmetrics Core Team wykorzystano powszechnie stosowane w ekonometrii modele AR(1)-GARCH(1,1) o specyfikacji $X_t = \mu + \theta X_{t-1} + \varepsilon_t$ dla członu AR oraz $Z_t = \sigma_t \varepsilon_t$, $\sigma_t^2 = c_0 + \alpha Z_{t-1}^2 + \beta \sigma_{t-1}^2$ dla członu GARCH i przy standardowych założeniach odnoście wartości parametrów (por. Fan, Yao, 2005).

MODEL 1: PIERWSZY REŻIM: składający się z dwóch modeli AR(1)-GARCH(1,1) z lokalnym liniowym trendem stochastycznym, pierwszy z parametrami AR($\mu = 5, \theta = 0.5$), GARCH($c_0 = 10^{-6}, \alpha = 0.1, \beta = 0.75$) i rozkładem warunkowym t-Studenta

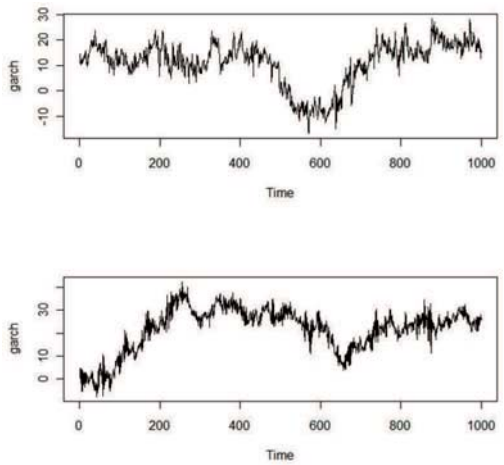
o czterech stopniach swobody; DRUGI REŻIM z parametrami AR($\mu = -5, \theta = -0.5$), GARCH($c_0 = 10^{-6}$,$\alpha = 0.6, \beta = 0.1$) i rozkładem warunkowym t-Studenta o czterech stopniach swobody (por. rys. 9-10 ).

MODEL2: PIERWSZY REŻIM: składający się z dwóch modeli AR(1)-GARCH(1,1) z lokalnym kwadratowym trendem stochastycznym, pierwszy z parametrami AR($\mu = 5, \theta = 0.5$), GARCH($c_0 = 10^{-6}, \alpha = 0.1, \beta = 0.75$) i rozkładem warunkowym t-Studenta o czterech stopniach swobody; DRUGI REŻIM z parameterami AR($\mu = -0.5, \theta = -0.8$), GARCH($c_0 = 10^{-6}$,$\alpha = 0.5, \beta = 0.1$) i rozkład warunkowy t-Studenta o czterech stopniach swobody.

MODEL 3: składający się z dwóch błądzeń przypadkowych z trendem $dB = mdt + sdX$, obserwowanych w równoodległych dyskretnych chwilach tzn. $B_t - B_{t-1} = m + \varepsilon_t$, $\varepsilon_t \sim N(0, s^2)$. Pierwszy z parametrami $m = 1, s = 1$, drugi z parametrami $m = -1, s = 2$.

Zmianą reżimów rządziła macierz przejścia $\boldsymbol{P}$ o kolumnach postaci $(0{,}99;0{,}01)^T$ i $(0{,}03;\ 0{,}97)^T$. Rozważano strumienie zawierające do 5% obserwacji odstających oraz strumienie bez jednostek odstających.

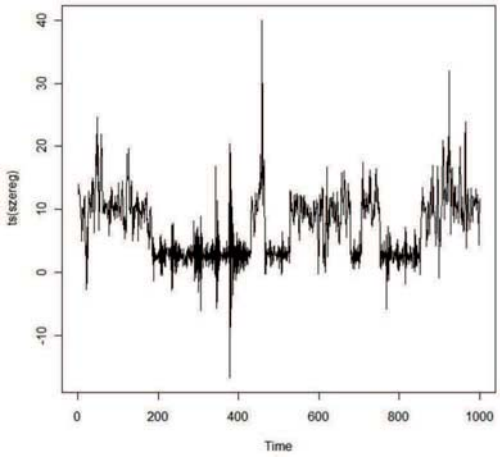
Dla każdego modelu generowano po 500 trajektorii składających się z 1000 obserwacji. Wykorzystując pierwszą propozycję obliczano jednookresową predykcję na podstawie ruchomego okna składającego się z 30 obserwacji. Przykład takiej predykcji zamieszczono na rysunkach 7-8. Za pomocą krzyży zaznaczono przewidywania na podstawie 5-elemntowego ruchomego okna, powierzchnie tarcz reprezentują błędy przewidywań. Własności propozycji porównywano przewidywaniami wykonywanymi za pomocą lokalnej[4] regresji najmniejszych kwadratów, lokalnego estymatora maksymalnej głębi regresyjnej, nieparametrycznej regresji Nadaraya-Watsona, algorytmu




Rysunek 9. Składowe modelu CHARME wykorzystywanego w symulacjach

Źródło: Obliczenia własne, {fGarch}.

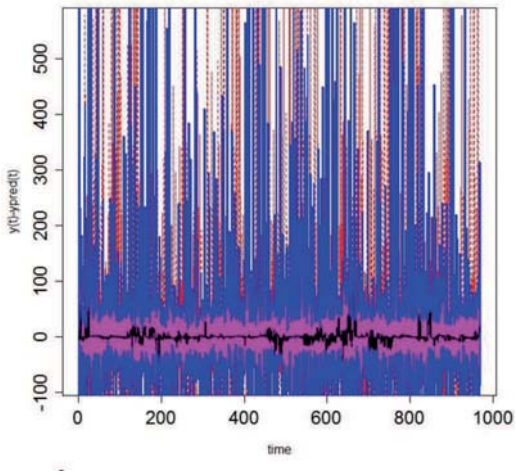
[4] Przez „lokalny" rozumiemy obliczany dla każdego okna.




Rysunek 10. Przykładowa trajektoria modelu CHARME o dwóch reżimach

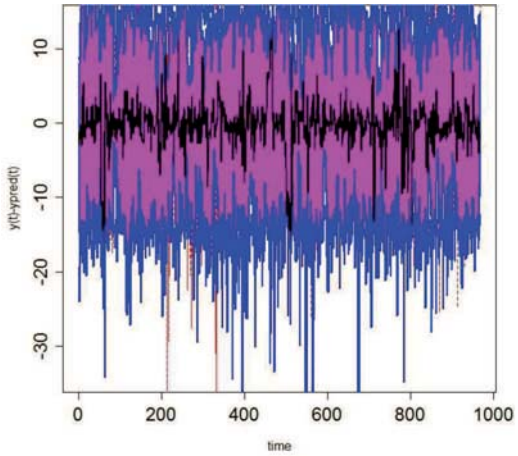
Źródło: Obliczenia własne, {fGarch}.

LOESS (najlepszej alternatywy). W tabelach 1-3 zawarto wyniki naszych symulacji w porównaniu z algorytmem LOESS. Rysunki 11-12 przedstawiają funkcjonalne wykresy ramka wąsy naszych symulacji (por. Ramsay i in., 2010). Obliczano także średniokwadratowy pierwiastek błędu prognozy (RMSEF), maksymalną z pięciuset średnich liczonych dla każdego z 1000 rozpatrywanych w symulacjach chwil $\max_i(\bar{x}_j)$ i średnią z pięciuset średnich liczonych dla każdego z 1000 rozpatrywanych w symulacjach chwil $średnia_i(\bar{x}_j)$. Wyniki przeprowadzonych symulacji (tab. 1-3) przemawiają na korzyść propozycji, zwłaszcza w sytuacji obecności obserwacji odstających.




Rysunek 11. Funkcjonalny rysunek ramka-wąsy – wyniki symulacji dla modelu 1 i algorytmu LOESS i 5% AO

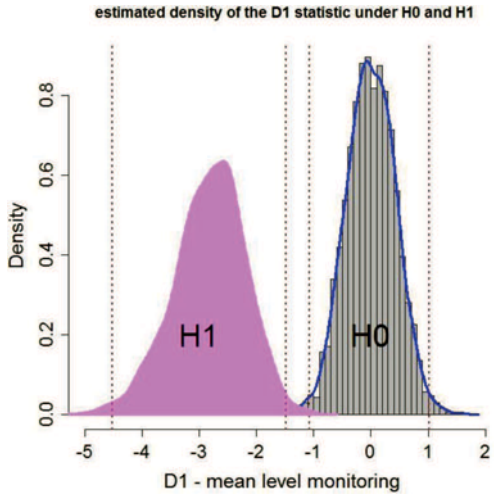
Źródło: Obliczenia własne, {fda}.




Rysunek 12. Funkcjonalny rysunek ramka-wąsy – wyniki symulacji dla modelu 1, pierwszej propozycji i 5% AO

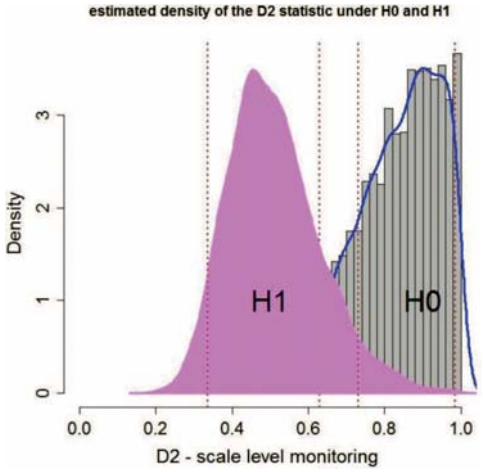
Źródło: Obliczenia własne, {fda}.

W celu sprawdzenia własności naszej drugiej propozycji – za pomocą symulacji oszacowano rozkład statystyk D1 i D2 pod warunkiem hipotezy zerowej głoszącej, że nie następuje zmiana charakterystyki położenia strumienia (w przypadku statystyki D1) oraz nie następuje zmiana charakterystyki rozrzutu strumienia (w przypadku statystyki D2) i kilku hipotez alternatywnych (następuje zmiana charakterystyki położenia albo charakterystyki rozrzutu strumienia) przy założeniu wyszczególnionych powyżej modeli 1, 2, i 3.




Rysunek 13. Propozycja druga – rozkład statystyki przy prawdziwości $H_0$ lub $H_1$, odpowiednio – położenie

Źródło: Obliczenia własne, {lsdepth}.




Rysunek 14. Propozycja druga – rozkład statystyki przy prawdziwości $H_0$ lub $H_1$ odpowiednio – rozrzut

Źródło: Obliczenia własne, {lsdepth}.

Rysunki 13-14 pokazują bardzo dobre własności naszych propozycji co do wykrywania zmian położenia centrum strumienia oraz zmian jego rozrzutu.

Tabela 1.

Wyniki symulacji własności pierwszej propozycji dla modelu 1

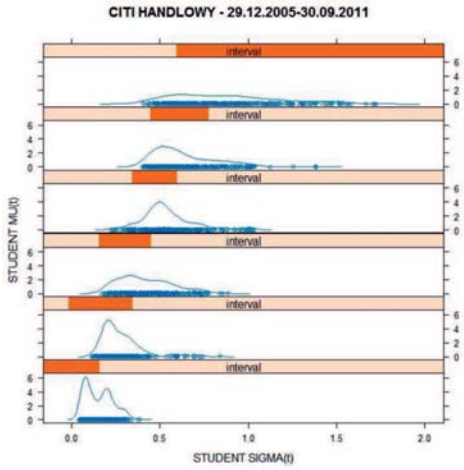
| **MODEL 1** | **RMS** | $\max_i(\bar{x}_j)$ | $średnia_i(\bar{x}_j)$ |
|---|---|---|---|
| **LOESS** | 4,223 | 0,591 | 0,105 |
| **LOESS+5\%** | 6507 | 34,067 | 0,204 |
| **STUDENT MED** | 36,158 | 1,77 | 0,554 |
| **STUDENT MED+5\% AO** | 2573 | 0,755 | -0,032 |

Rysunki 15-16 przedstawiają zastosowanie trzeciej propozycji do empirycznego szeregu cen akcji spółki Citi Handlowy notowanych na GPW pomiędzy 29.12.2005 a 30.09.2011 roku. Łatwo możemy zauważyć zależność pomiędzy poziomem położenia centrum procesu (zysk) a zmiennością procesu (ryzyko). W kontekście analizy skośności procesu zalecamy stosowanie wykresu konturowego głębi Studenta.

Tabela 2.

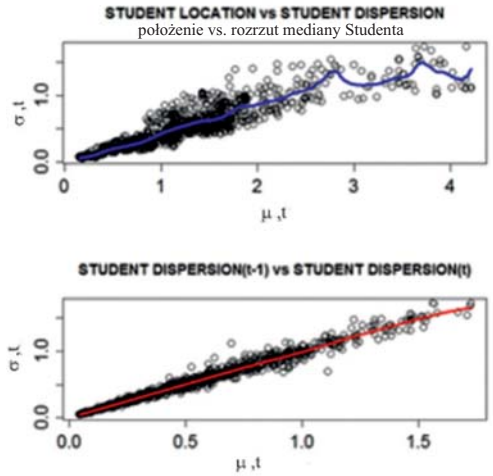
Wyniki symulacji własności pierwszej propozycji dla modelu 2

| **MODEL 2** | **RMS** | $\max_i(\bar{x}_j)$ | $średnia_i(\bar{x}_j)$ |
|---|---|---|---|
| **LOESS** | 7,95 | 0,52 | 0,027 |
| **LOESS+5\%** | 3991 | 17,84 | -0,12 |
| **STUDENT MED** | 20,47 | 0,83 | 0,12 |
| **STUDENT MED+5\% AO** | 1166,35 | 5,85 | 0,32 |




Rysunek 15. Przykład zastosowania trzeciej propozycji

Źródło: Obliczenia własne, {lsdepth}.




Rysunek 16. Przykład zastosowania trzeciej propozycji

Źródło: Obliczenia własne, {lsdepth.}

Tabela 3.

Wyniki symulacji własności pierwszej propozycji dla modelu 2

| **MODEL 3** | **RMS** | $\max_i(\bar{x}_j)$ | $średnia_i(\bar{x}_j)$ |
|---|---|---|---|
| **LOESS** | 1,389 | 0,203 | -0,016 |
| **LOESS+5\%** | 108528 | 230,2 | -0,473 |
| **STUDENT MED** | 15,775 | 0,693 | -0,227 |
| **STUDENT MED+5\% AO** | 14,809 | 0,24 | -0,222 |

## 6. PODSUMOWANIE

Warto podkreślić, że w przypadku ekonomicznych strumieni danych z racji tego, że dane muszą być przetwarzane na bieżąco oraz ich napływ nie ma końca – typowy sposób analizy danych statystycznych nie ma zastosowania. Nie możemy takich danych analizować za pomocą znanych procedur klasycznej statystyki wywodzących się z postulatów Fishera z lat dwudziestych ubiegłego wieku (por. Huber, 2011). Nie mamy bowiem do czynienia z dobrze zdefiniowanym eksperymentem, z danymi generowanymi przez precyzyjnie zdefiniowany model. Strumień danych niesie sygnałpojawiający się w losowych chwilach. Dodatkowo strumienie danych generowane są przez procesy niestacjonarne o zmiennym typie niestacjonarności.

W pracy przedstawiono trzy propozycje narzędzi przeznaczonych do odpornej analizy strumieni danych mogących zawierać obserwacje odstające. Badania symulacyjne wskazują na dobre własności statystyczne propozycji. Zdaniem autora odporna analiza strumieni danych ekonomicznych może przyczynić się do lepszego rozumienia zachowań uczestników rynku. Przedstawione w pracy propozycje stanowią punkt wyjścia dla dalszych studiów zagadnień analizy strumieni danych[5].

*Uniwersytet Ekonomiczny w Krakowie*

## LITERATURA

[1] Aggerwal Ch.C. (ed.), (2007), *Data Streams – Models and Algorithms*, Springer, New York.
[2] Ait-Sahalia Y., Jacod J., Li J., (2012), Testing for jumps in noisy high frequency data, *Journal of Econometrics*, 168, 207-222.
[3] Bocian, Kosiorowski, Węgrzynkiewicz, Zawadzki (2012), pakiet środowiska R {depthproc 1.0} https://r-forge.r-project.org/projects/depthproc/.
[4] Das T., Krishnan S., Venkatasubramanian S., Yi K., (2006), An Information-Theoretic Approach to Detecting Changes in Multi-Dimensional Data Streams. Proceedings of the 38th Symposium on the Interface of Statistics, Computing Science, and Applications (Interface '06)}, Pasadena, CA.
[5] Fan, J. Yao, Q. (2005), *Nonlinear time series: nonparametric and parametric methods*, Springer, New York.
[6] Genton M. G., Lucas A. (2003), Comprehensive Definitions of Breakdown Points for Independent and Dependent Observations, *Journal of the Royal Statistical Society Series* B 65(1), 81-84.
[7] Hall, P., Rodney, C. L. and Yao, Q. (1999). Methods for estimating a conditional distribution function. *Journal of the American Statistical Association*, 94, (445), 154-163.
[8] Hastie T., Tibshiriani R., Friedman J., (2009), The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Second Edition, Springer.
[9] Hahsler M., Dunhamr H. M., (2010), EMM: Extensible Markov Model for Data Stream Clustering in R, *Journal of Statistical Software*, 35(5), 2-31.
[10] Huber P., Ronchettii E. M., (2009), Robust Statistics. John Wiley & Sons. New York.
[11] Huber P., (2011) *Data Analysis: What Can Be Learned From the Past 50 Years*, John Wiley & Sons. New York.

[5] Autor uprzejmie dziękuje za szereg sugestii anonimowych Recenzentów, które w znaczącym stopniu poprawiły jakość niniejszego artykułu.

[12] Jacod J., Shiryaev A.N., 2003, *Limit Theorems for Stochastic Processes*, second ed., Springer-Verlag, New York.
[13] Kong L., Zuo Y., (2010), Smooth Depth Contours Characterize the Underlying Distribution, *Journal of Multivariate Analysis* 101, 2222-2226.
[14] Kosiorowski D., (2010), Depth Based Procedures for Estimation ARMA and GARCH Models, Y. Lechevallier, G. Saporta (ed.) Proceedings of COMPSTAT'2010 19th International Conference on Computational Statistics, Physica–Verlag, 1207-1214.
[15] Kosiorowski D., (2012), *Statystyczne funkcje głębi w odpornej analizie ekonomicznej*, Wydawnictwo UEK w Krakowie, Kraków.
[16] Kosiorowski D., (2012), Student depth in robust economic data stream analysis, Colubi A.(Ed.) Proceedings COMPSTAT'2012, The International Statistical Institute/International Association for Statistical Computing.
[17] Kosiorowski D., Snarska M., (2012), Robust monitoring of a multivariate data stream, LINSTAT 2012, artykuł złożony do *Communications in Statistics.*
[18] Maronna R.A., Martin R.D., Yohai V.J., (2006), *Robust Statistics - Theory and Methods.* Chichester: John Wiley & Sons Ltd.
[19] Mizera I., (2002), On Depth and Depth Poins: a Calculus. *The Annals of Statistics* (30), 1681-1736.
[20] Mizera I., C.H. Müller (2004), Location-scale Depth (with discussion), *Journal of the American Statistical Association* 99, 949-966.
[21] Ramsay J.O., Hooker G., Graves S., (2010), *Functional Data Analysis with R and Matlab*, Springer, New York.
[22] Shalizi C.R., Kontorovich A., (2007), *Almost None of the Theory of Stochastic Processes A Course on Random Processes, for Students of Measure-Theoretic Probability, with a View to Applications in Dynamics and Statistics*, http://www.stat.cmu.edu/˜cshalizi/almost-none/
[23] Serfling R., (2006). Depth Functions in Nonparametric Multivariate Inference, In: Liu R.Y., Serfling R., Souvaine D. L. (Eds.): Series in Discrete Mathematics and Theoretical Computer Science, AMS, vol. 72, 1-15.
[24] Stockis J.-P., Franke J., Kamgaing J.T., (2010). On geometric ergodicity of CHARME models, *Journal of the Time Series Analysis* 31, 141-152.
[25] Szewczyk W., (2010), Streaming data, *Wiley Interdisciplinary Reviews: Computational Statistics*, 3(1), (on-line journal).

## GŁĘBIA POŁOŻENIA-ROZRZUTU W STRUMIENIOWEJ ANALIZIE DANYCH EKONOMICZNYCH

S t r e s z c z e n i e

Z praktycznego punktu widzenia priorytetowym celem analizy ekonomicznego szeregu czasowego jest uzyskanie wglądu na podstawie stale uaktualnianej próby umiarkowanej długości w krótkookresowe właściwości probabilistyczne procesu generującego dane. Na podstawie takiej w ogólności nieprecyzyjnej wiedzy dokonywanych jest szereg decyzji ekonomicznych oraz prognoz. W praktyce bardzo ważną kwestią jest odpowiedź na pytanie co mówi nam większość danych o przyszłym zachowaniu większości uczestników pewnego rynku. Szczególnie trudno odpowiedzieć na takie pytanie w przypadku wielkich zbiorów danych generowanych przez zmieniający się wielowymiarowy model. W artykule prezentujemy wybrane zastosowania procedur indukowanych przez głębię położenia-rozrzutu Mizery i Müller w odpornej analizie strumienia danych ekonomicznych.

**Słowa kluczowe**: strumień danych, statystyczna funkcja głębi, wielowymiarowa mediana

## LOCATION-SCALE DEPTH IN ECONOMIC DATA STREAM ANALYSIS

A b s t r a c t

In this paper we study the properties of the location-scale depth procedures introduced by Mizera & Müller and look into the probabilistic information of the underlying time series model carried by them. We focus our attention on short term multivariate quantile based description of the possible time series model. We study robustness and utility of such the description in a decision making process. In particular we investigate properties of the moving Student median (two dimensional Tukey median in a location–scale problem).

**Key words**: Data Stream, Statistical depth function, multivariate median