B A D A N I A O P E R A C Y J N E I D E C Y Z J E

Nr 1 2008

Adam KUCHARSKI*

ALGORYTMY GENETYCZNE W PROGNOZOWANIU DANYCH GIEŁDOWYCH – USUWANIE OBSERWACJI NIETYPOWYCH

W pracy zaproponowano wykorzystanie algorytmu genetycznego do otrzymywania krótkookresowych prognoz instrumentów giełdowych. Użyty algorytm przypomina w swoim działaniu metodę naiwną z sezonowością, z tym że opóźnienie obserwacji stanowiącej prognozę może być różne dla kolejnych okresów. Dokonano przy tym wcześniejszej identyfikacji i usunięcia obserwacji nietypowych na podstawie macierzy rzutowania, znanej z estymacji odpornej.

Słowa kluczowe: *algorytm genetyczny, prognozowanie, szereg czasowy, estymacja odporna*

Wstęp

Inwestorzy giełdowi to bardzo zróżnicowana grupa. Różnią się preferencjami co do wybieranych instrumentów czy przyjmowanych strategii lokowania kapitału. Część z nich zainteresowana jest szybkim, spekulacyjnym zyskiem, część z kolei cierpliwie czeka, aż przemówią „fundamenty". Niezależnie od przyjętego kryterium podziału istnieje coś, co stanowi wspólny mianownik podejmowanych na parkiecie decyzji, a mianowicie oczekiwania co do przyszłego zachowania się danego instrumentu. Mówiąc krótko, wszyscy uczestnicy giełdy muszą ustalać prognozy. Prezentowana praca stanowi kontynuację badań tej tematyki, rozpoczętych w 2005 roku.

Punktem wyjścia są metody naiwne[1], stanowiące grupę najprostszych modeli występujących w prognozowaniu opartym na szeregach czasowych. Właśnie szeregi notowań i towarzyszących im wolumenów obrotów znajdują się w polu naszego za-

* Katedra Badań Operacyjnych, Uniwersytet Łódzki, ul. Rewolucji 1905 r. nr 41, 90-214 Łódź, e-mail: adamk@uni.lodz.pl

[1] Metody te bazują na założeniu, że wpływające do tej pory na badane zjawisko czynniki zachowają swój wpływ w najbliższej przyszłości na dotychczasowym poziomie i w niezmieniony sposób.

interesowania. Przypomnijmy w tym miejscu, iż u podstaw prognozowania niestrukturalnego leży założenie, że cała potrzebna informacja znajduje się w samym szeregu, a wydobywa się ją poprzez proces jego dekompozycji na poszczególne składowe[2]. Niektóre spośród spotykanych zachowań (np. układanie się obserwacji w serie lub pojawianie się punktów zwrotnych) potrafią jednak w znaczący sposób utrudnić wykorzystanie wspomnianej już dekompozycji. Jako metodę poszukiwania prognoz wykorzystamy algorytm genetyczny, specjalnie do tego celu zmodyfikowany. Ponieważ wcześniejsze badania wykazały[3], że problem stanowią obserwacje o nietypowo dużych wartościach, w prezentowanym artykule skupiono się zatem na próbie zniwelowania tej niedogodności.

Szczególną uwagę zwrócono na metodę naiwną bez zmian oraz metodę naiwną z sezonowością. W ich przypadku prognoza na dany okres kształtuje się na poziomie wartości zaobserwowanej w okresie poprzednim bądź wcześniejszym. Uznano, że dzięki temu dobrze nadają się one do zaimplementowania w algorytmie genetycznym.

Za twórcę nowoczesnej teorii dotyczącej algorytmów genetycznych uważa się Johna H. Hollanda, który już w latach sześćdziesiątych postanowił stworzyć ścisły model, opisujący ewolucyjne i adaptacyjne zjawiska zachodzące w przyrodzie. Z czasem okazało się, że zasugerowane rozwiązania mogą znaleźć wiele zastosowań w najróżniejszych dziedzinach. Zaproponowany przez Hollanda model nosi dziś nazwę klasycznego algorytmu genetycznego (KAG). Spuścizną po biologicznych korzeniach badań jest stosowana w tej dziedzinie do dziś nomenklatura.

W klasycznym algorytmie genetycznym operuje się na zbiorach łańcuchów, które mają stałą długość, a składają się z ciągów zer i jedynek. Łańcuchy te noszą nazwę chromosomów, a pojedyncze bity przyjmujące wartość „0" lub „1", nazywa się genami. Zbiór wszystkich chromosomów, które zostały utworzone w celu znalezienia rozwiązania problemu nazywa się populacją. Ponieważ algorytmy genetyczne skonstruowano z myślą o zagadnieniach optymalizacyjnych, występuje w nich oczywiście funkcja celu, zwana tu najczęściej (za naukami biologicznymi) funkcją przystosowania.

Algorytm zaprezentowany w pracy stanowi modyfikację KAG i w założeniu miałby tworzyć krótkookresowe prognozy szeregów czasowych, obserwowanych na giełdzie papierów wartościowych (takich jak notowania kursów akcji czy wolumenów obrotów), a zawierających wahania przypadkowe i stały poziom zmiennej lub trend liniowy[4].

[2] Dadzą się tu zauważyć pewne analogie do powszechnie używanej przez inwestorów analizy technicznej.

[3] Por. Kucharski [6].

[4] Cieślak M. [2001], Gajda J. [2001].

1. Sposób wyznaczania prognoz *ex post* i *ex ante*

Podstawą prognoz *ex post* wyznaczanych w pracy jest parametr, określany dalej jako *tm*. Jest to maksymalny dopuszczalny numer opóźnienia danej rzeczywistej, będącej prognozą dla aktualnego okresu. Taka definicja oznacza, że prognozę *ex post* dla wybranego okresu o numerze t może tworzyć dowolna obserwacja, nie starsza jednak niż $t - tm$. Sam parametr *tm* może przyjmować wartości z przedziału $<1, n - 2>$, gdzie n oznacza liczbę posiadanych obserwacji[5]. *tm* ustala się *a priori* na początku postępowania i pozostawia bez zmian do końca obliczeń. Pierwsza prognoza *ex post* powstaje dla trzeciego okresu. Wyniki predykcji w zaproponowanej metodzie niekoniecznie muszą sięgać w przeszłość na największą możliwą głębokość[6].

Prognozowana wartość równa się jednej z obserwacji pochodzących sprzed $t - tm$ okresów, np. w okresie trzecim będzie to opóźniona wartość y_{t-1} lub y_{t-2} (jeżeli przez $\boldsymbol{y}_{(t)}$ oznaczymy pewien hipotetyczny szereg czasowy). Poszukujemy takiej sekwencji opóźnień, która zagwarantuje jak najniższy błąd prognozy *ex post*. Dla dalszych niż trzeci okresów liczba możliwych prognoz rośnie aż do poziomu *tm*. Przykładowo, dla $n = 5$ i $tm = 3$ w okresach odpowiednio trzecim, czwartym i piątym mogą wystąpić m.in.: y_{t-1}, y_{t-1}, y_{t-2} lub y_{t-1}, y_{t-2}, y_{t-2}. Z uwagi na to, iż tworzonych w ten sposób potencjalnych szeregów złożonych z prognoz przybywa lawinowo, koniecznym stało się sięgnięcie po algorytm genetyczny. Jego zadanie to znalezienie takiego zbioru opóźnionych wartości, który zapewni jak najmniejszy błąd prognozy.

Ostatnia z wartości rzeczywistych nie bierze udziału w tworzeniu prognoz *ex post*. Wykorzystywana jest dopiero w chwili wyznaczania prognoz *ex ante*.

Mechanizm powstawania tych ostatnich przebiega dwuetapowo:

1. Wyznaczamy prognozę *ex post* na podstawie kryterium, które stanowi wybrana miara błędów prognoz *ex post*.

2. Określamy wartość opóźnienia wskazującego, które dane rzeczywiste wezmą udział w tworzeniu prognoz *ex ante*. Do tego celu może posłużyć dominanta lub mediana wartości opóźnień, wykorzystywanych podczas tworzenia prognoz *ex post*.

Jeżeli wrócimy do przykładowego szeregu, o którym była mowa wcześniej, w pierwszym przypadku mediana wynosi 1, a w drugim 2 okresy.

Na drugim etapie do głosu dochodzi dekompozycja szeregu czasowego. W razie wystąpienia stałego poziomu zmiennej postępowanie staje się analogiczne do metody naiwnej prostej lub naiwnej z sezonowością. Prognozą *ex ante* może być wartość rzeczywista z ostatniego okresu lub któregoś z okresów wcześniejszych, wskazanych

[5] W poprzednich wersjach algorytmu występował przedział $<1, n - 1>$.

[6] W praktyce rzadko okazywało się konieczne wykorzystanie opóźnień równych maksymalnej zakładanej wartości *tm*.

przez dominantę (względnie medianę opóźnień). Ostatnia obserwacja rzeczywista staje się ostatnią z prognoz *ex ante*.

Dla szeregów, w których stwierdzono trend liniowy naśladowane jest postępowanie znane z metody naiwnej z poprawką liniową. Ostatnia obserwacja rzeczywista podlega korekcie o przyrost między nią a obserwacją wskazaną przez dominantę lub medianę. Wprawdzie nie ma tutaj ograniczenia horyzontu prognozy, lecz należy pamiętać, że metody naiwne zaleca się stosować do prognoz krótkookresowych.

2. Modyfikacje klasycznego algorytmu genetycznego

W pracy skorzystamy ze zmodyfikowanego klasycznego algorytmu genetycznego (KAG)[7], przy czym uwzględnimy także wnioski płynące z poprzednich badań[8].

Pojedynczy chromosom odpowiadał jednej z możliwych prognoz *ex post* za okresy <3, *n*>. Wykorzystano w nim kodowanie rzeczywiste, w którym każdy gen zawierał informację o opóźnionym okresie, z którego pochodziła prognoza *ex post*. Takie podejście oznaczało, że przed obliczeniem wartości funkcji przystosowania należało „przetłumaczyć" opóźnienia na konkretne prognozy.

Funkcję przystosowania, a więc kryterium oceny jakości prognoz *ex post*, stanowił błąd RMSPE:

$$\mathrm{RMSPE} = \sqrt{\frac{1}{S}\sum_{i=1}^{S}\left(\frac{y_i - y_i^*}{y_i}\right)^2}\ , \tag{1}$$

gdzie:

y_i – wartość rzeczywista zmiennej y w okresie i,

y_i^* – prognoza *ex post* zmiennej y w okresie i,

S – liczba prognoz *ex post* dla szeregu.

KAG zakłada maksymalizację jako kierunek optymalizacji, zaś w naszym przypadku funkcja kryterium ma być minimalizowana. Jeżeli przez $F(x)$ oznaczymy funkcję przystosowania, to jej minimalizacja będzie równoważna z maksymalizacją pewnej funkcji $G(x)$, czyli

$$\min F(x) = \max G(x) = \max -F(x), \tag{2}$$

gdzie:

$F(x)$ – wartość funkcji przystosowania,

$G(x)$ – pewna funkcja, dla której zachodzi $G(x) = -F(x)$.

[7] Por. Michalewicz Z. [1996].

[8] Por. Kucharski A. [2005] oraz Kucharski [2006].

Konstrukcja chromosomów wymusiła zmiany w operatorach selekcji, krzyżowania i mutacji.

Selekcja chromosomów do reprodukcji odbywała się metodą elitarną i wartości oczekiwanej. Kiedy już wybrane chromosomy zostały dobrane w pary, wymiana genów dokonywała się w sposób analogiczny jak w przypadku reprezentacji binarnej. Chromosomy potomków powstawały przez wymianę genów rodziców za punktem krzyżowania.

Z kolei w operatorze mutacji należało uwzględnić użycie parametru *tm*. Wylosowane i zamieniane miejscami dwa geny nie mogło dzielić w czasie więcej niż *tm* okresów.

Wydłużanie się szeregu i wzrost parametru *tm* gwałtownie powiększa zbiór rozwiązań. Z punktu widzenia algorytmu genetycznego oznacza to konieczność operowania dużymi populacjami, co znacząco wydłuża czas obliczeń.

3. Transformacje przeprowadzane na danych

Jak wspomniano we wstępie, w pracy skupimy się na zagadnieniu pojawiających się nietypowych obserwacji. Ich występowanie znacząco pogarsza jakość prognoz otrzymywanych metodami ilościowymi, w skrajnych wypadkach wręcz je uniemożliwiając. Najpierw należy jednak określić, które z obserwacji uznamy za nietypowe. Do tego celu wykorzystamy narzędzia znane z estymacji odpornej. Pierwszym z nich jest macierz rzutowania:

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}. \tag{3}$$

Występującą w niej macierz obserwacji zmiennych objaśniających $\mathbf{X}$ zastąpimy wektorem obserwacji szeregu czasowego y_t, który obejmował okresy od 3 do n[9]. Stosowna macierz prezentuje się następująco:

$$\mathbf{H} = \mathbf{y}(\mathbf{y}^{\mathrm{T}}\mathbf{y})^{-1}\mathbf{y}^{\mathrm{T}}. \tag{4}$$

Dla szeregu wyznaczamy prognozy metodą średniej ruchomej prostej o stałej wygładzania $k = 2$. Na ich podstawie obliczamy reszty z prognozy *ex post*, które następnie standaryzujemy według formuły

$$e_t^* = \frac{e_t}{s\sqrt{1-h_t}}, \tag{5}$$

[9] Zostało to podyktowane przyjęciem założenia, że pierwszą prognozę *ex post* wyznaczono na trzeci okres.

gdzie:

e_t – reszta z prognozy *ex post*,

s – odchylenie standardowe reszt z prognozy,

h_t – element przekątnej głównej macierzy **H**, odpowiadający prognozie o numerze t.

Reszty wyznaczone według formuły (5) posłużyły do określenia, które obserwacje należy uznać za nietypowe. Jako kryterium potwierdzające tę tezę przyjmujemy wartość $|e_t^*| > 2$.

Obserwację uznaną za nietypową należy z wektora danych wykluczyć. Ponieważ jednak mamy do czynienia z szeregami czasowymi, konieczne jest zastąpienie jej inną wielkością. W pracy rozważono dwa podejścia:

1) zastąpienie przez prognozę na podstawie średniej ruchomej o $k = 2$,

2) zastąpienie przez średnią arytmetyczną, obliczoną z wartości poprzedzającej i następnej.

Drugi z wymienionych sposobów zazwyczaj dawał niższe wartości.

W dalszej części pracy porównamy wyniki obliczeń dla szeregów, wobec których zastosowano jedno i drugie podejście. Zestawimy je z prognozami uzyskanymi dla danych niepoddanych tego rodzaju transformacjom.

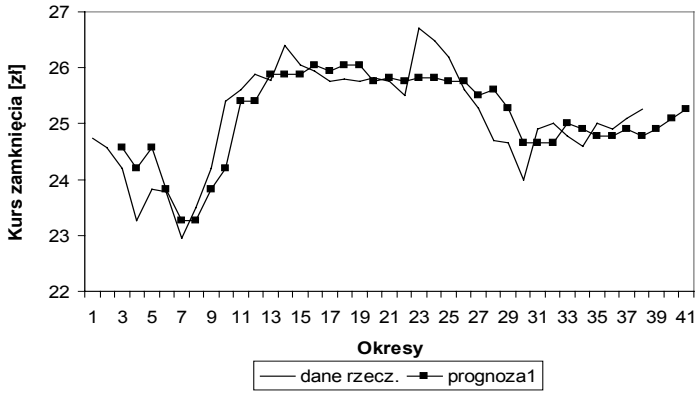
4. Wyniki obliczeń

Poziomy parametrów sterujących przebiegiem algorytmu przyjęliśmy na podstawie wniosków płynących z wcześniejszych badań oraz próbnych przebiegów, wykonanych dla zebranych danych. Liczebność populacji chromosomów wynosiła w każdym z przypadków 1000 osobników[10], prawdopodobieństwa krzyżowania i mutacji równały się odpowiednio: 0,3 oraz 0,1. Kryterium zatrzymania algorytmu była liczba pokoleń, którą określono jako równą 50. Wielkość populacji oraz liczba pokoleń są duże z uwagi na rozmiary zbioru, w którym poszukiwano rozwiązania. Wartość parametru *tm* kształtowała się na poziomie równym 5 dla kursów zamknięcia i 10 dla wolumenów obrotów.

Obliczenia wykonaliśmy dla szeregów notowań oraz wolumenów obrotów następujących walorów: WIG20, KGHM, TP SA, BZWBK. Dane pochodziły z okresu od 2.01.2007 do 22.02.2007 (38 obserwacji). Wybrane instrumenty miały reprezentować z jednej strony różne branże, a z drugiej odmienne warianty zachowań możliwych z punktu widzenia dekompozycji szeregów czasowych. Stały poziom zmiennej wystąpił dla kursu zamknięcia TP S.A. oraz wolumenów wszystkich analizowanych walorów. Dla pozostałych szeregów zaobserwowaliśmy trend liniowy. W danych nie występowała sezonowość.

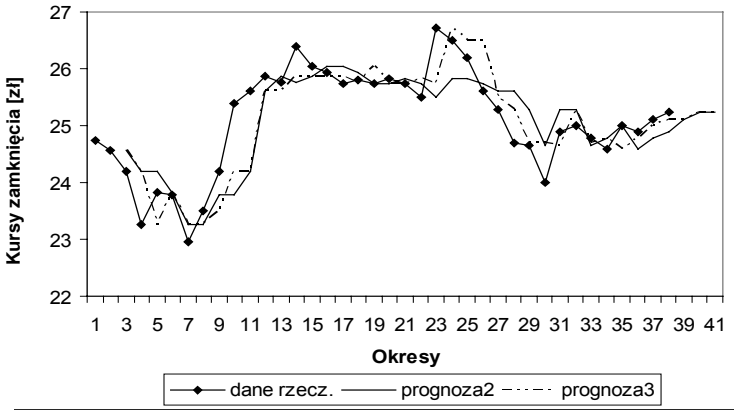
[10] Jeden chromosom równa się jednemu osobnikowi.

Jak wspomniano wcześniej, postanowiliśmy zbadać, jaki wpływ na prognozy generowane przez algorytm genetyczny ma zastąpienie obserwacji uznanych za nietypowe jednym z dwóch rodzajów uśrednień. Jako przykład zaprezentujemy wykresy przedstawiające kursy zamknięcia dla TP S.A. Przyjęliśmy następujące oznaczenia: *prognoza1* odnosi się do prognoz dla danych bez przekształceń, *prognoza2* powstała z wykorzystaniem średniej ruchomej o $k = 2$, zaś *prognoza3* w oparciu o średnią z obserwacji sąsiednich.



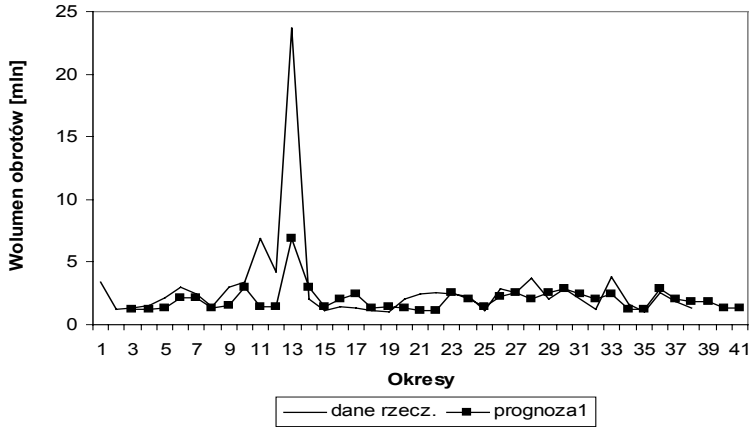
Rys. 1. Prognoza kursu zamknięcia TP S.A. bez przekształceń danych
Ź r ó d ł o: Opracowanie własne.

Analiza danych przedstawionych na rysunkach 1 i 2 wskazuje, że w szeregu nie występowały zbyt duże wahania przypadkowe. RMSPE dla prognozy przedstawionej na pierwszym z rysunków wyniosło 1,89%, co stanowi bardzo dobrą wartość. Błędy dla pozostałych prognoz notowań tej spółki równały się odpowiednio: 1,79% i 1,78%.

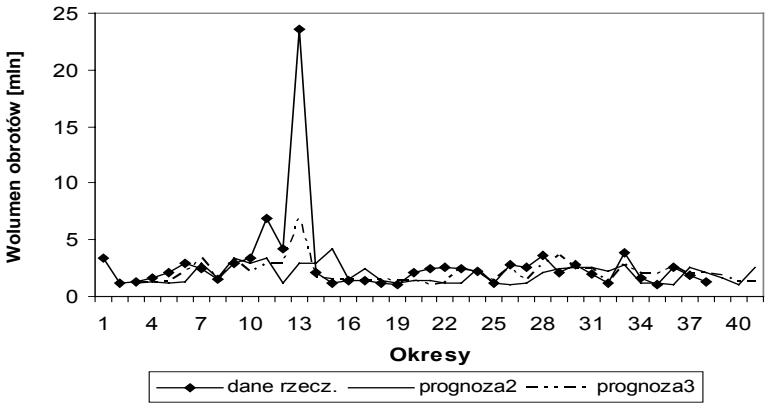


Rys. 2. Prognoza kursu zamknięcia TP S.A. z uśrednionymi obserwacjami nietypowymi
Ź r ó d ł o: Opracowanie własne.

Przyjrzyjmy się teraz graficznej ilustracji prognoz wolumenu obrotów dla tej samej spółki.



Rys. 3. Prognoza obrotów TP S.A. bez przekształceń na danych
Ź r ó d ł o: Opracowanie własne.



Rys. 4. Prognoza kursu zamknięcia TP S.A. z uśrednionymi obserwacjami nietypowymi
Ź r ó d ł o: Opracowanie własne.

Dla 13 obserwacji wystąpił wyjątkowo wysoki wolumen obrotów. Jak widać z rysunków 3 i 4, zastąpiliśmy go wartością mającą rząd wielkości odpowiadający pozostałym okresom. Błąd RMSPE, odpowiadający kolejnym prognozom, był równy: 38,82%, 44,17% i 35,6%. Zaznaczyła się tutaj cecha charakterystyczna dla wszystkich wyników: bardzo dobre prognozy otrzymywaliśmy dla kursów zamknięcia, natomiast w przypadku wszystkich walorów wystąpiło drastyczne pogorszenie rezultatów dla wolumenów obrotów. Wydaje się, że ma to związek ze skalą wahań przypadkowych,

która w drugim przypadku jest o wiele wyższa. Nie znaczy to oczywiście, że dla kursów nie występowały nietypowe obserwacje. Przykładowo, dla prezentowanych na rysunkach 1 i 2 notowań TP S.A. za nietypowe uznano dane: dziesiątą i trzynastą.

Analizując jakość prognoz dla szeregów po usunięciu obserwacji zaklasyfikowanych jako nietypowe, należy pamiętać, że RMSPE obliczyliśmy traktując przekształcony szereg jako dane rzeczywiste. Wykresy przedstawione na rysunkach 1–4 prezentują do porównań dane rzeczywiste bez transformacji.

W tabelach 1 i 2 znalazły się zbiorcze wyniki dla wszystkich walorów z podziałem na rodzaj prognozy.

Tabela 1. RMSPE dla kursów zamknięcia

Spółka	Dane bez wygładzenia [%]	Średnia ruchoma (k = 2) [%]	Średnia z sąsiednich [%]
WIG 20	1,51	1,53	1,36
BZWBK	2,48	2,30	2,31
KGHM	2,59	2,40	2,07
TP S.A.	1,89	1,79	1,78

Ź r ó d ł o: Opracowanie własne.

Tabela 2. RMSPE dla wolumenów

Spółka	Dane bez wygładzenia [%]	Średnia ruchoma (k = 2) [%]	Średnia z sąsiednich [%]
WIG 20	21,09	20,59	18,32
BZWBK	52,40	47,54	51,27
KGHM	32,69	32,14	32,14
TP S.A.	38,82	44,17	35,60

Ź r ó d ł o: Opracowanie własne.

Pierwszym, co rzuca się w oczy, jest dysproporcja jakości prognoz między wynikami uzyskanymi dla notowań kursów a wolumenów obrotów. Wcześniej wspomnieliśmy, że kursy zamknięcia charakteryzowały się znacznie niższymi wahaniami przypadkowymi, co przeniosło się na wartości RMSPE[11]. Nie zmienia to faktu, iż otrzymaliśmy w tym wypadku wyniki o bardzo zadowalającej jakości, ponieważ dla żadnej ze spółek nie przekroczono nawet 3%.

W przypadku wolumenów zmienność danych okazała się sprawiać poważne problemy i to nawet mimo odrzucenia kilku najbardziej odstających obserwacji.

Okazuje się, że dla obu analizowanych rodzajów notowań zastąpienie nietypowo zachowujących się danych prowadzi zazwyczaj do polepszenia wyników. Nie zawsze jednak jest to poprawa znacząca jak w przypadku wolumenu KGHM.

[11] Przypomnijmy, że miara ta podnosi resztę z prognozy *ex post* do kwadratu.

Wnioski

Przedstawione obliczenia prowadzą do dwóch podstawowych wniosków. Po pierwsze, usunięcie obserwacji nietypowej i zastąpienie jej przez wartość uśrednioną pozwala na poprawę jakości uzyskanych prognoz. Po drugie, średnia z obserwacji poprzedzającej i następnej dała nieco lepsze rezultaty.

Poprawa, choć zauważalna, nie jest jednak przesadnie duża. Można to zrozumieć w przypadku szeregów o małych wahaniach, dla których trudno poprawić prognozy uzyskane zaproponowaną metodą. Wciąż jednak problem stanowią duże wahania losowe. Otrzymane dla wolumenów obrotów wartości RMSPE pozostają dalekie od ideału. Oczywiście odpowiada za to sama natura zjawiska, lecz istnieje wciąż pole do popisu dla algorytmów genetycznych.

Niewątpliwie na uwagę zasługują rozmiary zbioru rozwiązań. Użyta metoda *de facto* sprawdza różne warianty, złożone z ułożonych w określony sposób obserwacji rzeczywistych. Dalsze badania, jak się wydaje, powinny iść w kierunku ulepszenia operatorów algorytmu oraz sposobów szybkiego znajdowania rozwiązań bliskich optymalnym.

Bibliografia

[1] CIEŚLAK M. (red.), *Prognozowanie gospodarcze. Metody i zastosowania*, PWN, Warszawa 2001.
[2] GAJDA J.B., *Prognozowanie i symulacje a decyzje gospodarcze*, C.H. Beck, 2001.
[3] GOLDBERG D., *Algorytmy genetyczne i ich zastosowania*, WNT, Warszawa 2001.
[4] KONARZEWSKA I., KARWACKI Z., *Planowanie i kontrola kosztów – wybrane problemy statystyczne*, *Zarządzanie organizacjami w świetle wyzwań XXI wieku – od teorii do praktyki*, Wydawnictwo Naukowe Wyższej Szkoły Kupieckiej, 2005, s. 445–459.
[5] KUCHARSKI A., *O pewnym zastosowaniu algorytmów genetycznych do prognozowania szeregów czasowych*, Prace Naukowe AE we Wrocławiu, Wrocław 2007, s. 143–153.
[6] KUCHARSKI A., *Wykorzystanie algorytmów genetycznych do krótkookresowych prognoz na giełdzie papierów wartościowych*, konferencja naukowa „Rynek Kapitałowy – skuteczne inwestowanie", Wydawnictwo Naukowe Uniwersytetu Szczecińskiego, Szczecin 2007, s. 135–145.
[7] MICHALEWICZ Z., *Algorytmy genetyczne + struktury danych = programy ewolucyjne*, WNT, Warszawa 1996.
[8] ZELIAŚ A., *Teoria prognozy*, PWE, Warszawa, 1997.

Genetic algorithms in stock forecasting – deleting outliers

This work presents a proposal of usage of genetic algorithm to short-term forecasting of price and volume quotations. Presented algorithm resembles the naive method with seasonality but a lag of observation used as predictor can change in order to achieve best adjustment of ex post prognosis to data.

The data were devoid of outliers with the help of hat matrix, taken from robust estimation.

The results confirmed the earlier assumptions and gave better ex post forecasts after removing outliers. Much better results were obtained for the prices, compared to those obtained for the volume, due to smaller in the case of prices, caused by smaller random fluctuations.

Keywords: *genetic algorithm, forecasting, time series, robust estimation*