Komputerowa analiza szeregów czasowych - raport II

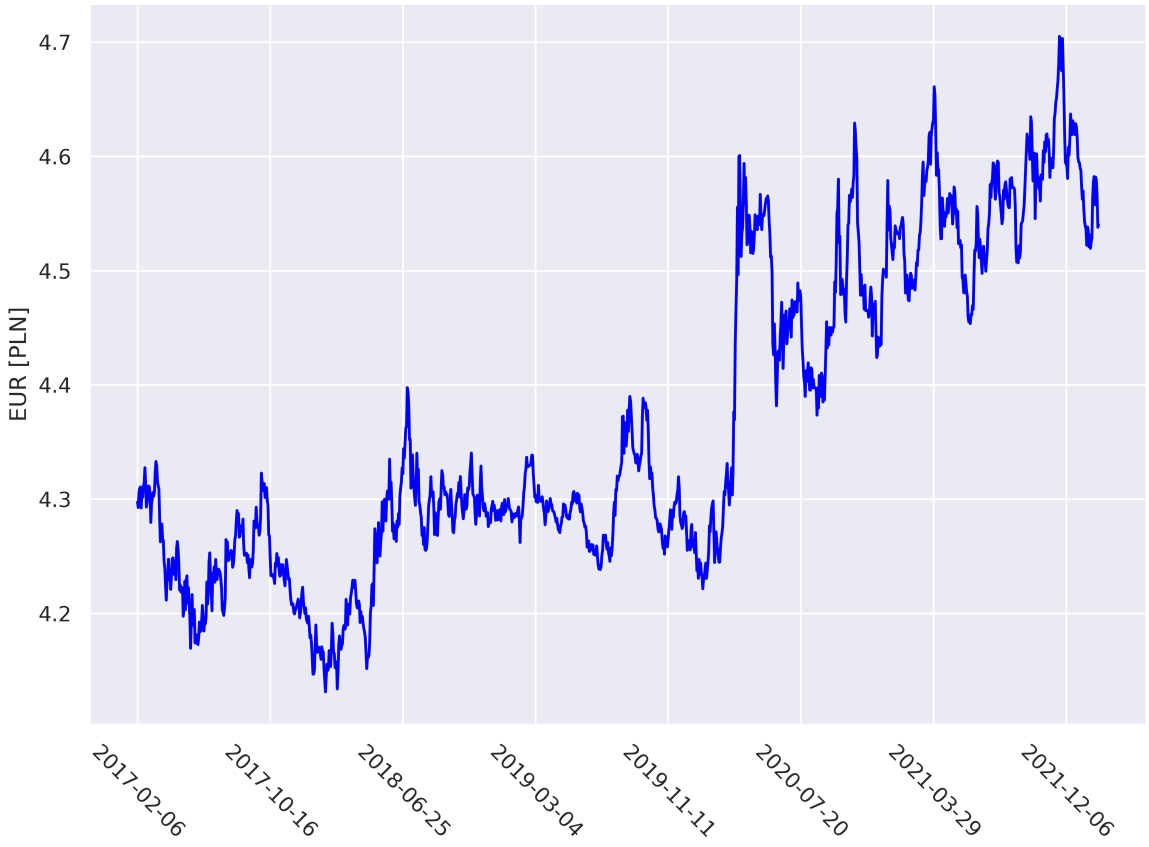
Analiza danych z wykorzystaniem modeli ARMA

Ada Majchrzak, Aleksander Jakóbczyk

7 lutego 2022

1 Wstęp

W poniższej pracy zajęliśmy się analizą danych pochodzących ze strony Yahoo Finance: EUR/PLN. Badane dane pochodzą z okresu od 2017-02-06 do 2022-02-04, zawierają 1305 obserwacji i dotyczą kursu euro w stosunku do złotego. Dla niektórych z dni cena nie została określona - w takim przypadku nieznaną cenę zastąpimy średnią z dwóch najbliższych dni. Głównym celem raportu będzie konstrukcja modelu ARMA, estymacja jego parametrów oraz analiza residuów.



Rysunek 1: Kurs EUR/PLN w okresie od 2017-02-06 do 2022-02-04

2 Statystyki opisowe

2.1 Średnia arytmetyczna

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\overline{x} \approx 4.36$$

2.2 Wariancja

$$S^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})^2$$

$$S^2 \approx 0.02$$

2.3 Odchylenie standardowe

$$S = \sqrt{S^2}$$

$$S \approx 0.14$$

2.4 Mediana

Dla posortowanej próby:

$$med = \text{Q2} = \begin{cases} x_{\frac{n+1}{2}}, & \text{gdy n nieparzyste} \\ \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}\right), & \text{gdy n parzyste} \end{cases}$$

$$\text{Q2} \approx 4.31$$

2.5 Kwartyle

Pierwszy kwartyl Q1 to mediana grupy obserwacji mniejszych od Q2.
Trzeci kwartyl Q3 to mediana grupy obserwacji większych od Q2.

$$\text{Q1} \approx 4.26$$

$$\text{Q3} \approx 4.50$$

2.6 Rozstęp międzykwartylowy

$$\text{IQR} = \text{Q3} - \text{Q1}$$

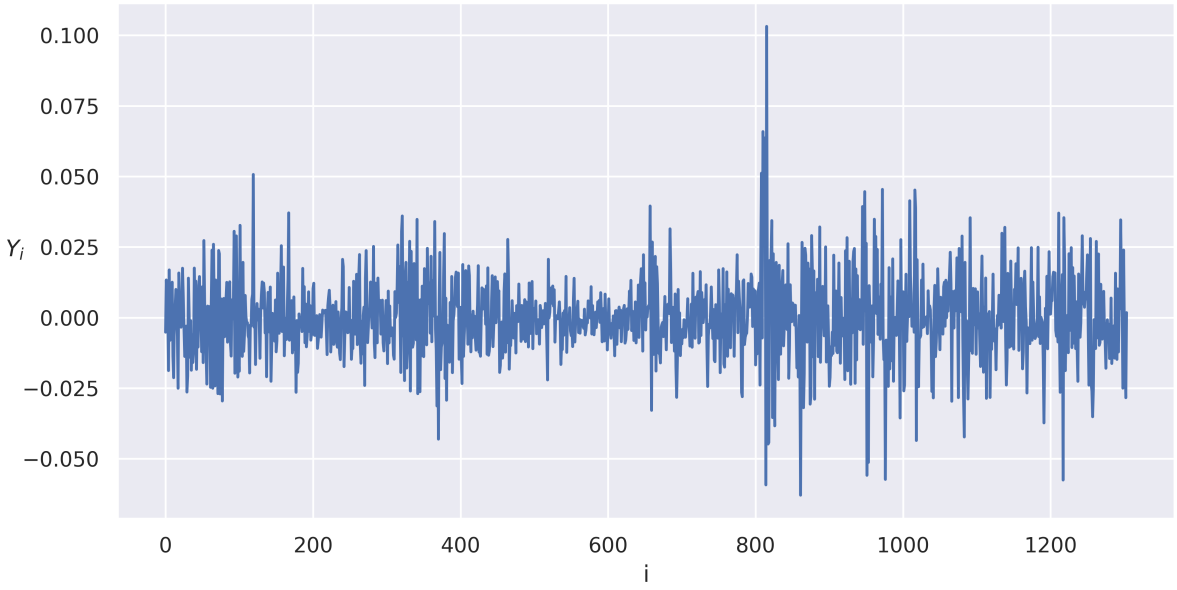
$$\text{IQR} \approx 0.24$$

3 Przygotowanie danych

Patrząc na Rysunek 1, po danych możemy spodziewać się istnienia pewnego trendu liniowego. W celu jego usunięcia skorzystamy z metody różnicowania danych:

$$Y_i = X_i - X_{i-1},$$

gdzie X_i jest i-tą wartością naszego kursu. Po takim przetransformowaniu naszych danych uzyskujemy nowy szereg czasowy Y_i z usuniętym trendem liniowym.



Rysunek 2: Dane po zastosowaniu transformacji różnicowej

Oczywiście transformacja danych zmienia ich statystyki opisowe - wartości kilku podstawowych statystyk umieściliśmy w poniższej tabeli:

mean	std	min	Q1	Median	Q3	max
0.000186	0.014416	-0.062600	-0.007915	-0.000255	0.007535	0.103460

Tabela 1: Tablica podstawowych statystyk przetransformowanych danych

Teraz pokażemy, jak dla naszych danych prezentują się empiryczne funkcje ACF i PACF, dane wzorami:

- ACF

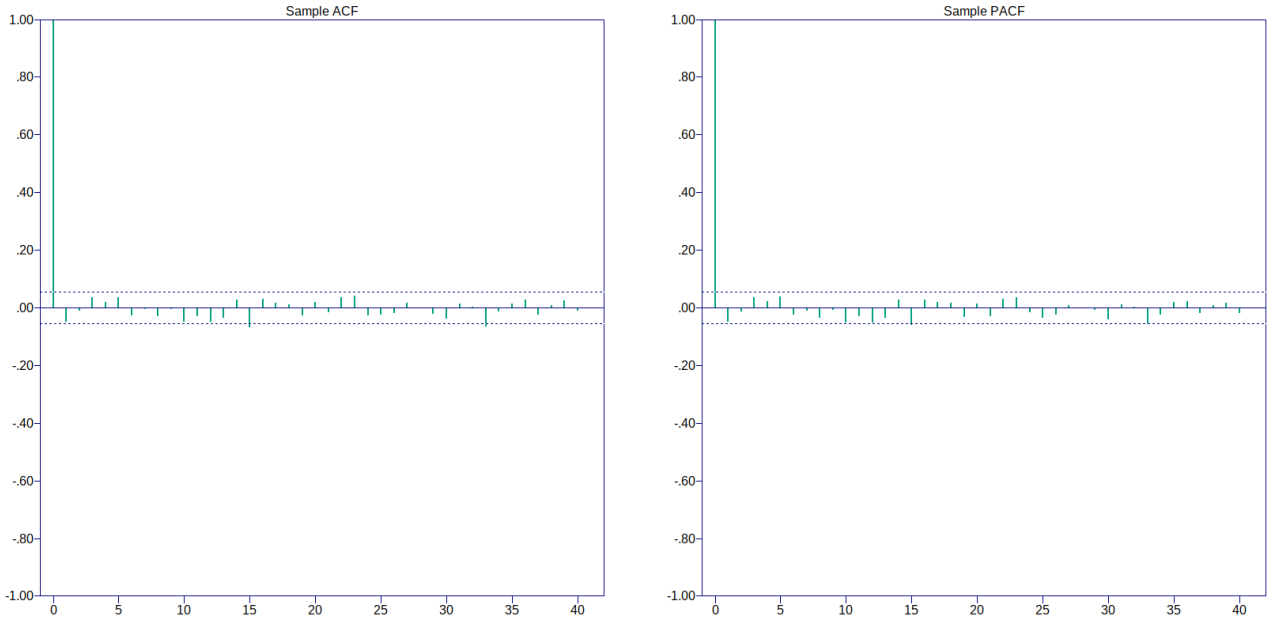
$$\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)} = \frac{1}{n} \sum_{t=1}^{n-|h|} (x_{t+|h|} - \overline{x})(x_t - \overline{x}),$$

  gdzie $\hat{\gamma}(h)$ - empiryczna funkcja autokowariancji, n - długość próby,

- PACF

$$\alpha(h) = \phi_{hh},$$

  gdzie ϕ_{hh} jest ostatnim współczynnikiem wektora $\phi_h = \hat{\Gamma}_h^{-1}\hat{\gamma}_h$, $\hat{\Gamma}_h = [\hat{\gamma}(i-j)]_{i,j=1}^{h}$.



Rysunek 3: Wykresy funkcji ACF i PACF

Z otrzymanych wykresów widzimy korelację na poziomie 1 tylko dla argumentu 0, w pozostałych przypadkach nasze funkcje oscylują wokół zera. Ponadto teraz średnia dla naszych danych jest stała w czasie i w przybliżeniu wynosi 0 - mamy więc dobrze przygotowane dane i możemy już przejść do doboru oraz analizy odpowiedniego modelu ARMA, czyli szeregu autoregresyjnego średniej ruchomej, zdefiniowanego następująco:

Definicja. *Szereg czasowy X_t jest szeregiem ARMA(p,q), jeśli jest stacjonarny w słabym sensie oraz spełnia równanie:*

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

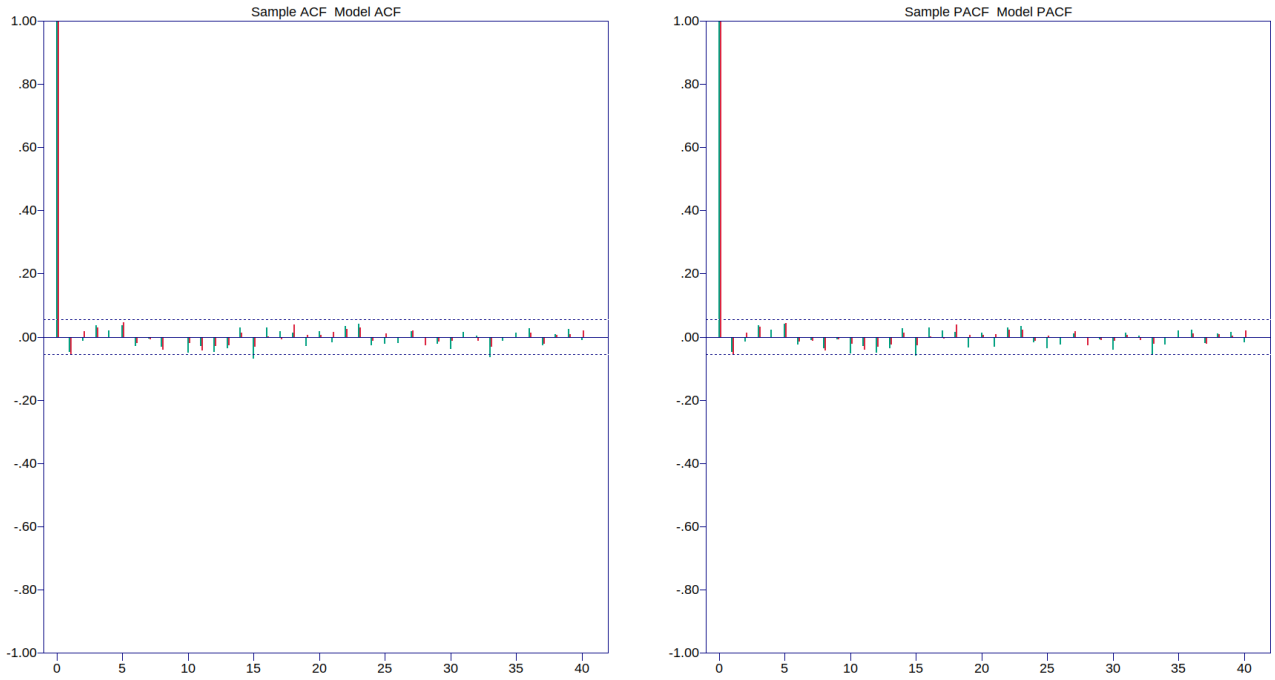
gdzie:

- $Z_t \sim WN(0, \sigma^2)$,
- *wielomiany* $\phi(z) = 1 - \phi_1 z - \ldots \phi_p z^p$, $\theta(z) = 1 + \theta_1 z + \ldots + \theta_q z^q$ *nie mają wspólnych pierwiastków.*

Model ARMA(p,q)

Dobieranie parametrów p i q w modelu ARMA opiera się na kryterium informacyjnym Akaike'go. Jest ono oszacowaniem tego, ile informacji tracimy, gdy zamiast rzeczywistego modelu wybierzemy rozważany. Wynika stąd, że im mniejsza wartość AIC, tym lepiej dobrany jest nasz model. Metoda "Likelyhood", której będziemy używać, opiera się na sprawdzeniu wszystkich kombinacji parametrów p i q modeli ARMA w zadanym przedziale, a następnie wybraniu tych, dla których wartość porównywanego kryterium informacyjnego jest najmniejsza. Wykorzystując powyższą metodę otrzymujemy, że model ARMA(6,7) najlepiej estymuje badany szereg czasowy względem kryterium informacyjnego AICC.

Aby zwizualizować poprawność dobranego przez nas modelu, porównamy empiryczne funkcje ACF i PACF naszych danych z teoretycznymi dla wyestymowanego modelu:



Rysunek 4: Porównanie funkcji ACF i PACF teoretycznych z empirycznymi

Widzimy, że w obu przypadkach wykresy funkcji teoretycznych (kolor czerwony) pokrywają się bardzo dobrze z empirycznymi (kolor zielony), co pozwala wnioskować, że model ARMA został przez nas poprawnie dobrany. Możemy zatem przejść do przedstawienia wyestymowanych parametrów modelu.

Estymowane parametry modelu ARMA(p,q)

Pierwsza tabela zawiera wartości kolejnych współczynników $\phi_k,\ k = 1, \ldots, 6$ (lewa strona równania w modelu ARMA),

ϕ_1	ϕ_2	ϕ_3	ϕ_4	ϕ_5	ϕ_6
0.178158	0.572810	-0.337375	0.632793	0.214153	-0.885537

Tabela 2: Parametry $\phi_1, ..., \phi_6$

z drugiej natomiast możemy odczytać wartości współczynników $\theta_k,\ k = 1, \ldots, 7$ (prawa strona równania).

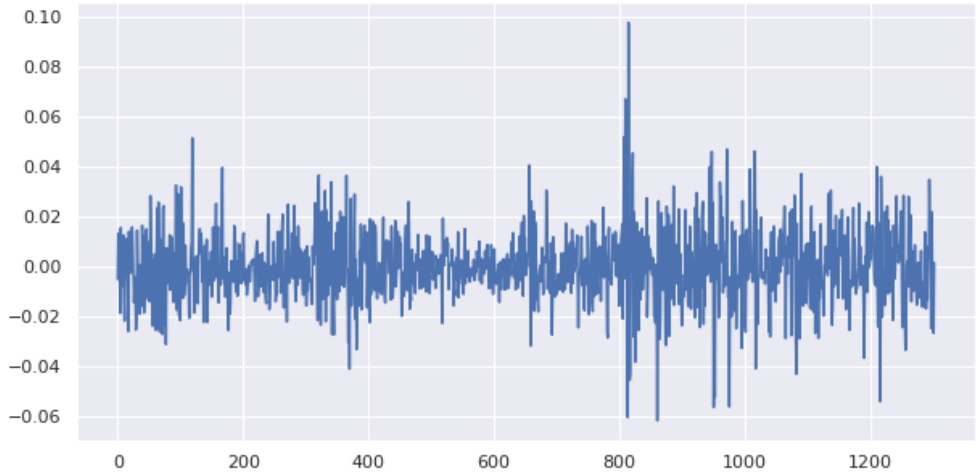
θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7
-0.235259	-0.542868	0.402289	-0.683107	-0.145044	0.889424	-0.109801

Tabela 3: Parametry $\theta_1, ..., \theta_7$

Oprócz tego, istotny dla nas jest również parametr σ^2, czyli wariancja białego szumu (szereg Z_t). W naszym przypadku wynosi ona $\sigma^2 = 0.000201$.

Analiza residuów

W poniższej części zajmiemy się analizą residuów, sprawdzimy czy są nieskorelowane i czy pochodzą z rozkładu normalnego.

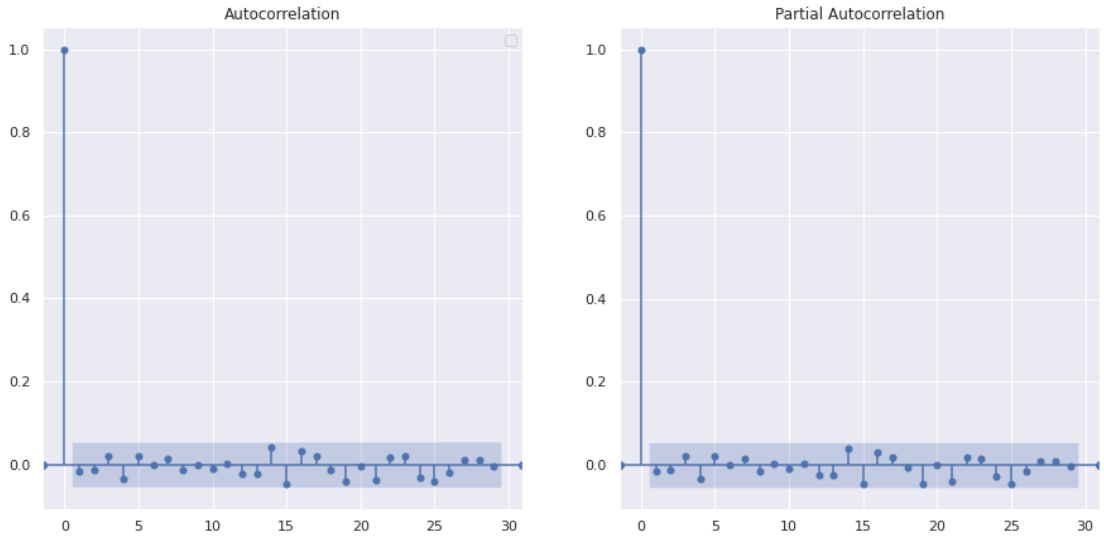


Rysunek 5: Wykres residuów

mean	std	min	Q1	Median	Q3	max
-0.00000050	0.00141866	-0.05641744	-0.00832161	-0.00026310	0.00729265	0.09383978

Tabela 4: Tablica podstawowych statystyk residuów.

Z Tabeli 4 widzimy, że residua mają średnią bardzo bliską 0. Jeśli zaś chodzi o wariancję, możemy zobaczyć jej wyraźną zmianę np. w okolicach osiemsetnej obserwacji, a co za tym idzie, zdecydowanie nie jest ona stała. Teraz chcemy sprawdzić, czy nasze residua są nieskorelowane - w tym celu przyjrzymy się wykresom funkcji ACF i PACF.



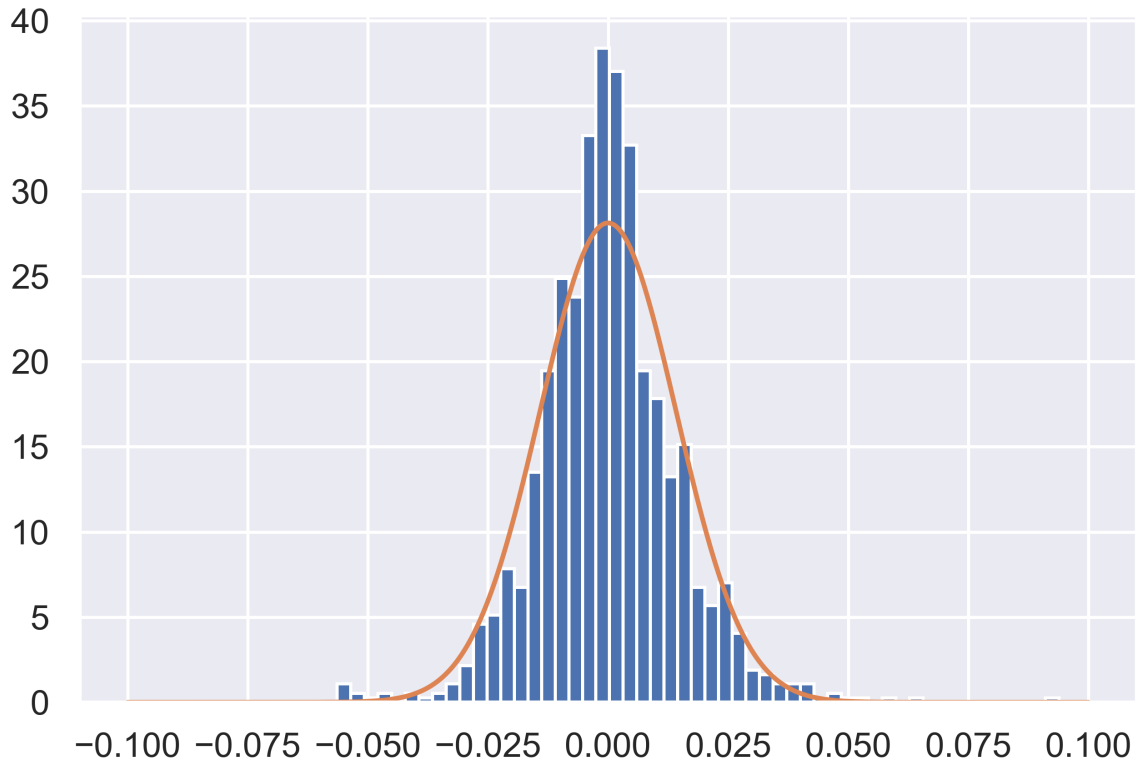
Rysunek 6: ACF i PACF residuów

Analizując powyższe wykresy możemy przypuszczać, że residua są nieskorelowane - aby jednak zweryfikować tę hipotezę, przeprowadzamy pięć testów losowości, których wyniki prezentują się następująco:

- Ljung - Box: p-value = 0.36053
- McLeod - Li: p-value = 0.00000
- Turning points: p-value = 0.29299
- Diff sign points: p-value = 0.11360
- Rank test statistic: p-value = 0.59416

Z pięciu przeprowadzonych testów, cztery zwracają p-wartość, która na poziomie istotności 0.05 nie daje podstaw do odrzucenia hipotezy zerowej o braku korelacji między residuami. Widzimy jednak, że test McLeod-Li zwraca p-wartość równą zeru, a zatem musimy odrzucić hipotezę zerową - wnioskujemy stąd, że nasze residua nie są nieskorelowane.

Następnym krokiem w analizie residuów będzie sprawdzenie, czy pochodzą one z rozkładu normalnego. Najpierw przeprowadzamy test normalności Jarque-Bera. W jego wyniku otrzymujemy p-wartość równą 0, zatem mamy podstawy do odrzucenia hipotezy zerowej o normalności residuów. Chcemy jednak zweryfikować ten wniosek, porównamy więc ich gęstość empiryczną z gęstością teoretyczną rozkładu $\mathcal{N}(0, S^2)$, gdzie S^2 jest empiryczną wariancją residuów.



Rysunek 7: Porównanie histogramu residuów z gęstością rozkładu $\mathcal{N}(0, S^2)$

Otrzymane wykresy potwierdzają, że powinniśmy odrzucić hipotezę zerową o normalności residuów - różnica między teoretyczną a empiryczną gęstością jest zbyt duża.

Podsumowanie

Ceny euro na przestrzeni ostatnich czterech lat zdecydowanie rosły, wykazując przy tym trend liniowy. Ich średnia w tym okresie wyniosła 4.36 PLN (bardzo podobnie do mediany, której wartość to 4.31 PLN), natomiast odchylenie od średniej to około 0.14 PLN. Ze względu na liniowy przebieg naszych danych, dokonaliśmy ich różnicowania, w wyniku czego otrzymaliśmy szereg czasowy stacjonarny w słabym sensie, co bardzo dobrze obrazują rysunki 2 i 3 - widzimy z nich stałą w czasie średnią (równą 0) oraz funkcję autokorelacji niezależną od czasu. Tak przetransformowane dane nadawały się już do doboru odpowiedniego modelu ARMA(p, q) oraz jego analizy. W tej części sprawozdania posługiwaliśmy się głównie programem **ITSM**. Przy jego pomocy otrzymaliśmy model o parametrach $p = 6$ i $q = 7$ (przy wartości AICC równej -7363.83). Mając już dobrany model, dokonaliśmy analizy jego residuów, w wyniku której wywnioskowaliśmy, że co prawda mają one średnią $\mu = 0$, ale za to ich wariancja jest zmienna w czasie. Dodatkowo residua wykazują pewną korelację między sobą i nie pochodzą z rozkładu normalnego.