Uniwersytet Ekonomiczny w Katowicach

Studia Ekonomiczne. Zeszyty Naukowe
Uniwersytetu Ekonomicznego w Katowicach
ISSN 2083-8611 Nr 265 · 2016

Michał Trzęsiok

Uniwersytet Ekonomiczny w Katowicach
Wydział Zarządzania
Katedra Analiz Gospodarczych i Finansowych
michal.trzesiok@ue.katowice.pl

IDENTYFIKACJA OBSERWACJI ODDALONYCH W SZEREGACH CZASOWYCH

Streszczenie: W artykule uwzględniono różne podejścia do zagadnienia identyfikacji obserwacji oddalonych: podejście dedykowane dla szeregów czasowych i modeli ARIMA, mierniki stopnia oddalenia obserwacji oraz metody klasyfikacyjne. Celem cząstkowym jest zestawienie istniejących metod, ze wskazaniem możliwości pewnych modyfikacji dla polepszenia wyników otrzymywanych z prowadzonej diagnostyki.

Słowa kluczowe: identyfikacja obserwacji oddalonych, klasyfikacja, szeregi czasowe.

Wprowadzenie

Anomalie, takie jak obserwacje odstające czy nagłe zmiany poziomu badanego zjawiska, często występują w rzeczywistych szeregach czasowych [Balke, 1993]. Te szczególne obserwacje występują pojedynczo lub tworzą krótkie ciągi obserwacji i są na ogół wynikiem zaistnienia pewnych wyjątkowych i rzadkich sytuacji, takich jak: wojny, kryzysy, strajki, zmiana regulacji prawnych itp. Wystąpienie w szeregu czasowym obserwacji odstających lub skokowych zmian poziomu rodzi wiele problemów w procesie modelowania [Chang, Tiao i Chen, 1988]. Niezależnie bowiem od tego jak bardzo wyrafinowana metoda zostanie wykorzystana do zbudowania modelu, jakość tego modelu zależy wprost od jakości danych. Przyczyny wystąpienia obserwacji odstających w szeregu czasowym mogą być różnego typu, podobnie jak i same obserwacje odstające mogą różnić się pod względem charakteru i konsekwencji dla dalszego sposobu analizy. Wprawdzie rozwija się dynamicznie obszar metod odpornych, które nie wymagają specjal-

nego traktowania danych zawierających obserwacje nietypowe, lecz szczególnie dla szeregów czasowych temat identyfikacji obserwacji odstających jest związany z pozyskiwaniem dodatkowej, cennej wiedzy o analizowanym zjawisku i pozostaje zagadnieniem ważnym i aktualnym.

W literaturze zagadnienie identyfikacji obserwacji odstających jest opisywane w różnych kontekstach i z tego względu występuje pod wieloma nazwami: wykrywanie anomalii, punktów zwrotnych, nadużyć/oszustw czy nietypowych zachowań, prognozowanie bankructwa (*outlier detection, anomaly/event/novelty/ /change point/fault/misuse detection*). Samo pojęcie obserwacji odstającej również nie jest definiowane jednoznacznie. W niniejszej pracy posłużono się dosyć ogólną definicją zaczerpniętą z pracy Hawkinsa [1980], który przez obserwację odstającą rozumie taką obserwację, która odchyla się tak bardzo od innych obserwacji, że rodzi to przypuszczenie, że powstała w wyniku działania innego mechanizmu.

Metody wykorzystywane do identyfikacji obserwacji odstających są bardzo różnorodne i pochodzą z odmiennych działów metodologii badań statystycznych, gdyż wśród nich znajdują się zarówno metody dyskryminacyjne, taksonomiczne, estymacji funkcji gęstości czy wizualizacji danych, jak i również przetwarzania sygnałów. Pierwsze rozległe prace badawcze nt. identyfikacji obserwacji odstających pojawiły się w latach 70. XX w. i nadal są rozwijane [Fox, 1972; Barnett i Lewis, 1978; Hawkins, 1980; Tsay, 1986, Chen i Liu, 1993, Rousseeuw i Leroy 2003].

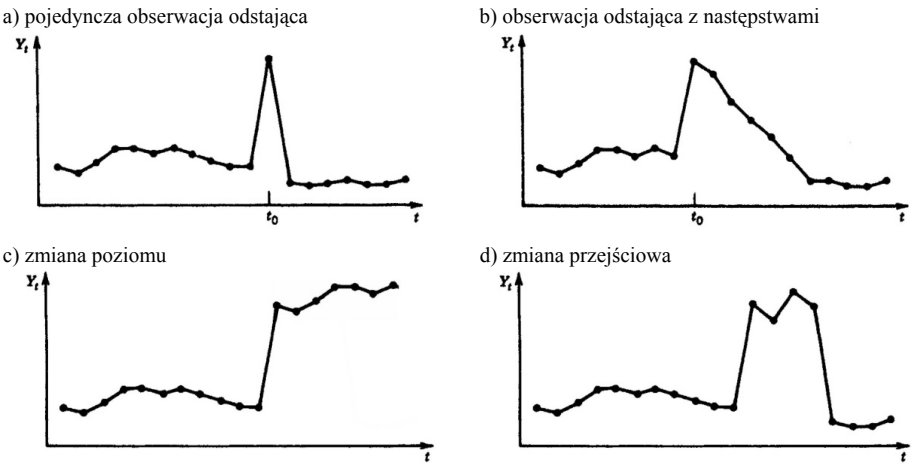
Przytoczona definicja obserwacji odstającej jest bardzo ogólna i można ją uszczegółowić podając kilka typów obserwacji odstających. Dla szeregów przekrojowych wyróżnia się trzy rodzaje obserwacji odstających:

1) obserwacje *nietypowe* (*outliers*), w których wyróżniona jest zmienna objaśniana Y i wartość tej zmiennej znacząco odchyla się od wartości dla innych obserwacji;
2) obserwacje *wysokiej dźwigni* (lub *dźwigniowe*; *leverage*), w których wartość przynajmniej jednej ze zmiennych objaśniających (**X**) znacząco odchyla się od wartości tej zmiennej dla innych obserwacji;
3) obserwacje *wpływowe* (*influential observations*), których wyłączenie ze zbioru danych powoduje istotną zmianę modelu [Rousseeuw i Leroy, 2003].

Uwzględniając specyfikę danych w postaci szeregów czasowych wyróżnia się nieco inne cztery typy obserwacji odstających:

a) pojedyncza obserwacja odstająca (AO – *Additive Outlier*),
b) obserwacja odstająca z następstwami (IO – *Innovation Outlier*),
c) zmiana poziomu (LS – *Level Shift*),
d) zmiana przejściowa (TC – *Temporary Change*) [Fox, 1972; Chen i Liu, 1993].

Zamiast przedstawiać formalne definicje przytoczonych czterech typów obserwacji odstających, zilustrowano je na rys. 1.



Rys. 1. Ilustracja czterech typów obserwacji odstających w szeregach czasowych

Źródło: Na podstawie [Rousseeuw i Leroy, 2003, rys. 6, s. 276].

Celem niniejszego artykułu jest przedstawienie trzech różnych metod identyfikacji obserwacji odstających, zestawienie ich własności i zilustrowanie na przykładzie danych rzeczywistych.

1. Wybrane metody identyfikacji obserwacji odstających w szeregach czasowych

1.1. Metoda Chena i Liu, dedykowana dla modeli ARIMA

Załóżmy, że dany jest pewien stacjonarny proces $\{x_t\}$ postaci:

$$x_t = f\left(x_{t-1}, x_{t-2}, \ldots, x_{t-p}; \varepsilon_{t-1}, \varepsilon_{t-2}, \ldots, \varepsilon_{t-s}\right) + \varepsilon_t, \tag{1}$$

gdzie $\varepsilon_t \sim N(0, \sigma^2), t \in \{1,2,\ldots,n\}$.

Przyjmujemy, że dane empiryczne to pewien obserwowany proces $\{y_t\}$ z zaburzeniem w momencie q [Battaglia i Orfei, 2005], przy czym dla pojedynczej obserwacji odstającej (AO) w momencie q ($1 < q < n$), proces $\{y_t\}$ można zapisać z wykorzystaniem delty Kroneckera:

$$y_t = x_t + \omega_q \delta_{t,q}, \quad \text{gdzie} \quad \delta_{t,q} = \begin{cases} 1, \text{ dla } t = q, \\ 0, \text{ dla } t \neq q, \end{cases} \tag{2}$$

zaś dla obserwacji odstającej z następstwami (IO) proces $\{y_t\}$ można wyrazić jako:

$$y_t = f\left(y_{t-1}, y_{t-2}, \ldots, y_{t-p}; \eta_{t-1}, \eta_{t-2}, \ldots, \eta_{t-s}\right) + \eta_t, \quad \eta_t = \varepsilon_t + \omega_q \delta_{t,q}. \tag{3}$$

Zakładamy przy tym, że wielkość ω_q jest nieznana (nazywana *wielkością odchylenia* obserwacji odstającej). Identyfikacja obserwacji odstających metodą Chena i Liu odbywa się według procedury iteracyjnej obejmującej trzy kroki przedstawione w tabeli 1.

Tabela 1. Iteracyjna procedura identyfikacji obserwacji odstających w metodzie Chena i Liu

Krok 1	Przeprowadź estymację parametrów modelu ARIMA dla procesu $\{y_t\}$
Krok 2	Mając dane parametry modelu z kroku 1, załóż, że w każdym momencie q wystąpiła obserwacja odstająca y_q i oszacuj dla niej wielkość odchylenia ω_q. Jeśli wielkość ta przekracza ustaloną wartość progową (np. 3,5 *SE*, gdzie *SE* to błąd standardowy), to przyjmij, że jest to obserwacja odstająca i przejdź do kroku 3
Krok 3	Usuń efekt wystąpienia obserwacji odstającej przez odjęcie od y_q oszacowanej wielkości odchylenia ω_q i skoryguj wszystkie kolejne obserwacje zgodnie z modelem zbudowanym w kroku 1 i powróć do kroku 2

W przypadku zidentyfikowania pewnej obserwacji w momencie q^* jako obserwacji odstającej, zachowanie kolejnych obserwacji i wielkość ewentualnego odchylenia dla tych obserwacji decyduje o tym, jakiego typu obserwacją odstającą jest y_{q^*}.

Przedstawiona metoda Chena i Liu zostanie zestawiona z dwiema innymi, nieco bardziej uniwersalnymi metodami identyfikowania obserwacji odstających. Jedną z tych metod jest znana z wielowymiarowej analizy regresji metoda wykorzystująca odległość Mahalonobisa.

1.2. Metoda wykorzystująca odległość Mahalanobisa

Przede wszystkim w ekonometrii stosuje się metody identyfikacji obserwacji oddalonych wykorzystujące kryterium bazujące na odległości Mahalonobisa [Healy, 1968]:

$$MD^2(\mathbf{x}) = (\mathbf{x} - \hat{\boldsymbol{\mu}})\hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{x} - \hat{\boldsymbol{\mu}}), \tag{4}$$

gdzie $\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}}$ jest wartością przeciętną, a $\hat{\boldsymbol{\Sigma}} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^{\mathrm{T}}(\mathbf{x}_i - \hat{\boldsymbol{\mu}})$ macierzą wariancji i kowariancji.

Ze względu na podkreślenie ogólności omawianej metody pozostawiono oryginalny zapis macierzowy miary Mahalanobisa (przypadek regresji wielorakiej), choć oczywiście w sytuacji stosowania jej do szeregów czasowych mamy przypadek jednowymiarowy. Identyfikacja obserwacji odstających odbywa się przez porównanie kwadratu odległości Mahalanobisa dla każdej obserwacji z wartościami krytycznymi odczytanymi z rozkładu χ^2. W przypadku wystąpienia dużych różnic (na przyjętym poziomie istotności) daną obserwację traktuje się jako odstającą. To podejście ma jednak tę podstawową wadę, że wartość samego kryterium (4) w bezpośredni sposób zależy od klasycznych statystyk, które są bardzo wrażliwe na występowanie wartości oddalonych. W celu wyeliminowania tej wady zaproponowano modyfikacje wyliczania wartości miernika (4) poprzez zastąpienie średniej $\hat{\boldsymbol{\mu}}$ przez odporny parametr położenia. Jedna z propozycji to wykorzystanie estymatora *MVE* (*Minimum Volume Ellipsoid Estimator*), tj. estymatora o minimalnej objętości elipsoidy [Rousseeuw, 1984]. Druga z propozycji to wyznaczenie parametru położenia $\hat{\boldsymbol{\mu}}$ we wzorze (4), wykorzystując estymator o minimalnym wyznaczniku macierzy kowariancji (*MCD – Minimum Covariance Determinant Estimator*) [Rousseeuw, 1984]. Trzecie podejście zasugerowane w pracy [Filzmoser, Maronna, Werner, 2008] wykorzystuje analizę głównych składowych i identyfikuje obserwacje oddalone właśnie po przekształceniu wszystkich obserwacji w przestrzeń głównych składowych, przez wyznaczenie w tej przestrzeni wartości kwadratu odległości Mahalanobisa. Nadmienić należy, że w tym podejściu wykorzystuje się nieco zmodyfikowany, odporny wariant metody głównych składowych, w którym na etapie przygotowania danych do analizy występuje standaryzacja zmiennych z wykorzystaniem mediany jako parametru położenia oraz *MAD*, czyli medianowego odchylenia bezwzględnego, jako parametru rozproszenia. Po zastosowaniu takiej standaryzacji, obliczanie odległości euklidesowej w przestrzeni głównych składowych jest równoważne obliczaniu odpornego wariantu odległości Mahalanobisa. W części empirycznej tego artykułu wykorzystano właśnie ów trzeci wariant, tj. metodę *MD**, opartą na odległości Mahalanobisa z poprawkami zaproponowanymi przez Filzmosera, Maronnę i Wernera [2008] (identyfikacja obserwacji odstających w przestrzeni głównych składowych).

1.3. Metoda SVM identyfikacji obserwacji odstających

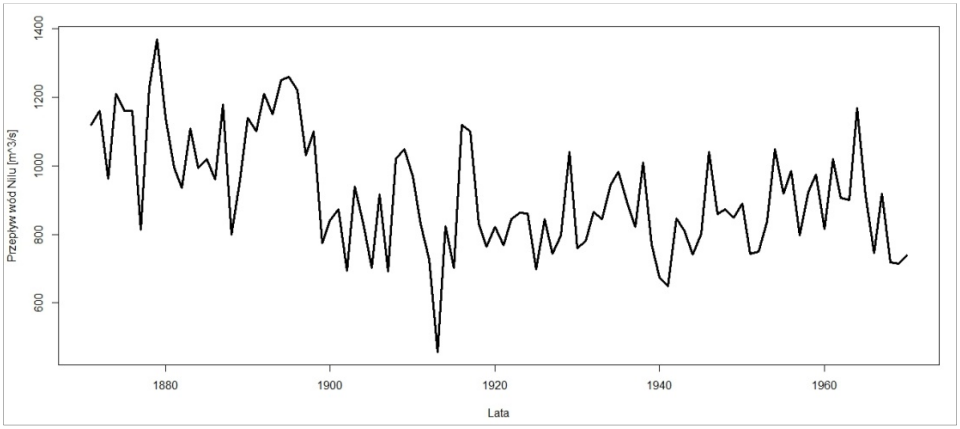
Trzecia z porównywanych metod identyfikacji obserwacji odstających należy do nowej generacji statystycznych metod automatycznego uczenia się. Metoda wektorów nośnych *SVM* (*Support Vector Machines*) ma wiele wariantów, które można wykorzystać do różnych zagadnień (w dyskryminacji, regresji, tak-

sonomii). Jeden z wariantów metody SVM pozwala na wyznaczenie *uogólnionego wielowymiarowego kwantyla rozkładu* generującego dane z analizowanego zbioru. Przez uogólniony kwantyl rozkładu należy rozumieć taki obszar $Q \subset \mathbf{R}^k$ wielowymiarowej przestrzeni danych, w którym z jednej strony niemal wszystkie obserwacje wygenerowane z rozkładu należą do Q, a z drugiej niemal wszystkie obiekty niepochodzące z rozkładu generującego dane należą do dopełnienia zbioru Q. Podobnie jak to było w metodzie bazującej na odległości Mahalanobisa, metoda ta oryginalnie została zaproponowana do rozwiązywania problemów w wielowymiarowych przestrzeniach danych, ale można również wykorzystać ją w przypadku jednowymiarowym (dla szeregu czasowego). Szczegółowy opis metody SVM zastosowanej do wyznaczania uogólnionego kwantyla rozkładu znaleźć można w pracach [Ben-Hur i in., 2001; Trzęsiok, 2007]. W tym miejscu ograniczono opis metody do podania jej głównej idei. Mianowicie, poprzez wykorzystanie pewnej wybranej funkcji jądrowej, określającej nieliniowe przekształcenie przestrzeni danych, standardową technikę stosowaną w metodzie wektorów nośnych, poszukiwanie rozwiązania problemu zostaje przeniesione w przestrzeń $\mathbf{Z}$ o znacznie większym wymiarze i w tej nowej przestrzeni wyznaczana jest optymalna hiperkula, zawierająca obrazy obserwacji ze zbioru uczącego. Poszukiwana jest hiperkula o najmniejszym możliwym promieniu, tzw. hiperkula Czebyszewa. Tej hiperkuli w przestrzeni $\mathbf{Z}$ odpowiada (jako przeciwobraz) pewien zbiór w pierwotnej przestrzeni danych. Jest nim poszukiwany uogólniony kwantyl Q. Ze względu na uelastycznienie metody, na wypadek wystąpienia w zbiorze danych potencjalnych błędów pomiaru lub obserwacji nietypowych, wyznaczona hiperkula Czebyszewa nie musi zawierać obrazów wszystkich obserwacji z analizowanego zbioru danych. Obiekty, które znalazły się poza tą hiperkulą, można łatwo zidentyfikować. Są to obserwacje, które znajdują się poza uogólnionym kwantylem rozkładu i potencjalnie pochodzą z innego rozkładu, czyli mogą zostać potraktowane jako obserwacje odstające.

Przy wykorzystywaniu metody SVM do modelowania użytkownik musi podać wartości dwóch parametrów (γ – parametr funkcji jądrowej Gaussa oraz parametr regularyzacji $\nu \in [0,1]$, określający kompromis między dopasowaniem modelu a jego zdolnością do uogólniania – [por. Trzęsiok, 2008]). Wybór wartości tych parametrów ma kluczowe znaczenie dla liczby obserwacji zidentyfikowanych jako odstające. W badaniach empirycznych przedstawionych w dalszej części artykułu wybrano strategię, w której zbudowano wiele modeli SVM identyfikujących obserwacje odstające przy różnych kombinacjach obu kluczowych parametrów, a ostatecznie uznano za odstające tylko te obserwacje, które co najmniej dwukrotnie otrzymały takie wskazanie (przez przynajmniej dwa modele ze zbioru modeli SVM).

2. Ilustracja i porównanie wyników działania przedstawionych metod

Przedstawione metody identyfikacji obserwacji odstających w szeregach czasowych zostaną zilustrowane na zbiorze danych rzeczywistych `Nile` wykorzystywanym do badania i porównywania własności metod statystycznego modelowania dla szeregów czasowych. Dane w zbiorze `Nile` dotyczą przepływu wody w Nilu w okolicach Asuanu (dane w [m^3/s] od 1871 r. do 1970 r.). Dane z analizowanego szeregu czasowego przedstawiono na rys. 2.



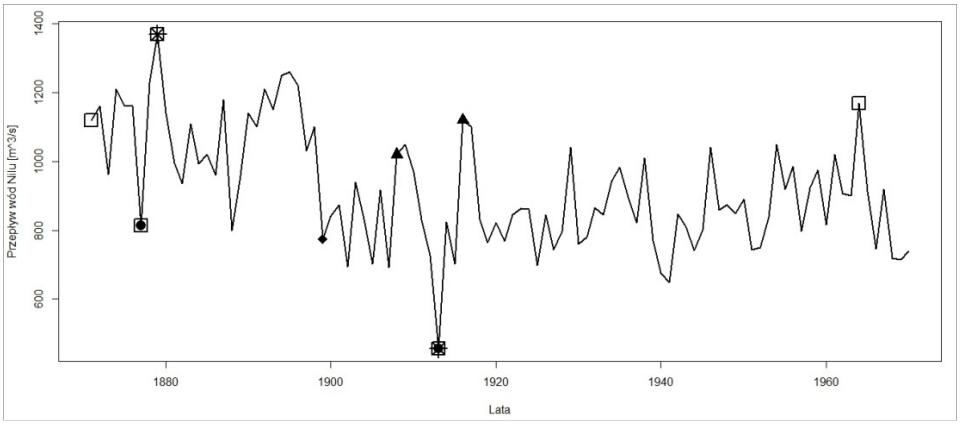
Rys. 2. Przepływ wód Nilu w okolicach Asuanu

Wszystkie obliczenia zostały zrealizowane z wykorzystaniem programu statystycznego **R** i jego dodatkowych pakietów [tsoutliers, mvoutlier, e1071 oraz własnych funkcji i procedur napisanych w języku programu **R**].

W pierwszej kolejności przystąpiono do zidentyfikowania obserwacji odstających metodą Chena i Liu [1993], dedykowaną dla szeregów czasowych. Z powodu prawostronnej asymetrii rozkładu przepływu wód Nilu (zestandaryzowany moment centralny trzeciego rzędu równy 0,318) poddano analizie dane zlogarytmowane. Sprawdzono stacjonarność szeregu rozszerzonym testem Dickeya–Fullera (testem ADF) z hipotezą alternatywną postaci „badany szereg jest stacjonarny". Obliczenia wskazały, że na poziomie istotności $\alpha = 0{,}05$ należy odrzucić hipotezę zerową na rzecz alternatywnej, czyli badany szereg ma własność stacjonarności [wartość statystyki Dickeya–Fullera = –3,3657, przy rzędzie opóźnienia równym 4, prawdopodobieństwo testowe *p-value* = 0,04724]. Nie było więc potrzeby wykorzystania operacji różnicowania szeregu czasowego. W dalszej części przeprowadzono identyfikację obserwacji odstających metodą Chena i Liu otrzymując wyniki przedstawione w tabeli 2 i zaznaczono na rys. 3.

Tabela 2. Obserwacje odstające zidentyfikowane metodą Chena i Liu

Lp.	Rok	Typ obserwacji odstającej
1	1877	pojedyncza obserwacja odstająca (AO; na rys. 3: '⌠')
2	1899	zmiana poziomu (LS; na rys. 3: '⌊')
3	1908	tymczasowa zmiana (TC; na rys. 3: '∫')
4	1913	pojedyncza obserwacja odstająca (AO; na rys. 3: '⌠')
5	1916	tymczasowa zmiana (TC; na rys. 3: '∫')



Rys. 3. Przepływ wód Nilu w okolicach Asuanu z zaznaczonymi obserwacjami odstającymi zidentyfikowanymi metodami: Chena i Liu (symbole wypełnione), SVM (kwadraty puste), wykorzystującą odległość Mahalanobisa (gwiazdki)

Wyniki identyfikacji obserwacji odstających metodą wykorzystującą odległość Mahalonobisa oraz metodą SVM przedstawiono w tabeli 3 oraz na rys. 3.

Tabela 3. Obserwacje odstające zidentyfikowane metodą wykorzystującą odległość Mahalanobisa (*MD**) oraz metodą SVM

Lp.	Metoda *MD**	Metoda SVM
1	1879 (na rys. 3: '✳')	1871 (na rys. 3: '〈')
2	1913 (na rys. 3: '✳')	1877 (na rys. 3: '〈')
3		1879 (na rys. 3: '〈')
4		1913 (na rys. 3: '〈')
5		1964 (na rys. 3: '〈')

Wygenerowane zbiory obserwacji odstających różnią się dla każdej z metod. Można zauważyć, że tylko jedna obserwacja (z roku 1913) została zidentyfikowana jako odstająca przez wszystkie trzy przedstawione metody. Sytuacja ta jest jednak zgodna z intuicją, gdyż metody te wykorzystują bardzo różniące się

podejścia – zarówno pod względem traktowania tego, czym jest obserwacja odstająca, jak i przestrzeni, w której są one poszukiwane, czy też samej metodologii.

Podsumowanie

Przedstawiono i zilustrowano na przykładzie danych rzeczywistych trzy metody identyfikacji obserwacji odstających dla szeregów czasowych. Przedstawione metody znacząco różnią się w sposobie rozwiązania postawionego problemu, a w konsekwencji również wyznaczają odmienne zbiory obserwacji odstających.

Dedykowana dla szeregów czasowych metoda Chena i Liu ma najmniejszy zakres stosowalności, gdyż jest ściśle związana z konkretnym typem modelowania szeregów czasowych i wymaga spełnienia najbardziej restrykcyjnych założeń. Ma jednak zdecydowaną przewagę nad pozostałymi dwiema metodami w tym, że nie tylko identyfikuje czy dana obserwacja jest odstająca czy nie, ale również wskazuje typ obserwacji odstającej, co jest bardzo cenną, dodatkową wiedzą pozyskaną o badanym zjawisku.

Metoda wykorzystująca odległość Mahalanobisa jest uniwersalna i nadaje się do identyfikowania obserwacji odstających nie tylko w szeregach czasowych, ale również w wielowymiarowych danych przekrojowych. Wydaje się jednak, że metoda ta potrafi dla szeregów czasowych zidentyfikować tylko pojedyncze obserwacje odstające. Metoda ta raczej nie radzi sobie z wykryciem anomalii typu „zmiana poziomu" lub „tymczasowa zmiana". Dla szeregów czasowych o znaczących fluktuacjach metoda ta generuje albo bardzo liczny zbiór obserwacji odstających, który wymaga użycia dodatkowych heurystyk do odfiltrowania tych „najważniejszych", albo zbiór obserwacji odstających jest mało liczny i są to wyłącznie wyraźnie odstające pojedyncze obserwacje odstające (łatwe do zidentyfikowania również przez wizualizację szeregu czasowego).

Metoda SVM jest najbardziej elastyczną z przedstawionych metod. Nie nakłada niemal żadnych założeń na badany szereg czasowy. Mechanizm działania tej metody ma jednak charakter czarnej skrzynki, przez co trudno jakkolwiek interpretować i zestawiać wyniki działania tej metody z wynikami innych metod identyfikacji obserwacji odstających. Podobnie jak metoda *MD** metoda SVM wykrywa głównie pojedyncze obserwacje odstające.

Każda z przedstawionych metod w nieco odmienny sposób wykrywa anomalie w szeregu czasowym. Nie można jednoznacznie ocenić, która z metod lepiej nadaje się do tego celu, bo sam problem identyfikacji obserwacji odstają-

cych nie ma jednoznacznego rozwiązania. W przypadku braku jednoznaczności rozwiązania dobrze jest zapoznać się z wynikami analizy z wykorzystaniem metod różniących się podejściem. Daje to analitykowi więcej wiedzy o badanym zjawisku i pozwala na lepsze dostosowanie dalszej procedury badawczej do rozpatrywanego problemu.

Literatura

Balke N.S. (1993), *Detecting Level Shifts in Time Series*, "Journal of Business & Economic Statistics", Vol. 11(1), s. 81-92.

Barnett V., Lewis T. (1978), *Outliers in Statistical Data*, John Wiley & Sons, New York.

Battaglia F., Orfei L. (2005), *Outlier Detection and Estimation in Nonlinear Time Series*, "Journal of Time Series Analysis", Vol. 26(1), s. 107-121.

Ben-Hur A., Horn D., Siegelman H.T., Vapnik V. (2001), *Support Vector Clustering*, "Journal of Machine Learning Research", Vol. 2, s. 125-137.

Chang I., Tiao G.C., Chen C. (1988). *Estimation of Time Series Parameters in the Presence of Outliers*, "Technometrics", Vol. 30(2), s. 193-204.

Chen C., Liu L.M. (1993), *Joint Estimation of Model Parameters and Outlier Effects in Time Series*, "Journal of the American Statistical Association", Vol. 88(421), s. 284-297.

Filzmoser P., Maronna R.A., Werner M. (2008), *Outlier Identification in High Dimensions*, "Computational Statistics & Data Analysis", Vol. 52, s. 1694-1711.

Fox A.J. (1972), *Outliers in Time Series*, "Journal of the Royal Statistical Society. Series B (Methodological)", s. 350-363.

Hawkins D. (1980), *Identification of Outliers*, Chapman and Hall, London – New York.

Healy M.J.R. (1968), *Multivariate Normal Plotting*, "Applied Statistics", Vol. 17, s. 157-161.

Rousseeuw P.J. (1984), *Least Median of Squares Regression*, "Journal of the American Statistical Association", Vol. 79, s. 871-880.

Rousseeuw P.J., Leroy A.M. (2003), *Robust Regression and Outlier Detection*, John Wiley & Sons, New York.

Trzęsiok M. (2007), *Identyfikacja obserwacji oddalonych z wykorzystaniem metody wektorów nośnych* [w:] K. Jajuga, M. Walesiak (red.), *Taksonomia 14. Klasyfikacja i analiza danych – teoria i zastosowania*, Prace Naukowe, nr 1169, Wydawnictwo Akademii Ekonomicznej, Wrocław, s. 350-357.

Trzęsiok M. (2008), *Wybór wartości parametrów przez walidację wyników klasyfikacji taksonomicznej metody wektorów nośnych* [w:] K. Jajuga, M. Walesiak (red.), *Taksonomia 15. Klasyfikacja i analiza danych – teoria i zastosowania*, Prace Naukowe, nr 7 (1207), Wydawnictwo Uniwersytetu Ekonomicznego, Wrocław, s. 354-363.

Tsay R.S. (1986), *Time Series Model Specification in the Presence of Outliers*, "Journal of the American Statistical Association", Vol. 81(393), s. 132-141.

DETECTION OF OUTLIERS IN TIME SERIES

Summary: The paper presents three different methods for detecting anomalies in time series. The first one is dedicated for time series analysis and ARIMA models. Two other two come from very different background: one is associated with measuring the distance from the given observation to the remaining objects in dataset. The other one belongs to the family of classification methods within machine learning framework. The goal of the paper is to present, compare and illustrate these three different approaches on a real world dataset.

Keywords: outliers detection, classification, time series.