

Ewa Chodakowska
Politechnika Białostocka

APLIKACJA WIELOWYMIAROWYCH MODELI ARIMA W PAKIECIE SCA

1. Wstęp

Modele MARIMA są jedną z klas modeli szeregów czasowych, która dzięki swoim własnościom może znaleźć praktyczne zastosowanie w wielu dziedzinach. Modele te nie są jednak często stosowane. W większości popularnych programów komputerowych do statystycznej analizy danych nie ma zaimplementowanych algorytmów identyfikacji modeli MARIMA oraz estymacji ich parametrów. SCA Statistical System jest jednym z nielicznych, który oferuje takie możliwości. W artykule przedstawiono metodykę budowy modeli MARIMA z wykorzystaniem programu SCA. Na przykładzie analizy szeregu czasowego wartości sprzedaży przedyskutowano kolejne etapy tworzenia modeli: identyfikację, estymację, diagnostykę i prognozowanie.

2. Postać modelu MARIMA

Modele MARIMA, czyli scałkowane wielowymiarowe modele autoregresji i średniej ruchomej (ang. *multivariate autoregressive integrated moving average*), pozwalają opisać relację pomiędzy jedną zmienną a jej przeszłymi wartościami oraz obecnymi i przeszłymi wartościami innych zmiennych. Są odpowiednie w sytuacji, gdy wartości analizowanego szeregu czasowego zależą od kilku różnych czynników, przy czym reakcja na te czynniki jest opóźniona w czasie, jednocześnie w szeregu występuje znacząca autokorelacja [DeLurgio 1998].

Modele MARIMA w literaturze są określane także jako modele funkcji przenoszenia [Box, Jenkins 1983; Liu, Hudak 2000] lub modele regresji dynamicznej

[Makridakis, Wheelwright, Hyndman 1998]. Pod nazwą MARIMA pojawiają się też modele VARIMA (ang. *vector ARIMA*) [Makridaki, Wheelwright, Hyndman 1998]. Niniejsza publikacja dotyczy modeli, w których jedna endogeniczna zmienna objaśniana jest funkcją swoich opóźnionych wartości oraz obecnych i opóźnionych wartości egzogenicznych zmiennych objaśniających. W tym sensie modele MARIMA można traktować jako rozszerzenie modeli wielowymiarowej regresji i modeli ARIMA [Liu, Hudak 2000].

Model MARIMA dla dwóch stacjonarnych szeregów czasowych Y i X składa się z funkcji przenoszenia TF (ang. *transfer function*) i modelu szumu N_t (ang. *noise model*) [DeLurgio 1998]. Można go zapisać równaniem [Liu, Hudak 2000]:

$$Y_t = C + \frac{\omega(B)}{\delta(B)} X_t + N_t, \quad (1)$$

gdzie: $\delta(B) = 1 - \delta_1 B - \delta_2 B^2 - \dots - \delta_r B^r$,

$$\omega(B) = (\omega_0 - \omega_1 B - \omega_2 B^2 - \dots - \omega_s B^s) B^b,$$

Y_t – zmienna objaśniana,

X_t – zmienna objaśniająca,

B – operator przesunięcia wstecz,

C – stała,

N_t – model szumu; przy czym jest to model $ARIMA(p,d,q)(P,D,Q)$ dla zmiennej Y zawierający łączny efekt innych czynników oddziałujących na zmienną prognozowaną.

Model MARIMA można przedstawić także w postaci zlinearyzowanej, korzystając z operatora $v(B)$ [Liu, Hudak 2000]:

$$Y_t = C + v(B) X_t + N_t, \quad (2)$$

gdzie: $v(B) = \frac{\omega(B)}{\delta(B)} = v_0 + v_1 B + v_2 B^2 + \dots + v_k B^k$.

Rozszerzeniem modelu na wiele zmiennych objaśniających jest model [3]:

$$Y_t = C + \frac{\omega_1(B)}{\delta_1(B)} X_{1t} + \frac{\omega_2(B)}{\delta_2(B)} X_{2t} + \dots + \frac{\omega_m(B)}{\delta_m(B)} X_{mt} + N_t. \quad (3)$$

Na podstawie przedstawionych wzorów można zauważyć, że kształt funkcji przenoszenia jest zdeterminowany przez wartości trzech parametrów: r , s , i b lub w postaci liniowej przez wybór właściwych v_i . Wartość b wskazuje opóźnienie reakcji wartości zmiennej objaśnianej na zmianę wartości zmiennej objaśniającej. Liczba parametrów licznika s opisuje początkową reakcję na wejście. Liczba parametrów mianownika r odpowiada za sposób zanikania efektu wywołanego przez zmienną objaśniającą. Wagi v_i mierzą wpływ wartości zmiennej wejściowej z okresu $t - i$ na zmienną wyjściową w okresie t [Liu, Hudak 2000].

3. Modele MARIMA – zastosowanie

W związku z omówionymi własnościami modele MARIMA są dosyć uniwersalnym narzędziem modelowania i prognozowania szeregów czasowych. Mogą znaleźć praktyczne zastosowania w dziedzinach takich, jak np.: marketing, energetyka, ekonomia, hydrologia, meteorologia.

Modele te nadają się np. do analizowania wielkości, podlegających wpływom tzw. efektu dni handlowych (lub efektu dni świątecznych). Problem ten dotyczy takich obserwacji, jak np.: sprzedaż, liczba rozmów telefonicznych czy natężenie ruchu. Bez rozważenia wpływu liczby dni wolnych w miesiącu, świąt, czy innych wydarzeń, powodujących zmianę zachowań konsumentów (klientów), można w interpretacji danych przedstawionych jako skumulowane wartości miesięczne wysnuć błędne wnioski.

Wybrane przykłady wykorzystania modeli MARIMA zawarto w tab. 1.

Tabela 1. Wybrane możliwości zastosowania modeli ARIMA

Dziedzina	Przykład
Marketing	$\text{sprzedaż} = f(\text{reklama, cena, pora roku})$
Energetyka	$\text{zużycie energii} = f(\text{temperatura, opady atmosferyczne, siła wiatru})$
Ekonomia	$\text{popyt} = f(\text{cena, cena konkurentów, wydatki na reklamę})$
Hydrologia	$\text{powódź} = f(\text{opady deszczu, nasycenie gleby, poziom dopływów})$
Finanse	$\text{indeks giełdowy} = f(\text{oprocentowanie długoterminowych rządowych obligacji, podstawowe wskaźniki gospodarcze})$

Źródło: [DeLurgio 1998, s. 534].

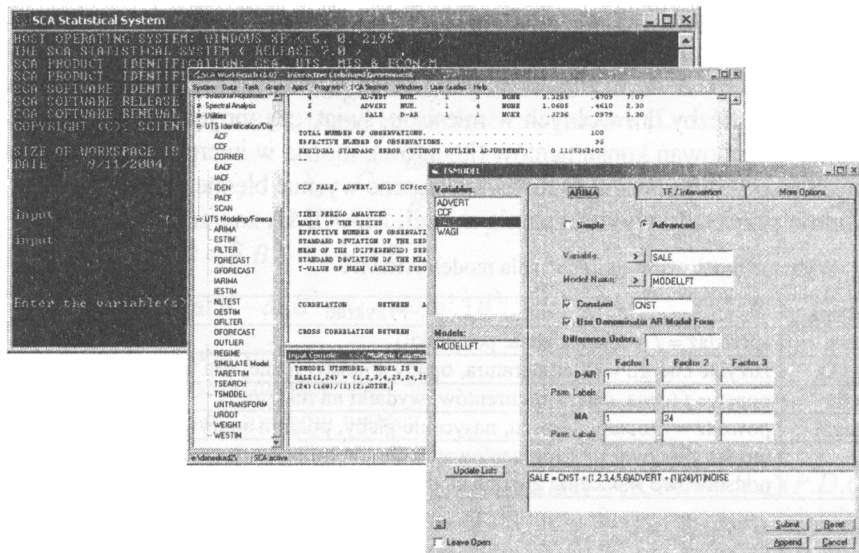
Pewne ograniczenie stosowania modeli wynika z tego, że wymagają one jednokierunkowej zależności pomiędzy zmienną objaśnianą a zmienną objaśniającą (zmiennymi objaśniającymi) [DeLurgio 1998; 3 Liu, Hudak 2000]. W procesach przemysłowych warunk ten jest stosunkowo łatwy do spełnienia. W danych ekonomicznych czasami trudno jest jednoznacznie określić, czy dana zmienna jest na pewno egzogeniczna, jednak założenie to nie wyklucza całkowicie możliwości otrzymania efektywnych modeli MARIMA również wówczas, gdy relacja między wielkościami jest dwukierunkowa [Liu, Hudak 2000].

Kolejnym warunkiem jest stabilność badanego systemu. Sprowadza się to do wymogu, aby zmienne zarówno objaśniające, jak i objaśniane były stacjonarne, a liczba wag v_i skończona. Stacjonarność szeregów można uzyskać, wykorzystując te same operacje co w modelach ARIMA: różnicowanie i transformowanie. Zmienna wejściowa powinna być też niezależna od szumu opisanego modelem N_t [Makridaki, Wheelwright, Hyndman 1998].

4. SCA Statistical System

W większości popularnych programów komputerowych do statystycznej analizy danych, takich jak: MATLAB, SPSS, PcGive, STATISTICA, Mathematica,

System został opracowany przez L. Liu; przy jego tworzeniu współpracowali światowej sławy eksperci w dziedzinie analizy szeregów czasowych – G.E.P. Box, G.C. Tiao, M.E. Muller [Liu, Hudak 2000].



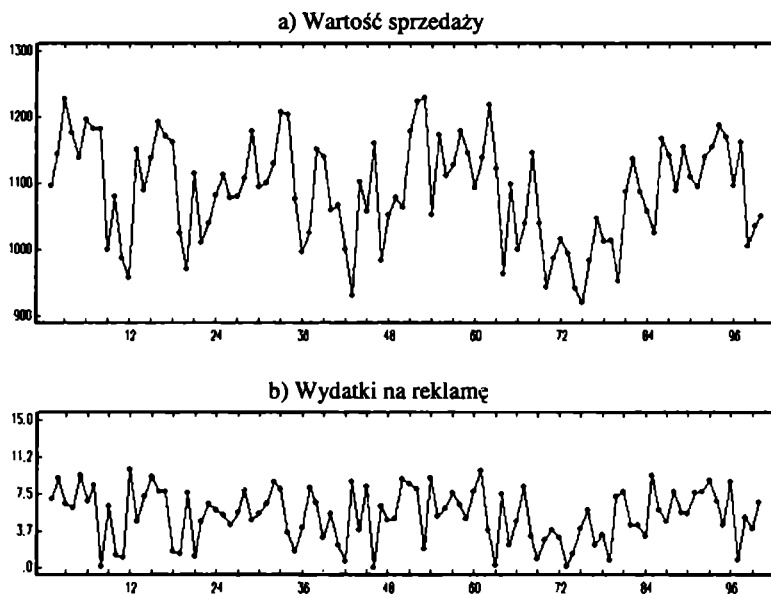
Rys. 1. Program SCA Statistical System:
program, SCA Workbench, okno dialogowe funkcji przenoszenia
Źródło: opracowanie własne.

Niestety program nie jest łatwy w obsłudze. Korzystanie z niego wymaga znajomości języka poleceń SCA. Dostępny graficzny interfejs (SCA Workbench) i wbudowane okna dialogowe tylko w pewnym stopniu ułatwiają dostęp do systemu (rys. 1). Program ma też ubogi i mało intuicyjny system graficznej prezentacji danych, a ograniczona do 60 znaków ścieżka dostępu do plików może zaskakiwać, szczególnie mniej wprawionych użytkowników komputerów.

5. Przykład budowy modelu

W celu zaprezentowania możliwości programu wykonano analizę danych opisujących wartość sprzedaży tarcicy, zrealizowaną przez przedsiębiorstwo budow-

lane Building Materials Company, oraz wydatki na reklamę w jednostkach umownych (j.u.) w okresie 100 kolejnych miesięcy. Szereg wartości sprzedaży jest procesem autoregresji, szereg wydatków na reklamę można zaś traktować jako biały szum (rys. 2). Dane zaczerpnięto z książki [DeLurgio 1998]. Taki wybór danych miał na celu skonfrontowanie wyników podręcznikowych z wynikami uzyskanymi za pomocą systemu SCA.



Rys. 2. Wykresy a) wartości zrealizowanej sprzedaży, b) wartości wydatków na reklamę
Źródło: opracowanie własne przy wykorzystaniu programu SCA.

Budowa modelu MARIMA, analogicznie do jednowymiarowego ARIMA, obejmuje cztery fazy upowszechnionych przez Boxa oraz Jenkinsa, tj. identyfikację, estymację, ocenę jakości (diagnozę) oraz prognozowanie [Box, Jenkins 1983], niemniej jednak omawiany proces jest bardziej skomplikowany.

Identyfikację modelu MARIMA za pomocą programu SCA można przeprowadzić na dwa sposoby. Pierwszy polega na określeniu funkcji przenoszenia na podstawie analizy przebiegu funkcji korelacji krzyżowej (CCF – ang. *cross correlation function*). Wartości CCF oblicza się dla reszt uprzednio zbudowanego modelu ARIMA dla zmiennej objaśniającej oraz dla reszt identycznego modelu, ale dla zmiennej objaśnianej. Model szumu określa się po uwzględnieniu reszt modelu funkcji przenoszenia.

Metodę tę opisał i zaimplementował DeLurgio, który otrzymał następujący model [DeLurgio 1998]:

$$Y_t = 872,63 + \frac{24,92B}{1 - 0,39B} X_t + \frac{1}{1 - 0,32B} e_t, \quad (4)$$

gdzie: Y_t – wartość sprzedaży,

X_t – wydatki na reklamę.

Parametry modelu estymowane w systemie SCA z wykorzystaniem algorytmu największej wiarygodności według Hilmera i Tao [Liu, Hudak 2000] różnią się nieznacznie:

$$Y_t = 873,54 + \frac{24,86B}{1 - 0,38B} X_t + \frac{1}{1 - 0,31B} e_t. \quad (5)$$

W rzeczywistych danych trudno jest czasami jednoznacznie określić kształt funkcji przenoszenia na podstawie funkcji CCF. Co więcej, opisana metoda zastosowana dla więcej niż jednej zmiennej objaśniającej może być kłopotliwa, gdyż wymaga oszacowania modeli ARIMA i MARIMA oddzielnie dla każdej zmiennej [DeLurgio 1998].

Znacznie lepsze jest skorzystanie z metody identyfikacji opartej na liniowej postaci funkcji przenoszenia (LTF – ang. *linear transfer function*). Polega ona na wyborze dostatecznie dużej liczby współczynników funkcji przenoszenia (wag odpowiedzi impulsowej) i wskazaniu wstępnej postaci modelu szumu. Następnie tak zbudowany model jest poddawany automatycznemu, iteracyjnemu procesowi estymacji i redukcji, podczas którego odrzuca się nieistotne współczynniki. Reszty końcowego modelu służą do identyfikacji ostatecznej postaci N_t . To podejście do identyfikacji usprawnia wybór modelu i nie wymaga żmudnej analizy funkcji CCF. Po ich wykorzystaniu otrzymano następujący model dla analizowanego szeregu:

$$Y_t = 881,05 + (24,83B + 9,78B^2 + 3,32B^3 + 1,06B^4) X_t + \frac{1}{1 - 0,32B} e_t. \quad (6)$$

Model opracowany za pomocą liniowej funkcji przenoszenia ma więcej parametrów, a wkład pracy i czas potrzebny do jego otrzymania jest znacznie mniejszy. Analizy wartości współczynnika determinacji R^2 oraz pierwiastka ze średniego kwadratowego błędu – RMSE – nie wskazują na znaczącą różnicę w jakości dopasowania modeli uzyskanych dwoma opisanymi metodami identyfikacji (tab. 2). Dalszemu sprawdzaniu diagnostycznemu poddano model z liniową funkcją przenoszenia (LTF).

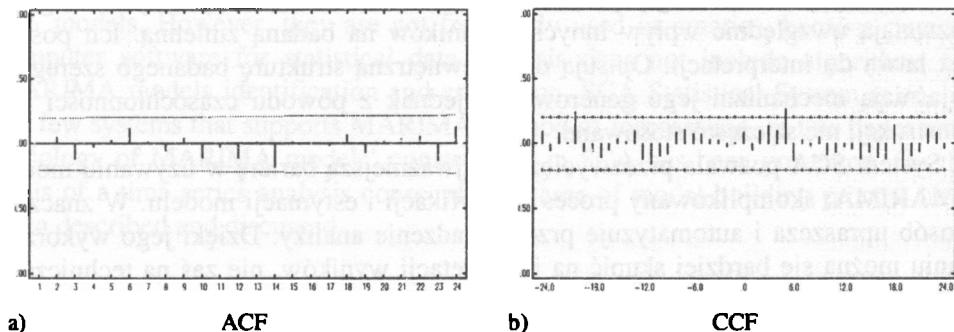
Tabela 2. Dopasowanie modeli MARIMA

Wskaźnik dopasowania	Metoda identyfikacji		
	model DeLurgio	analiza CCF	LTF
R^2	0,975	0,976	0,976
RMSE	12,114	11,842	11,847

Źródło: opracowanie własne.

Przed wszystkim modele MARIMA powinny mieć wszystkie parametry istotne [DeLurgio 1998; Liu, Hudak 2000]. W systemie SCA weryfikacja istotności odbywa się automatycznie i nieistotne współczynniki mogą być odrzucane na etapie estymacji na podstawie statystyki t . Dodatkowo reszty modelu powinny być białym szumem, co najczęściej sprawdza się przez zbadanie, czy wartości funkcji autokorelacji reszt różnią się od autokorelacji białego szumu [DeLurgio 1998; Liu, Hudak 2000]. W badanych danych ocena wizualna korelogramu reszt modelu (6) (rys. 3a)

oraz statystyka Q nie dają podstaw do odrzucenia hipotezy o losowości reszt modelu. W poprawnym modelu MARIMA nie powinna też występować korelacja pomiędzy stacjonarną zmienną objaśniającą a resztami [DeLurgio 1998; Liu, Hudak 2000]. W celu sprawdzenia tego warunku policzono wartości CCF pomiędzy szeregiem wydatków na reklamę a resztami modelu MARIMA. Wynik analizy, zilustrowany na rys. 3b, nie wskazuje na występowanie zależności.

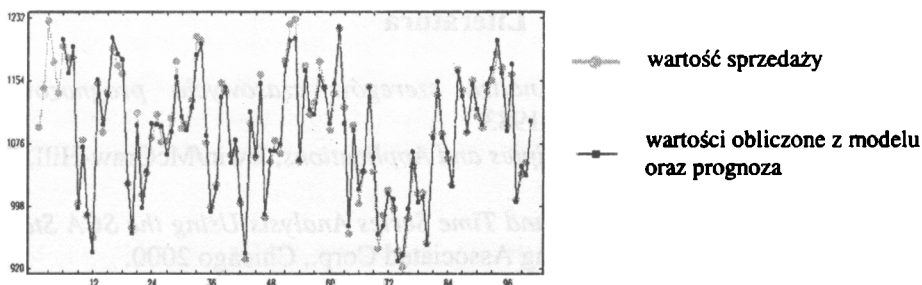


Rys. 3. Funkcja: a) autokorelacji reszt modelu ACF, b) korelacji krzyżowej CCF
Źródło: opracowanie własne przy wykorzystaniu programu SCA.

System SCA szybko i sprawnie wykonuje numeryczną analizę danych, jednak nie należy na pewno do najlepszych systemów pod względem możliwości graficznej prezentacji danych. Przeprowadzenie analizy wizualnej w systemie SCA wymaga albo korzystania z rysunków wykonanych w trybie tekstowym, albo też zapamiętania wyników analiz w plikach i późniejszego narysowania wykresów za pomocą modułu SCAGRAF.

Diagnostykę numeryczną warto też uzupełnić o dyskusję nad merytorycznym sensem modelu [DeLurgio 1998]. W badanym przypadku nie widać sprzeczności między racjonalnym przewidywaniem a postacią modelu: malejące wagi funkcji przenoszenia wskazują, iż wraz z upływem czasu wpływ reklamy na sprzedaż słabnie.

Pozytywne wyniki sprawdzania diagnostycznego pozwalają na wykorzystanie oszacowanego modelu do obliczenia prognozy. Na rysunku 4 pokazano rzeczywiste wartości, obliczone z modelu oraz prognozę jednokrokową.



Rys. 4. Rzeczywista realizacja zmiennej, wartości obliczone z modelu (6) oraz prognoza
Źródło: opracowanie własne przy wykorzystaniu programu SCA.

Opracowanie prognozy sprzedaży na dłuższy czas wymagałoby najpierw oszacowania przewidywanych wydatków na reklamę.

6. Podsumowanie

Modele MARIMA mogą sprawdzić się w analizowaniu różnych zjawisk, gdyż pozwalają uwzględnić wpływ innych czynników na badaną zmienną. Ich postać jest łatwa do interpretacji. Opisują one wewnętrzną strukturę badanego szeregu i objaśniają mechanizm jego generowania, jednak z powodu czasochłonności ich konstrukcji nie są często stosowane.

System SCA pozwala przezwyciężyć najważniejszą barierę w używaniu modeli MARIMA: skomplikowany proces identyfikacji i estymacji modelu. W znaczny sposób upraszcza i automatyzuje przeprowadzenie analizy. Dzięki jego wykorzystaniu można się bardziej skupić na interpretacji wyników, nie zaś na technicznej stronie analizy.

Zdaniem autorki, SCA jest wygodnym i jednocześnie bardzo silnym narzędziem (także w porównaniu z programem SAS). SAS System jest dużą i bardzo rozbudowaną aplikacją do wszechstronnej analizy danych. Budowa modeli szeregów czasowych klasy MARIMA w tym programie wymaga albo pisania własnych procedur w modelu SAS/ETS, albo korzystania z graficznego interfejsu Time Series Forecasting System (TSFS), jednak TSFS pozwala na estymację i diagnostykę jedynie najprostszych modeli.

SCA jest przeznaczony do budowy modeli szeregów czasowych. Pozwala na zapis modeli w postaci zarówno symbolicznej, jak i w postaci matematycznych równań. Wprowadzanie procedur ułatwiają okienka dialogowe funkcji.

Istotną zaletą programu SCA jest przede wszystkim możliwość automatycznej budowy modelu z wykorzystaniem liniowej funkcji przenoszenia. Doświadczenia autorki w pracy z programem SCA wskazują, że jest to dogodne narzędzie do analizy rozbudowanych modeli szeregów czasowych.

Literatura

- Box G.E.P., Jenkins G.M., *Analiza szeregów czasowych: prognozowanie i sterowanie*, PWN, Warszawa 1983.
- DeLurgio S.A., *Forecasting Principles and Applications*, Irwin/McGraw-Hill, Boston 1998.
- Liu L., Hudak G.B., *Forecasting and Time Series Analysis Using the SCA Statistical System*, Scientific Computing Associated Corp., Chicago 2000.
- Makridakis S., Wheelwright S.C., Hyndman R.J., *Forecasting. Methods and Applications*, John Wiley & Sons Inc., New York 1998.

AN APPLICATION OF MULTIVARIATE ARIMA MODELS IN SCA STATISTICAL SYSTEM

Summary

MARIMA models can be used in a variety of applications. This class of models effectively connects the basic concepts of the regression model with these of ARIMA models. However, they are not frequently used in practice, because popular computer software for statistical data analysis does not include algorithms for MARIMA models identification and estimation. SCA Statistical System is one of the few systems that supports MARIMA method of forecast-ing. In this paper methodology of MARIMA models' construction in SCA has been presented. On the basis of a time series analysis consecutive phases of model-building process have been described and discussed.