

Analiza danych rzeczywistych natężenia prądu plazmy przy pomocy modelu ARMA

Dominika Lewandowska, Nikodem Drelak

6 lutego 2026

Spis treści

1	Wstęp	2
1.1	Cel pracy	2
1.2	Opis i źródło danych	2
1.3	Charakterystyka próby i wizualizacja	2
2	Przygotowanie danych do analizy	4
2.1	Analiza jakości danych	4
2.1.1	Analiza autokorelacji surowych danych	4
2.2	Dekompozycja szeregu czasowego	4
2.2.1	Różnicowanie danych	4
2.2.2	Ocena autokorelacji danych po dekompozycji	6
3	Modelowanie przy pomocy ARMA	8
3.1	Dobranie rzędu modelu i parametrów	8
3.2	Ocena dopasowania modelu	9
3.2.1	Porównanie empirycznych i teoretycznych ACV i PACV	9
3.2.2	Analiza trajektorii ARMA w porównaniu do badanego szeregu	9
4	Weryfikacja założeń dotyczących szumu	12
4.1	Założenie dotyczące średniej	12
4.2	Założenie dotyczące wariancji	12
4.3	Założenie dotyczące niezależności	13
4.3.1	Autokorelacja reszt	13
4.4	Założenie dotyczące normalności rozkładu	13
5	Wnioski	13
6	Podsumowanie	16
	Literatura	18

1 Wstęp

1.1 Cel pracy

Celem niniejszej pracy jest analiza statystyczna oraz modelowanie dynamiki szeregu czasowego pochodzącego z rzeczywistego eksperymentu fizycznego. Przedmiotem badań jest natężenie prądu plazmy (ang. *Plasma Current*) zarejestrowane w urządzeniu typu tokamak.

Głównym zadaniem jest weryfikacja hipotezy o możliwości opisu fluktuacji prądu w fazie stabilnej (tzw. *flat-top*) za pomocą liniowego modelu stochastycznego klasy ARMA. Analiza obejmuje zbadanie stacjonarności procesu, identyfikację rzędu modelu, estymację parametrów oraz weryfikację założeń dotyczących reszt.

1.2 Opis i źródło danych

Dane wykorzystane w projekcie pochodzą z reaktora fuzyjnego MAST (Mega Ampere Spherical Tokamak), znajdującego się w Culham Centre for Fusion Energy w Wielkiej Brytanii. Zostały pobrane za pośrednictwem otwartego interfejsu API udostępnionego w ramach projektu [FAIR-MAST](#).

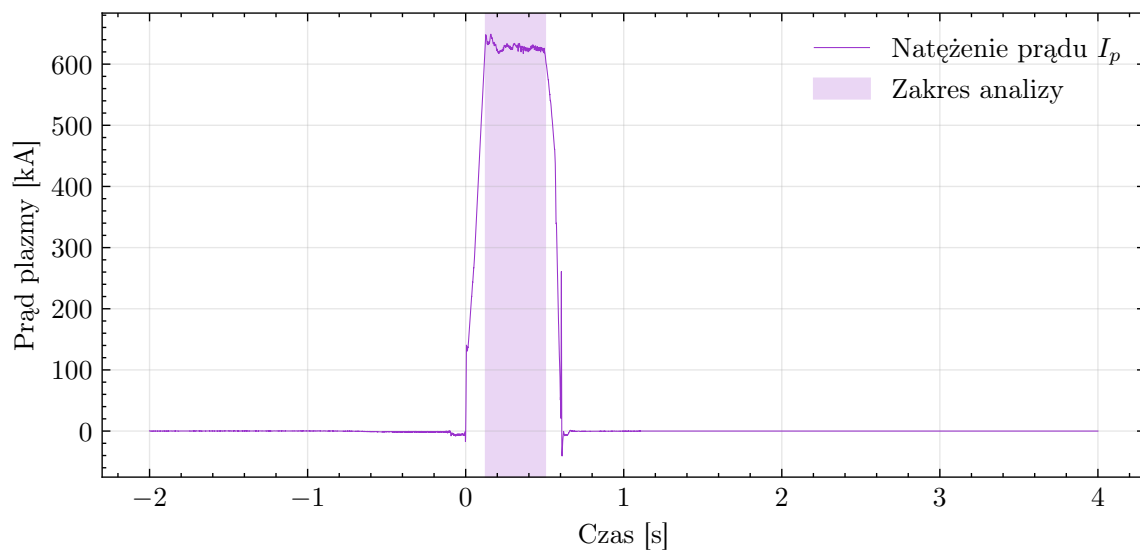
Interpretacja fizyczna zmiennej: Analizowana zmienna to natężenie prądu plazmy (I_p), wyrażone w kiloamperach (kA). W uproszczeniu, parametr ten jest kluczowym wskaźnikiem „życia” eksperymentu:

- Wzrost prądu oznacza formowanie się plazmy.
- Utrzymywanie stałej wartości (plateau) oznacza fazę stabilną, w której przeprowadza się właściwe eksperymenty.
- Nagły spadek wartości do zera może sygnalizować niekontrolowaną utratę stabilności (tzw. disruption).

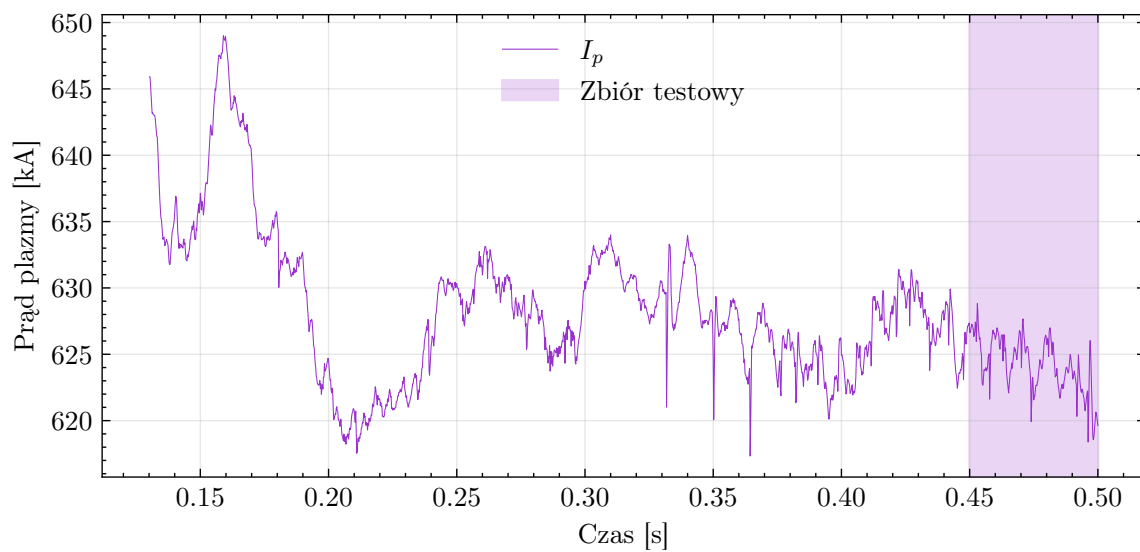
Dla potrzeb analizy szeregów czasowych, fluktuacje tego prądu w fazie stabilnej traktujemy jako proces stochastyczny, wynikający z turbulencji wewnątrz gorącego gazu oraz działania systemów sterowania reaktora.

1.3 Charakterystyka próby i wizualizacja

Do analizy wybrano eksperyment (tzw. *shot*) o numerze ID: 30421. Pełny przebieg eksperymentu (przedstawiony na rysunku 1) trwa około 6 sekund. Ze względu na niestacjonarny charakter całego procesu (faza rozruchu i wygaszania), do modelowania ARMA wyodrębniono wycinek czasowy odpowiadający fazie stabilnej (rysunek 2). Zbiór analityczny odpowiada 0,37 sekundy obserwacji i zawiera 1850 próbek. W celu późniejszej weryfikacji poprawności modelowania ARMA dla przyszłych wartości, został on podzielony na zbiór treningowy - pierwsze 0,32 sekundy (1500 próbek) oraz zbiór testowy - ostatnie 0.05 sekund (250 próbek).



Rysunek 1: Pełny przebieg natężenia prądu plazmy w czasie eksperymentu. Półprzezroczystym obszarem zaznaczono obszar fazy stabilnej wybrany do analizy.



Rysunek 2: Wyodrębniony fragment szeregu czasowego (faza *flat-top*) poddany modelowaniu ARMA. Półprzezroczystym obszarem zaznaczono obszar zbioru testowego.

2 Przygotowanie danych do analizy

2.1 Analiza jakości danych

Na wstępnym etapie analizy dokonano oceny jakości danych pomiarowych. Sprawdzono obecność braków danych, wartości odstających oraz nieciągłości czasowych. Analizowany szereg czasowy nie zawiera brakujących obserwacji ani duplikatów, a odstępy czasowe pomiędzy kolejnymi próbkami są jednorodne.

Wartości natężenia prądu mieszczą się w zakresie fizycznie uzasadnionym dla pracy tokamaka i nie zaobserwowano anomalii mogących wskazywać na błędy pomiarowe. Na tej podstawie dane uznano za poprawne i odpowiednie do dalszej analizy statystycznej.

2.1.1 Analiza autokorelacji surowych danych

W celu wstępnej identyfikacji struktury zależności czasowych w szeregu obliczono funkcję autokorelacji (ACF) oraz funkcję częściowej autokorelacji (PACF).

Funkcja autokorelacji ACF opisuje stopień liniowej zależności pomiędzy obserwacjami oddzielnymi o h kroków czasowych i dana jest wzorem:

$$\rho(h) = \frac{Cov(X_t, X_{t-h})}{Var(X_t)}. \quad (1)$$

Z kolei funkcja PACF mierzy korelację pomiędzy X_t i X_{t-h} po wyeliminowaniu wpływu opóźnień pośrednich $1, 2, \dots, h-1$, co pozwala na lepszą identyfikację rzędu części autoregresyjnej modelu. Rysunek 3 przedstawia wykres empirycznej funkcji autokorelacji ACF dla surowych danych ze zbioru treningowego. Widoczna jest bardzo wolno malejąca autokorelacja oraz istotne wartości współczynnika dla dużych opóźnień, znacząco przekraczające granice przedziału ufności. Taki kształt ACF jest charakterystyczny dla procesów niestacjonarnych.

Rysunek 4 przedstawia wykres funkcji częściowej autokorelacji PACF. Również w tym przypadku obserwuje się istotne wartości dla wielu opóźnień, co potwierdza brak stacjonarności oraz obecność silnej struktury trendowej w danych.

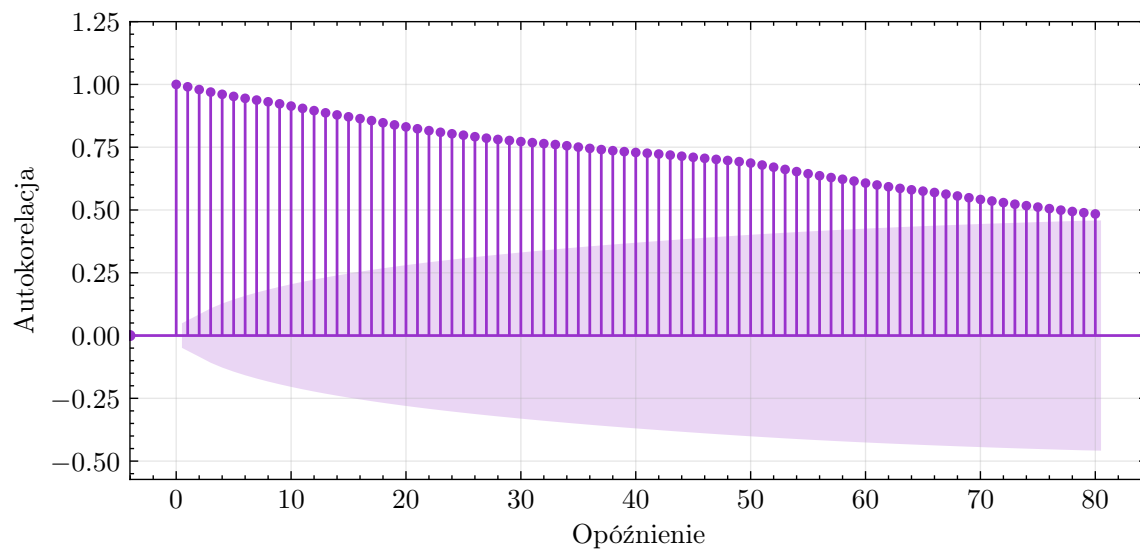
Dodatkowo przeprowadzono test Augmented Dickey–Fullera (ADF), weryfikujący hipotezę zerową o istnieniu pierwiastka jednostkowego. Otrzymano statystykę testową równą 0.665 oraz wartość p równą 0.427, co oznacza brak podstaw do odrzucenia hipotezy o niestacjonarności szeregu.

2.2 Dekompozycja szeregu czasowego

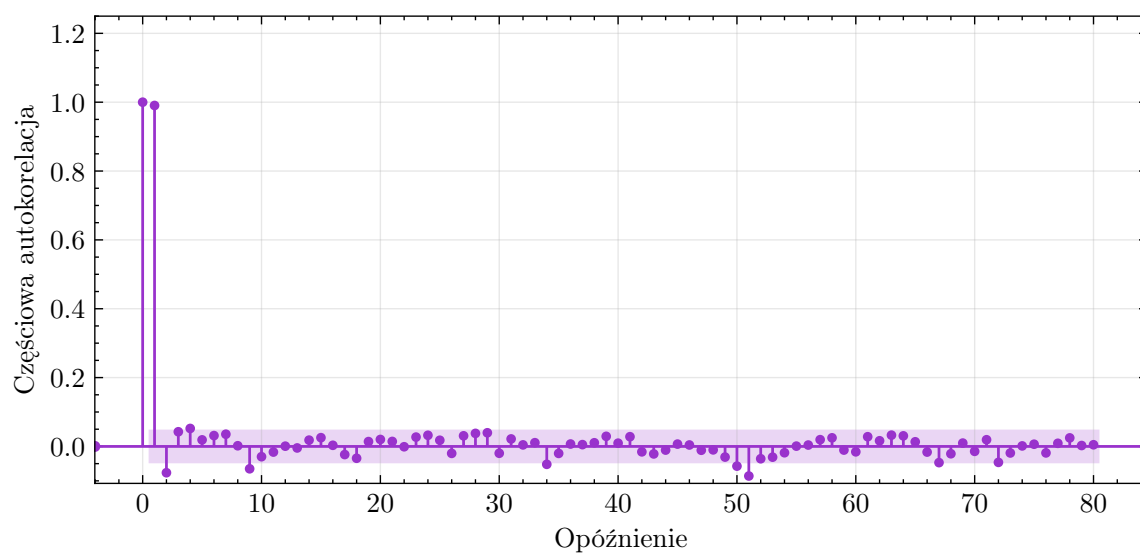
2.2.1 Różnicowanie danych

W celu uzyskania stacjonarności szeregu zastosowano różnicowanie pierwszego rzędu. Operacja ta polega na zastąpieniu oryginalnego szeregu X_t nowym szeregiem:

$$Y_t = X_t - X_{t-1}, \quad (2)$$

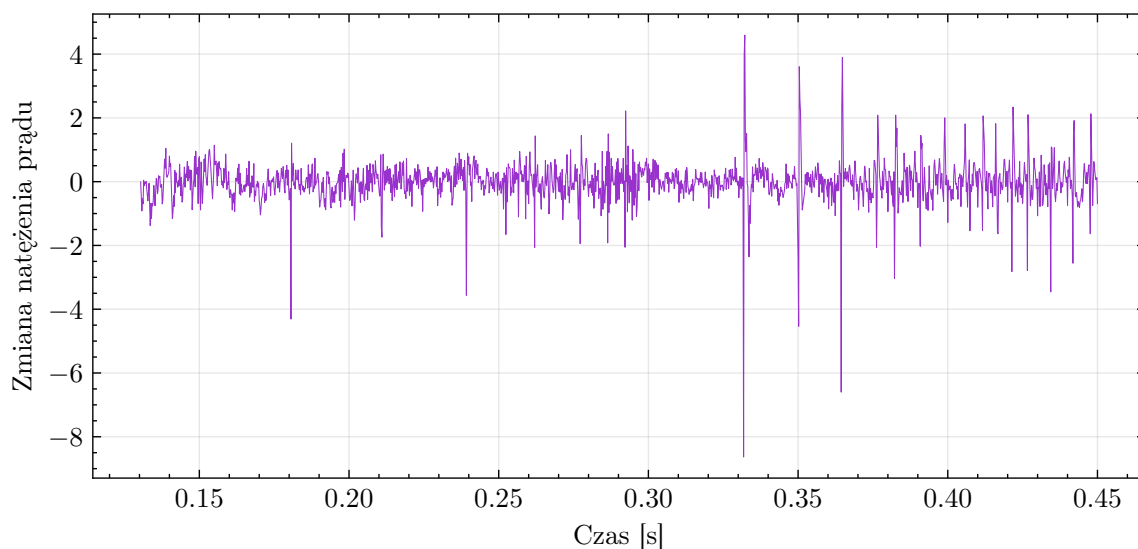


Rysunek 3: Wykres współczynnika autokorelacji surowych danych ze zbioru treningowego w zależności od opóźnienia (parametru h). Półprzezroczystym obszarem zaznaczono przedział ufności o poziomie 95% dla hipotezy o braku korelacji (szum biały).



Rysunek 4: Wykres współczynnika częściowej autokorelacji surowych danych ze zbioru treningowego w zależności od opóźnienia (parametru h). Półprzezroczystym obszarem zaznaczono przedział ufności o poziomie 95% dla hipotezy o braku korelacji (szum biały).

który opisuje zmiany wartości natężenia prądu pomiędzy kolejnymi chwilami czasu. Różnicowanie jest standardową metodą eliminacji trendu oraz wolnozmiennych składowych deterministycznych w analizie szeregów czasowych. Rysunek 5 przedstawia przebieg szeregu po różnicowaniu. Widoczne jest usunięcie trendu oraz oscylowanie wartości wokół zera, co wskazuje na poprawę własności stacjonarnych.

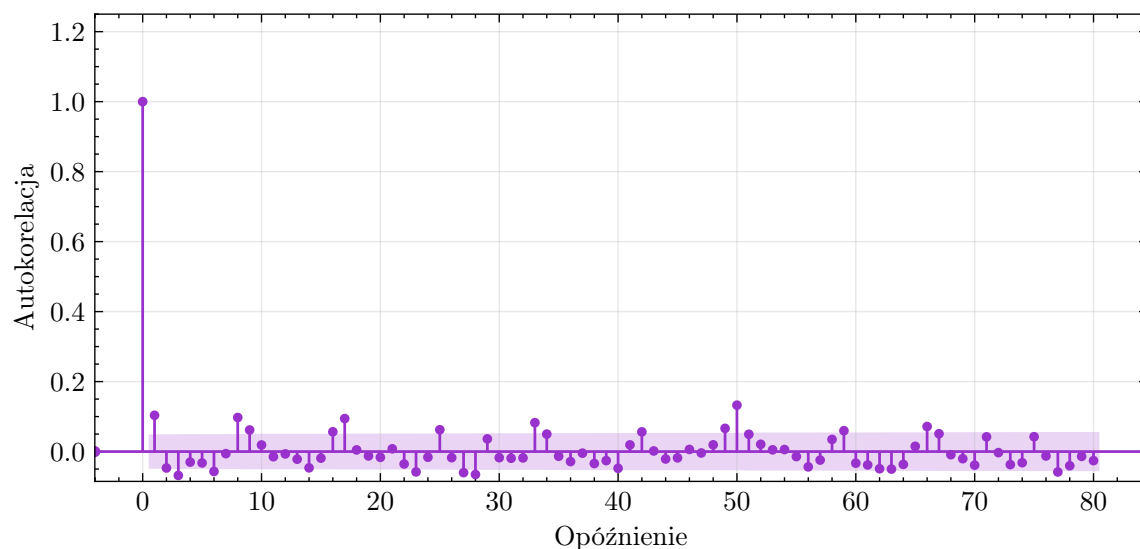


Rysunek 5: Szereg czasowy ze zbioru treningowego poddany różnicowaniu. Przedstawia zmianę natężenia prądu w zależności od czasu.

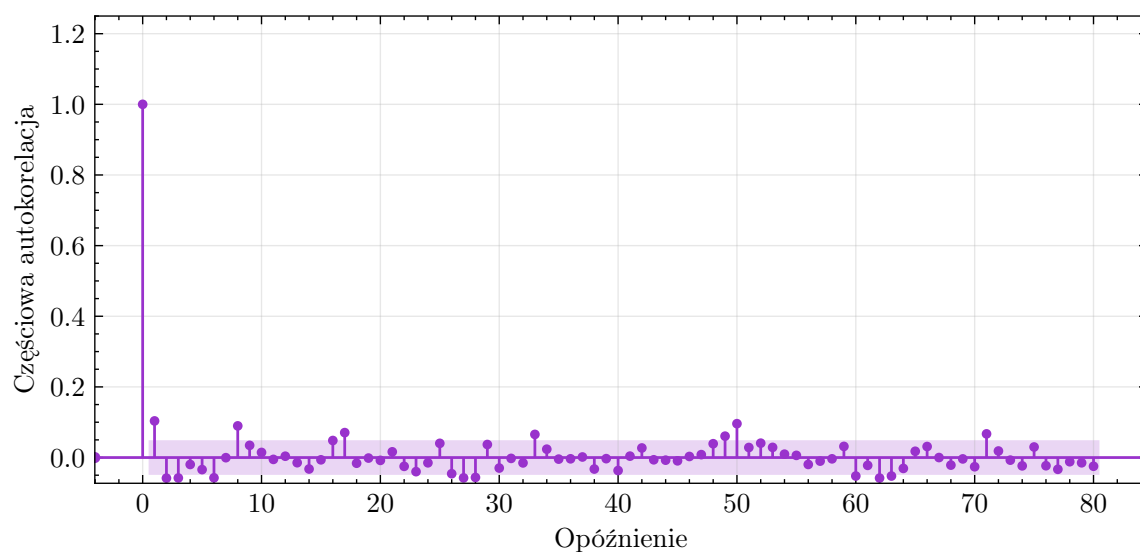
2.2.2 Ocena autokorelacji danych po dekompozycji

Na rysunkach 6 i 7 przedstawiono odpowiednio wykresy ACF oraz PACF dla danych po różnicowaniu. W przeciwieństwie do danych surowych, autokorelacje szybko zanikają i w większości mieszczą się w granicach przedziału ufności. Świadczy to o skutecznym usunięciu niestacjonarności.

Test ADF przeprowadzony dla szeregu po dekompozycji dał statystykę 14.148 oraz wartość p równą 0.0, co jednoznacznie prowadzi do odrzucenia hipotezy o niestacjonarności. Oznacza to, że uzyskany szereg jest stacjonarny i może być modelowany przy użyciu modeli klasy ARMA.



Rysunek 6: Wykres współczynnika autokorelacji danych ze zbioru treningowego poddanych różnicowaniu w zależności od opóźnienia (parametru h). Półprzezroczystym obszarem zaznaczono przedział ufności o poziomie 95% dla hipotezy o braku korelacji (szum biały).



Rysunek 7: Wykres współczynnika częściowej autokorelacji danych ze zbioru treningowego poddanych różnicowaniu w zależności od opóźnienia (parametru h). Półprzezroczystym obszarem zaznaczono przedział ufności o poziomie 95% dla hipotezy o braku korelacji (szum biały).

3 Modelowanie przy pomocy ARMA

3.1 Dobranie rzędu modelu i parametrów

Model $\text{ARMA}(p, q)$ łączy w sobie składnik autoregresyjny rzędu p oraz składnik średniej ruchomej rzędu q . Jego postać ogólna dana jest wzorem:

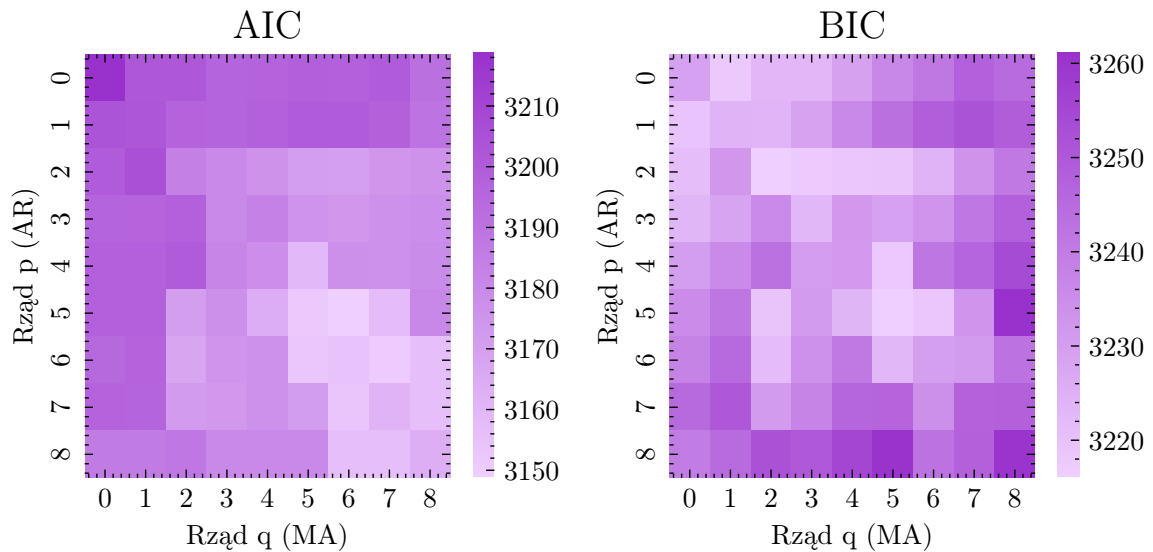
$$X_t = \sum_{i=1}^p \Phi_i X_{t-1} + \sum_{j=1}^q \theta_j \epsilon_{t-1} + \epsilon_t, \quad (3)$$

gdzie ϵ_t jest białym szumem.

Dobór rzędów p i q przeprowadzono na podstawie kryteriów informacyjnych Akaikego (AIC) oraz Bayesowskiego (BIC), zdefiniowanych jako:

$$AIC = -2 \ln(L) + 2k, BIC = -2 \ln(L) + k \ln(n), \quad (4)$$

gdzie L jest wartością funkcji wiarygodności, k liczbą parametrów, a n liczbą obserwacji. Na podstawie heatmap przedstawionych na rysunku 8 stwierdzono, że minimum obu kryteriów osiągane jest dla modelu $\text{ARMA}(5,5)$. Parametry modelu zostały wyznaczone metodą największej wiarygodności (MLE), a ich wartości zestawiono w tabeli 1.

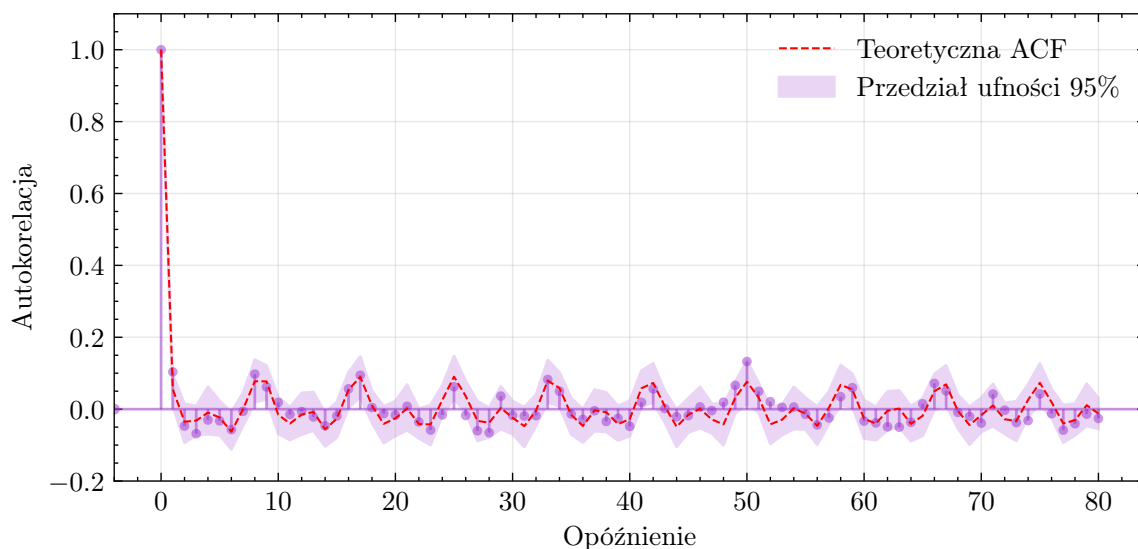


Rysunek 8: Heatmapa wartości kryteriów AIC oraz BIC w zależności od rzędów p i q modelu $\text{ARMA}(p, q)$ badanego szeregu. Jaśniejszy kolor oznacza niższą wartość kryterium.

3.2 Ocena dopasowania modelu

3.2.1 Porównanie empirycznych i teoretycznych ACV i PACV

Rysunki 9 i 10 przedstawiają porównanie empirycznych oraz teoretycznych funkcji ACF i PACF dla dopasowanego modelu ARMA(5,5). Widoczna jest bardzo dobra zgodność przebiegów, a większość wartości empirycznych mieści się w granicach przedziałów ufności. Świadczy to o poprawnym odwzorowaniu struktury zależności czasowych przez model.



Rysunek 9: Wykres empirycznego i teoretycznego współczynnika autokorelacji modelu ARMA(5, 5) dopasowanego do zbioru treningowego, w zależności od opóźnienia (parametru h). Półprzezroczystym obszarem zaznaczono przedział ufności o poziomie 95% wyznaczony metodą Monte Carlo ($N=1000$).

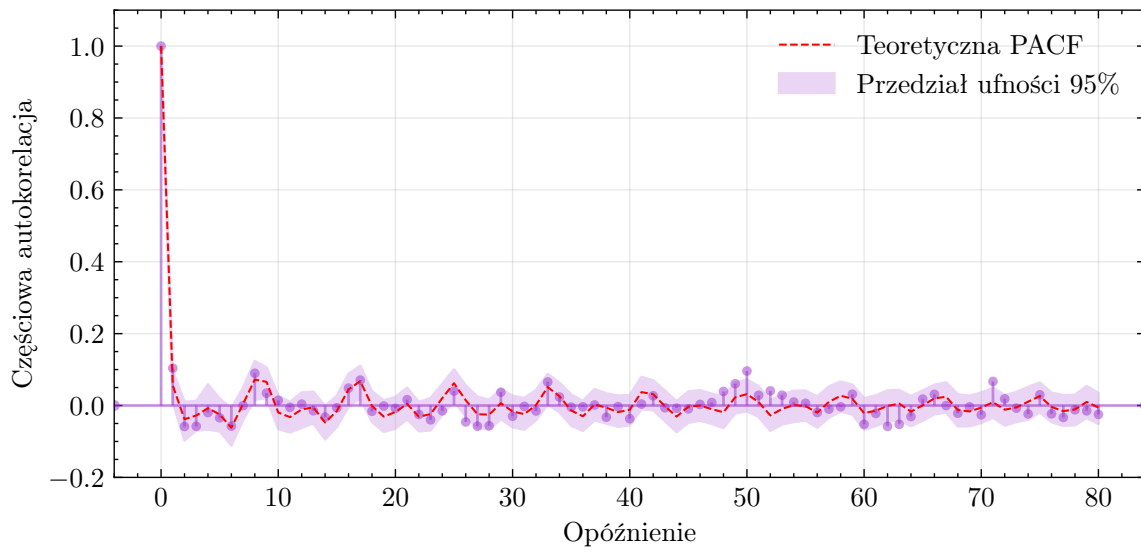
3.2.2 Analiza trajektorii ARMA w porównaniu do badanego szeregu

Rysunek 11 przedstawia porównanie trajektorii rzeczywistego szeregu po różnicowaniu z trajektorią generowaną przez model ARMA(5,5). Model poprawnie odwzorowuje dynamikę zmian oraz amplitudę fluktuacji.

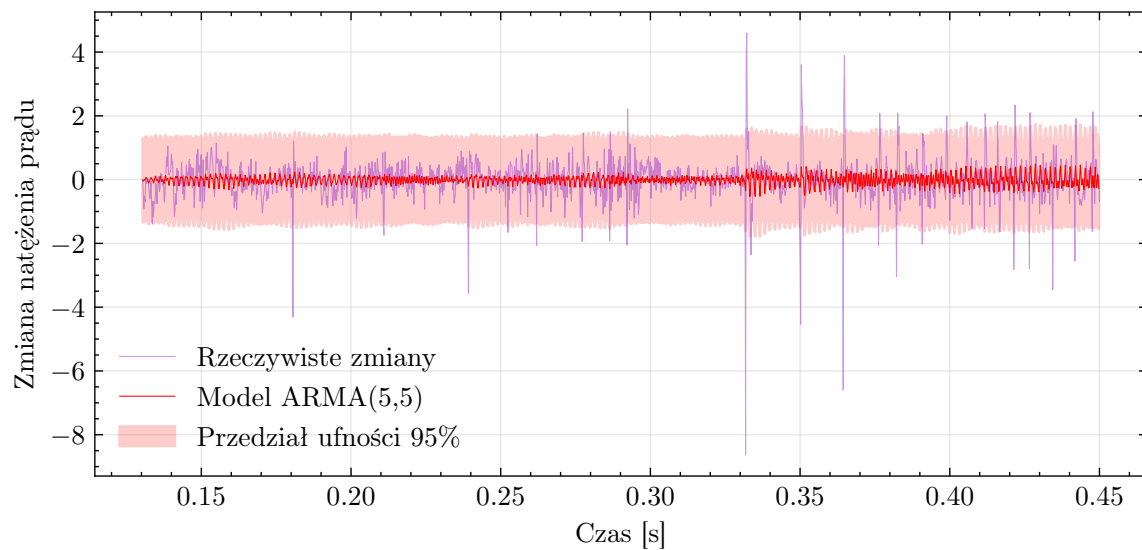
Na rysunku 12 zaprezentowano prognozę modelu dla zbioru testowego. Większość obserwacji rzeczywistych mieści się w 95% przedziałach ufności, co potwierdza dobrą jakość predykcyjną modelu w krótkim horyzoncie czasowym.

Tabela 1: Parametry modelu ARMA(5, 5) dopasowanego do szeregu ze zbioru treningowego, wyznaczone metodą największej wiarygodności (MLE).

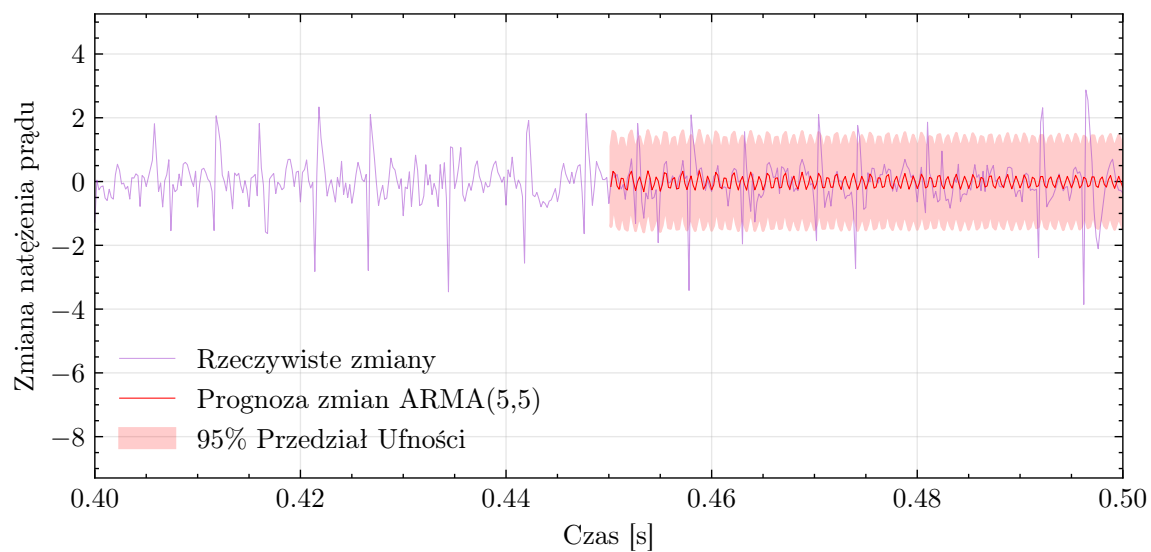
AR(5)	ϕ_1	ϕ_2	ϕ_3	ϕ_4	ϕ_5
	-0.57475292	0.61155579	0.5790856	-0.55870469	0.97329654
MA(5)	θ_1	θ_2	θ_3	θ_4	θ_5
	-0.53449985	0.5807027	0.601362	-0.55516097	0.9431987



Rysunek 10: Wykres empirycznego i teoretycznego współczynnika częściowej autokorelacji modelu ARMA(5, 5) dopasowanego do zbioru treningowego, w zależności od opóźnienia (parametru h). Półprzezroczystym obszarem zaznaczono przedział ufności o poziomie 95% wyznaczony metodą Monte Carlo ($N=1000$).



Rysunek 11: Trajektoria szeregu czasowego ze zbioru treningowego poddanego różnicowaniu oraz trajektoria dopasowanego modelu ARMA(5, 5) wraz z przedziałami ufności o poziomie 95%.

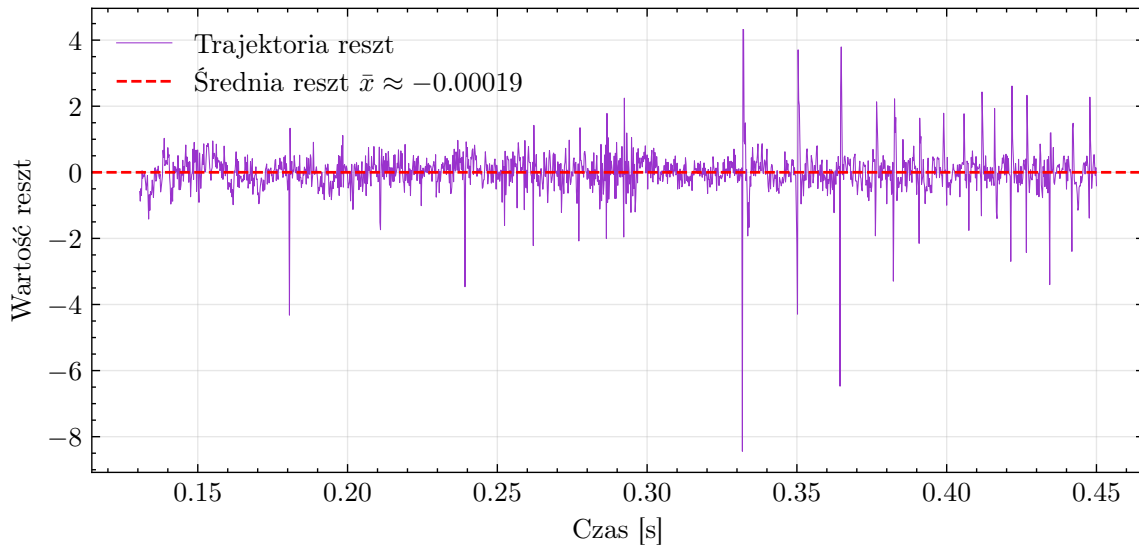


Rysunek 12: Trajektoria szeregu czasowego ze zbioru treningowego i ze zbioru testowego poddanych różnicowaniu, oraz prognoza trajektorii dopasowanego modelu ARMA(5, 5) dla zbioru testowego wraz z przedziałami ufności o poziomie 95%.

4 Weryfikacja założeń dotyczących szumu

4.1 Założenie dotyczące średniej

Przeprowadzono test t-Studenta dla średniej reszt. Otrzymana wartość $p = 0.9389$ nie daje podstaw do odrzucenia hipotezy zerowej, co oznacza, że średnia reszt jest statystycznie równa zero (rysunek 13).



Rysunek 13: Wykres wartości reszt modelu ARMA(5, 5) dopasowanego do zbioru treningowego. Czerwoną przerywaną linią zaznaczono średnią reszt o wartości około -0.00123 .

4.2 Założenie dotyczące wariancji

W celu weryfikacji założenia o stałej wariancji reszt zastosowano test ARCH (Autoregressive Conditional Heteroskedasticity). Test ten sprawdza, czy wariancja reszt w danym momencie czasu zależy od kwadratów reszt z poprzednich chwil, co jest charakterystyczne dla procesów o zmiennej zmienności.

Hipoteza zerowa testu ARCH zakłada brak efektu ARCH, czyli stałą wariancję reszt w czasie. Odrzucenie hipotezy zerowej oznacza występowanie heteroskedastyczności warunkowej, czyli zmiennej wariancji zależnej od przeszłych zaburzeń losowych.

Otrzymana wartość $p = 0.0000$ prowadzi do jednoznacznego odrzucenia hipotezy zerowej, co wskazuje na obecność efektu ARCH w resztach modelu ARMA. Oznacza to, że wariancja procesu nie jest stała w czasie, a model ARMA nie opisuje w pełni struktury zmienności danych.

W celu lokalizacji zmiany reżimu wariancji wykorzystano sumy skumulowanych kwadratów

reszt:

$$C_j = \sum_{i=1}^j \epsilon_i^2, \quad (5)$$

oraz funkcję błędu:

$$V(j) = \frac{j}{n} C_n - C_j. \quad (6)$$

Rysunek 14 wskazuje punkt l , w którym zmiana wariancji jest największa.

4.3 Założenie dotyczące niezależności

Do weryfikacji założenia o niezależności reszt zastosowano test Ljunga–Boxa, który bada, czy grupa autokorelacji do zadanego opóźnienia różni się istotnie od zera. Test ten sprawdza hipotezę zerową, że reszty są nieskorelowane, czyli mają charakter białego szumu.

Otrzymana wartość $p = 0.2137$ nie daje podstaw do odrzucenia hipotezy zerowej, co oznacza, że w resztach nie występuje istotna autokorelacja, a model poprawnie opisuje zależności czasowe w średniej procesie.

4.3.1 Autokorelacja reszt

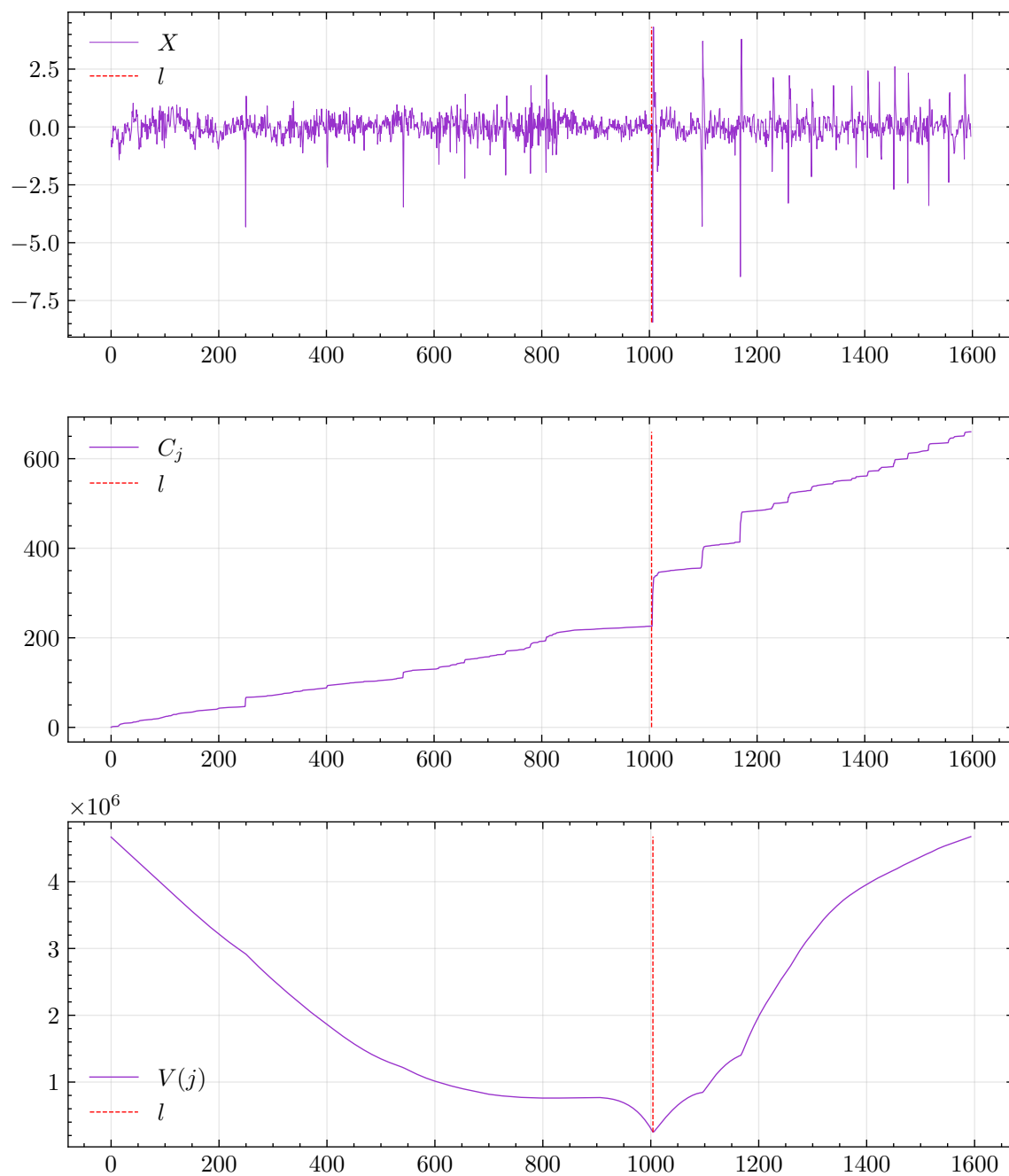
Reszty (błędy) w modelu ARMA/ARIMA reprezentują różnicę między obserwowanymi wartościami szeregu czasowego a wartościami przewidywanymi (dopasowanymi) przez model. Poprawne reszty powinny zachowywać się jak „biały szum”, czyli mieć stałą wariancję, zerową średnią i brak autokorelacji. Rysunki 15 i 16 przedstawiają wykresy ACF i PACF reszt. Wszystkie współczynniki mieszczą się w granicach przedziałów ufności, co potwierdza brak autokorelacji i spełnienie założenia białego szumu.

4.4 Założenie dotyczące normalności rozkładu

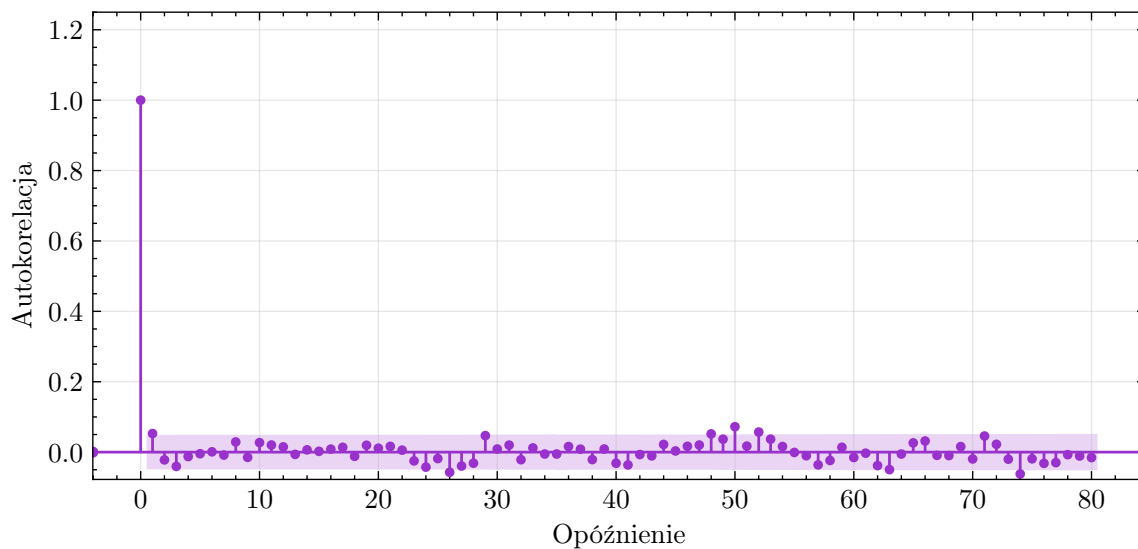
Normalność rozkładu reszt zweryfikowano przy użyciu testu Jarque–Bera, który opiera się na analizie skośności i kurtozy rozkładu. Hipoteza zerowa testu zakłada zgodność rozkładu reszt z rozkładem normalnym. Otrzymana wartość $p = 0.0000$ prowadzi do odrzucenia hipotezy zerowej, co oznacza, że rozkład reszt istotnie odbiega od normalnego. Tego typu zachowanie jest często obserwowane w danych pochodzących z procesów fizycznych o charakterze turbulentnym. Histogram oraz wykres QQ (Rysunek 17) wskazują na cięższe ogony rozkładu, co jest typowe dla danych fizycznych pochodzących z procesów turbulentnych.

5 Wnioski

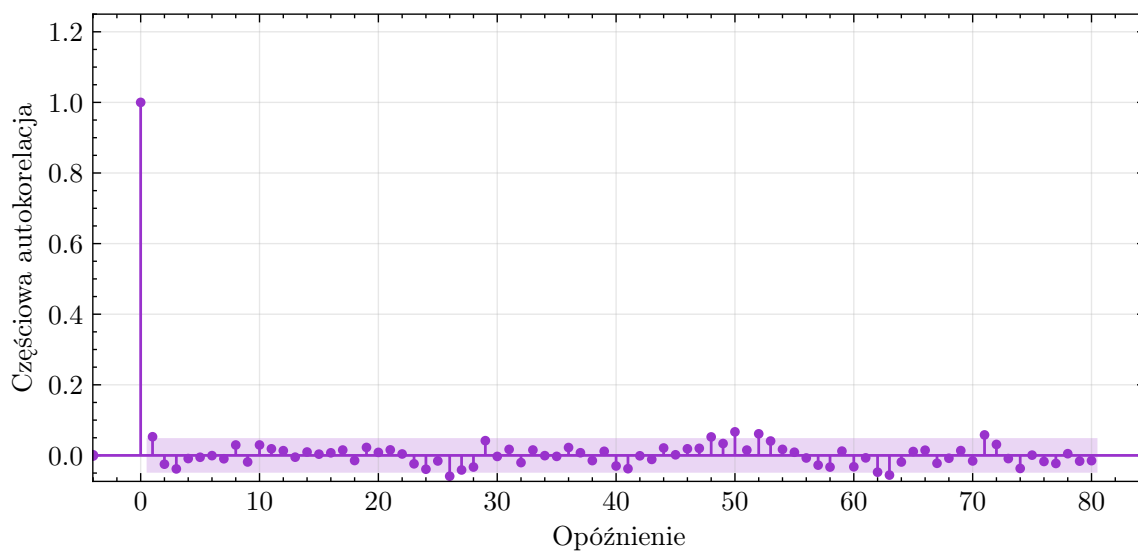
Przeprowadzona analiza natężenia prądu plazmy w fazie stabilnej eksperymentu tokamakowego pozwoliła na ocenę możliwości opisu badanego szeregu czasowego przy użyciu modelu ARMA. Wykazano, że surowe dane są niestacjonarne, jednak zastosowanie różnicowania



Rysunek 14: Wykresy przedstawiające kolejno: trajektorię reszt modelu ARMA(5, 5) dopasowanego do zbioru treningowego, sumę skumulowanych kwadratów reszt C_j oraz sumę błędów kwadratowych $V(j)$, w zależności od czasu. Czerwoną linią przerywaną zaznaczono punkt l , czyli punkt największej zmiany reżimu wariancji.



Rysunek 15: Wykres współczynnika autokorelacji reszt modelu ARMA(5, 5) w zależności od opóźnienia (parametru h). Półprzezroczystym obszarem zaznaczono przedział ufności o poziomie 95% dla hipotezy o braku korelacji (szum biały).



Rysunek 16: Wykres współczynnika częściowej autokorelacji reszt modelu ARMA(5, 5) w zależności od opóźnienia (parametru h). Półprzezroczystym obszarem zaznaczono przedział ufności o poziomie 95% dla hipotezy o braku korelacji (szum biały).

pierwszego rzędu skutecznie doprowadziło do uzyskania szeregu stacjonarnego, co potwierdziły zarówno analiza ACF i PACF, jak i test Augmented Dickey–Fullera.

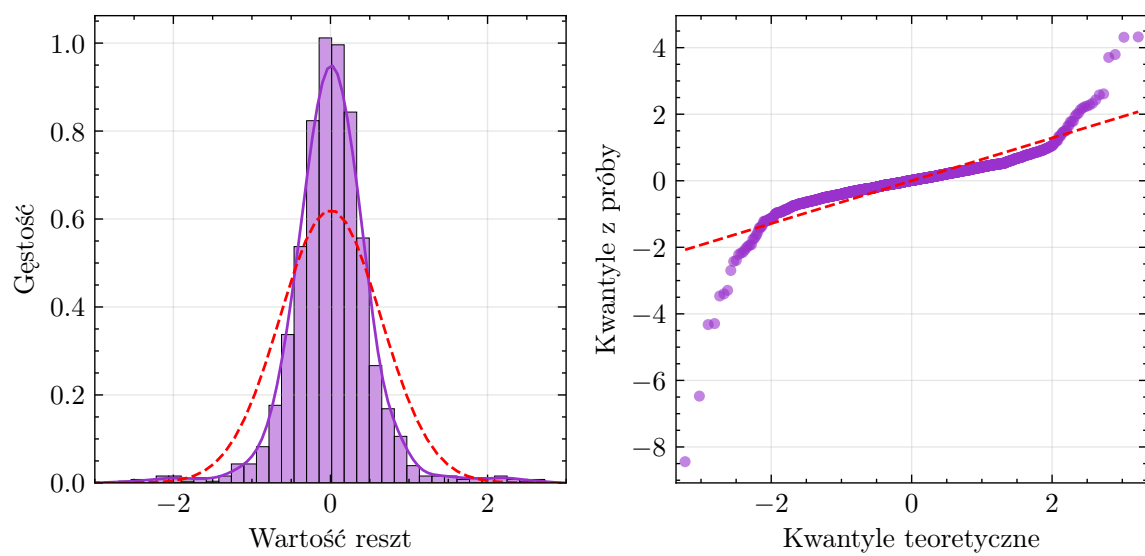
Na podstawie kryteriów informacyjnych AIC i BIC dobrano model ARMA(5,5), który dobrze odwzorowuje strukturę zależności czasowych obecnych w danych. Analiza dopasowania oraz prognoz krótkoterminowych wykazała, że model poprawnie opisuje średnią dynamikę procesu, a obserwacje rzeczywiste w większości mieszczą się w wyznaczonych przedziałach ufności.

Jednocześnie weryfikacja założeń modelowych ujawniła istotne ograniczenia. Pomimo spełnienia założeń o zerowej średniej i braku autokorelacji reszt, stwierdzono występowanie zmiennej wariancji oraz odstępstwa od normalności rozkładu. Wskazuje to, że model ARMA nie w pełni opisuje statystyczne własności fluktuacji prądu plazmy.

6 Podsumowanie

W pracy zastosowano klasyczne metody analizy szeregów czasowych do modelowania rzeczywistych danych pochodzących z eksperymentu fizycznego. Uzyskane wyniki pokazują, że modele ARMA mogą być użyteczne do opisu krótkoterminowej dynamiki badanego procesu, jednak ich zastosowanie jest ograniczone w przypadku danych charakteryzujących się heteroskedastycznością i nienormalnością rozkładu.

W związku z tym dalsze badania powinny obejmować wykorzystanie modeli uwzględniających zmienność wariancji, takich jak ARCH lub GARCH, które lepiej odpowiadają naturze analizowanych danych plazmowych.



Rysunek 17: Histogram i empiryczna gęstość oraz wykres kwantylowy (QQ-plot) wartości reszt w porównaniu do rozkładu normalnego (czerwona, przerywana linia) o parametrach wyznaczonych na podstawie badanej próby reszt (średniej i wariancji z próby).

Literatura

- [1] J. Gajda, G. Sikora, and A. Wyłomańska. Regime variance testing — a quantile approach. *Acta Physica Polonica B*, 44(5):1015, 2013.
- [2] Samuel Jackson, Saiful Khan, Nathan Cummings, James Hodson, Shaun de Witt, Stanislas Pamela, Rob Akers, and Jeyan Thiyagalingam. An Open Data Service for Supporting Research in Machine Learning on Tokamak Data. *IEEE Transactions on Plasma Science*, 2025.
- [3] Samuel Jackson, Saiful Khan, Nathan Cummings, James Hodson, Shaun de Witt, Stanislas Pamela, Rob Akers, Jeyan Thiyagalingam, and The MAST Team. Fair-mast: A fusion device data management system. *SoftwareX*, 27:101869, 2024.
- [4] Skipper Seabold and Josef Perktold. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*, 2010.