predictive SOLUTIONS

Co trzeba wiedzieć korzystając z modelu ARIMA i które parametry są kluczowe?

PROGNOZOWANIE

MARTA PŁONKA
Predictive Solutions

W trzecim już artykule dotyczącym szeregów czasowych przyjrzymy się modelom ARIMA. Dzisiaj skupimy się na metodzie doboru odpowiedniego modelu do naszego szeregu oraz postawimy prostą prognozę w poznanym w poprzednim artykule kreatorze. Przejdziemy krok po kroku przez proces, który w kolejnych artykułach pozostawimy do zrobienia programowi. Pozwoli nam to zrozumieć etap budowania modeli oraz raportowane wartości w tabelach wynikowych. Dzięki temu będziemy mogli w dalszej części serii skupić się na praktycznym wykorzystaniu modeli ARIMA.

Zacznijmy może od odszyfrowania nazwy modelu. Z języka angielskiego modele ARIMA biorą swój skrót od *Autoregressive integrated moving average model*, czyli po polsku autoregresyjny zintegrowany model średniej ruchomej. Zarówno nazwa jak i model składa się z trzech elementów:

- procesu autoregresyjnego AR (ang. *autoregressive*)
- procesu średniej ruchomej MA (ang. *moving averages*)
- stopnia integracji I (ang. *integrated*).

Czym jest proces autoregresyjny?

Proces autoregresyjny bazuje na podstawie znanego nam już z poprzednich artykułów pojęcia autoregresji. Proces autoregresyjny to proces, w którym każda wartość jest liniową kombinacją poprzednich wartości – wykorzystuje pamięć procesu. Procesy autoregresyjne mogą być różnego rzędu. Rząd określa ile wcześniejszych wartości ma wpływ na bieżącą wartość. Procesy autoregresyjne oznacza się symbolem AR(p) gdzie *p* to rząd autoregresji. Na przykład proces AR(1) możemy przedstawić, jako:

$$y_t = \varphi\, y_{t-1} + \varepsilon_t$$

y_t – wartość szeregu w chwili t
y_{t-1} – wartość szeregu w chwili t-1
ε_t – składnik losowy, zaburzenie z chwili t

Parametr φ szacujemy, aby ocenić siłę wpływu wartości poprzedniej procesu na wartość bieżącą. Tak więc na przykład, jeżeli t oznacza kolejne miesiące model AR(1) dla wielkości sprzedaży wyglądałby następująco:

$$sprzedaz_{sierpien} = \varphi\, sprzedaz_{lipiec} + \varepsilon_{sierpien}$$

Jeżeli rząd wynosiłby 2, to proces AR (2) wyglądałby następująco:

$$sprzedaz_{sierpien} = \varphi_1\, sprzedaz_{lipiec} + \varphi_2\, sprzedaz_{czerwiec} + \varepsilon_{sierpien}$$

A parametry φ_i wskazują na wpływ kolejnych, co raz starszych wartości szeregu na bieżącą wartość.

Czym jest proces średniej ruchomej?

Proces średniej ruchomej jest trochę podobny do procesu autoregresyjnego. Zakładamy, że wartość szeregu zależy od zaburzeń w chwili obecnej i wcześniejszych. Jak dużo wcześniej zaburzeń bierzemy określa nam parametr *q* – rząd procesu MA. Proces MA(1) możemy zapisać jako:

$$y_t = \beta\, \varepsilon_{t-1} + \varepsilon_t$$

y_t – wartość szeregu w chwili t
ε_{t-1} – składnik losowy, zaburzenie z chwili t–1
ε_t – składnik losowy, zaburzenie z chwili t

Na przykład, jeżeli t oznacza kolejne miesiące model MA(1) dla wielkości sprzedaży wyglądałby następująco:

$$sprzedaz_{sierpien} = \beta\, \varepsilon_{lipiec} + \varepsilon_{sierpien}$$

Jeżeli rząd wynosiłby 2, to proces MA (2) wyglądałby następująco:

$$sprzedaz_{sierpien} = \beta_1\, \varepsilon_{lipiec} + \beta_2\, \varepsilon_{czerwiec} + \varepsilon_{sierpien}$$

A parametry β_i wskazują na wpływ kolejnych, co raz starszych zaburzeń szeregu na bieżącą wartość.

Można powiedzieć, że proces MA jest średnią ważoną z ostatnich zaburzeń, składników losowych modelu. AR używa średniej ważonej z ostatnich wartości.

Czym jest integracja procesu?

Klasyczne modele ARMA miały zastosowanie do szeregów stacjonarnych. Modele ARIMA można już stosować do procesów niestacjonarnych, ale oczywiście nie do wszystkich. Stosujemy je wyłącznie do takich procesów, które możemy sprowadzić do postaci stacjonarnej.

Przypomnijmy, proces stacjonarny ma stały poziom w czasie (średnią) oraz odchylenie od średniego poziomu (wariancje). Wiemy również, że stacjonarność procesu możemy sprawdzać na wykresie autokorelacji: możemy przyjąć, że szeregi niestacjonarne mają więcej niż 6 opóźnień istotnie różnych od zera. Korzystając z modeli ARIMA model sam sprowadza dane do postaci stacjonarnej. Ale jak samemu możemy sprowadzić szereg do stacjonarności?

Jedną z metod jest różnicowanie szeregu (ang. *Differencing*). Różnicowanie polega na obliczeniu pierwszych lub dalszych różnic między kolejnymi wartościami. Pierwsze różnice obliczamy, jako:

$$\Delta y_t = y_t - y_{t-1}$$

Będą one reprezentować zmiany wartości w badanym szeregu np. spadki i wzrosty sprzedaży z okresu na okres. Tak więc przykładowy szereg może mieć przebieg wykazujący trend a po różnicowaniu otrzymamy wartości, które (bardzo często) będą już stacjonarne, a przynajmniej będą miały stały w czasie średni poziom.

Nie zawsze taki zabieg wystarczy, czasami rozpatrujemy różnice różnic. To którego stopnia różnice obliczamy określa nam stopień zintegrowania szeregu – parametr *d*. Podsumowując, szereg zintegrowany to szereg niestacjonarny, który poprzez odpowiednie różnicowanie można doprowadzić do stacjonarności. Możemy również, zastosować przekształcenia w celu ustabilizowania wariancji takie jak logarytmowanie całego szeregu.

Łącząc wszystkie trzy składowe razem otrzymujemy model ARIMA z trzema parametrami: ARIMA (*p, d,q*)

Gdzie:
p to parametr autoregresji,
d to stopień integracji szeregu,
q to parametr średniej ruchomej.

Może się zdarzyć, że parametry będą równe zero i uzyskamy np. model ARIMA(1,0,0) czyli po prostu procesu AR(1).

Jak dobrać odpowiednie parametry modelu?

Klasyczna procedura doboru odpowiedniego modelu ARIMA, została opracowana w 1976 roku przez Boxa i Jenkinsa. Składa się ona z trzech kroków: identyfikacji, estymacji i diagnozy. Innym sposobem identyfikacji modeli jest skorzystanie z kryteriów informacyjnych.

Etap I. Identyfikacja

Etap ten polega na wybraniu wstępnego modelu, który przejdzie do następnych etapów. Kandydata podajemy z wybranymi ekspercko parametrami *p*, *d* i *q*. W tym pomoże nam analiza wykresów i funkcji autokorelacji.
Zacznijmy od integracji – musimy spróbować ocenić stacjonarność szeregu. Jeżeli analiza potwierdzi nam stacjonarność, nie mamy problemu, nie musimy przekształcać szeregu, a stopień integracji będzie równy 0. Jeżeli szereg uznamy za niestacjonarny, możemy spróbować go przekształcić. Innymi słowy sprawdzimy czy szereg da się sprowadzić do stacjonarności i dzięki temu określimy stopień integracji szeregu.

Jeżeli chodzi o parametry *p* i *q* sprawa jest trudniejsza. Możemy jednak pomóc sobie opracowanymi przez badaczy wzorcami funkcji autokorelacji (ACF) i cząstkowej autokorelacji (PACF) dla najpopularniejszych wariantów tych modeli.

Funkcje autokorelacji już znamy z pierwszego artykułu. Musimy się przyjrzeć teraz funkcji autokorelacji cząstkowej. Autokorelacja cząstkowa PACF jest analogiczną miarą do korelacji cząstkowej. Wyznacza autokorelacje wartości szeregu z opóźnioną wartością, bez uwzględnienia autokorelacji wynikającej z pośrednich wartości. Innymi słowy czy nadal zachodzi związek między wartością bieżącą a opóźnioną o dwa dni, jeżeli usuniemy wpływ wartości opóźnionej o dzień.

Wiedząc już, co to jest ACF i PACF możemy przytoczyć jedną z zasad, które pomogą wybrać nam odpowiednie parametry *p* i *q*:

- W modelu ARIMA występuje „składowa" MA(*d*), gdy dla modelowanej zmiennej wartość PACF wygasa wykładniczo (co do wartości bezwzględnych). Konkretną wartość dla parametru *d* określa nam natomiast liczba pierwszych *peaków* na wykresie ACF.

Połączenie procesów AR(p), MA(q) komplikuje wspomniane wzorce, stąd często ten etap powtarzamy aż do uzyskania satysfakcjonującego modelu.

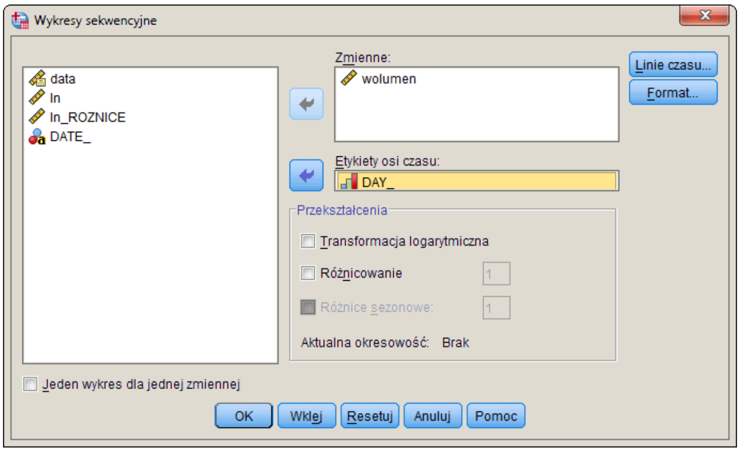
Etap II. Estymacja

Etap estymacji polega na dobraniu odpowiedniej metody do uzyskania oszacowania parametrów modelu (φ_i,β_i), do czego może być wykorzystywana np. metoda największej wiarygodności (ML) albo CLS (ang. *Conditional Least Squares*).

Etap III. Diagnoza

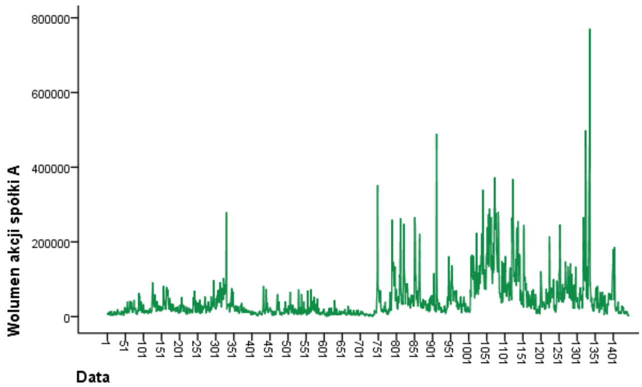
Etap diagnozy polega na ocenie naszego kandydata na model. Możemy oprzeć się na wyliczonych dla modelu statystykach. Oprócz tego możemy sprawdzić reszty przy pomocy ACF i PACF. Dla dobrze dobranego modelu obydwie funkcje nie powinny przyjmować wartości istotnie różnych od zera.
Na koniec dodajmy, że modele ARIMA również doczekały się swoich modyfikacji pozwalających modelować efekt sezonowości (modele SARIMA).

Otwórzmy teraz zbiór danych *ARIMA.sav*. Zbiór składa się z szeregu przedstawiającego wolumen akcji spółki „A" (*wolumen*), zmiennych etykietujących czas (*data*, *DAY_*, *DATE_*) oraz dwie zmienne, które utworzymy w przykładzie (*ln*, *ln_ROZNICE*). Jednostką czasu jest dzień. Załóżmy, że analityk opracował wskaźnik do wyliczenia, którego potrzebuje 10 dniowej prognozy wspomnianego szeregu. Dokładniej, potrzebna mu jest górna granica 90% prognozy przedziałowej (przedziału ufności dla prognozy). Zacznijmy od przedstawienia szeregu w procedurze ANALIZA ▶ PROGNOZOWANIE ▶ WYKRESY SEKWENCYJNE.
Zmienną *DAY_* wykorzystamy jako ETYKIETY OSI CZASU, a *wolumen* przenieśmy na listę ZMIENNE. Kliknijmy OK.



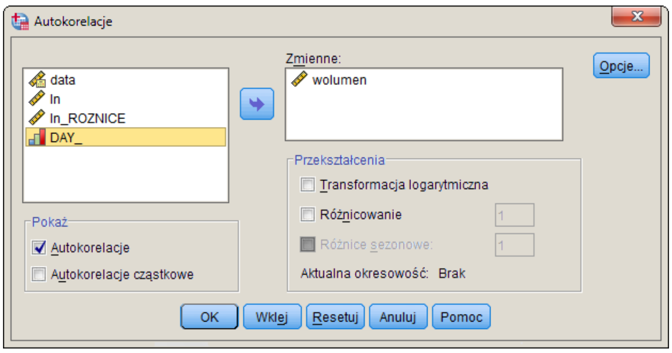
RYSUNEK 1.
Tworzenie wykresu sekwencyjnego

Poniższy wykres przedstawia nasz szereg. Widzimy, że przebieg serii charakteryzuje duża zmienność w niektórych okresach i występowanie wartości skrajnych.



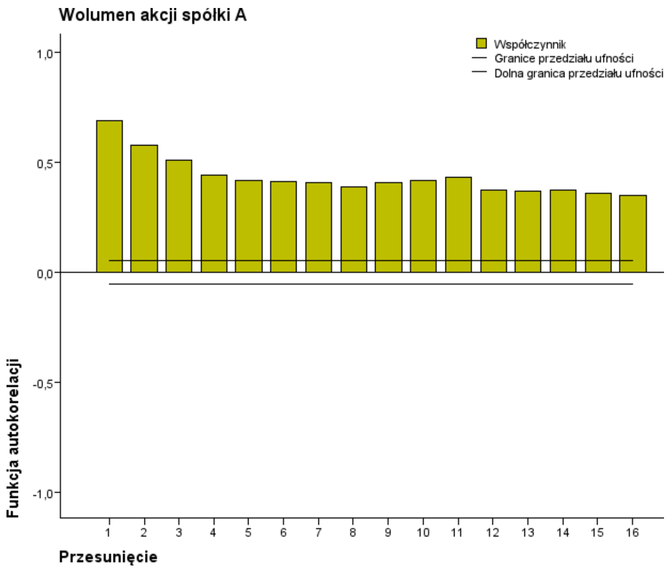
RYSUNEK 2.
Wykres sekwencyjny

Przejdziemy teraz krok po kroku przez klasyczną procedurę identyfikacji modelu ARIMA. Zacznijmy od funkcji autokorelacji. Wybierzmy z menu ANALIZA ▶ PROGNOZOWANIE ▶ AUTOKORELACJA. Dzięki niej sprawdzimy czy konieczne jest przekształcenie naszego szeregu – sprawdzimy czy nasz szereg jest stacjonarny.



RYSUNEK 3.
Obliczenie funkcji autokorelacji

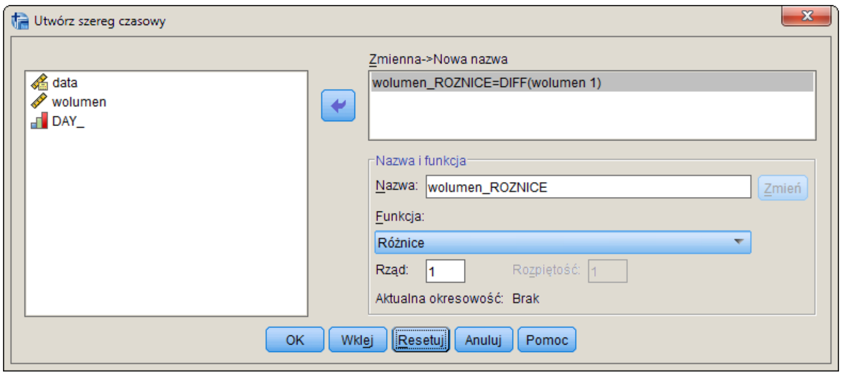
Przenieśmy *wolumen* na listę ZMIENNE, odznaczmy na razie AUTOKORELACJE CZĄSTKOWĄ i kliknijmy OK.



RYSUNEK 4.
Autokorelacja szeregu

Zgodnie z zasadą przytoczoną w pierwszym artykule (szeregi niestacjonarne mają funkcję autokorelacji istotnie większą od zera w przypadku przynajmniej 6 opóźnień), możemy stwierdzić, że szereg jest niestacjonarny. Spróbujmy więc różnicowania.

Przekształcanie szeregów czasowych jest bardzo często wykorzystywane i dla tego znajdziemy je w kilku miejscach w programie. Na przykład, w trakcie tworzenia wykresu sekwencyjnego możemy wybrać opcję, aby wykres przedstawiał wartości przekształcone – okno dialogowe z Rysunku 1. Tak samo możemy postąpić przy wyliczaniu funkcji autokorelacji – okno dialogowe z Rysunku 3. My jednak skorzystajmy z możliwości wyliczenia nowej kolumny w naszym zbiorze. Wybierzmy z menu PRZEKSZTAŁCENIA ▶ UTWÓRZ SZEREG CZASOWY.

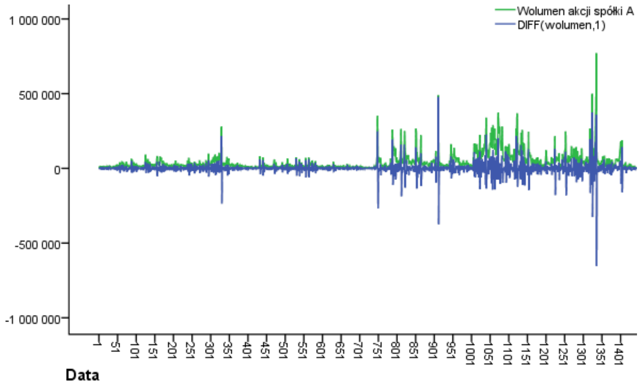


RYSUNEK 5.
Utwórz szereg czasowy

W oknie tym, mamy możliwość przekształcić nasz szereg na różne sposoby. Między innymi możemy tam obliczyć RÓŻNICE szeregu. Przenieśmy *wolumen* na listę po prawej stronie. Możemy zmienić nazwę nowego szeregu na *wolumen_ROZNICE* (zmianę potwierdzamy przyciskiem ZMIEŃ). Wybierzmy odpowiednią funkcję, wprowadźmy RZĄD równy 1 i zatwierdźmy OK.

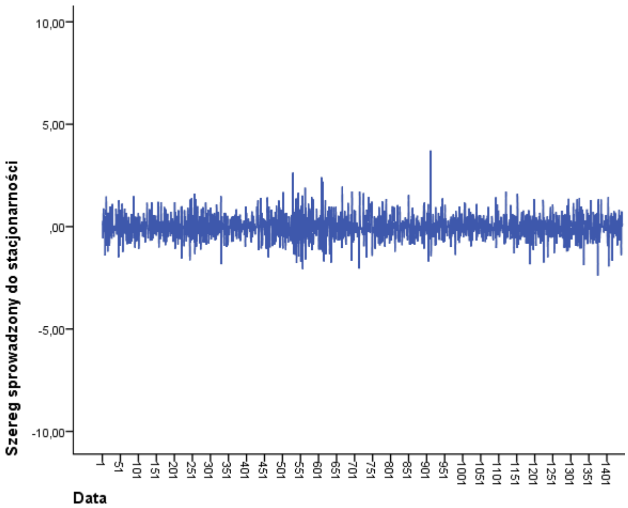
W raporcie pojawi się podsumowanie dotyczące utworzonego szeregu. Zauważmy, że nowy szereg będzie zaczynał się od drugiej obserwacji, ponieważ procedura nie obliczy $\Delta y_1 = y_1 - y_0$.

Powinniśmy teraz sprawdzić czy nie jest konieczne dalsze różnicowanie. Nie będziemy tego już jednak prezentować, zakładamy, że różnicowanie pierwszego stopnia wystarczy, a więc nasz model ARIMA będzie miał parametr *d* równy 1. Jak wiemy różnicowaniem szeregu zapewniliśmy stały poziom procesu. Ale co z jego wariancją? Zobaczmy jak teraz wygląda nasz przekształcony szereg. Przywołajmy znowu okno wykresów sekwencyjnych i dodajmy nowo wyliczoną zmienną do wykresu.



RYSUNEK 6.
Zobrazowanie wyniku operacji różnicowania (pierwszego stopnia)

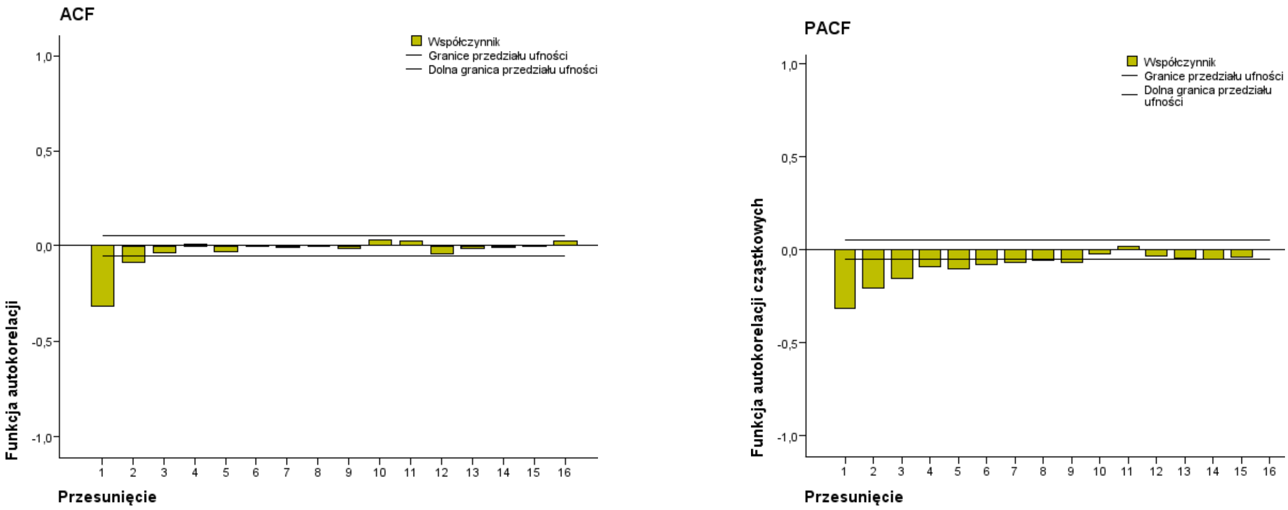
Na wykresie szereg zielony prezentuje nasze dane wejściowe, a szereg niebieski jego pierwsze różnice (DIFF(zmianna, stopień)). Tak jak mogliśmy podejrzewać, szereg przekształcony nie ma stałego odchylenia od poziomu (np. mała zmienność w dniach od 600–750, duża 1000–1200). Możemy więc spróbować ustabilizować wariancję procesu np. poprzez logarytmowanie. Ponieważ logarytm naturalny obliczamy dla liczb dodatnich, musimy cofnąć się do szeregu *wolumen* i przekształcenia wykonać w kolejności: obliczyć logarytm zmiennej *wolumen*, a następnie pierwsze różnice zmiennej zlogarytmowanej. Ostatecznie uzyskamy szereg przedstawiony na rysunku 7.



RYSUNEK 7.
Szereg wolumen sprowadzony do stacjonarności

Wiemy już więc jakie przekształcenia są potrzebne, aby spełnić założenia modelu co do stacjonarności procesu: zlogarytmowanie szeregu oraz różnicowanie pierwszego stopnia. Dzięki temu wiemy już, że wybierzemy model ARIMA(p, 1,q) z transformacją logarytmiczną. Pozostaje wybrać nam teraz parametr p oraz q.

Przywołajmy ponownie okno AUTOKORELACJI, tym razem zaznaczmy AUTOKORELACJE i AUTOKORELACJE CZĄSTKOWĄ i wykonajmy procedurę dla szeregu *ln_ROZNICE*.



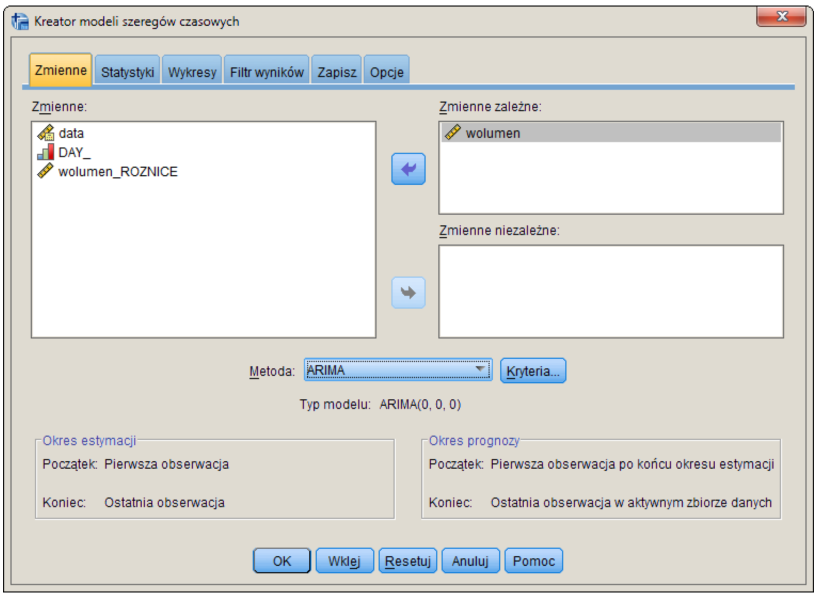
RYSUNEK 8.
Funkcja ACF i PACF dla przekształconego szeregu czasowego

Przypomnijmy sobie teraz zasadę pomagającą w identyfikacji z początku artykułu:

- W modelu ARIMA, występuje „składowa” MA (*d*), gdy dla modelowanej zmiennej wartość PACF wygasa wykładniczo (co do wartości bezwzględnych). Konkretną wartość dla parametru *d* określa nam natomiast liczba pierwszych *peaków* na wykresie ACF.

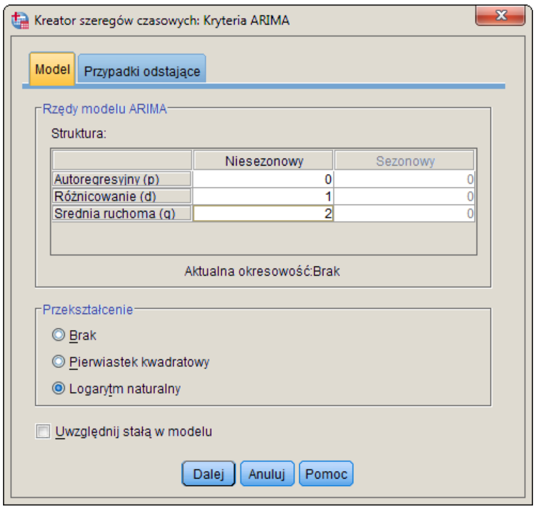
Na wykresie PACF widzimy spadek wartości autokorelacji cząstkowej. Funkcja ACF ma jeden wyraźny *peak* na pierwszym opóźnieniu. Na drugim opóźnieniu autokorelacja jest mniejsza, ale również przekracza poziom przedziałów ufności.

Podsumowując wszystkie zebrane przez nas informacje możemy stwierdzić, że dobrą pierwszą propozycją na model ARIMA powinien być model ARIMA (0,1,2), bądź ARIMA (0,1,1). Wybierzmy model ARIMA (0,1,2) i spróbujmy wyznaczyć prognozę na 10 dni wprzód. Przejdźmy więc do znanego nam już kreatora modeli szeregów czasowych: ANALIZA ▶ PROGNOZOWANIE ▶ UTWÓRZ MODELE... . Przenieśmy nasz szereg *wolumen* do listy ZMIENNE ZALEŻNE . Z listy wybierzmy metodę ARIMA.



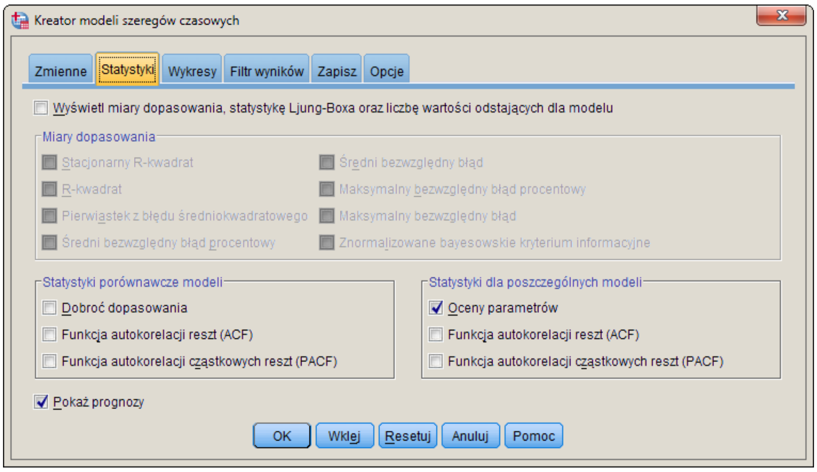
RYSUNEK 9.
Utwórz model – Wybór zmiennych

Pod przyciskiem KRYTERIA wprowadźmy ustalone przez nas parametry. Odznaczmy UWZGLĘDNIJ STAŁĄ W MODELU oraz zaznaczmy LOGARYTM NATURALNY w obszarze PRZEKSZTAŁCENIA .



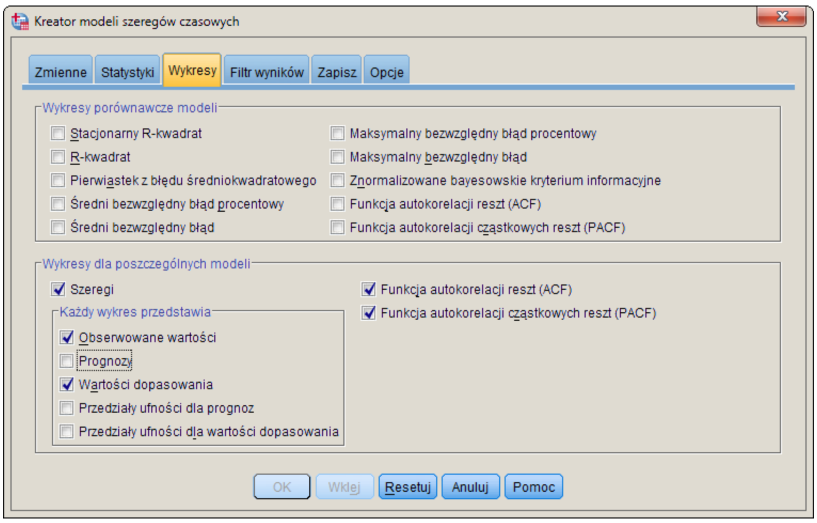
RYSUNEK 10.
Utwórz Model – Kryteria modeli ARIMA

Na zakładce STATYSTYKI zaznaczmy, tak jak robiliśmy to na przykładzie z poprzedniego artykułu POKAŻ PROGNOZY i OCENY PARAMETRÓW. Resztę opcji odznaczmy.



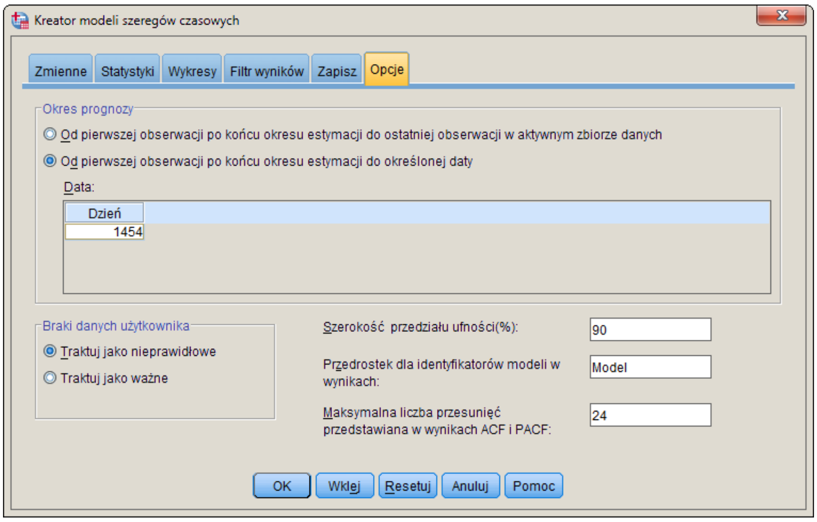
RYSUNEK 11.
Utwórz model – Statystyki

Na zakładce WYKRESY pozostawmy opcje zgodnie z oknem dialogowym prezentowanym na rysunku 12.



RYSUNEK 12.
Utwórz modele – Wykresy

Przejdźmy na ostatnią już zakładkę OPCJE. Ponieważ chcemy uzyskać przedział prognozy na 10 dni wprzód, zaznaczmy opcje OD PIERWSZEJ OBSERWACJI PO KOŃCU OKRESU i wprowadźmy wartość 1454, czyli wartość o 10 większą od liczby obserwacji w analizowanym zbiorze. Zgodnie z wymogami prognozy musimy również zmniejszyć szerokość przedziału ufności do 90%. Teraz możemy nacisnąć OK.



RYSUNEK 13.
Utwórz model – Opcje

Zbudowany model mamy podsumowany w trzech tabelach. Pierwsza z nich zawiera informacje o wybranym przez nas modelu, a druga zawiera informacje o oszacowanych parametrach. W tabeli parametry modelu ARIMA widzimy, że mamy dwa parametry dla procesu średnia ruchoma: opóźnienie 1 (β_1) i opóźnienie 2 (β_2). Oszacowana wartość parametru β_1 wynosi 0,477, a parametru β_2 0,180. Wartość istotność możemy interpretować tak samo jak w przypadku modeli regresji. Wartość równa zero informuje nas o tym, że oszacowane parametry są statystycznie istotne.

OPIS MODELU

			TYP MODELU
IDENTYFIKATOR MODELU	akcje	Model_1	ARIMA(0,1,2)

PARAMETRY MODELU ARIMA

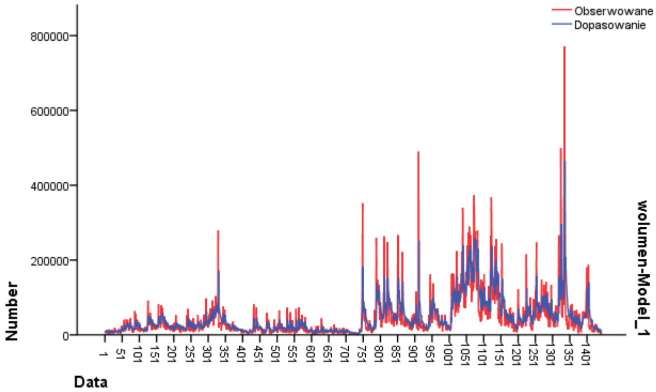
			WARTOŚĆ OSZACOWANIA	BŁĄD STANDARDOWY	T	ISTOTNOŚĆ
LOGARYTM NATURALNY	Różnicowanie		1			
	Średnia ruchoma (MA)	Opóźnienie 1	,477	,026	18,409	,000
		Opóźnienie 2	,180	,026	6,931	,000

WARTOŚCI PROGNOZOWANE

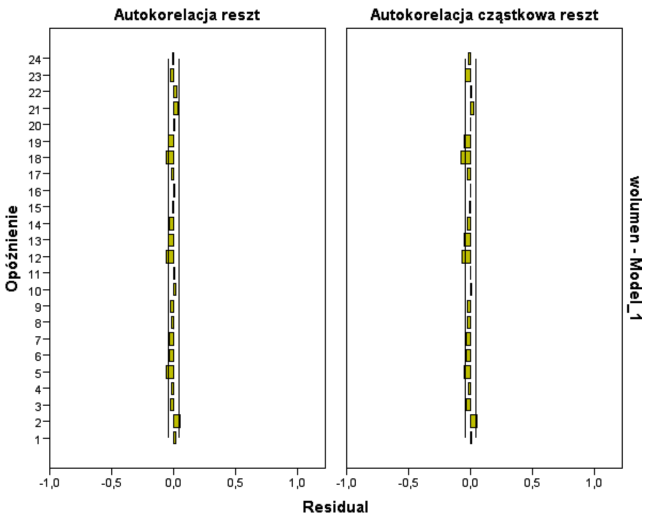
MODEL		1454
WOLUMEN AKCJI SPÓŁKI A-MODEL_1	Wartości prognozowane	8340
	Górna granica przedziału ufności	23449
	Dolna granica przedziału ufności	1453

TABELE 1–3.
Tabele wynikowe utworzonego modelu

W trzeciej tabeli widzimy poszukiwaną przez nas prognozę (tabela w raporcie będzie zawierać prognozy na kolejne dni aż do dnia 1454). Zgodnie z zastosowanym przez nas modelem ARIMA dla wolumenu akcji spółki A, poszukiwana górna granica prognozy na 10 dni wprzód wynosi w przybliżeniu 23,5 tys. Na koniec raportu dostaniemy również wykres przedstawiający prognozowany przez nas szereg razem z dopasowanym do niego modelem oraz ACF i PACF dla reszt. Na podstawie tych ostatnich możemy stwierdzić, że dobrze dobraliśmy nasz model. Możemy to również potwierdzić sprawdzając, jaki model wybierze procedura Automatycznego doboru modeli.



RYSUNEK 14.
Wartości historyczne szeregu (zaobserwowane) oraz teoretyczne wartości modelu ARIMA (0,1,2) (Dopasowanie)



RYSUNEK 15.
Funkcje autokorelacji i autokorelacji cząstkowej dla reszt

W praktyce do wyznaczania prognoz szeregów o takim przebiegu jak w dzisiejszym przykładzie, możemy dodatkowo korzystać z modeli dedykowanych do prognozowania *zmienności*. Głównym celem prezentowanego przykładu było zapoznanie czytelnika z metodą identyfikacji modeli ARIMA. W praktyce wybierając model, możemy skorzystać z opcji AUTOMATYCZNEGO WYBORU MODELI dostępnej w IBM SPSS Statistics. Warto jednak wiedzieć, dlaczego niektóre przekształcenia zostały zastosowane i dlaczego oraz co oznaczają oszacowane parametry. W następnym artykule, przyjrzymy się bliżej ocenie jakości zbudowanych przez nas modeli. W kolejnych, wrócimy do modeli ARIMA i skupimy się na możliwość zastosowania ich do modelowania sezonowości. Zapoznamy się również z modelem ARIMA pozwalającym uwzględnić wpływ dodatkowych zmiennych na prognozę.

Więcej o prognozowaniu szeregów czasowych można się dowiedzieć na szkoleniu **AN2 – Analiza szeregów czasowych i prognozowanie**