Komputerowa analiza szeregów czasowych - raport 2

"Analiza danych rzeczywistych przy pomocy modelu ARMA"

Przedmiot i prowadzący:
Komputerowa analiza szeregów czasowych,
piątki 11.15 - 13.00 (grupa nr 5),
Mgr inż. Wojciech Żuławiński

Aleksandra Hodera (268733) Aleksandra Polak (268786)

Spis treści

1 Opis danych, cel

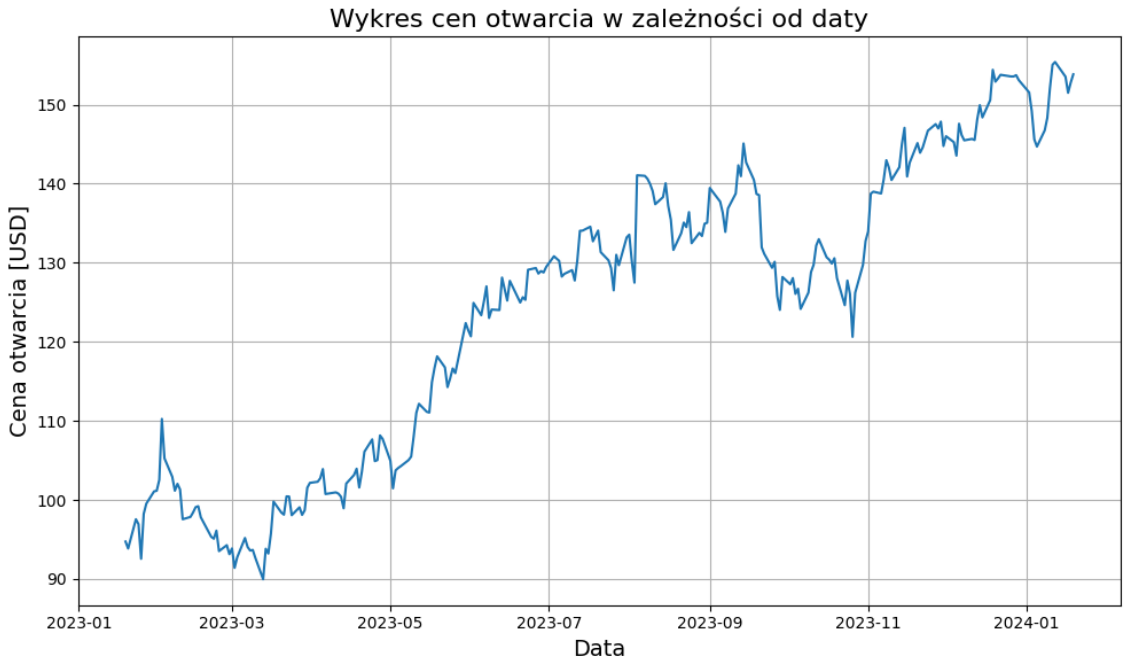
Celem niniejszego raportu jest analiza danych rzeczywistych za pomocą modelu ARMA. Dane te można znaleźć pod poniższym linkiem: *Akcje Amazon*. Dotyczą one cen akcji przedsiębiorstwa Amazon w okresie 19.01.2023 r. do 19.01.2024 r. i zawierają 252 obserwacje. Giełda była nieczynna w weekendy i święta. Dla każdego dnia zostały zebrane informacje dotyczące:

- ceny otwarcia (Open),
- maksymalnej ceny akcji (High),
- minimalnej ceny akcji (Low),
- ceny zamknięcia (Close),
- skorygowanej ceny zamknięcia (Adj Close),
- liczby akcji danej spółki, które zmieniły właściciela w danym dniu (Volume).

Wszystkie ceny zostały podane w dolarach amerykańskich. W naszym raporcie analizowałyśmy pierwszą z informacji, czyli cenę otwarcia w zależności od czasu. Zależność ta została przedstawiona na wykresie 1.

Skupiłyśmy się na zbadaniu surowych danych oraz odpowiednim ich przekształceniu, tak aby można dopasować model ARMA. Następnie, z pomocą kryteriów informacyjnych wyestymowałyśmy parametry modeli oraz oceniłyśmy ich dopasowanie do danych rzeczywistych. Sprawdziłyśmy również założenia dotyczące szumu. Dodatkowo wyodrębniłyśmy zbiór testowy, aby móc wykonać predykcję.

Działania te miały na celu ocenę czy ceny akcji amerykańskiego przedsiębiorstwa Amazon można modelować za pomocą modeli ARMA.



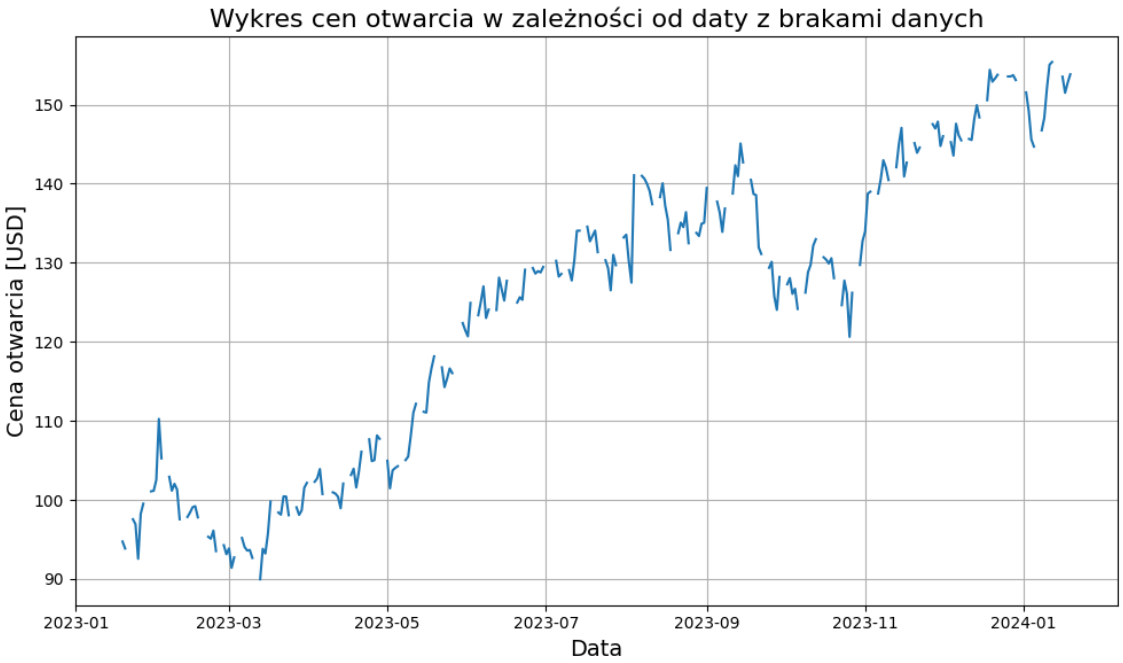
Rys. 1

2 Przygotowanie danych do analizy

Ponieważ chcemy analizować dane rzeczywiste za pomocą modelu ARMA musimy w pierwszej kolejności odpowiednio przygotować nasze dane.

2.1 Zbadanie jakości danych

Ze względu na to, że giełda była zamknięta w weekendy i święta w naszym zestawie zamiast 366 danych są 252 obserwacje. Łatwo zauważyć, że brakuje 114 obserwacji. Na wykresie 2 zwizualizowana jest zależność ceny otwarcia od czasu z widocznymi brakującymi danymi.



Rys. 2

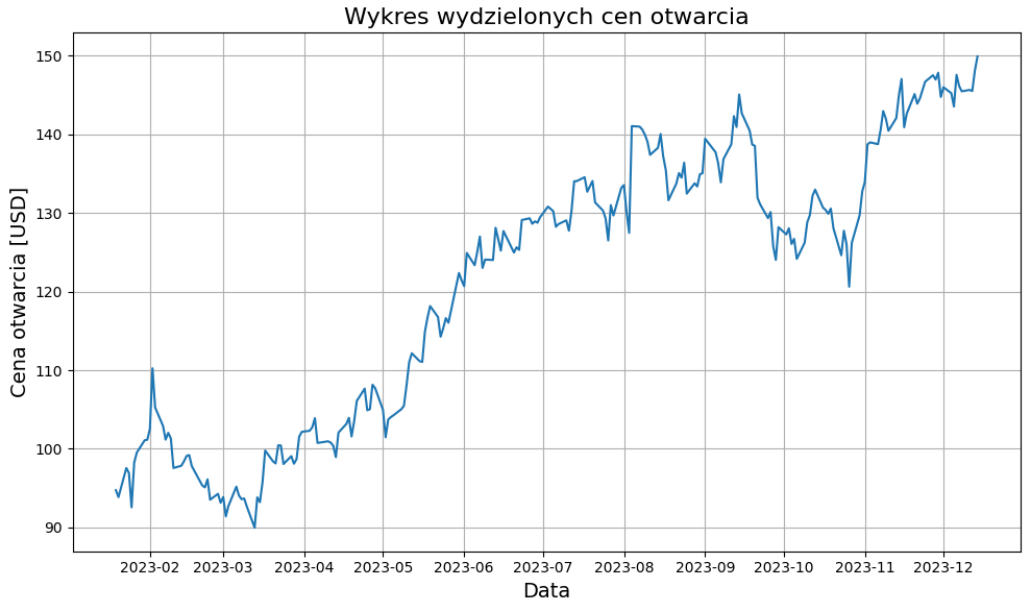
Wiemy również, że aby móc analizować dane przy pomocy modelu ARMA muszą być one równo oddalone w czasie od siebie. Brakujące daty uzupełniłyśmy metodą interpolacji liniowej, która polega na wykorzystaniu istniejących danych do przybliżenia braków na podstawie równania prostej łączącej znane daty. W tym celu wykorzystałyśmy funkcję interp z biblioteki numpy wbudowanej w języku Python. Na wykresie 3 zwizualizowane zostały ceny otwarcia w zależności od daty z uzupełnionymi danymi na podstawie interpolacji liniowej.

Rys. 3

Analizując wykres 3 widzimy, że uzupełnione ceny nie są ani zbyt niskie, ani zbyt wysokie, nie wykazują znaczących odchyłów. Pozwala nam to stwierdzić, że wyestymowane ceny zachowują się podobnie do rzeczywistych.

2.2 Wyodrębnienie zbioru testowego

Zanim przeszłyśmy do dalszej analizy wyodrębniłyśmy 10 % najnowszych danych. Zabieg ten miał na celu podzielenie zestawu na zbiór treningowy i testowy. Na wykresie 4 widoczna jest zależność ceny otwarcia od daty dla pierwszych 90 % danych. W dalszej części raportu wykorzystamy te dane do predykcji pozostałych 10 %.



Rys. 4

2.3 Transformacja danych i ich weryfikacja

2.3.1 Analiza funkcji ACF i PACF dla surowych danych

Chcąc analizować dane za pomocą modelu ARMA musimy zweryfikować założenie o stacjonarności. W tym celu, za pomocą wbudowanych funkcji, zwizualizujemy ACF i PACF.

ACF to funkcja wyrażająca korelację między dwoma obserwacjami z szeregu $\{X_t\}$, oddalonymi od siebie o $h \in \mathrm{Z}$, we wzorze wyraża się to jako $corr(X_t, X_{t+h})$. Estymuje się ją w następujący sposób:

$$\hat{\rho} = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)},$$

gdzie $\hat{\gamma}(h)$ to empiryczna funkcja autokowariancji, którą wyraża się wzorem:

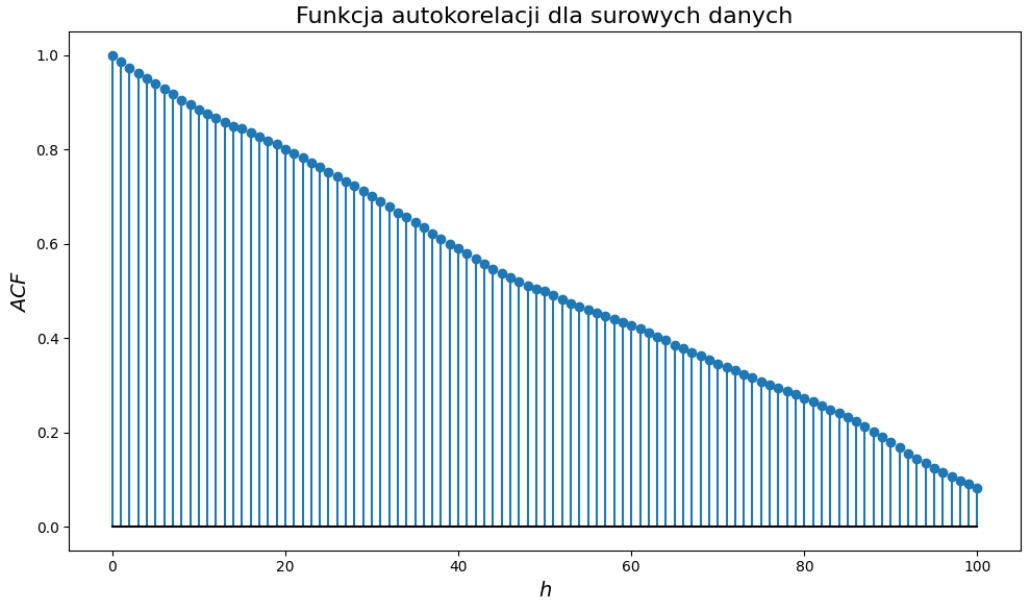
$$\hat{\gamma}(h) = \frac{1}{n} \sum_{t=1}^{n-|h|} (x_{t+|h|} - \overline{x})(x_t - \overline{x}),$$

gdzie $x_1, x_2, ..., x_n$ to realizacje szeregu czasowego $\{X_t\}$, a $\overline{x}$ to średnia równa $\frac{1}{n}\sum_{i=1}^{n} x_i$. Wyniki dla $h = 0, 1, ..., 100$ zostały przedstawione na wykresie 5.

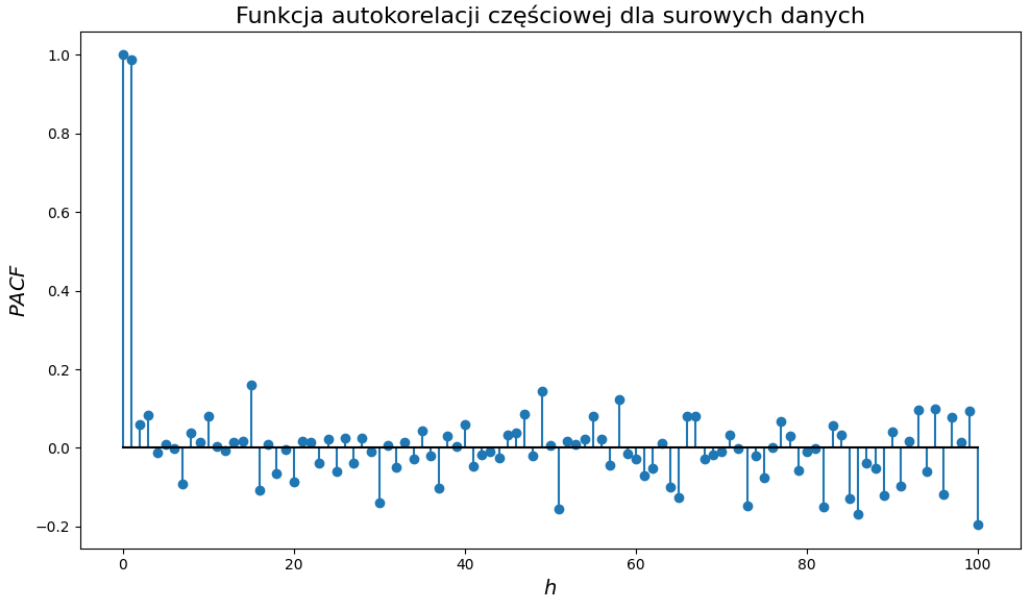
PACF to funkcja częściowej autokorelacji. Używa się ją w celu określenia bezpośredniej zależności pomiędzy X_t, a X_{t+h}. Definiuje się ją w następujący sposób:

$$\alpha(h) = \begin{cases} 1, & \text{gdy } h = 0 \\ \phi_{hh}, & \text{gdy } h \geq 1 \end{cases},$$

gdzie ϕ_{hh} to ostatni składnik wektora $\Gamma_h^{-1}\gamma(h)$, gdzie $\Gamma_h = [\gamma(1), \gamma(2), ..., \gamma(h)]'$. Wyniki dla $h = 0, 1, ..., 100$ zostały przedstawione na wykresie 6. Wykresy 5 (autokorelacja) i 6 (częściowa autokorelacja) pomogą nam także ocenić czy w naszych danych istnieje liniowość bądź sezonowość.



Rys. 5



Rys. 6

Jak możemy zauważyć na wykresie 5, funkcja autokorelacji nie zbiega szybko do 0. Analizując wykres 6 widzimy, że funkcja częściowej autokorelacji odchyla się od 0. Oba te fakty pozwalają nam wnioskować, że nasze dane najprawdopodobniej nie są stacjonarne. Dodatkowo, z wykresu 5 możemy wywnioskować, że dane wykazują liniowość i sezonowość. Intuicyjnie moglibyśmy przypuszczać, że dane finansowe nie będą wykazywać sezonowości, jednak jak pokazuje ten przykład intuicja czasami może zawieść.

2.3.2 Sprawdzenie stacjonarności danych - test ADF

W celu potwierdzenia hipotezy o niestacjonarności szeregu przeprowadziłyśmy test ADF, czyli Augmented Dickey-Fuller Test. Wykorzystuje on następujące hipotezy zerowe i alternatywne:

$$H_0 : \text{Szereg czasowy nie jest stacjonarny } H_1 : \text{Szereg czasowy jest stacjonarny}$$

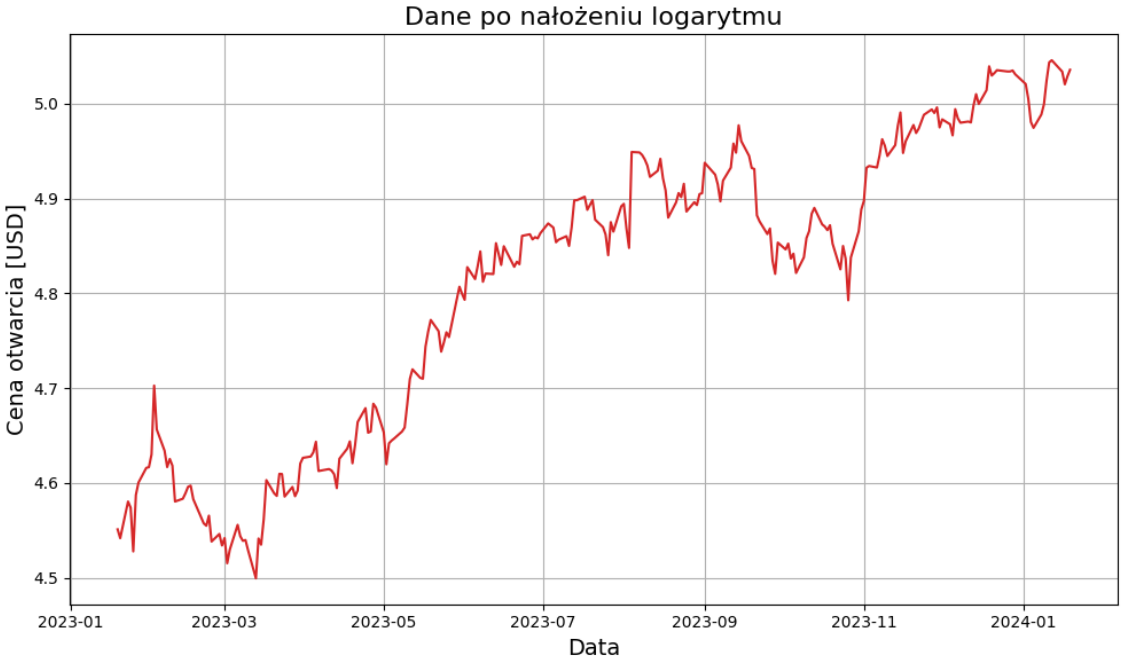
Jeśli wartość p z testu jest mniejsza niż pewien poziom istotności α możemy odrzucić hipotezę zerową i stwierdzić, że szereg czasowy jest stacjonarny. Do zaimplementowania tego testu użyjemy wbudowanej funkcji adfuler z biblioteki statsmodels.tsa.stattools. Zakładamy poziom istotności $\alpha = 0.05$. W związku z tym, jeżeli otrzymamy p-wartość powyżej poziomu α przyjmiemy hipotezę zerową i stwierdzimy, że dany szereg nie jest stacjonarny. W naszym przypadku otrzymaliśmy p-wartość równą $0.779 > 0.05$. Potwierdza to naszą hipotezę na temat niestacjonarności danych.

2.3.3 Identyfikacja trendów deterministycznych - transformacja Boxa-Coxa

Tak jak pokazaliśmy wyżej, dane wykazują liniowość oraz sezonowość. Aby móc w późniejszej części podjąć próbę dopasowania modelu ARMA musimy je usunąć. W tym celu, w pierwszej kolejności wykorzystamy transformacje Boxa-Coxa. Jest to rodzaj transformacji używanej głównie do stabilizacji wariancji. Działa ona na danych nieujemnych.

$$y(\lambda) = \begin{cases} \frac{y^\lambda - 1}{\lambda}, & \text{gdy } \lambda \neq 0 \\ \log(y), & \text{gdy } \lambda = 0 \end{cases}$$

W naszym przypadku użyjemy uproszczonej wersji wyżej opisanej transformaty, czyli nałożymy logarytm na dane. Na wykresie 7 widoczne są dane po tej transformacji.



Rys. 7

Jak możemy zauważyć na wykresie 7, otrzymaliśmy szereg o dużo mniejszym rozstępie. Do dalszej analizy będziemy wykorzystywać dane po transformacji.

2.3.4 Identyfikacja trendów deterministycznych - dekompozycja Walda

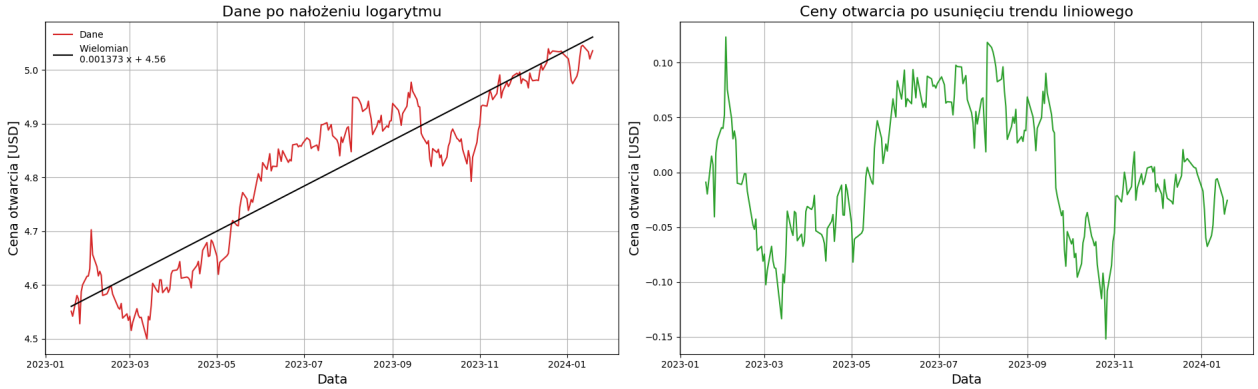
W celu weryfikowania trendu i sezonowości danych stosuje się dekompozycję Walda. Jest to metoda polegająca na podziale szeregu czasowego na składniki opisujące liniowość oraz sezonowość:

$$Y_t = m(t) + s(t) + X_t,$$

gdzie:

- $m(t)$ opisuje deterministyczny trend,
- $s(t)$ to funkcja okresowa,
- X_t to szereg czasowy stacjonarny w słabym sensie.

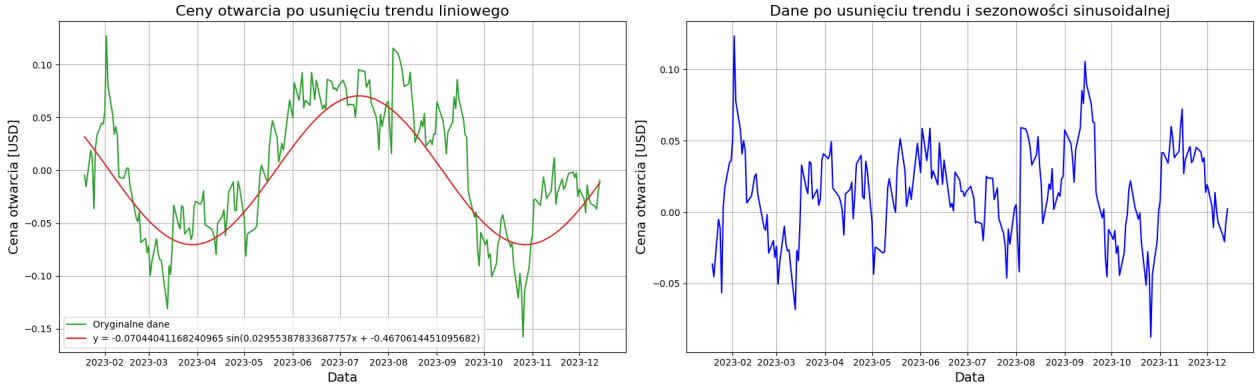
W pierwszej kolejności metodą interpolacji spróbowałyśmy dopasować do danych funkcję liniową, żeby móc stwierdzić czy nasze dane wykazują liniowość. W tym celu wykorzystałyśmy funkcje wbudowane: polyfit i poly1d z biblioteki numpy. Wykres cen otwarcia wraz z dopasowaną funkcją oraz wykres po usunięciu trendu liniowego widoczne są poniżej, na rysunku 8.



Rys. 8

Jak możemy zauważyć na rysunku 8, na wykresie po lewej stronie, dane wykazywały liniowość. Natomiast wykres po prawej stronie pokazuje, że została ona usunięta.

Kolejnym krokiem było usunięcie sezonowości. W tym celu dopasowałyśmy do danych funkcję sinus, a następnie za jej pomocą usunęłyśmy z danych sezonowość. Wyniki zostały przedstawione na rysunku 9.

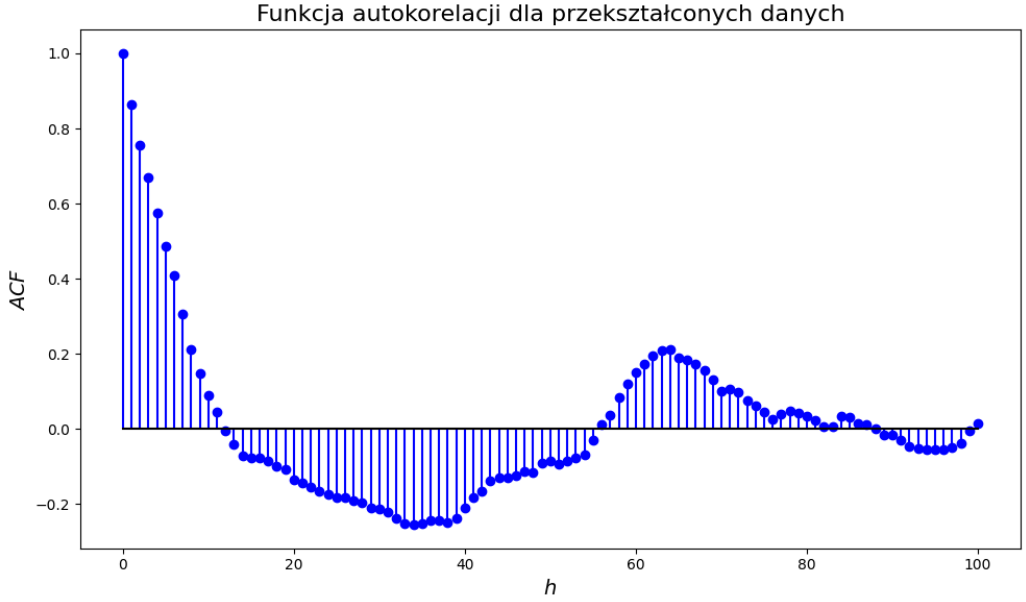


Rys. 9

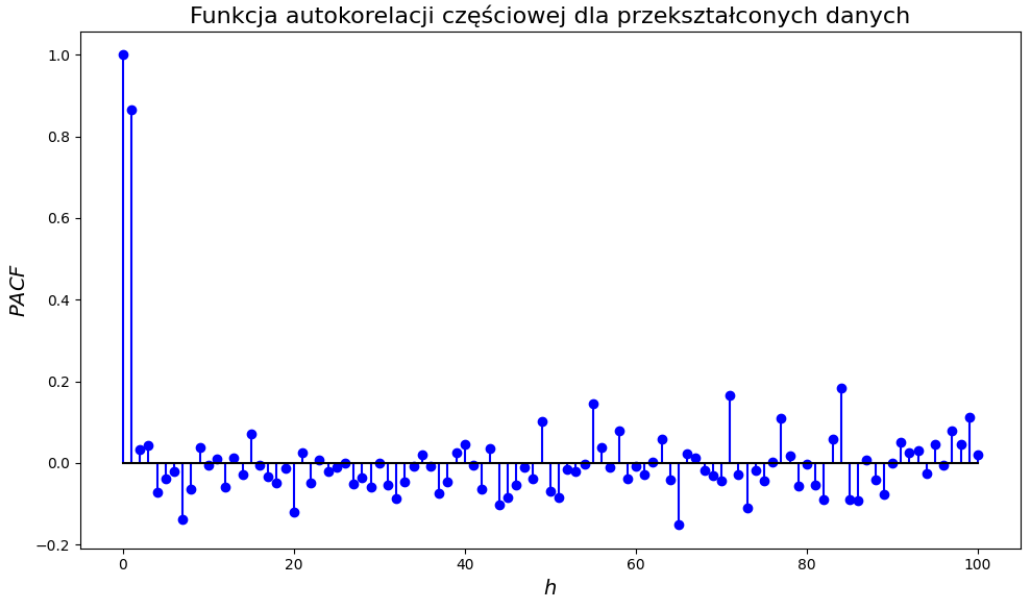
Jak widzimy na rysunku 9, na wykresie po lewej stronie, dane wykazywały sezonowość. Analizując kształt wykresy widocznego po prawej stronie widzimy, że z danych zostały usunięte trend oraz sezonowość, czyli że dekompozycja najprawdopodobniej przebiegła pomyślnie.

2.3.5 Analiza funkcji ACF i PACF dla uzyskanych szeregów

W celu weryfikacji poprawności usunięcia trendu oraz sezonowości z danych ponownie zwizualizowałyśmy funkcję ACF i PACF, tym razem dla danych po dekompozycji.



Rys. 10



Rys. 11

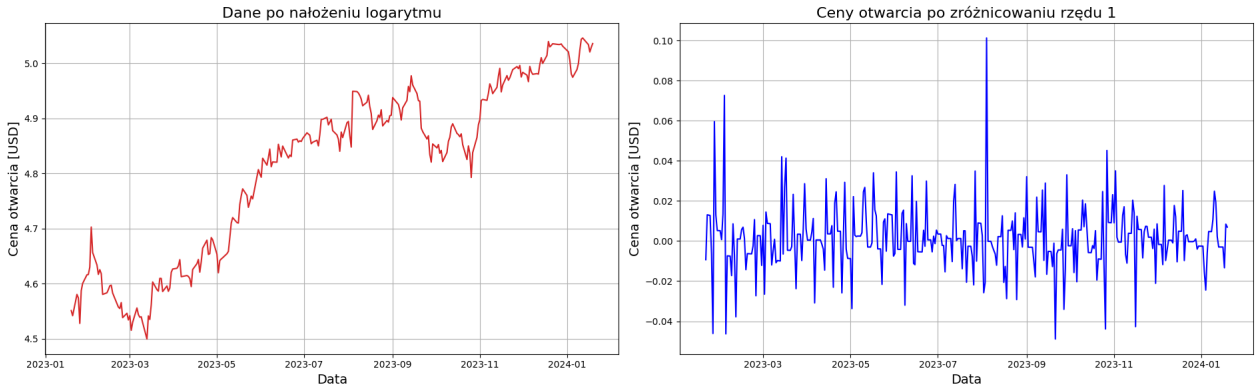
Jak możemy zauważyć na wykresie 10, funkcja autokorelacji w 0 ma 1, później zbiega dość szybko do 0, co pozwala nam wnioskować, że najprawdopodobniej udało się uzyskać szereg stacjonarny. Dodatkowo widzimy, że mniej więcej dla $h < 17$ dane wykazują silną zależność. Pozwala nam to stwierdzić, że modelem, który będzie najlepiej przybliżał nasze dane będzie model zawierający część AR(p). Analizując zachowanie funkcji PACF, na wykresie 11 widzimy, że dla $h \neq 0, 1$ funkcja częściowej autokorelacji oscyluje wokół 0, widoczne są jedynie pojedyncze wartości odstające. Fakt ten również pozwala nam stwierdzić, że dekompozycja Walda najprawdopodobniej przebiegła pomyślnie. Widzimy także, że funkcja, dla $h = 1$ przyjmuje wartość około 0.9, w dalszej części sprawdzimy czy wartość ta pokrywa się z dobranym parametrem modelu.

2.3.6 Identyfikacja trendów deterministycznych - różnicowanie

Innym możliwym sposobem doprowadzenia danych do postaci stacjonarnej jest różnicowanie. Do tej metody także użyłyśmy danych po transformacji Boxa-Coxa. W naszym przypadku wykorzystałyśmy różnicowanie rzędu 1. Metoda ta polega na utworzeniu z próbki $\{x_t\}_{t\in T}, T = \{1, 2, ..., n\}$ nowej, będącej postaci:

$$\{\delta x_t : \delta x_t = x_t - x_{t-1}, t \in T \backslash \{1\}\}.$$

W ten sposób uzyskałyśmy szereg $Y_t = X_t - X_{t-1}$. Ceny otwarcia po zróżnicowaniu widoczne są na wykresie 12.

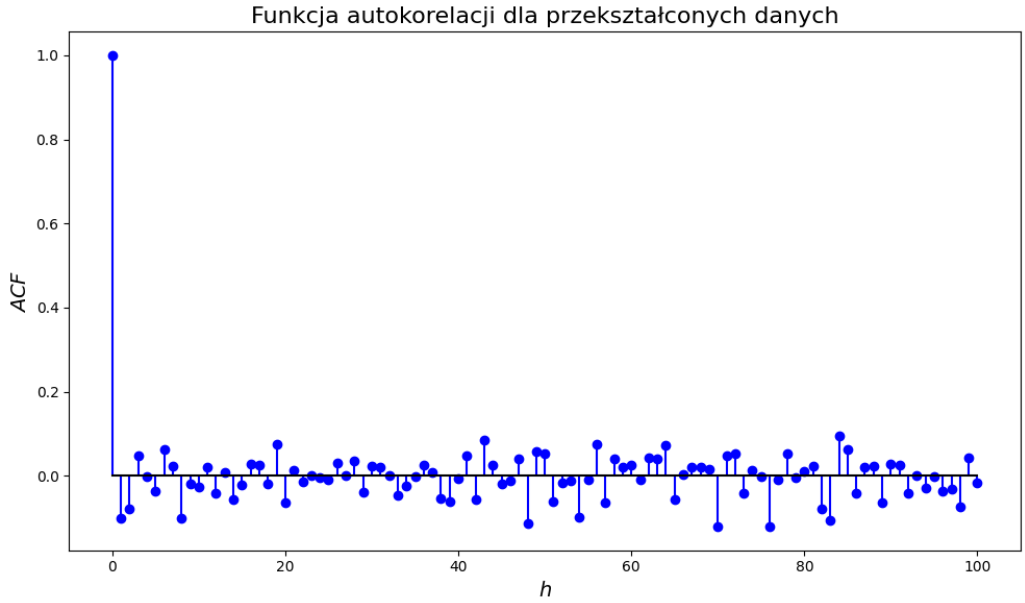


Rys. 12

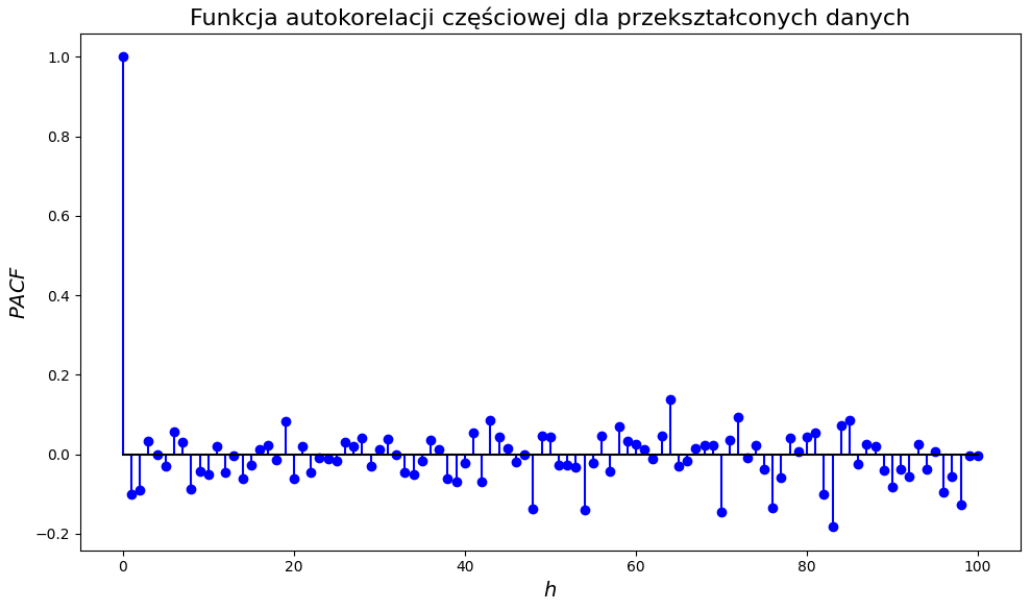
Jak możemy zauważyć, w tej metodzie także udało się usunąć z danych trend oraz sezonowość.

2.3.7 Analiza funkcji ACF i PACF dla uzyskanych szeregów

Dla zróżnicowanych danych także wyznaczyłyśmy funkcje ACF oraz PACF. Wyniki są widoczne na rysunkach 13 (ACF) i 14 (PACF).



Rys. 13



Rys. 14

Tutaj, podobnie jak dla danych po dekompozycji Walda, także widzimy, że funkcja autokorelacji szybko zbiega do 0. Dodatkowo, możemy zauważyć, że w danych prawie wcale nie występuje zależność lub jeśli występuje to jest ona ukryta. Pozwala nam to wnioskować, że dane te może najlepiej przybliżać model ARMA(p,q) lub MA(q) z parametrami bliskimi 0. Hipotezę tą zweryfikujemy w dalszej części. Wykres częściowej autokorelacji potwierdza nam powyższe wnioski.

2.3.8 Sprawdzenie zachowania niezależności uzyskanych szeregów - estymatory odporne

Ze względu na fakt, że w danych może występować ukryta zależność, której funkcje ACF i PACF mogą "nie zauważać"postanowiłyśmy przenalizować odporne estymatory funkcji autokorelacji. W tym celu wykorzystamy następujące estymatory:

1. Quadrant correlation - mierzy zależność między danymi poprzez analizę, w którym kwadrancie układu współrzędnych leży dana para punktów.

$$\hat{\rho}_{x*}(h) = \frac{1}{n-h}\sum_{i=1}^{n-h} sign\left((x_i - \hat{\mu})(x_{i+h} - \hat{\mu})\right)$$

$$\hat{\rho}_x(h) = sin\left(\frac{\pi\hat{\rho}_{x*}(h)}{2}\right),$$

   gdzie funkcja $signx$ określa znak x, n to liczba obserwacji, h to opóźnienie między wartościami szeregu, a $\hat{\mu}$ to mediana.

2. Korelacja Spearmana - mierzy siłę i kierunek zależności rangowej między danymi, poprzez przydzielenie odpowiednich rang obserwacjom, a następnie obliczenie korelacji pomiędzy tymi rangami.

$$\hat{\rho}_{x*}(h) = \frac{\sum_{i=1}^{n-h}(r_t - \overline{r})(r_{t+h} - \overline{r})}{\sum_{i=1}^{n}(r_t - \overline{r})^2}$$

$$\hat{\rho}_x(h) = 2sin\left(\frac{\pi\hat{\rho}_{x*}(h)}{6}\right),$$

   gdzie funkcja r_t to ranga X_t, $\overline{r}$ to średnia z rang, n to liczba obserwacji, h to opóźnienie między wartościami szeregu.

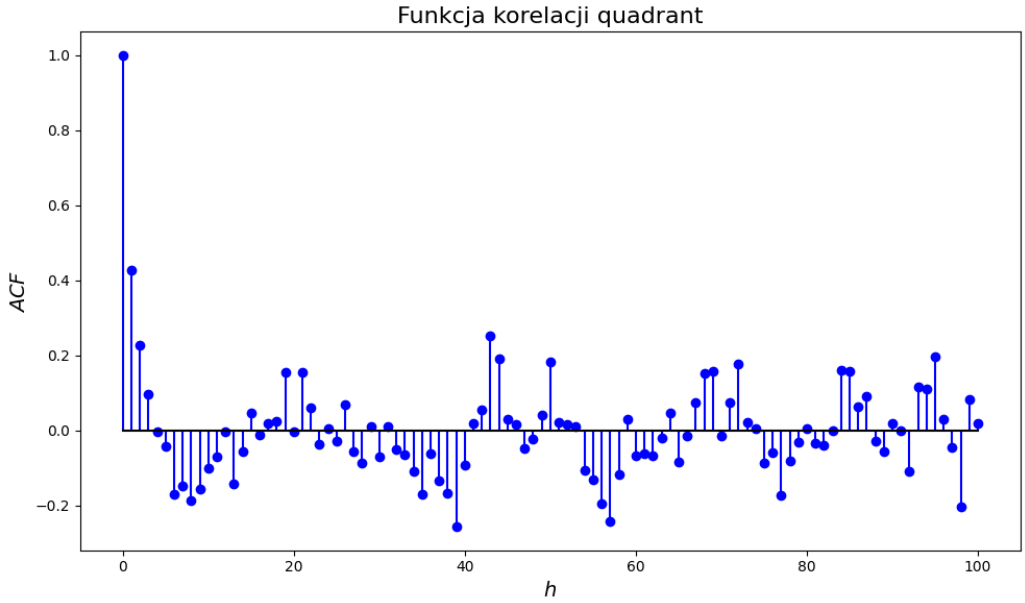
3. Korelacja Kendalla - bada proporcję par punktów, które mają taką samą kolejność rang, do wszystkich punktów.

$$\hat{\rho}_{x*}(h) = \frac{1}{(n-h)(n-h-1)}\sum_{j=1}^{n-h-1}\sum_{i=j+1}^{n-h} sign\left((x_i - x_j)(x_{i+h} - x_{j+h})\right)$$

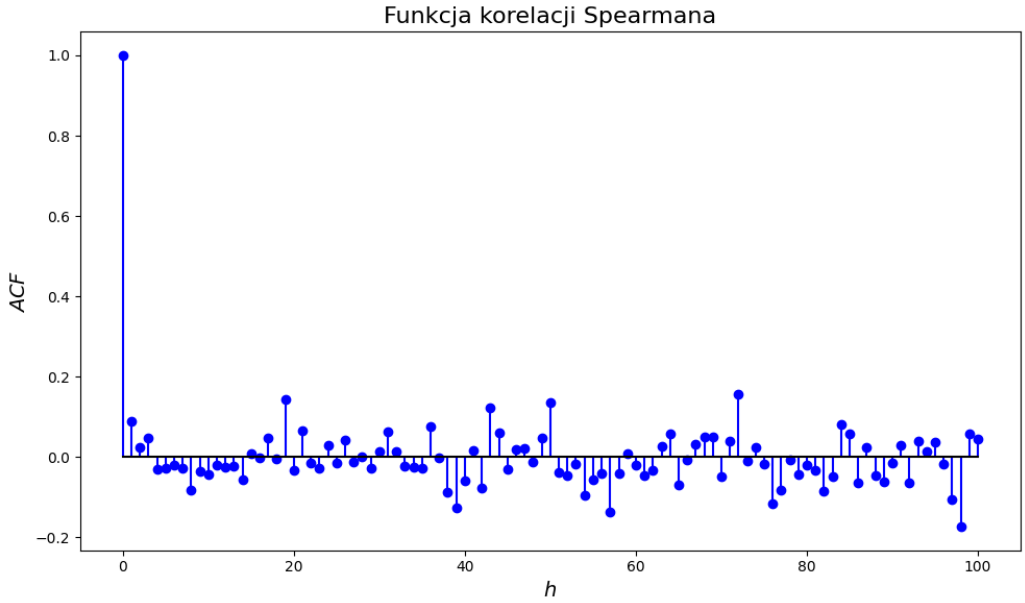
$$\hat{\rho}_x(h) = sin\left(\frac{\pi\hat{\rho}_{x*}(h)}{2}\right),$$

   gdzie funkcja $signx$ określa znak x, n to liczba obserwacji, h to opóźnienie między wartościami szeregu.

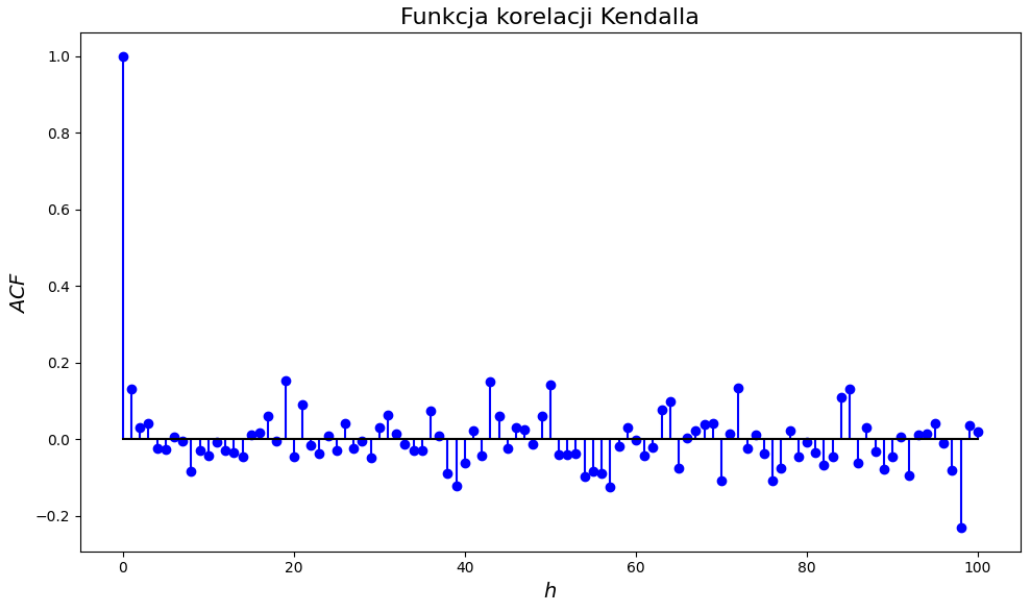
Wszystkie wyżej opisane estymatory mogą przyjmować wartości z przedziału $[-1; 1]$, gdzie wartość bliska 1 oznacza dodatnią zależność, bliska -1 ujemną, natomiast bliska 0 wskazuje na brak zależności. Wyniki zostały przedstawione na wykresach 15 (korelacja quadrant), 16 (korelacja Spearmana), 17 (korelacja Kendalla).


Funkcja korelacji quadrant
ACF
h
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
0
20
40
60
80
100

Rys. 15


Funkcja korelacji Spearmana
ACF
h
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
0
20
40
60
80
100

Rys. 16



Rys. 17

Na rysunku 15 widzimy, że funkcja korelacji quadrant wykazuje zależność między danymi, może się jednak zdarzyć, że ten estymator zawodzi, dlatego analizując wykres 16 i 17 możemy stwierdzić, że dane po różnicowaniu najprawdopodobniej są niezależne.

2.3.9 Sprawdzenie stacjonarności uzyskanych szeregów - test ADF

W celu potwierdzenia naszych hipotez na temat doprowadzenia szeregu do postaci stacjonarnej metodami dekompozycji Walda oraz różnicowania ponownie przeprowadziłyśmy test ADF. Tym razem także przyjęłyśmy poziom istotności $\alpha = 0.05$. Zarówno w przypadku dekompozycji Walda, jak i różnicowania otrzymaliśmy p-wartość bliską 0. Pozwala nam to stwierdzić, że uzyskane dane najprawdopodobniej są stacjonarne. Dzięki temu, w dalszej części będziemy mogły podjąć próbę dopasowania modelu ARMA(p,q) do danych opisujących ceny otwarcia amerykańskiego przedsiębiorstwa Amazon.

3 Modelowanie danych przy pomocy modelu ARMA

Model ARMA(p,q) (ang. *autoregressive moving average*) to autoregresyjny model średniej ruchomej. Ma on zastosowanie w opisie szeregów czasowych. Można go opisać równaniem:

$$X_t - \phi_1 X_{t-1} - ... - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + ... + \theta_q Z_{t-q},$$

gdzie:

- $\phi(z) = 1 - \phi_1 z - ... - \phi_p z^p$ - wielomian autoregresyjny szeregu czasowego X_t,
- $\theta(z) = 1 + \theta_1 z + ... + \theta_q z^q$ - wielomian średniej ruchomej szeregu,
- Z_t - biały szum, czyli ciąg nieskorelowanych zmiennych o średniej 0 i wariancji σ^2,
- p - rząd modelu AR, odpowiadający części autoregresyjnej,

- q - rząd modelu MA, odpowiadający części dotyczącej średniej ruchomej.

Wielomiany $\phi(z)$ i $\theta(z)$ nie mają wspólnych pierwiastków. W celu znalezienia optymalnych parametrów wykorzystałyśmy bibliotekę statsmodels.tsa.arima.model.ARIMA.

3.1 Dobranie rzędu modelu - kryteria informacyjne

Pierwszym krokiem w dobraniu modelu ARMA jest znalezienie rzędu p i q. Do tego celu wykorzystałyśmy 3 różne kryteria informacyjne:

1. Kryterium informacyjne AIC (*Akaike information criterion*)

   $$AIC(p,q) = -2ln(L) + 2(p+q),$$

   gdzie L to funkcja największej wiarygodności, a p i q to parametry modelu ARMA(p,q). Kryterium to uwzględnia zarówno poprawność dopasowania do danych, jak i liczbę parametrów.

2. Kryterium informacyjne BIC/kryterium Schwarza (*Bayesian information criterion/Schwarz information criterion*)

   $$BIC(p,q) = -2ln(L) + (p+q)ln(n),$$

   gdzie L to funkcja największej wiarygodności, n to liczba obserwacji, a p i q to parametry modelu ARMA(p,q). Kryterium to opiera się na podejściu bayesowskim. Jest ono bardziej rygorystyczne w kwestii zbyt skomplikowanych modeli, niż AIC.

3. Kryterium informacyjne HQIC (*Hannan-Quinn information criterion*)

   $$HQIC(p,q) = -2ln(L) + 2(p+q)ln(ln(n)),$$

   gdzie L to funkcja największej wiarygodności, n to liczba obserwacji, a p i q to parametry modelu ARMA(p,q). Kryterium to jest kompromisem pomiędzy AIC i BIC.

W przypadku każdego z wyżej wymienionych kryteriów, im mniejsza jego wartość, tym lepsze dopasowanie.

3.1.1 Dla szeregu po dekompozycji Walda

Dla szeregu po dekompozycji Walda, analizując wszystkie wyżej wspomniane kryteria informacyjne otrzymałyśmy parametry: $p = 1, q = 0$. Czyli znaleziony model to AR(1). Zatem rozważany szereg jest postaci:

$$X_t - \phi_1 X_{t-1} = Z_t$$

Potwierdza to nasze obserwacje, że w szeregu otrzymanym tą metodą występuje zależność.

3.1.2 Dane szeregu po różnicowaniu

Dla szeregu po różnicowaniu również przeanalizowałyśmy kryteria informacyjne. W przypadku AIC i HQIC otrzymałyśmy wartości $p = 0, q = 1$, czyli model MA(1). Zatem rozważany szereg jest postaci:

$$X_t = Z_t + \theta_1 Z_{t-1}.$$

Natomiast analizując kryterium BIC otrzymałyśmy $p = q = 0$, co odpowiada modelowi ARMA(0,0), czyli białemu szumowi. W tym przypadku także potwierdzają się nasze wcześniejsze wnioski.

3.2 Estymacja parametrów modelu

Kiedy już wiemy ile parametrów będzie w modelach możemy przejść do ich estymacji. W tym celu wykorzystamy funkcję ARIMA.fit, która dopasuje parametry do danych.

3.2.1 Model AR(1)

Dla szeregu po dekompozycji Walda otrzymałyśmy wartość $\phi_1 \approx 0.865$, czyli nasz szereg jest postaci:

$$X_t - 0.865X_{t-1} = Z_t$$

3.2.2 Model MA(1)

Dla szeregu po różnicowaniu, dla parametrów wyznaczonych kryterium AIC otrzymałyśmy wartość $\theta_1 \approx -0.118$, czyli nasz szereg jest postaci:

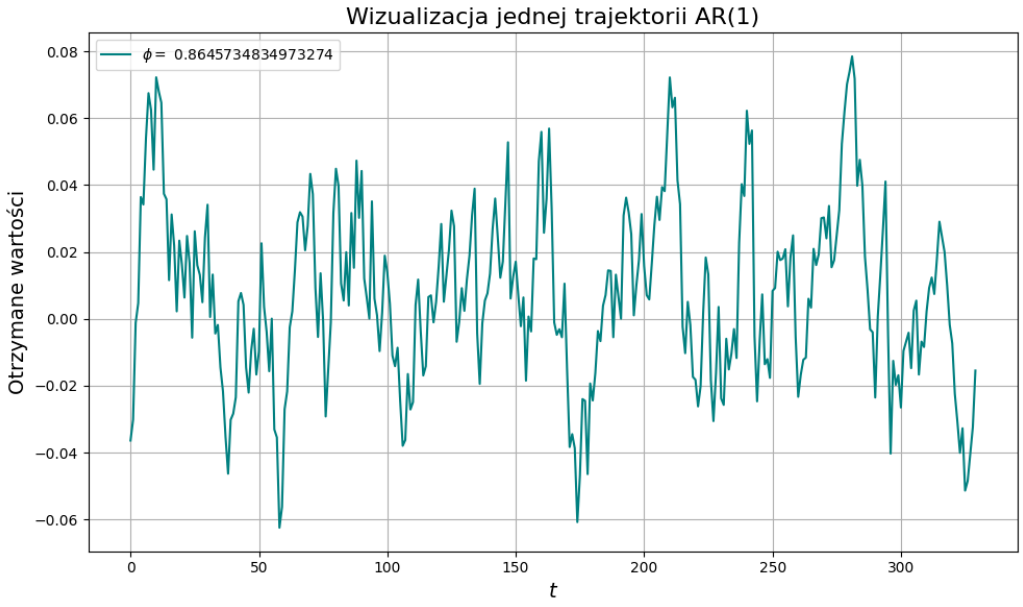
$$X_t = Z_t - 0.118Z_{t-1}$$

3.2.3 Model ARMA(0,0)

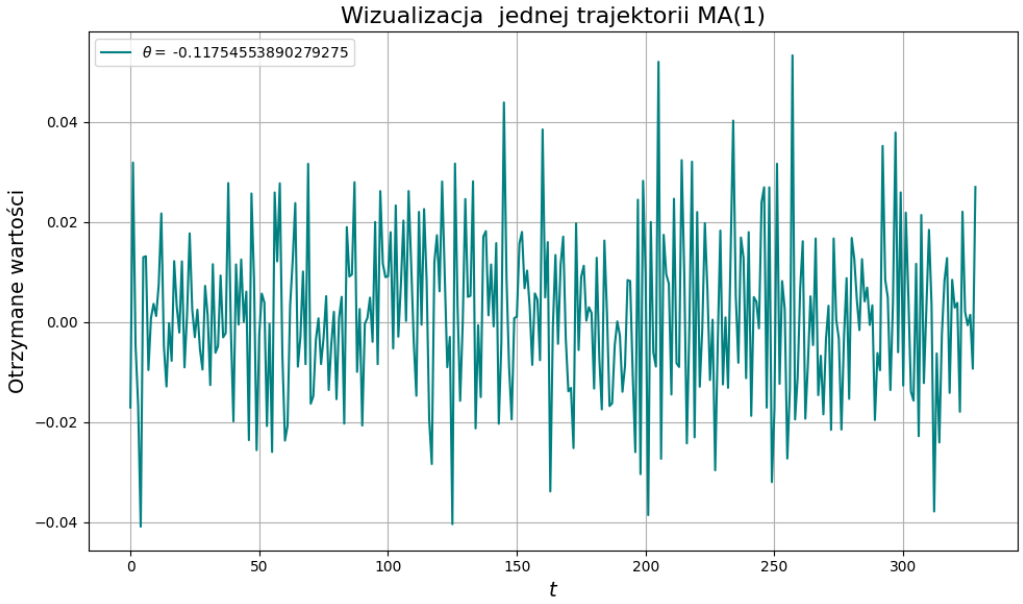
Dla szeregu po różnicowaniu, dla parametrów wyznaczonych kryterium BIC i HQIC otrzymałyśmy biały szum, czyli nieskorelowaną zmienną losową o zerowej średniej i stałej wariancji.

3.3 Porównanie trajektorii dobranych modeli

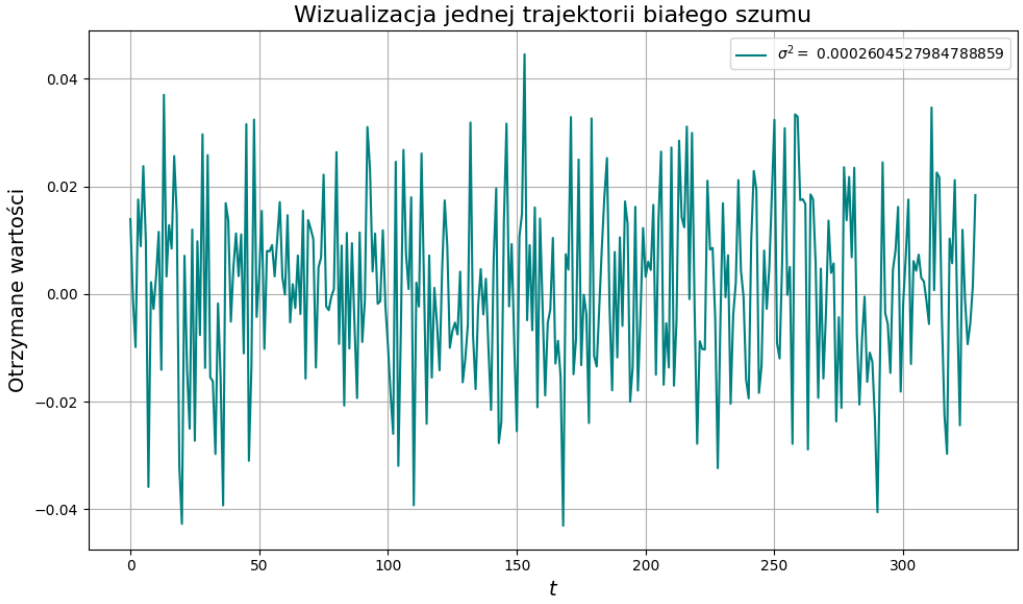
W celu porównania wyglądu otrzymanych szeregów narysowałyśmy po jednej trajektorię modeli AR(1), MA(1) i ARMA(0,0) z odpowiednimi parametrami. Są one widoczne na wykresach numer 18 (AR(1)), 19 (MA(1)), 20 (ARMA(0,0)) .



Rys. 18


Wizualizacja jednej trajektorii MA(1)
θ = -0.11754553890279275
Otrzymane wartości
t
0.04
0.02
0.00
−0.02
−0.04
0
50
100
150
200
250
300

Rys. 19


Wizualizacja jednej trajektorii białego szumu
σ² = 0.0002604527984788859
Otrzymane wartości
t
0.04
0.02
0.00
−0.02
−0.04
0
50
100
150
200
250
300

Rys. 20

4 Ocena dopasowania modeli

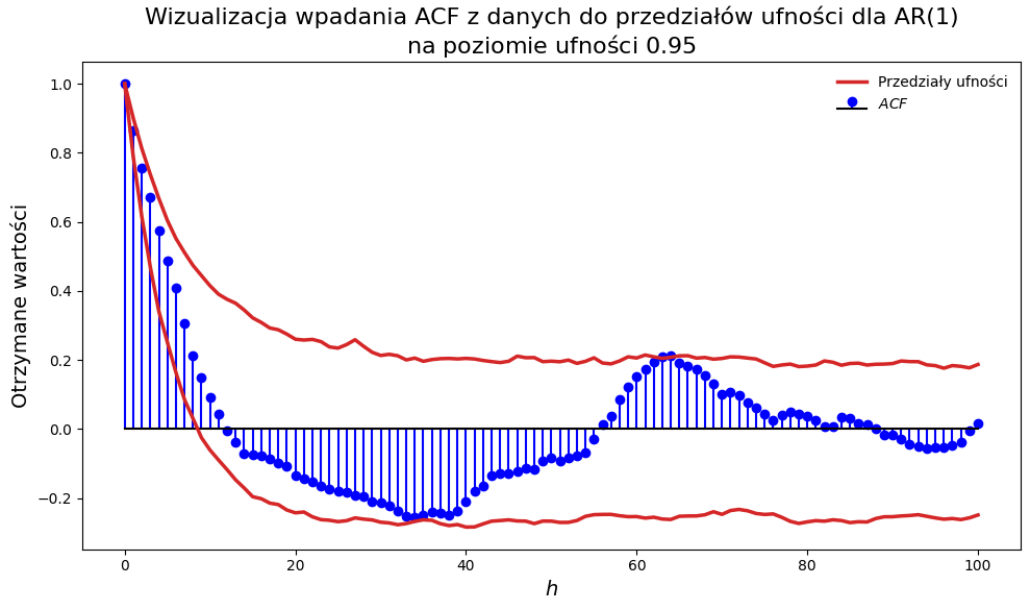
Następnym krokiem, który wykonałyśmy było sprawdzenie jakości dopasowania modelu do danych dotyczących cen otwarcia. W tym celu wyznaczyłyśmy przedziały ufności dla funkcji ACF i PACF, porównałyśmy linie kwantylowe z trajektoriami oraz sprawdziłyśmy jak wyznaczone modele prognozują przyszłe obserwacje.

4.1 Przedziały ufności dla funkcji PACF i ACF

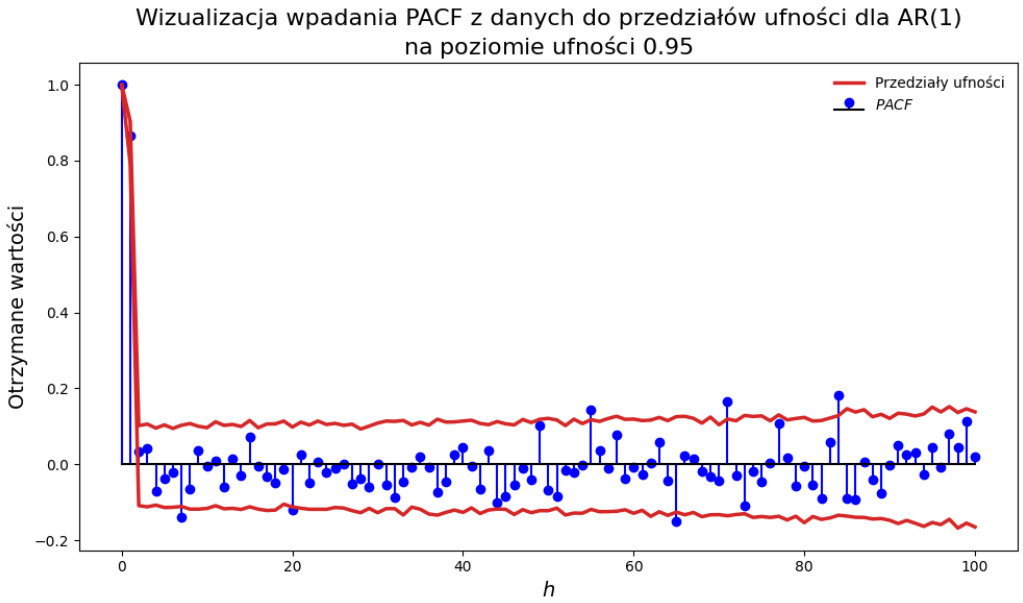
Przedziały ufności wyznaczyłyśmy dla $\alpha = 0.05$, za pomocą metody Monte Carlo dla 100 kroków. Dla każdej z nich policzyłyśmy funkcję autokorelacji i częściowej autokorelacji dla $h = 100$. Następnie sprawdziłyśmy czy wykresy funkcji ACF i PACF wpadają do wyznaczonych przedziałów.

4.1.1 Model AR(1)

Na wykresie 21 (ACF) i 22 (PACF) zostało zwizualizowane wpadanie odpowiednich funkcji do przedziałów ufności dla szeregu AR(1) otrzymanego dla danych po dekompozycji Walda.



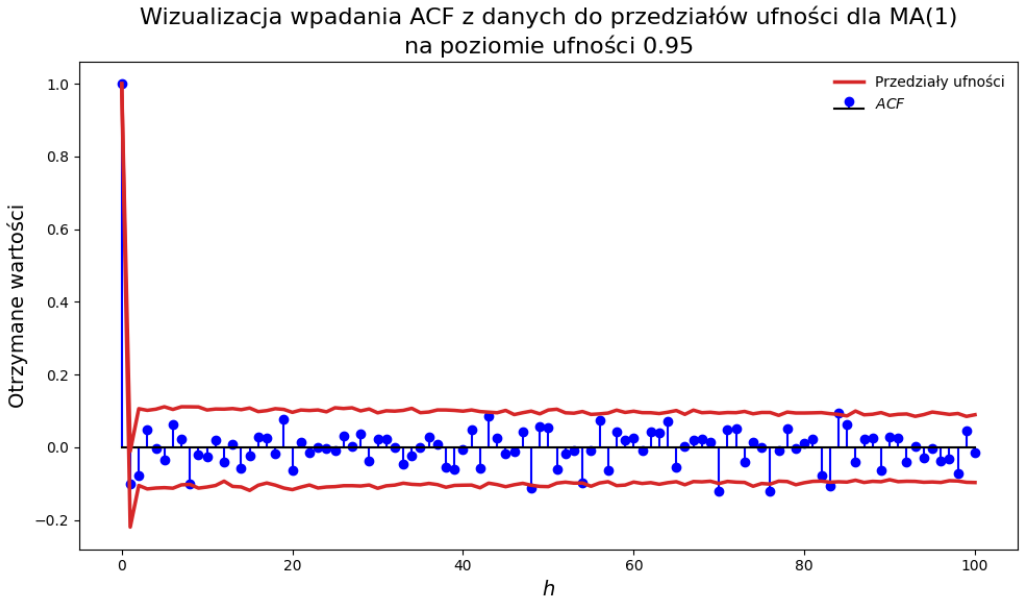
Rys. 21



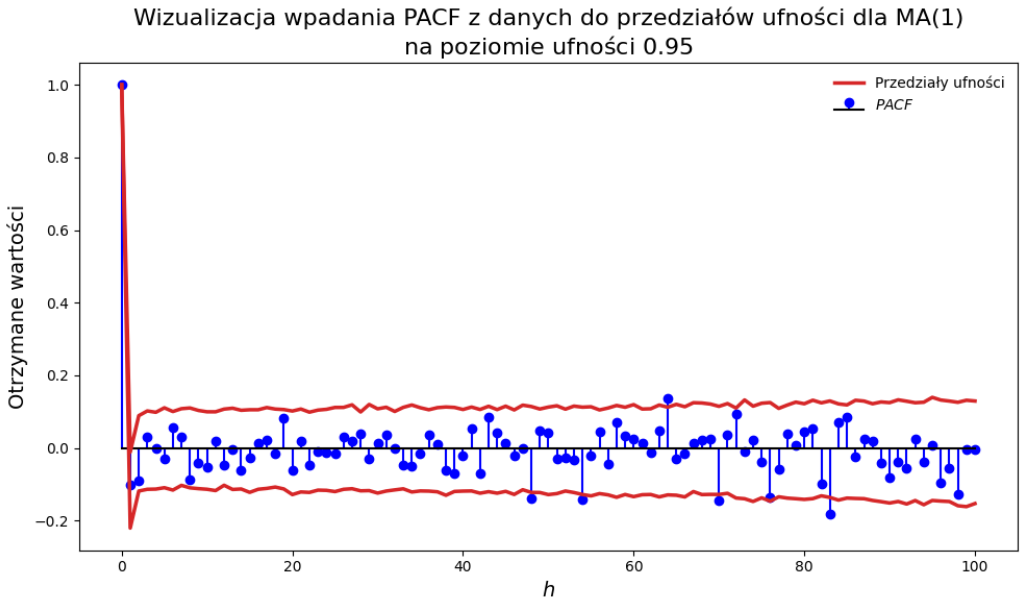
Rys. 22

4.1.2 Model MA(1)

Na wykresie 23 (ACF) i 24 (PACF) zostało zwizualizowane wpadanie odpowiednich funkcji do przedziałów ufności dla szeregu MA(1) otrzymanego dla danych po różnicowaniu.



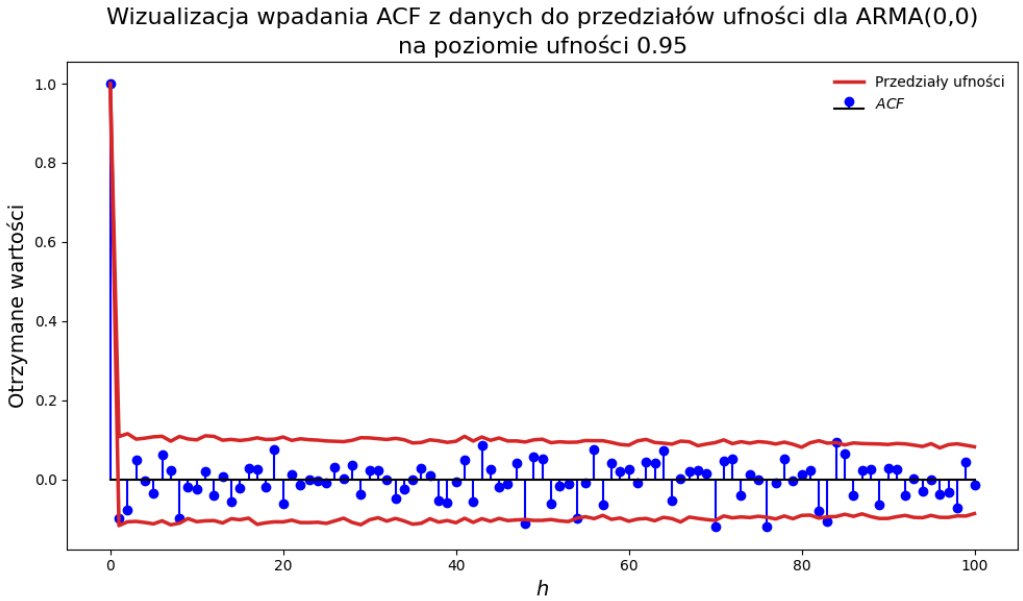
Rys. 23



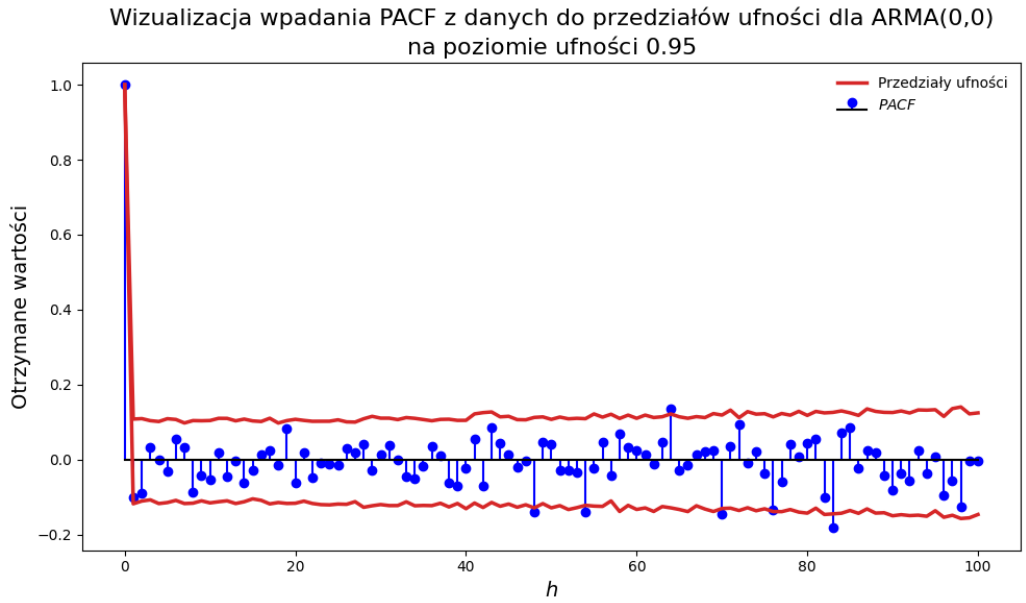
Rys. 24

4.1.3 Model ARMA(0,0)

Na wykresie 25 (ACF) i 26 (PACF) zostało zwizualizowane wpadanie odpowiednich funkcji do przedziałów ufności dla szeregu ARMA(0,0) otrzymanego dla danych po różnicowaniu.



Rys. 25



Rys. 26

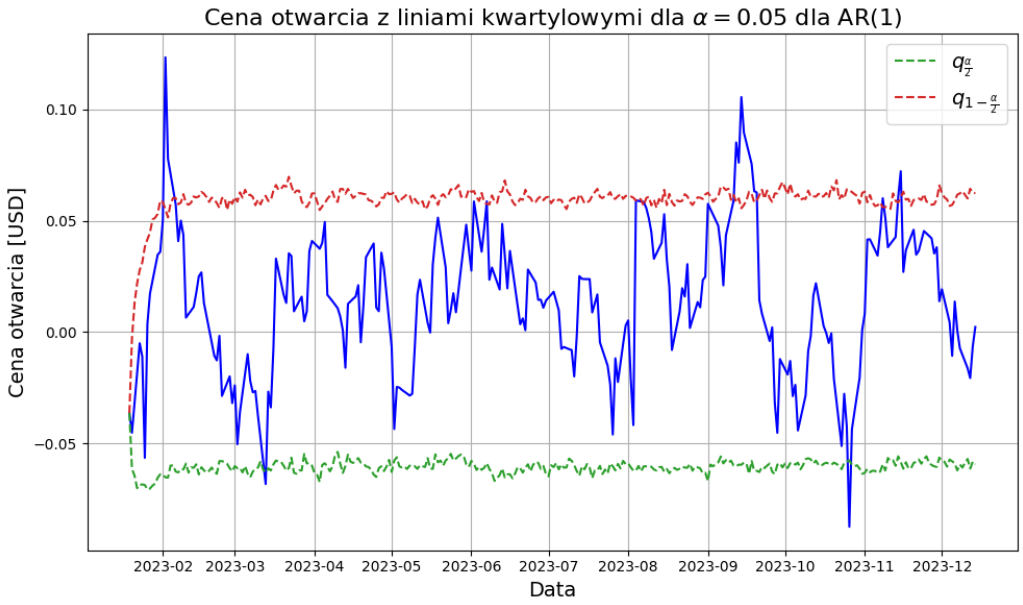
Jak możemy zauważyć na wykresach 21 - 26 prawie wszystkie wartości funkcji autokorelacji znajdują się w wyznaczonych przedziałach. Pozwala nam to uznać, że w języku obu metryk nasze dane dobrze wpasowują się w wyznaczone modele. Dodatkowo możemy zauważyć, że model MA(1) zachowuje się bardzo podobnie do modelu ARMA(0,0).

4.2 Porównanie linii kwantylowych z trajektoriami

Kolejnym krokiem oceny jakości dopasowania modeli było porównanie trajektorii wyznaczonych szeregów z liniami kwantylowymi dla odpowiednich modeli. W tym celu także wykorzystaliśmy metodę Monte Carlo, dla liczby kroków równej 1000.

4.2.1 Model AR(1)

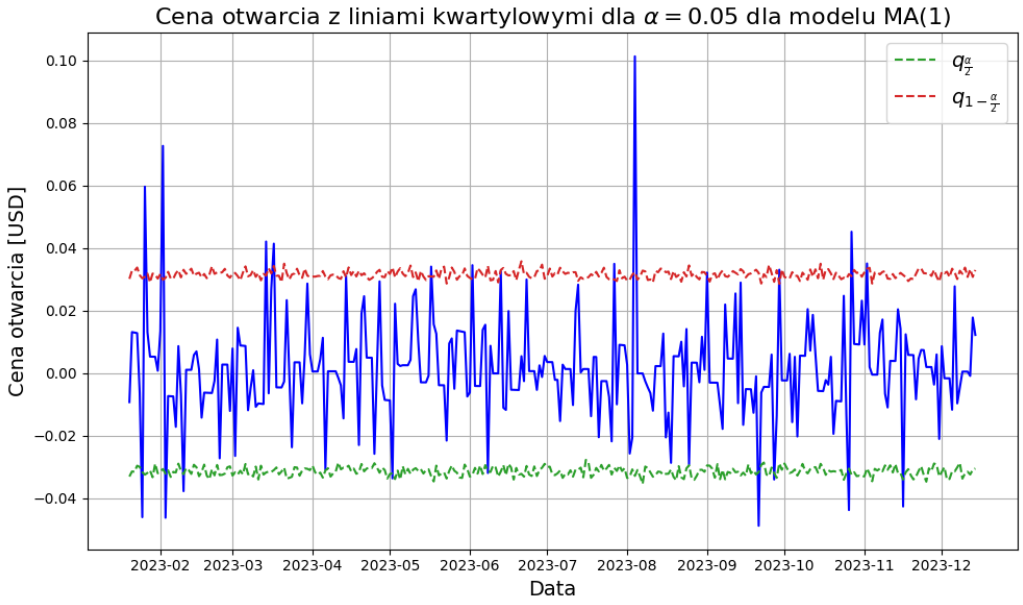
Na wykresie 27 zostały zwizualizowane ceny otwarcia wraz z liniami kwantylowymi dla modelu AR(1).



Rys. 27

4.2.2 Model MA(1)

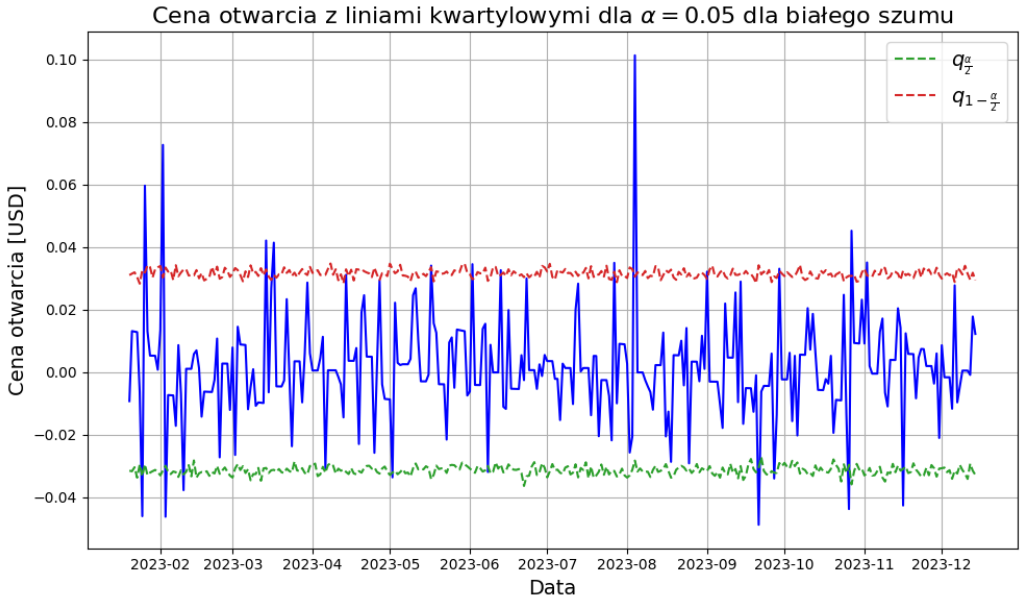
Na wykresie 28 zostały zwizualizowane ceny otwarcia wraz z liniami kwantylowymi dla modelu MA(1).



Rys. 28

4.2.3 Model ARMA(0,0)

Na wykresie 29 zostały zwizualizowane ceny otwarcia wraz z liniami kwantylowymi dla modelu ARMA(0,0).



Rys. 29

Jak możemy zauważyć, prawie wszystkie ceny otwarcia mieszczą się w przedziałach wyznaczonych przez linie kwantylowe dla $\alpha = 0.05$. Pojedyncze wyskoki nie przeszkadzają w stwierdzeniu, że modele najprawdopodobniej zostały dobrze dobrane.

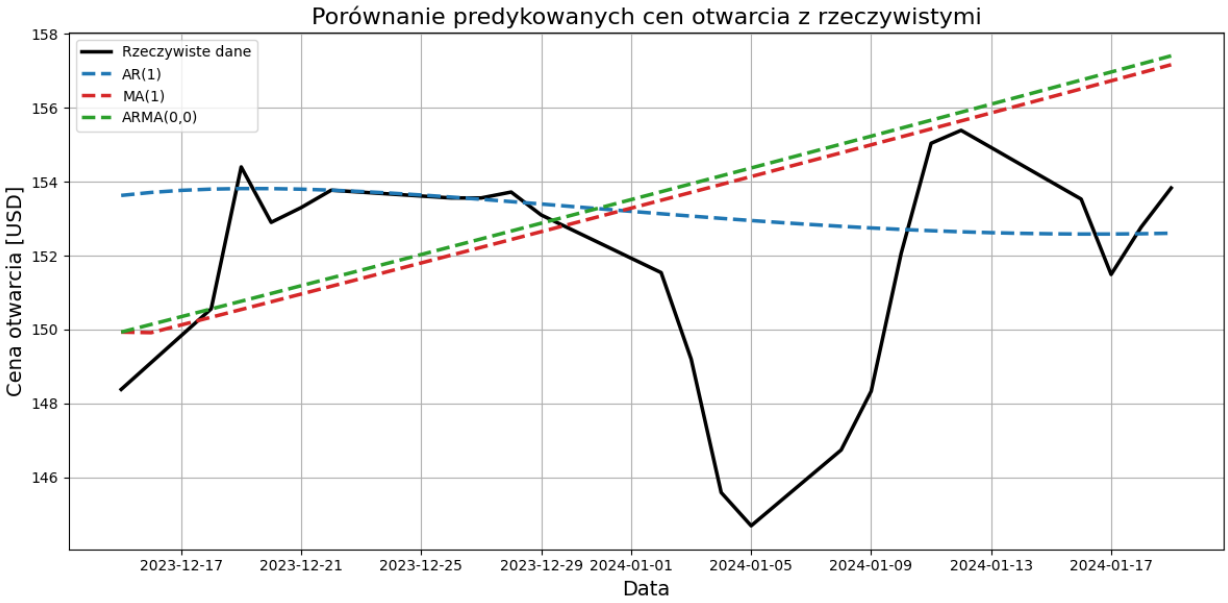
Dodatkowo widzimy, że model MA(1) zachowuje się bardzo podobnie do modelu ARMA(0,0). Potwierdza to fakt, że zależność w modelu MA(1) jest niewielka (na co wskazuje również wartość parametru $\theta_1 \approx -0.118$, która jest dość mała).

4.3 Prognoza dla przyszłych obserwacji i porównanie ich z rzeczywistymi danymi

W punkcie 2.2 wyodrębniłyśmy 10% danych do zbioru testowego. Następnie na podstawie pozostałych 90% prognozowałyśmy liniowo najnowsze obserwacje i porównywałyśmy je z tymi ze zbioru testowego. W tym celu wykorzystałyśmy funkcję wbudowaną get_forecast. Następnie dla predykowanych danych wyznaczyłyśmy przedziały ufności na poziomie $\alpha = 0.05$. Wykonałyśmy także transformacje odwrotne predykowanych danych, odpowiednio:

- dla modelu $AR(1)$ wykonałyśmy odwrotność dekompozycji, czyli dodałyśmy trend oraz liniowość, a następnie nałożyłyśmy funkcję *exp* z biblioteki numpy (odwrotność logarytmowania),
- dla modelu $MA(1)$ oraz $ARMA(0,0)$ wykonałyśmy procedurę odwrotnego różnicowania, następnie także nałożyłyśmy funkcję *exp*.

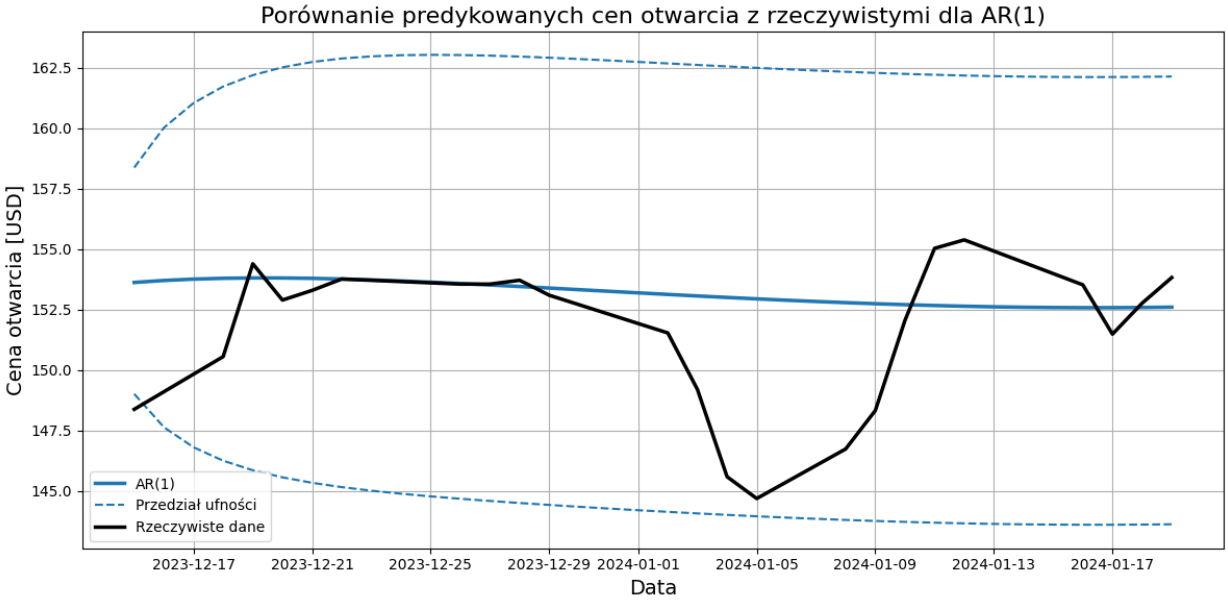
Przekształciłyśmy również odpowiednio wyliczone wcześniej przedziały ufności. Na wykresie 30 widoczne jest porównanie predykowanych cen otwarcia z rzeczywistymi. Natomiast na wykresach 31-33 zostało ukazane wpadanie predykowanych wartości do przedziałów ufności dla odpowiednich modeli.



Rys. 30

Jak możemy zauważyć dane predykowane w modelu MA(1) zachowują się bardzo podobnie do tych z modelu ARMA(0,0). Potwierdza to nasze wcześniejsze obserwacje.

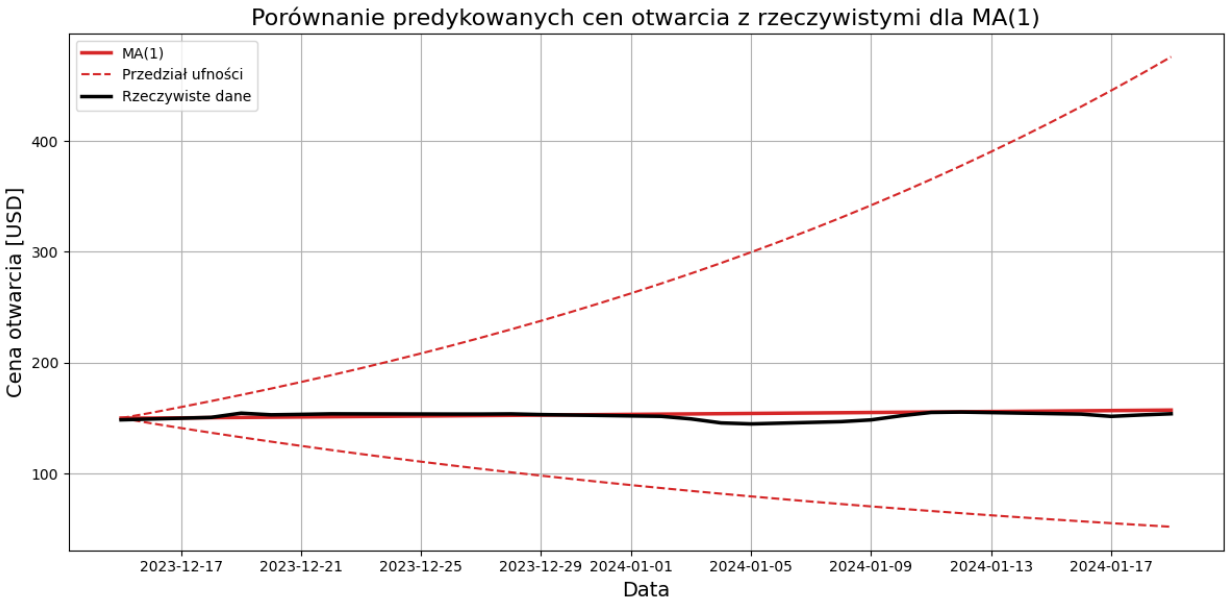
4.3.1 Model AR(1)



Rys. 31

Na wykresie 31 możemy zauważyć, że dla modelu AR(1), czyli tego po dekompozycji Walda, predykowane wartości mieszczą się w przedziałach ufności.

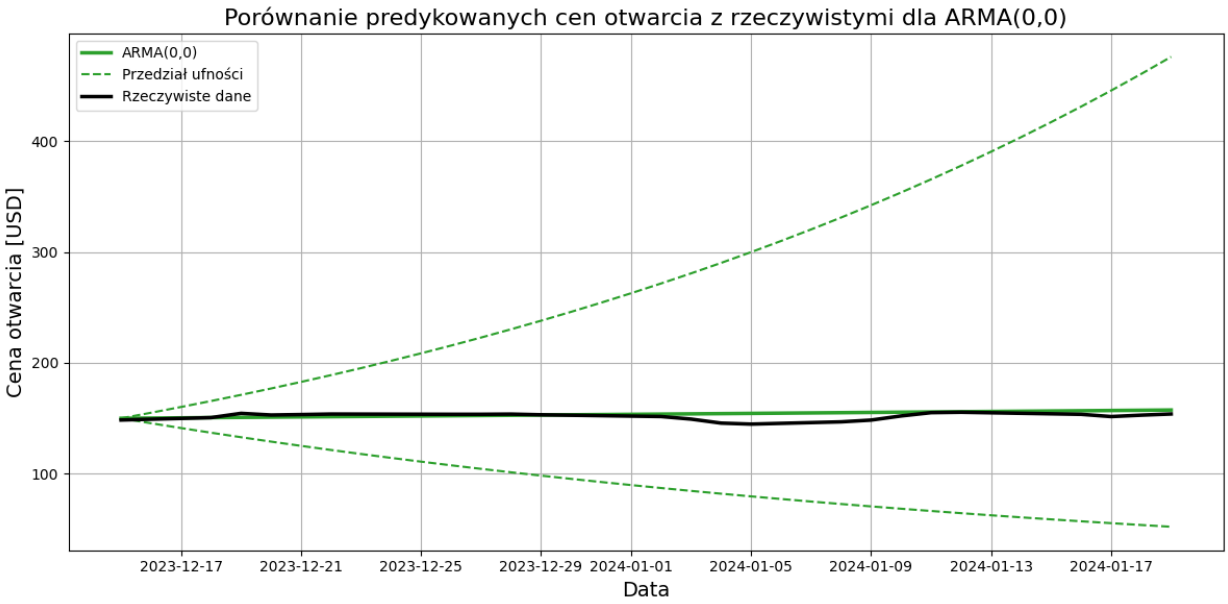
4.3.2 Model MA(1)



Rys. 32

Na wykresie 32 możemy zauważyć, że dla modelu MA(1), czyli tego po różnicowaniu, predykowane wartości także mieszczą się w przedziałach ufności. Jednak w przypadku tego modelu wyznaczone przedziały ufności są dość szerokie, co pozwala nam wnioskować, że model AR(1) lepiej przybliża dane dotyczące cen otwarcia.

4.3.3 Model ARMA(0,0)



Rys. 33

Na wykresie 33 możemy zauważyć, że dla modelu ARMA(0,0), czyli tego po różnicowaniu, predykowane wartości także mieszczą się w przedziałach ufności. W tym przypadku także przedziały się szybko rozszerzają (podobnie jak w przypadku modelu MA(1)), co potwierdza, że model AR(1) najlepiej przybliża nasze dane.

Podsumowując, chcemy, aby przedziały ufności były jak najwęższe i jednocześnie, żeby wpadało do nich jak najwięcej wartości. Zważywszy na ten fakt i analizując wygląd wykresów 31-33 możemy stwierdzić, że modelem, który najlepiej przybliża dane dotyczące cen otwarcia jest model AR(1).

5 Weryfikacja założeń dotyczących szumu

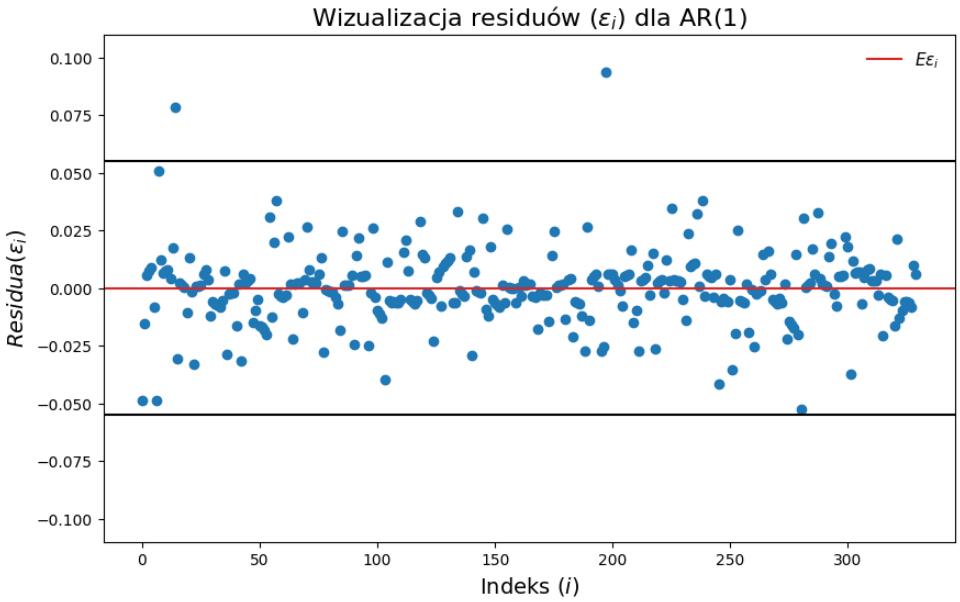
Kolejnym krokiem, który wykonałyśmy w celu zbadania poprawności dopasowanych modeli jest sprawdzenie założeń dotyczących szumu. Założenia te to:

1. $E\epsilon_i = 0, \quad \forall i = 1, ..., n,$
2. $Var\epsilon_i = \sigma^2, \quad \forall i = 1, ..., n,$
3. $\epsilon_1, ...\epsilon_n$ - niezależne,
4. $\epsilon_i \sim \mathrm{N}(0, \sigma^2), \quad \forall i = 1, ..., n.$

5.1 Założenie dotyczące średniej oraz wariancji

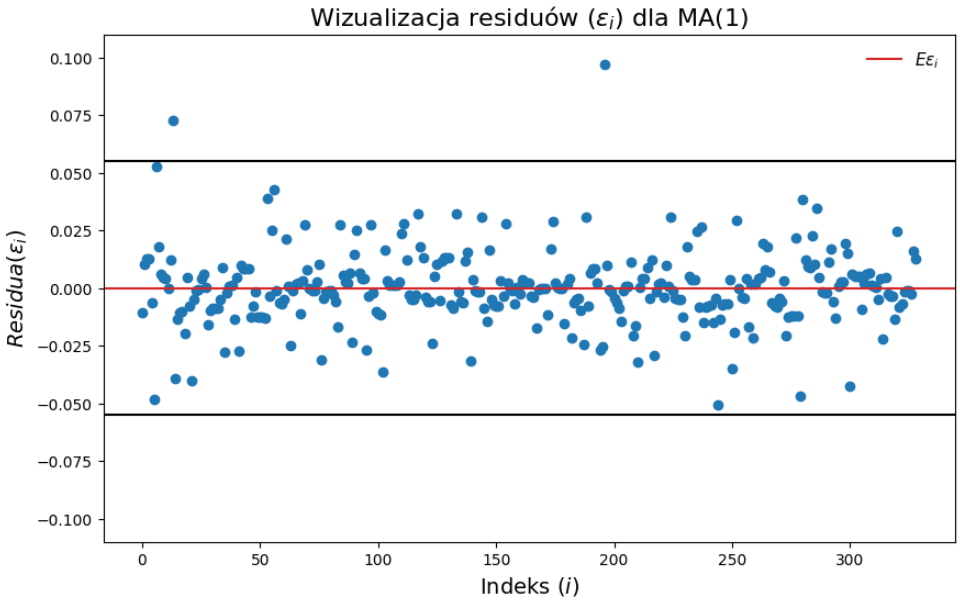
W pierwszej kolejności zbadałyśmy założenia dotyczące średniej i wariancji. W tym celu sporządziłyśmy wykresy residuów w zależności od indeksów czasowych dla każdego z modeli. Wyniki są widoczne na wykresach 34 (model AR(1)), 35 (model MA(1)), 36 (model ARMA(0,0)).

5.1.1 Model AR(1)



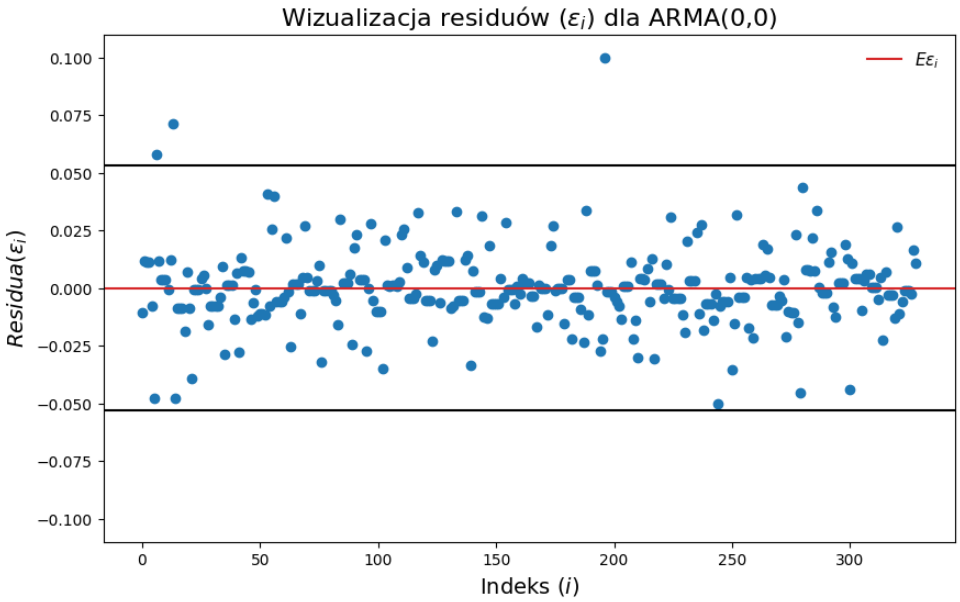
Rys. 34

5.1.2 Model MA(1)



Rys. 35

5.1.3 Model ARMA(0,0)



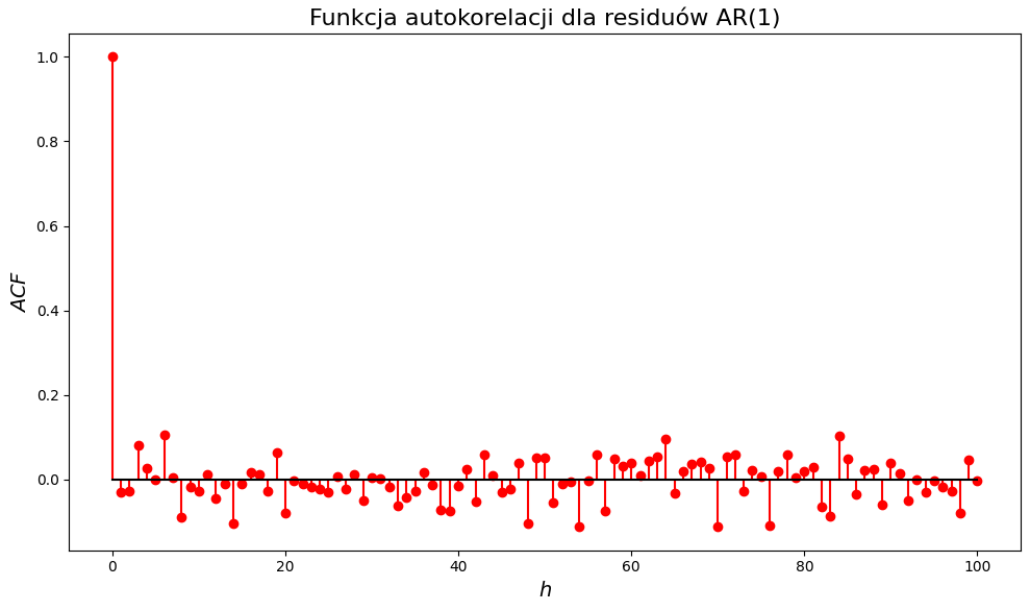
Rys. 36

Jak możemy zauważyć na wykresach 34-36, dla każdego z modeli średnia oscyluje wokół 0, a wariancja jest stała. Są zauważalne pojedyncze wartości odstające, jednak jest ich na tyle mało i na tyle niewiele odstają, że możemy je pominąć w weryfikacji poprawności założeń.

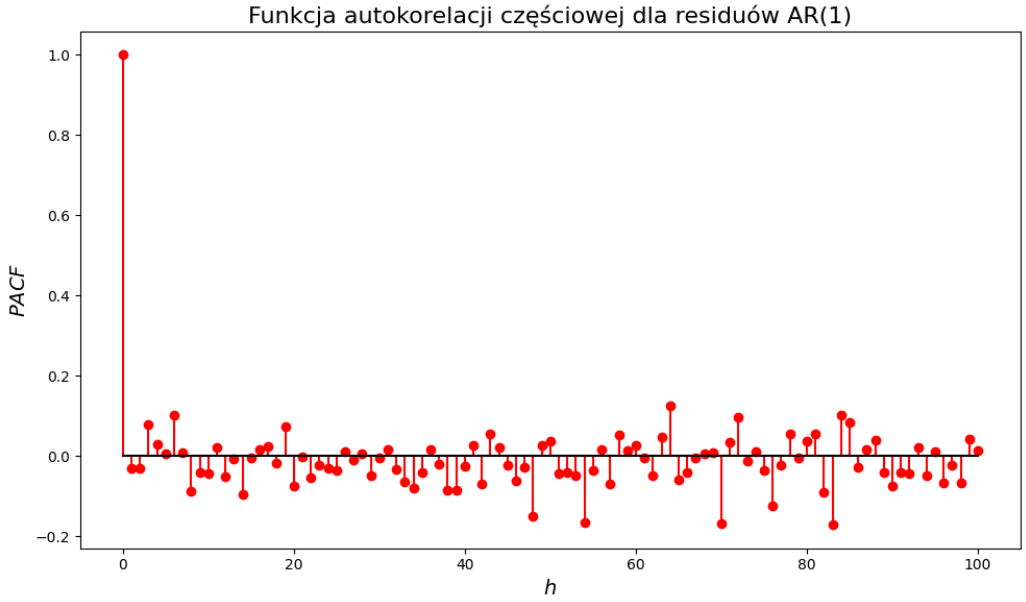
5.2 Założenie dotyczące niezależności

W celu zbadania niezależności residuów wykonałyśmy wykresy funkcji autokorelacji (wykresy 37 (AR(1)), 39 (MA(1)), 41 (ARMA(0,0))) oraz częściowej autokorelacji (wykresy 38 (AR(1)), 40 (MA(1)), 42 (ARMA(0,0))).

5.2.1 Model AR(1)

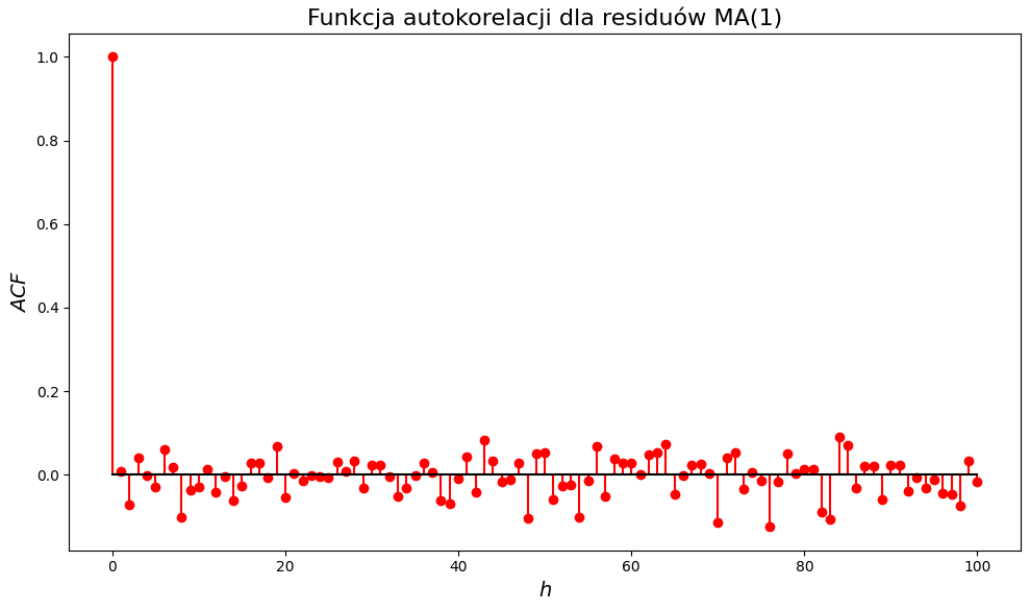


Rys. 37

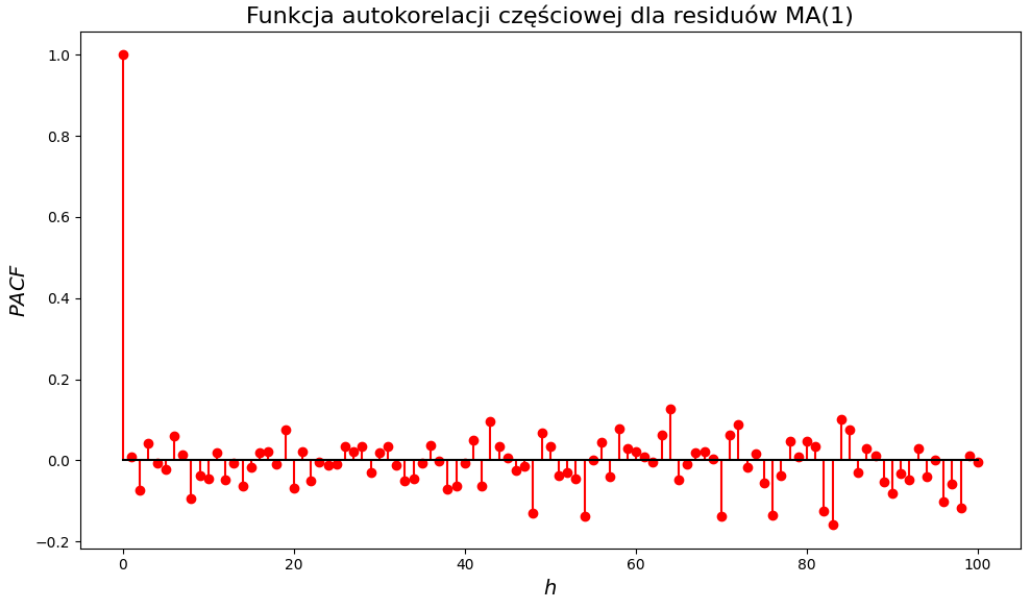


Rys. 38

5.2.2 Model MA(1)

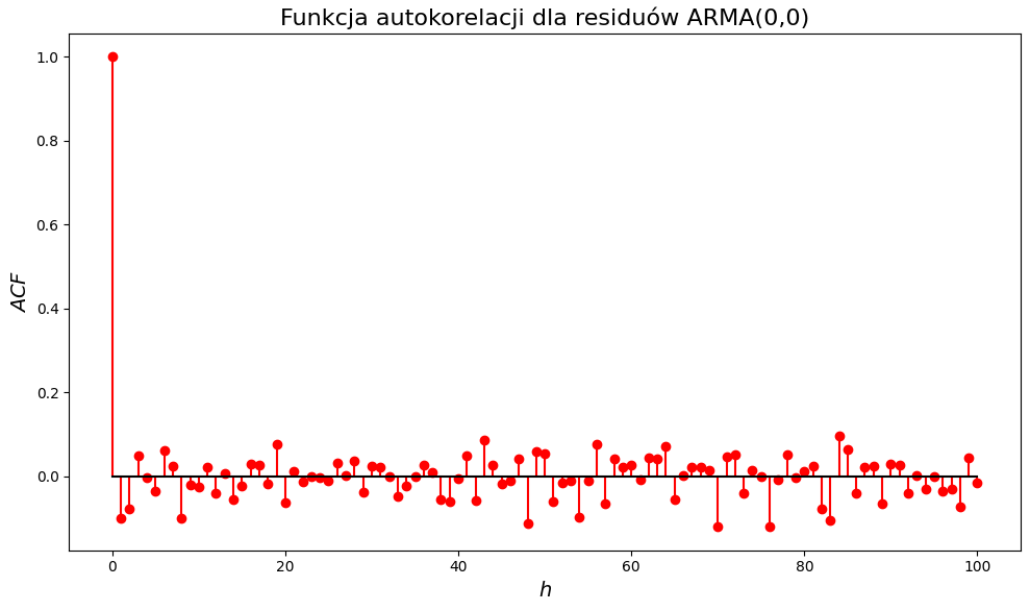


Rys. 39

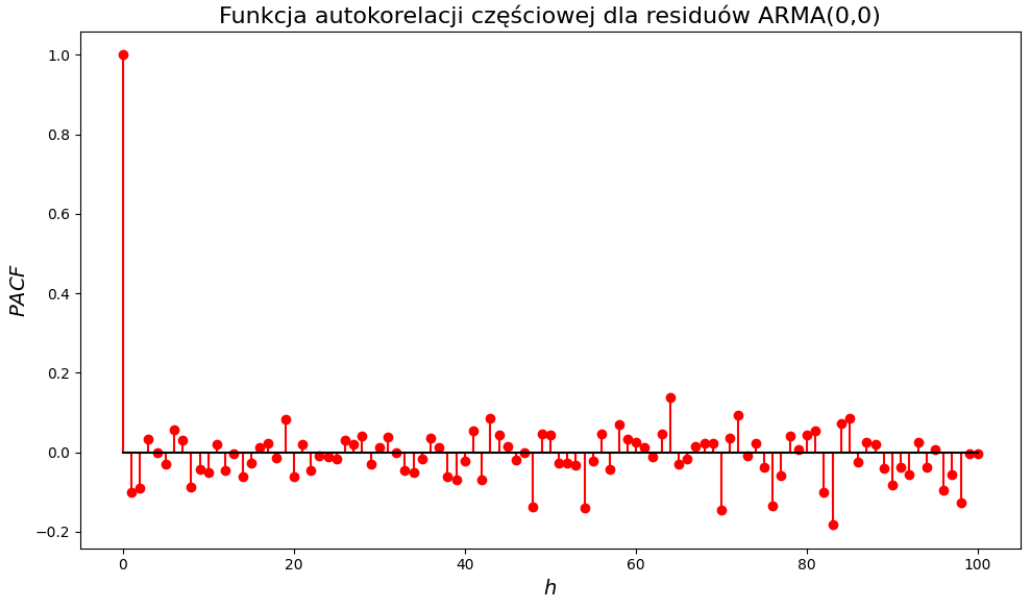


Rys. 40

5.2.3 Model ARMA(0,0)



Rys. 41



Rys. 42

Jak możemy zauważyć na wykresach 37, 39, 41 funkcja autokorelacji, wraz ze wzrostem h oscyluje ciągle wokół 0. Na wykresach 38, 40, 42 widzimy dodatkowo, że funkcja częściowej autokorelacji także oscyluje wokół 0. Fakty te pozwalają nam stwierdzić, że założenie dotyczące niezależności residuów jest spełnione.

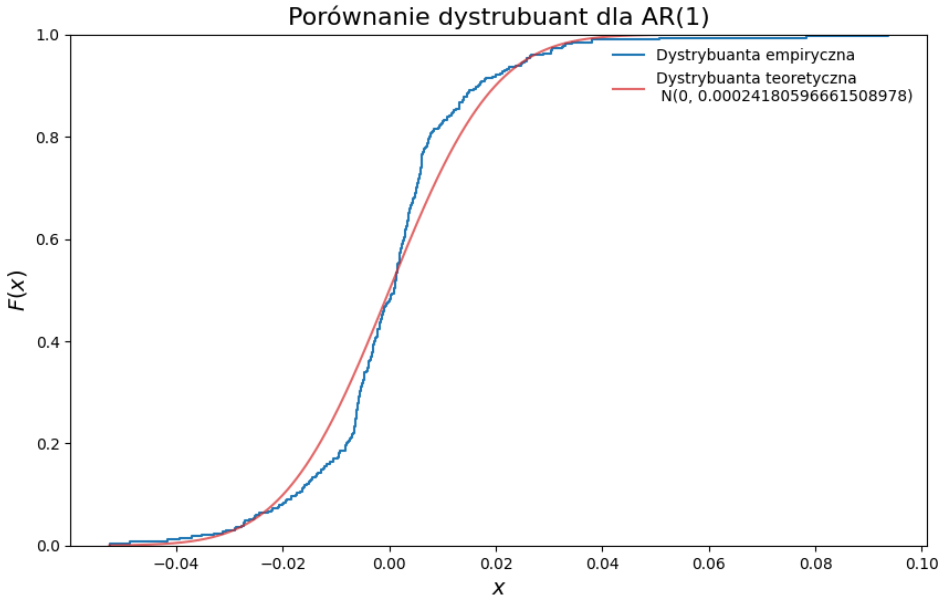
5.3 Założenie dotyczące normalności rozkładu

Ostatnim założeniem dotyczącym szumu jest założenie dotyczące normalności rozkładu. W celu sprawdzenia czy residua mają rozkład normalny wykonałyśmy wykresy porównujące dystrybuanty teoretyczne i empiryczne dla każdego z modeli. Wyniki są widoczne na wykresach 43 (AR(1)), 44 (MA(1)), 45 (ARMA(0,0)). Wykonałyśmy także testy Kołmogorowa-Smirnowa, Shapiro-Wilka oraz Jarque-Bera. W tym celu wykorzystałyśmy funkcje wbudowane z biblioteki scipy.stats. Ich wyniki zostały przedstawione w tabelach 1 (AR(1)), 2 (MA(1)), 3 (ARMA(0,0))

5.3.1 Model AR(1)

Test	Przybliżona p-wartość
Kołmogorowa-Smirnowa	0.0001819
Shapiro-Wilka	$1.767 \cdot 10^{-12}$
Jarque-Bera	$1.227 \cdot 10^{-110}$

Tabela 1

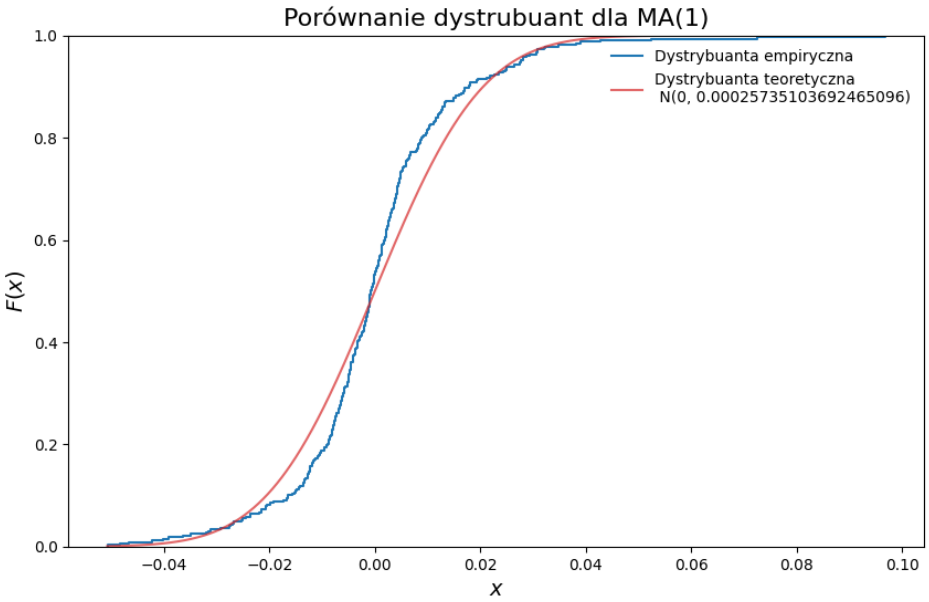


Rys. 43

5.3.2 Model MA(1)

Test	p-wartość
Kołmogorowa-Smirnowa	0.0004439
Shapiro-Wilka	$6.249 \cdot 10^{-12}$
Jarque-Bera	$2.719 \cdot 10^{-99}$

Tabela 2

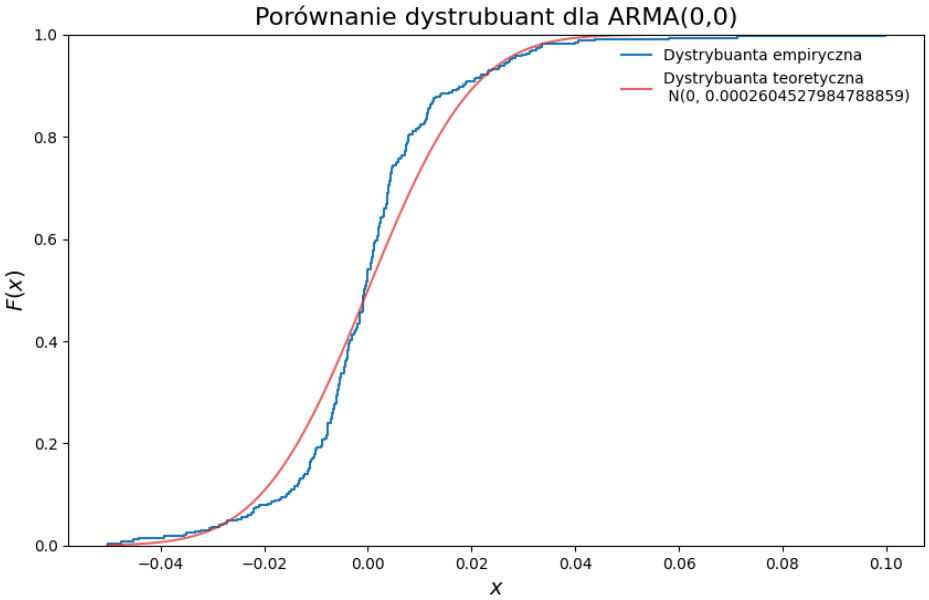


Rys. 44

5.3.3 Model ARMA(0,0)

Test	p-wartość
Kołmogorowa-Smirnowa	$2.726 \cdot 10^{-5}$
Shapiro-Wilka	$4.723 \cdot 10^{-13}$
Jarque-Bera	$2.073 \cdot 10^{-119}$

Tabela 3



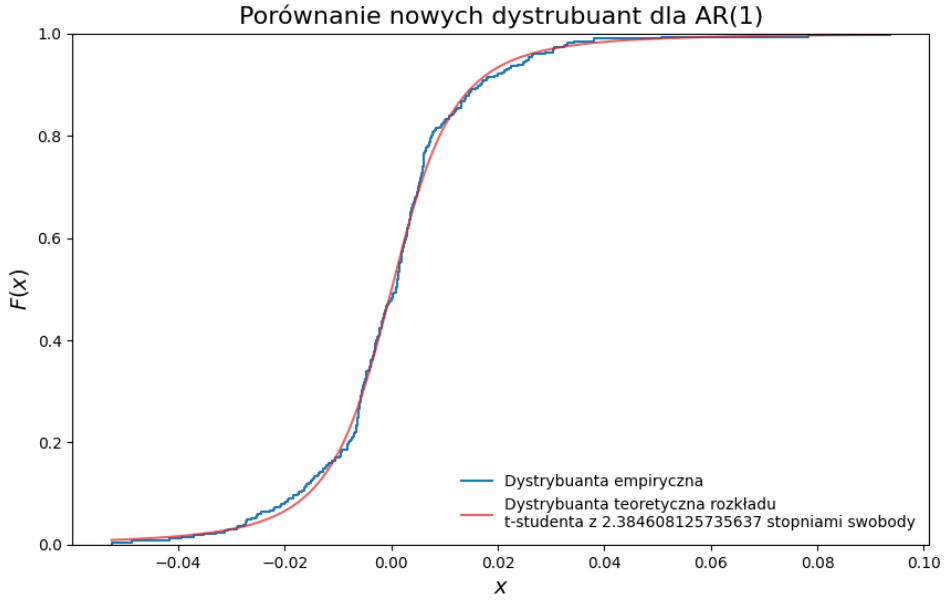
Rys. 45

Jak możemy zauważyć, dla wszystkich modeli dystrybuanty empiryczne nie pokrywają się z dystrybuantami teoretycznymi rozkładu normalnego z odpowiednio dobranymi parametrami. Dodatkowo widzimy, że dla każdego z modeli, p-wartości we wszystkich testach są znacznie mniejsze od $\alpha = 0.05$. Korzystając z wiedzy, że hipoteza zerowa dla każdego z testów mówi nam o tym, że jeśli otrzymamy p-wartość $> \alpha$ to mamy do czynienia z rozkładem normalnym, możemy stwierdzić, że residua w każdym z dobranych przez nas modeli nie mają rozkładu normalnego. Czyli widzimy, że założenie dotyczące normalności rozkładu residuów nie jest spełnione.

5.4 Założenie dotyczące normalności rozkładu - próba dobrania innego rozkładu

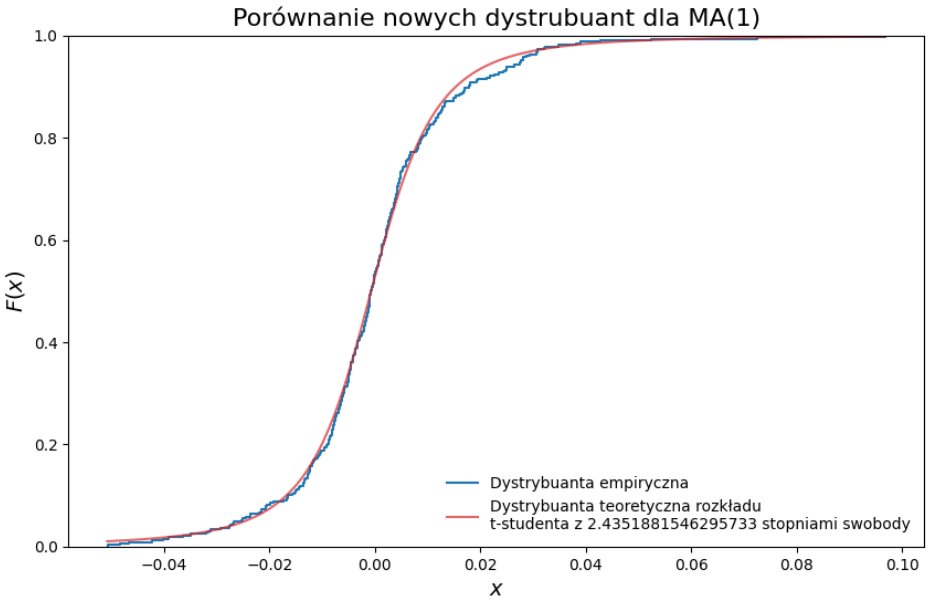
Wiedząc, że residua nie mają rozkładu normalnego postaramy się znaleźć inny, który będzie je jak najdokładniej przybliżał. W tym celu skorzystamy z rozkładu t-Studenta, dla którego dobierzemy odpowiednie parametry, a następnie porównamy dystrybuanty empiryczne i teoretyczne. Wyniki są widoczne na wykresach 46 (AR(1)), 47 (MA(1)), 48 (ARMA(0,0)).

5.4.1 Model AR(1)



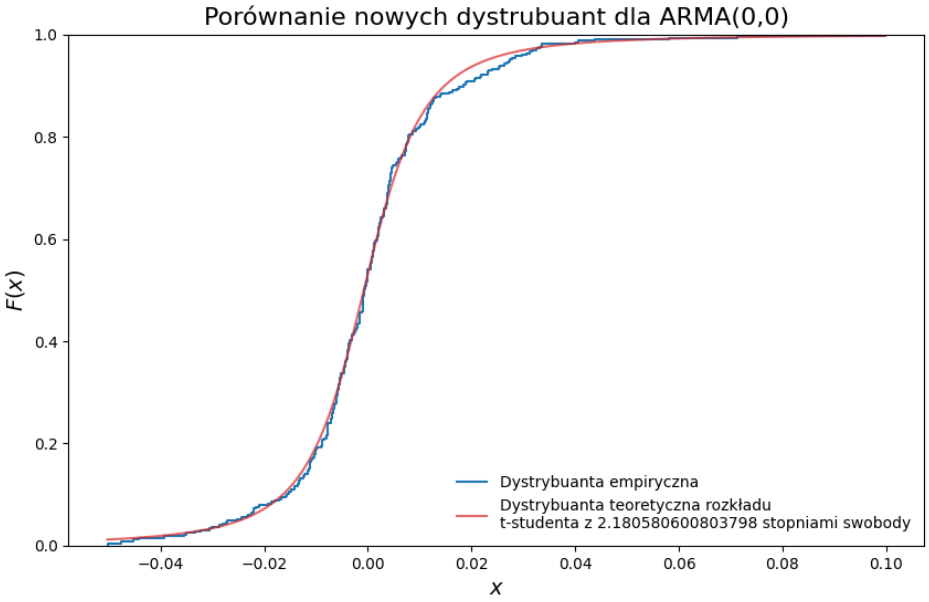
Rys. 46

5.4.2 Model MA(1)



Rys. 47

5.4.3 Model ARMA(0,0)



Rys. 48

Jak możemy zauważyć, dla każdego z modeli dystrybuanty teoretyczne rozkładów t-Studenta pokrywają się z dystrybuantami empirycznymi. Pozwala nam to wnioskować, że residua dla każdego z modeli mają najprawdopodobniej rozkład t-Studenta.

5.5 Weryfikacja założeń dotyczących szumu - podsumowanie

Podsumowując, nie wszystkie założenia dotyczące szumu są spełnione. Zarówno dla modelu po dekompozycji Walda (AR(1)), jak i dla modeli po różnicowaniu (MA(1), ARMA(0,0)) spełnione są założenia dotyczące zerowej średniej oraz stałej wariancji residuów, a także to mówiące o ich niezależności. Nie jest spełnione natomiast założenie dotyczące normalności rozkładów - we wszystkich modelach residua mają najprawdopodobniej rozkład t-Studenta. Wiemy jednak, że rozkład ten również ma średnią 0 i stałą wariancję, czyli jest białym szumem. Fakt ten pozwala nam przyjąć, że dobrane przez nas modele poprawnie przybliżają dane dotyczące cen otwarcia.

6 Podsumowanie, wnioski

Celem powyższego raportu była analiza cen akcji na otwarcie amerykańskiego przedsiębiorstwa Amazon za pomocą modelu ARMA.

W celu przygotowania danych do analizy, w pierwszej kolejności uzupełniłyśmy braki danych wynikające z faktu zamknięcia giełdy w weekendy i święta za pomocą metody interpolacji liniowej. Wyodrębniłyśmy także zbiór testowy, który posłużył nam do późniejszej predykcji.

Kolejnym krokiem, jaki wykonałyśmy było sprawdzenie stacjonarności szeregu. W tym celu przeanalizowałyśmy wykresy funkcji autokorelacji i częściowej autokorelacji oraz wykonałyśmy test ADF. Na wykresach widoczne były liniowość i sezonowość co przemawiało za brakiem stacjonarności. Test ADF to potwierdził.

Wiedząc, że aby móc analizować szereg przy pomocy modelu ARMA musi on być stacjonarny, postanowiłyśmy nałożyć na dane logarytm, a następnie wykonać dekompozycję Walda oraz różnicowanie rzędu 1. Następnie ponownie wykonałyśmy wykresy funkcji ACF i PACF, zarówno dla danych po dekompozycji, jak i tych po różnicowaniu. Analiza wyglądu wyżej wspomnianych wykresów pozwoliła nam stwierdzić, że w danych po dekompozycji Walda występuje znacząca zależność, natomiast w tych po różnicowaniu albo występuje niewielka zależność, albo nie ma jej wcale. Dodatkowo, aby upewnić się co do braku ukrytej zależności wykorzystałyśmy estymatory odporne: Quadrant correlation, korelację Spearmana, korelację Kendalla. Wykonałyśmy także ponownie test ADF. Wszystkie wspomniane metody potwierdziły, że przekształcone dane są stacjonarne, czyli transformacje przyniosły zamierzone efekty.

Następnym krokiem, który wykonałyśmy było dobranie rzędu modelu za pomocą kryteriów informacyjnych AIC, BIC, HQIC odpowiednio dla szeregów po dekompozycji i różnicowaniu. W przypadku pierwszej transformacji wszystkie kryteria wskazały model AR(1), w przypadku drugiej kryteria AIC i HQIC wskazały model MA(1), natomiast BIC - model ARMA(0,0). Znając już modele, które najlepiej opisują nasze dane przeszłyśmy do estymacji ich parametrów z wykorzystaniem funkcji wbudowanej arima. Dla modelu AR(1) otrzymałyśmy $\phi_1 \approx 0.865$, co jest zgodne z naszymi wcześniejszymi przypuszczeniami wynikającymi z analizy wyglądu funkcji PACF. Dla modelu MA(1) otrzymałyśmy $\theta_1 \approx -0.118$, co także zgadza się z wyglądem funkcji PACF. Natomiast model ARMA(0,0) to biały szum.

Po dobraniu modeli przeszłyśmy do oceny ich dopasowania. W tym celu zwizualizowałyśmy wpadanie funkcji ACF i PACF do przedziałów ufności. Zauważyłyśmy, że prawie wszystkie wartości mieszczą się w wyznaczonych przedziałach, co pozwala wnioskować, że modele są najprawdopodobniej poprawnie dobrane. Tezę tą potwierdzają także wykresy porównujące linie kwantylowe z trajektoriami. Wykonałyśmy także prognozę przyszłych obserwacji na podstawie wyodrębnionych wcześniej danych, a następnie porównałyśmy je

z wartościami rzeczywistymi. Możemy tutaj zauważyć, że predykowane wartości mieszczą się w przedziałach ufności w przypadku wszystkich modeli, jednak dla modelu AR(1) przedziały te są znacznie węższe, co pozwala nam wnioskować, że to właśnie ten model lepiej przybliża dane dotyczące cen otwarcia. Ostatnim wykonanym przez nas krokiem w celu oceny dopasowania modeli było sprawdzenie założeń dotyczących szumu. Wszystkie modele spełniają założenia dotyczące zerowej średniej i stałej wariancji oraz niezależności. Nie spełniają natomiast założenia dotyczącego normalności rozkładu. Pokazałyśmy jednak, że residua mają rozkład t-Studenta, który, podobnie jak rozkład normalny, jest białym szumem, co pozwala nam stwierdzić, że modele są dobrze dobrane.

Po podsumowaniu wszystkich aspektów możemy stwierdzić, że poprawnie dobrałyśmy modele opisujące ceny otwarcia amerykańskiego przedsiębiorstwa Amazon.

7 Źródła

- Wykłady dr hab. inż. Krzysztofa Burneckiego oraz laboratoria dr inż. Aleksandry Grzesiek z przedmiotu „Statystyka stosowana".
- Wykłady dr hab. inż. Agnieszki Wyłomańskiej oraz laboratoria mgr inż. Wojciecha Żuławińskiego i Justyny Witulskiej z przedmiotu „Komputerowa analiza szeregów czasowych".
- Wykłady dr hab. Janusza Szwabińskiego z przedmiotu „Metody numeryczne".
- https://docs.python.org/pl/3/
- https://www.kaggle.com/code/purvasingh/time-series-analysis-with-arma-and-arima
- https://alkaline-ml.com/pmdarima/tips_and_tricks.htmltips-and-tricks
- Delucia, J. (1997). "The Brutality of Kickboxing: A Medical Perspective."Clinics in Sports Medicine.