PRACE NAUKOWE AKADEMII EKONOMICZNEJ WE WROCŁAWIU
Nr 1167 — 2007
Współczesne tendencje rozwojowe badań operacyjnych

Adam Kucharski
Uniwersytet Łódzki

O PEWNYM ZASTOSOWANIU ALGORYTMÓW GENETYCZNYCH DO PROGNOZOWANIA SZEREGÓW CZASOWYCH

1. Wstęp

Chyba nikogo nie trzeba przekonywać o przydatności wynikającej z wykorzystania prognoz we współczesnej rzeczywistości ekonomicznej. Obecnie znane metody można podzielić na cztery grupy [Gajda 2001]:

1) metody prognozowania na podstawie szeregów czasowych,
2) metody prognozowania na podstawie modeli ekonometrycznych,
3) metody analogowe,
4) metody heurystyczne.

W niniejszej pracy skoncentrowano się na pierwszej z wymienionych kategorii metod zwanych też niekiedy niestrukturalnymi. W jej obrębie występuje wiele konkretnych metod stosowanych w zależności od potrzeb i własności analizowanego szeregu. Pomimo tego zdarzają się sytuacje, kiedy okazuje się, iż w żaden sposób nie można uzyskać satysfakcjonujących prognoz. Wiąże się to najczęściej z wystąpieniem elementów dekompozycji szeregu czasowego. Jeżeli nie istnieje szansa na wykorzystanie którejś z pozostałych grup metod, sytuacja staje się nie do pozazdroszczenia. Dlatego warto poszukiwać innych rozwiązań, które mogłyby się sprawdzić w tego typu okolicznościach, na przykład sięgając po algorytm genetyczny.

Punktem wyjścia rozważań stały się metody naiwne. Przyczyna tego była dwojaka. Po pierwsze, mimo swej nazwy, potrafią one znaleźć zastosowanie, dając czasem wyniki nie gorsze od bardziej skomplikowanych podejść. Po drugie, prostota stosowania wydawała się wręcz stworzona do wykorzystania ich w ramach algorytmu genetycznego, choćby ze względu na konstrukcję chromosomów.

Postanowiono więc zbudować, a następnie przetestować algorytm genetyczny oparty na metodach naiwnych, który tworzyłby prognozy szeregów czasowych zawierających wahania przypadkowe i stały poziom zmiennej lub trend liniowy.

2. Zastosowany algorytm genetyczny

Algorytm, którego użyto w pracy, stanowi modyfikację klasycznego algorytmu genetycznego (KAG). Różnica ta jest najbardziej widoczna w przypadku budowy pojedynczego osobnika. Po pierwsze zamiast reprezentacji haploidalnej wykorzystano diploidalną. Po drugie zastosowano kodowanie rzeczywiste w miejsce binarnego. Wartości, jakie przyjmuje para genów, odpowiadają jednemu okresowi prognozy *ex post*. Pierwszy zawiera konkretną prognozę na dany okres, drugi zaś informację o okresie, z którego wywodzi się dana wartość.

Operatory krzyżowania i mutacji różnią się od tych znanych z KAG o tyle, o ile różni się również konstrukcja chromosomów. Tak więc w pierwszym przypadku losujemy punkt krzyżowania[1] wybranych wcześniej osobników, a następnie dokonujemy wymiany informacji między chromosomami w sposób znany m.in. z reprezentacji binarnej KAG. Oznacza to, że geny rodzica 1 znajdujące się za punktem krzyżowania były zastępowane genami rodzica 2, leżącymi również za punktem krzyżowania. Losowanie osobników do reprodukcji odbywa się metodą koła ruletki [Goldberg 1995].

Nieco więcej zmian zastosowano w przypadku mutacji. Wprawdzie nadal polega ona na zamianie miejscami dwóch wybranych genów, lecz teraz jej zasięg ogranicza parametr opisujący maksymalne opóźnienie przyjęte dla prognozy (szerzej o nim będzie mowa w dalszej części pracy).

Funkcję przystosowania stanowi wybrana miara błędów prognoz *ex post*. W trakcie obliczeń był to najpierw *mean absolute percentage error* – MAPE, a następnie *root mean square perentage error* – RMSPE. Oznacza to minimalizację rzeczonej funkcji, co z kolei wymusza sięgnięcie po przekształcenie funkcji kryterium. Tak się bowiem składa, że w KAG funkcja kryterium jest zawsze maksymalizowana, nam z kolei zależy na możliwie najniższej wartości średniego błędu prognoz *ex post*. Ponieważ oba wymienione mierniki przyjmują tylko wartości dodatnie, wartość funkcji przystosowania ($G(x)$) obliczano wg formuły [Michalewicz 1996]:

$$G(x) = -F(x)+C \tag{1}$$

gdzie: $F(x)$– wartość uśrednionego błędu prognoz *ex post*,
C – stała, określana *a priori* na wejściu.

Wartość C ma spore znaczenie, jeśli chodzi o zastosowaną funkcję przystosowania. Aby mieć gwarancję, że nie pojawią się dla niej wartości ujemne, przyjęto $C = 100$. Tak przekształcona funkcja była następnie maksymalizowana.

Sam przebieg algorytmu właściwie nie odbiega od standardowego podejścia. Po wygenerowaniu populacji początkowej, przez kolejne pokolenia dokonywana jest selekcja metodą ruletki, a następnie krzyżowanie i mutacja według zasad opisanych wcześniej.

[1] W pracy przeanalizowano przypadek jednego oraz dwóch punktów krzyżowania.

Podczas testowania algorytmu wprowadzono jeszcze jedną modyfikację. W danym pokoleniu zachowywano najlepszego osobnika. Jeżeli w następnym występowało pogorszenie wartości funkcji, zastępowano najgorszego obecnie osobnika zapamiętanym wcześniej.

Jako kryterium zatrzymania algorytmu przyjęto ustalaną na początku obliczeń liczbę pokoleń.

3. Metoda wyznaczania prognoz

Czas wreszcie przyjrzeć się sposobom wyznaczania prognoz *ex post* i *ex ante* użytym w postępowaniu. Przed ich wykonaniem należy przyjąć wartość parametru określającego maksymalne dopuszczalne opóźnienie prognozy *ex post* w danym okresie. Jest to wielkość ustalana *a priori* na początku obliczeń i nie ulegająca zmianie do ich końca. Przyjęto przy tym, że punktem startowym będzie prognoza wykonana na okres drugi. Z tego względu pierwsze predykcje niekoniecznie muszą sięgać w przeszłość na możliwie największą głębokość. Załóżmy, że szereg czasowy liczy sobie 5 obserwacji, przyjęte zaś opóźnienie (oznaczmy je przez *tm*) wynosi 3. W tabeli 3 zaprezentowano zbiór prognoz *ex post*, które mogą wystąpić w danym okresie.

Tabela 1. Zestaw przykładowych wartości prognoz *ex post*

Wartość rzeczywista	Zbiór prognoz *ex post*
y_1	
y_2	y_1
y_3	y_1, y_2
y_4	y_1, y_2, y_3
y_5	y_2, y_3, y_4

Źródło: opracowanie własne.

Jak wspomniano wcześniej, pojedynczego osobnika opisują dwa chromosomy. W pierwszym kolejne geny zawierają prognozowane wartości, w drugim informację o tym, z którego okresu w przeszłości pochodzi dana prognoza. Nietrudno więc się domyślić, w jaki sposób generowane są chromosomy populacji początkowej. Dla danego momentu losowano jedną spośród tworzących zbiór wartości. Powstają w ten sposób zestawy prognoz o opóźnieniu nie przekraczającym założonego z góry poziomu.

Analiza tabeli 1 wykazuje, że ostatnia dostępna wartość rzeczywista nie jest wykorzystywana w procesie wyznaczania prognoz *ex post*. Służy ona za to do obliczania prognoz *ex ante*. Warto w tym miejscu przypomnieć, z uwagi na występujące w jej przypadku podobieństwa do opisywanego w pracy postępowania, w jaki sposób wyznacza się prognozy metodą naiwną prostą. Otóż przyjmuje się w niej, iż prognoza na dany okres równa jest wartości zmiennej pochodzącej z okresu poprzedniego [*Prognozowanie...* 2001]. Nieco podobne założenie leży u podstaw metody naiwnej z sezonowością, tyle że w jej przypadku mamy do czynienia z wartościami pochodzącymi z okresu odległego w czasie o „sezon".

Mechanizm tworzenia prognoz *ex ante* w naszym przypadku przebiega dwuetapowo:

1. Najpierw wyznaczamy prognozę *ex post* na podstawie chromosomu wybranego z populacji osobnika, wykorzystując w tym celu wybraną miarę błędów prognozy.
2. Następnie (wykorzystując drugi chromosom) określamy wartość opóźnienia wskazującego, które wartości wezmą udział w tworzeniu prognoz *ex ante*.

Od tego miejsca zaczynają się rozbieżności w tworzeniu prognoz w zależności od dekompozycji szeregu czasowego. Otóż w przypadku szeregów o stałym poziomie nasze postępowanie staje się analogiczne do tego z metody naiwnej prostej (ewentualnie naiwnej z sezonowością), tyle że to na podstawie wyników algorytmu ustalamy korespondujący z prognozowanym okres oraz same wartości prognoz[2]. Ostatnia dostępna obserwacja rzeczywista będzie w tym momencie stanowić granicę tych prognoz.

Z kolei dla szeregów zawierających trend liniowy postąpimy podobnie jak w metodzie naiwnej z poprawką liniową. Konieczne jest więc obliczenie przyrostu, który w naszym przypadku tworzyć będą ostatnia prognoza *ex post* i wartość wskazana przez wyznaczony poziom opóźnienia. Tak obliczony przyrost służy do korekty, już w tradycyjny sposób, ostatniej dostępnej wartości rzeczywistej. Rzecz jasna, nie wystąpi tutaj ograniczenie horyzontu prognozy, lecz należy pamiętać, że zaleca się stosowanie metod naiwnych do predykcji krótkookresowej.

Niezależnie od rodzaju dekompozycji szeregu pojawia się kwestia określenia wartości opóźnienia wskazującej, z którego okresu pochodzić będą wartości używane do obliczania prognoz *ex ante*. Należy bowiem także pamiętać, iż stosowny chromosom może zawierać kilka różnych wariantów. W tej sytuacji autor proponuje skorzystać z mediany ewentualnie dominanty tych opóźnień.

3. Prezentacja wyników obliczeń

Postanowiono przetestować opisaną powyżej metodę na kilku sztucznie wygenerowanych szeregach o zadanych z góry własnościach i w tym celu wykorzystano następujące szeregi:

- o stałym poziomie i małych wahaniach przypadkowych,
- o stałym poziomie i dużych wahaniach przypadkowych,
- o małych wahaniach przypadkowych i trendzie liniowym,
- o stałym poziomie i nietypowo zachowujących się wartościach,
- z trendem i nietypowymi wartościami.

Dla pierwszych trzech szeregów przyjęto liczbę dwunastu obserwacji. Wykonując prognozy *ex post*, ograniczono się do dziesięciu obserwacji, aby możliwe było sprawdzenie również jakości prognoz *ex ante*. Szeregi te posłużyły także do określenia zestawu parametrów algorytmu genetycznego i ich wpływu na otrzymywane wyniki. Pozostałe dwa szeregi posłużyły jako sposób na weryfikację algorytmu w trudniejszych warunkach, w których nie da się stosować tradycyjnych metod.

[2] Wykorzystując wartości tworzące chromosom odpowiedzialny za obliczenie funkcji kryterium.

Jednym z rozpatrywanych czynników wpływających na wyniki była liczba punktów krzyżowania. Pod uwagę wzięto jeden oraz dwa punkty krzyżowania. W toku przeprowadzonych obliczeń okazało się, że wybór danego wariantu nie ma większego znaczenia dla wyników. Chodzi tu zarówno o czas potrzebny na wykonanie obliczeń, jak i wartość optymalizowanej funkcji. Okazało się bowiem, że przy odpowiednio dużej populacji różnica w wymianie informacji między jednym a dwoma punktami krzyżowania przestaje mieć znaczenie.

Wzrost liczby osobników w populacji pozwala na poprawę rezultatu. Podobnie rzecz ma się w razie powiększania się liczby pokoleń. W obu przypadkach wytłumaczenie jest podobne i ma związek ze sposobem tworzenia chromosomu. Poszczególne osobniki różnią się bowiem między sobą ułożeniem genów. Zwiększanie liczby osobników lub pokoleń daje w rezultacie potencjalnie większą liczbę rozpatrywanych wariantów prognoz. Zdecydowano się na przyjęcie populacji złożonej z 2000 osobników oraz 20 pokoleń.

W ramach prawdopodobieństwa krzyżowania postanowiono postawić na sporą wymianę informacji. Pod tym kątem analizowano wzrost tej wartości w czasie obliczeń. Wniosek z tego płynie taki, że lepsze warianty daje prawdopodobieństwo 0,3 do 0,5. W tym drugim przypadku czas, jakiego potrzebował algorytm na zakończenie postępowania, był nawet o 50% dłuższy. Wartość 0,3 uznana została za wystarczającą do wymiany informacji, jednocześnie zapewniając akceptowalny czas obliczeń.

Prawdopodobieństwo mutacji we wszystkich wariantach obliczeń przyjęto na tym samym poziomie, równym 0,1.

Za swego rodzaju parametr można również przyjąć rodzaj optymalizowanej funkcji. Przypomnijmy, że pod uwagę brano miary: MAPE i RMSPE. Wyniki wstępne, przeprowadzone dla populacji złożonej z 1000 osobników (przy liczbie pokoleń równej 20), wykazały, że MAPE stwarza pewne problemy. Kilkukrotne uruchomienie procedury dawało w pewnych sytuacjach osobniki o tej samej wartości funkcji przystosowania, lecz o różnym składzie. Nie stanowiłoby to jeszcze dużego problemu, wszak naszym celem jest otrzymywanie prognoz *ex ante*, lecz mediany (a również i dominanty) opóźnień w kolejnych podejściach przyjmowały różne wartości. Wyniki otrzymywane przy użyciu jako kryterium RMSPE okazały się bardziej stabilne. W końcowym rozrachunku zdecydowano się właśnie na to rozwiązanie[3], dodatkowo powiększając populację.

Jako pierwsze zaprezentowane zostaną wyniki otrzymane dla szeregu o stałym poziomie i małych wahaniach przypadkowych. Cały szereg liczył 12 obserwacji, przy czym prognozy *ex post* wyznaczono dla pierwszych 10 danych.

Do obliczenia prognoz *ex ante* wykorzystano medianę opóźnień okresów *ex post*. W tabeli 2 zamieszczono dodatkowo prognozy metodą naiwną prostą (w celu porównania znalazły się tam również błędy MPE).

[3] Innym wyjściem z powyższej sytuacji byłoby obliczenie dla spornych chromosomów dodatkowej miary błędu (np. mean percentage error – MPE) jako dodatkowego kryterium.

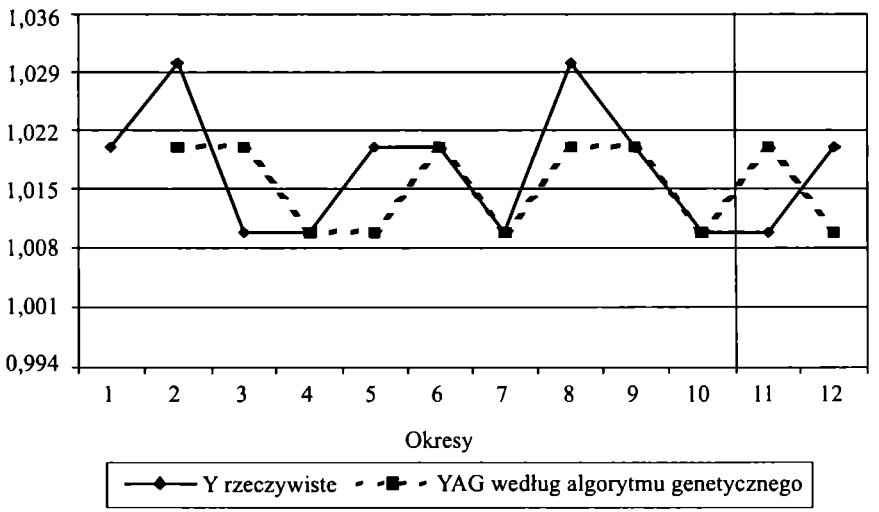
Tabela 2. Wyniki prognoz *ex ante* dla stałego poziomu zmiennej i małych wahań losowych (*tm* = 3)

	Prognozy AG	Metoda naiwna
Okres *t*+1	1,02	1,01
Okres *t*+2	1,01	1,01
RMSPE	0,0065	0,0118
MPE	0,0011	–0,012

Źródło: obliczenia własne.

Z kolei wykres na rys. 1 prezentuje całość prognoz (szereg oznaczony YAG) w porównaniu do danych rzeczywistych.

Przyjęte maksymalne dopuszczalne opóźnienie (*tm*) wyniosło 3 okresy. Mediana opóźnień równa się 2, co pozwala otrzymać prognozy na tyle właśnie okresów. Należy przypomnieć, iż metody naiwnej prostej nie zaleca się stosować dalej niż jeden okres poza próbę. Jakość prognoz *ex post* dla algorytmu genetycznego jest wyraźnie wyższa, prognozy *ex ante* zaś nieco się różnią.



Rys. 1. Dopasowanie prognoz do danych – stały poziom, małe wahania

Źródło: opracowanie własne.

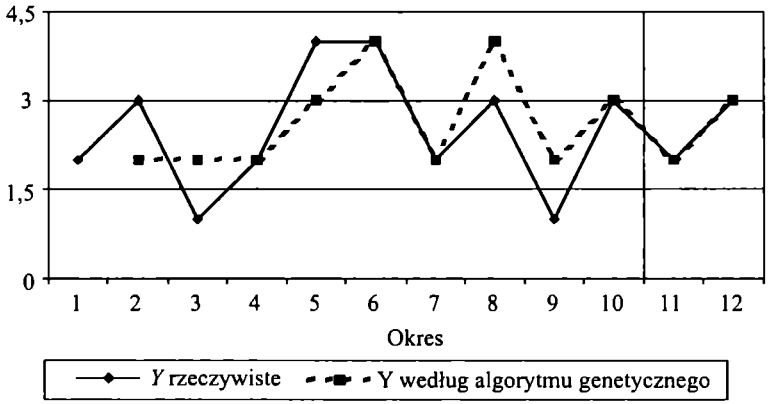
W trakcie obliczeń dla *tm* = 4 otrzymano nieznacznie lepszą (0,0056) wartość funkcji celu, jednak mediana w tym przypadku była taka sama, a co za tym idzie, takie same prognozy *ex ante*.

W następnej kolejności zaprezentowane zostaną wyniki dla szeregu o stałym poziomie zmiennej i większych niż poprzednio wahaniach przypadkowych. Należy zwrócić uwagę, iż metoda naiwna prosta słabo sprawdza się w tego typu szeregach, co potwierdza wartość błędu RMSPE.

Tabela 3. Wyniki prognoz *ex ante* dla stałego poziomu zmiennej i dużych wahań losowych ($tm = 3$)

	Prognozy AG	Metoda naiwna
Okres $t+1$	2	3
Okres $t+2$	3	3
RMSPE	0,5038	1,0628
MPE	–0,1944	–0,2963

Źródło: obliczenia własne.



Rys. 2. Dopasowanie prognoz do danych – stały poziom, duże wahania

Źródło: opracowanie własne.

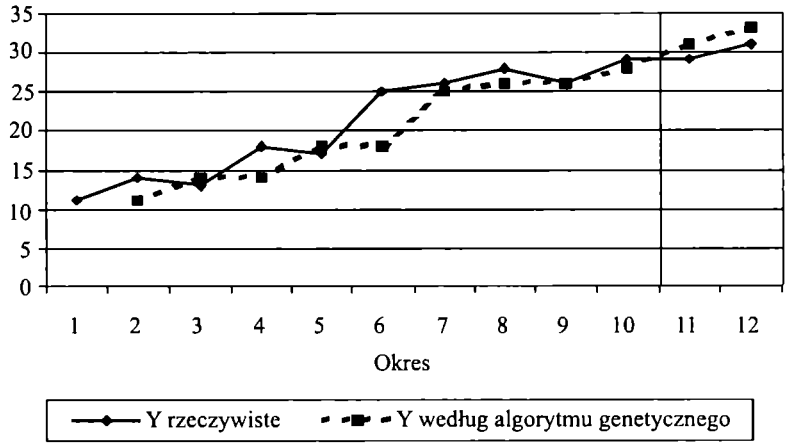
Przyjęte opóźnienie *ex post* wyniosło 3 okresy. Z punktu widzenia przyjętego kryterium dalsze sięganie w przeszłość dało w rezultacie takie same wyniki. Mediana (jak w poprzednim przypadku) wyniosła 2 okresy.

Kolejny szereg da się zdekomponować na rosnący trend liniowy oraz małe wahania przypadkowe. RMSPE dla algorytmu genetycznego jest dwa razy niższe od przewidzianej dla takich szeregów metody naiwnej z poprawką liniową. Rozpatrując następujące warianty opóźnień: 3, 4 i 5 okresów, otrzymywano w rezultacie tę samą wartość funkcji przystosowania oraz mediany (1 okres).

Tabela 4. Wyniki prognoz *ex ante* dla trendu liniowego i małych wahań losowych ($tm = 3$)

	Prognozy AG	Metoda naiwna
Okres $t+1$	31	32
Okres $t+2$	33	35
RMSPE	0,1456	0,2709
MPE	0,0801	–0,0228

Źródło: obliczenia własne.



Rys. 3. Dopasowanie prognoz do danych – trend liniowy, małe wahania

Źródło: opracowanie własne.

Wartość mediany oznacza, że sposób tworzenia prognoz w obu metodach jest podobny. Różnica bierze się stąd, iż w przypadku algorytmu genetycznego opieramy się na prognozach *ex post,* obliczając poprawkę liniową ostatniej dostępnej obserwacji. Powoduje to mniejszą wartość przyrostu korygującego ostatnią dostępną obserwację.

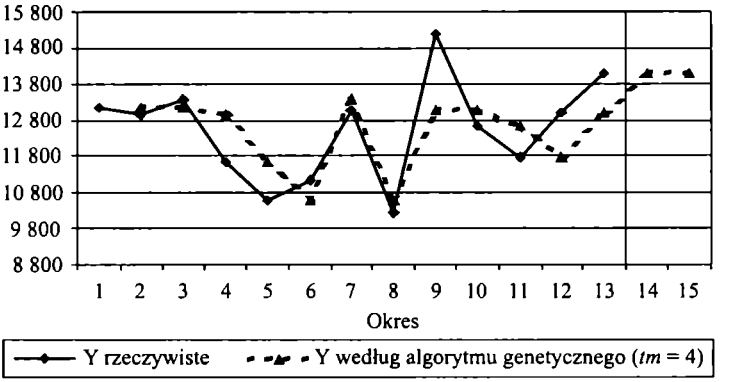
Dotychczasowe przykłady przedstawiały szeregi charakteryzujące się „podręcznikowymi" własnościami. Poniższe dwa przypadki stanowią już trudniejszy orzech do zgryzienia. W pierwszym z nich obok stałego poziomu zmiennej i wahań przypadkowych występują nietypowo zachowujące się obserwacje. Tego typu szereg sprawia sporo kłopotu metodom prognozowania niestrukturalnego. Ponieważ nie spełniają one założeń stawianych metodom naiwnym, zdecydowano ograniczyć się do prognoz jedynie przy użyciu algorytmu genetycznego.

Dla tych samych co poprzednio parametrów algorytmu genetycznego okazało się, iż większego znaczenia zaczyna nabierać (z punktu widzenia kryterium optymalizacji) kwestia opóźnienia w prognozach *ex post.*

Tabela 5. Wyniki prognoz *ex ante* dla stałego poziomu i nietypowych wartości

	Prognozy AG ($tm = 3$)	Prognozy AG ($tm = 4$)
Okres t+1	12628,3	14095,1
Okres t+2	14095,1	14095,1
RMSPE	0,0806	0,0755
Me	2	1,5

Źródło: obliczenia własne.



Rys. 4. Dopasowanie prognoz do danych – stały poziom, nietypowe wahania

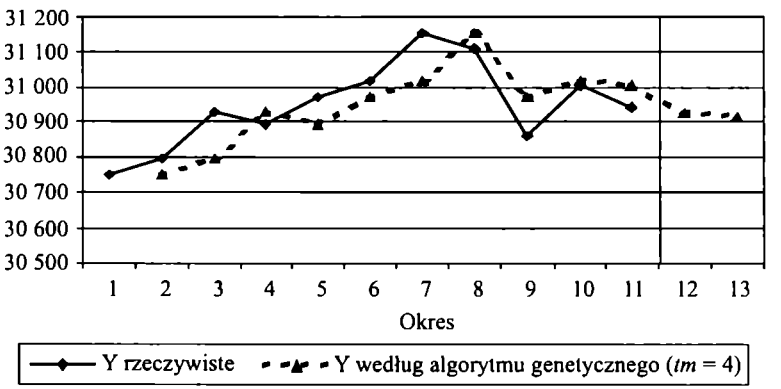
Źródło: opracowanie własne.

Tabela 6. Wyniki prognoz *ex ante* dla trendu i nietypowych wartości

	Prognozy AG (*tm* = 3)	Prognozy AG (*tm* = 4)
Okres *t*+1	30835	30927
Okres *t*+2	30729	30913
RMSPE	0,003	0,0026
Me	1	1

Źródło: obliczenia własne.

Tym razem zrezygnowano ze skracania szeregu i prognozy wykonano dla całej dostępnej informacji. Mimo stosunkowo niewielkiej różnicy błędów RMSPE występującej między *tm* = 3 a *tm* = 4, należy zwrócić uwagę, iż pojawiła się różnica w wartościach mediany. Powoduje ona różnice prognoz *ex ante*. Wobec powyższego wybrano model o niższym wariancie funkcji celu.



Rys. 5. Dopasowanie prognoz do danych – trend, nietypowe wahania

Źródło: opracowanie własne.

Kolejny szereg różni się od poprzedniego tym, iż pojawia się w nim trend. Ponownie wystąpiły różnice optymalizowanej funkcji dla różnych opóźnień. W tym przypadku mediany dla obu wariantów są jednakowe.

Wykres na rys. 5 prezentuje dopasowanie dla lepszego wariantu. Jego analiza prowadzi do wniosku, że użycie algorytmu genetycznego dla prognoz *ex post* pozwala zachować w niej te elementy dekompozycji, które zaobserwowano dla danych rzeczywistych.

4. Wnioski

Pierwszą rzeczą, jaka zwraca uwagę po serii obliczeń, jest różnica w jakości prognoz *ex post* między algorytmem genetycznym a metodami naiwnymi. Praktycznie zawsze prezentowana w pracy metoda okazywała się lepsza. Nadal jednak mamy do czynienia z kolejnym wariantem mechanicznego generowania prognoz. Dlatego ważne stają się prognozy *ex ante*. Te zaś okazują się być nie gorsze (a niekiedy lepsze) w razie użycia algorytmu genetycznego. Wydaje się więc, że można sięgnąć po tę metodę bez większego ryzyka. Co ciekawsze, sprawdziła się ona również w przypadku szeregu o stałym poziomie zmiennej i wahaniach sezonowych, mimo iż na początku nie czyniono takiego założenia.

Wciąż pozostają jednak kwestie, które należy rozwiązać. Pierwsza z nich to rozmiary zbioru, w którym poszukujemy rozwiązania. Wyprowadzona formuła na liczbę możliwych w algorytmie szeregów *ex post* (a tym samym na maksymalną liczbę różnych osobników populacji) jest następująca:

$$LO = tm^{n-tm}(tm-1)! \tag{2}$$

gdzie: LO – liczba osobników,
tm – maksymalne dopuszczalne opóźnienie,
n – liczba obserwacji szeregu czasowego.

Tabela 7 zawiera wartości wyznaczone z wzoru (2) dla przykładowych liczebności szeregu oraz przyjętych opóźnień.

Tabela 7. Maksymalna liczba osobników w populacji w zastosowanym algorytmie

Liczba obserwacji	*tm* = 3	*tm* = 4	*tm* = 5
10	4 374	24 576	75 000
11	13 122	98 304	375 000
12	39 366	393 216	1 875 000
13	118 098	1 572 864	9 375 000

Źródło: obliczenia własne.

Na podstawie tab. 7 da się ocenić przydatność algorytmu w zależności od przyjętych parametrów. Oznacza to na przykład, że dla dłuższych szeregów selekcja metodą ruletki może okazać się nie dość wydajna, gdyż wymagać będzie bardzo dużych populacji, aby zagwarantować dobre wyniki.

Oprócz zastosowania innych metod selekcji warto rozważyć także zmianę mechanizmu krzyżowania czy użycie jeszcze innych operatorów poprawiających wydajność oraz wyniki prognoz.

Literatura

Gajda J.B., *Prognozowanie i symulacje a decyzje gospodarcze*, C.H. Beck 2001.
Goldberg D., *Algorytmy genetyczne i ich zastosowania*, WNT, Warszawa 1995.
Michalewicz Z., *Algorytmy genetyczne + struktury danych = programy ewolucyjne*, WNT, Warszawa 1996.
Prognozowanie gospodarcze. Metody i zastosowania, red. M. Cieślak, PWN, Warszawa 2001.
Zeliaś A., *Teoria prognozy*, PWE, Warszawa 1997.

ON THE USE OF GENETIC ALGORITHMS TO TIME SERIES FORECASTING

Summary

This paper presents how genetic algorithms can make forecasts of pretty complicated time series using simple naive methods. Moreover, it is possible to achieve adjustement of some kind for different types of time series. All this can be done thanks to flexibility of genetic algorithms. There had been carried out modifications in classic genetic algorithm. Some operators, like crossing or mutation, had been changed in order to adapt them to a certain parameter called *tm*. *tm* expresses maximum allowable lag of observation making current ex post forecast. This parameter, included in algorithm, had become the basis of presented prediction method.