

# Komputerowa analiza szeregów czasowych

<b>Kierunek, nazwa wydziału</b> Matematyka stosowana, Wydział Matematyki	<b>Grupa ćwiczeniowa, termin zajęć</b> T00-79a Wtorek 15:15
<b>Imię, nazwisko, numer albumu</b> Małgorzata Kowalczyk 262295, Julia Mazur 262296	<b>Data oddania</b> 09.02.2023 r.
<b>Tytuł</b> Analiza danych rzeczywistych przy pomocy modelu ARMA	<b>Prowadzący</b> Dr inż. Aleksandra Grzesiek



## Spis treści

<b>1 Wstęp</b>	<b>2</b>
1.1 Cel pracy . . . . .	2
1.2 Opis danych . . . . .	2
1.3 Wizualizacja danych . . . . .	2
1.4 Podstawowe statystyki . . . . .	4
<b>2 Przygotowanie danych do analizy</b>	<b>5</b>
2.1 Zbadanie jakości danych . . . . .	5
2.2 Podział danych na zbiór testowy oraz treningowy . . . . .	5
2.3 Dekompozycja szeregu czasowego . . . . .	6
<b>3 Modelowanie danych przy pomocy ARMA</b>	<b>13</b>
3.1 Dobranie rzędu modelu . . . . .	13
3.2 Estymacja parametrów modelu wybraną metodą . . . . .	15
<b>4 Ocena dopasowania modelu</b>	<b>16</b>
4.1 Przedziały ufności dla PACF/ACF . . . . .	16
4.2 Porównanie linii kwantylowych z trajektorią . . . . .	18
4.3 Prognoza dla przyszłych obserwacji . . . . .	19
4.4 Wnioski . . . . .	19
<b>5 Weryfikacja założeń dotyczących szumu</b>	<b>20</b>
5.1 Założenie dotyczące średniej oraz wariancji . . . . .	20
5.2 Założenie dotyczące niezależności . . . . .	21
5.3 Założenie dotyczące normalności rozkładu . . . . .	23
5.4 Wnioski . . . . .	25
<b>6 Zakończenie</b>	<b>26</b>

# **1 Wstęp**

## **1.1 Cel pracy**

Niniejszy raport dotyczy analizy danych rzeczywistych, które opisują średnie temperatury powietrza w wybranym mieście, dla konkretnego przedziału czasowego. Celem pracy jest modelowanie tych obserwacji, z wykorzystaniem modelu ARMA, tak aby dobrać odpowiedni rząd modelu oraz wyestymować jego parametry. W kolejnym etapie uzyskane dopasowanie zostanie sprawdzone, a następnie zweryfikowane zostaną założenia dotyczące szumu.

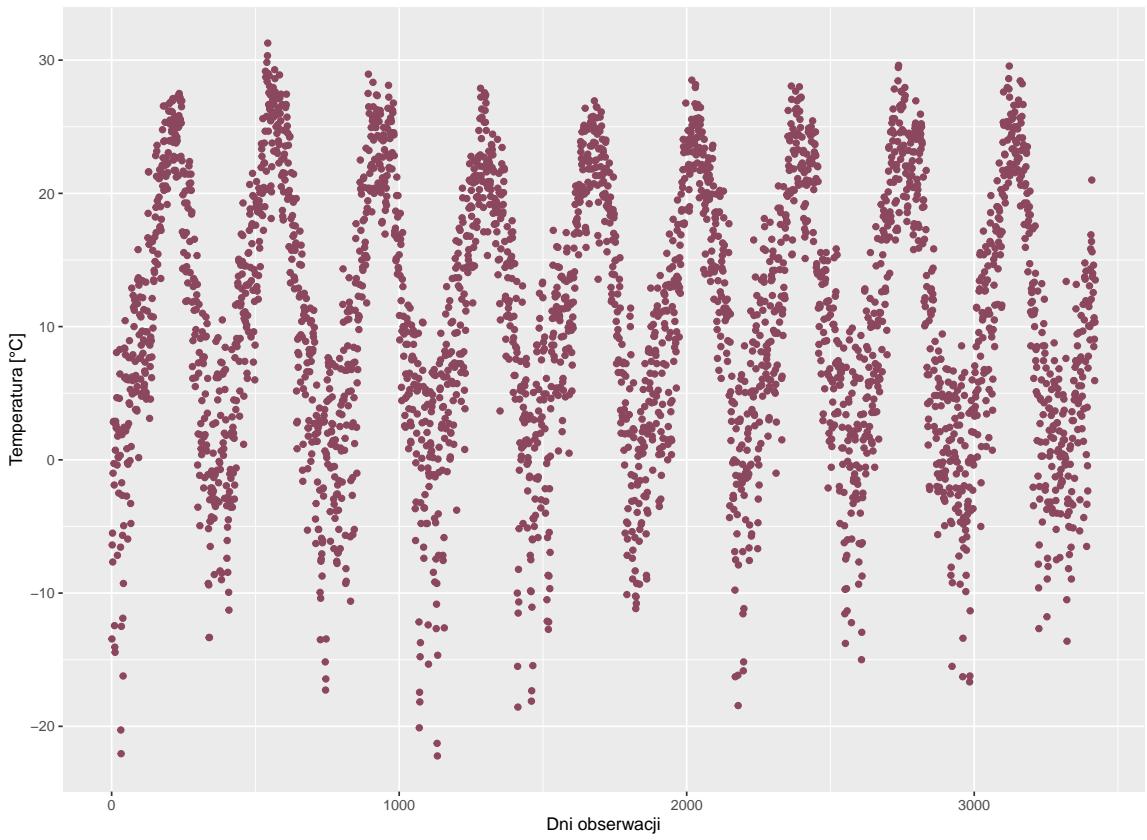
Wszystkie wyniki zaprezentowane w pracy otrzymano przy pomocy języka R. Ponadto, w trakcie realizacji sprawozdania korzystano z wykładów dr hab. inż. Agnieszki Wyłomańskiej.

## **1.2 Opis danych**

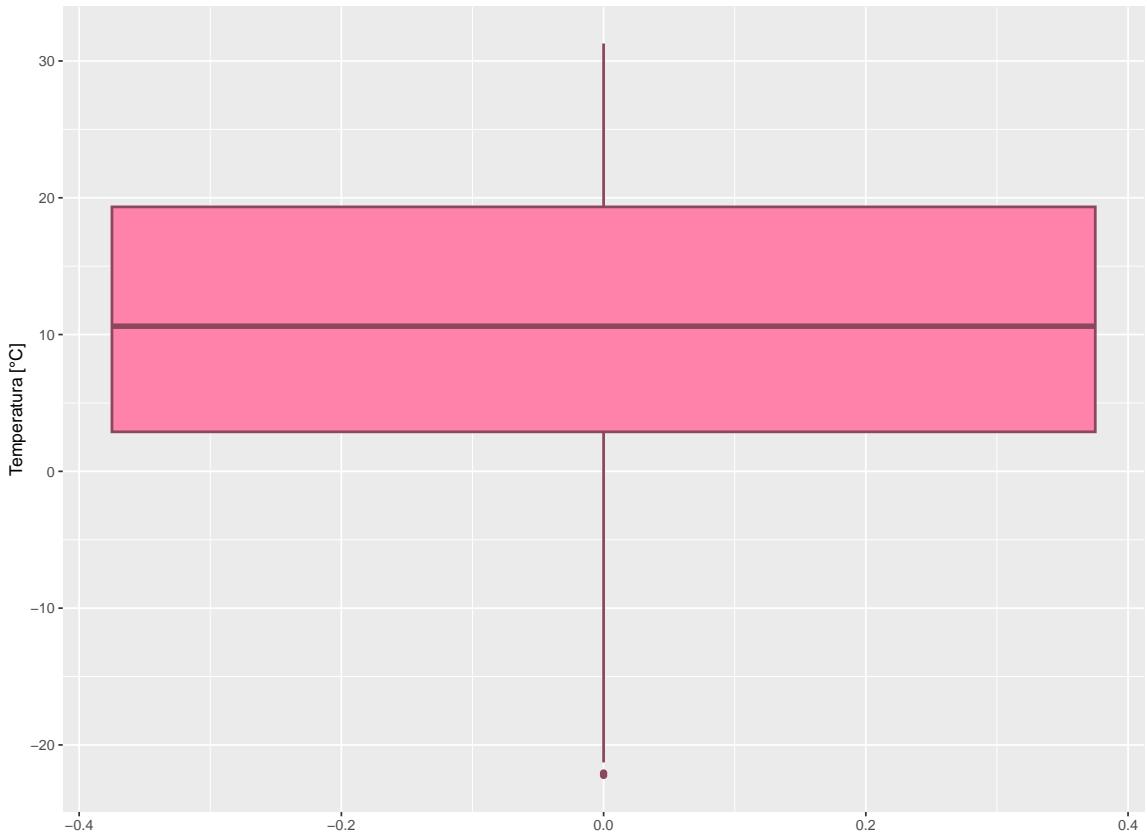
Dane, które zostały wykorzystane w pracy, pobrano ze strony <https://www.kaggle.com/>. Jest to obszerny zbiór zawierający informacje o średniej temperaturze w danym dniu z wielu różnych miast na świecie. Każdy pomiar dotyczy jednego dnia, zatem punkty te są równooddalone w czasie, co spełnia założenia szeregu czasowego. Do przeprowadzonej analizy zostały poddane temperatury powietrza w Denver występujące od 1 stycznia 2011 roku do 13 maja 2020 roku. Otrzymano 3421 takich obserwacji. W celu przejrzystości wykresów postanowiono także wszystkie temperatury ukazywać w stopniach Celsjusza. Dokonano tego poprzez odpowiednie przemnożenie każdego rekordu.

## **1.3 Wizualizacja danych**

W celu zobrazowania analizowanych danych stworzono wykres przedstawiający wartości temperatur w Denver oraz wykres pudełkowy.



**Wykres 1.1:** Rozkład badanej temperatury



**Wykres 1.2:** Boxplot dla badanej temperatury

Na wykresie 1.1 łatwo zauważyc, że temperatury co pewien okres powtarzają się, co jest zrozumiałe, ponieważ wpływa na nie występująca aktualnie pora roku. Wykres pudełkowy 1.2 informuje, że zakres występujących temperatur w Denver wynosi około  $50^{\circ}C$ , ale głównie dominują wartości dodatnie. Rozkład jest delikatnie lewostronnie skośny, ponieważ obserwacje odstające, znajdują się pod pudełkiem. Oznacza to, że w Denver zdarzyły się dni, w których temperatura była wyjątkowo niska.

#### 1.4 Podstawowe statystyki

W poniższej sekcji wyliczono podstawowe statystyki dla temperatury.

**Tabela 1.1:** Miary położenia temperatury

Średnia	Minimum	Kwartyl pierwszy	Kwartyl drugi	Kwartyl trzeci	Maksimum
10.50	-22.22	2.89	10.61	19.33	31.28

**Tabela 1.2:** Miary rozproszenia temperatury

Rozstęp międzykwartylowy	Rozstęp	Wariancja	Odcylenie standardowe	Współczynnik zmienności
16.44	53.50	103.41	10.17	0.97

**Tabela 1.3:** Miary asymetrii i miary spłaszczenia temperatury

Współczynnik skośności	Kurtoza
-0.24	-0.66

Wartości przedstawione w tabelach 1.1, 1.2, 1.3 pozwalają wysunąć następujące wnioski.

- Wartość minimalna temperatury w badanym czasie wynosi  $-22.22^{\circ}C$ , natomiast wartość maksymalna jest równa  $31.28^{\circ}C$ . Średnia temperatura w Denver to  $10.50^{\circ}C$ . Mediana, która dzieli dane na dwa podzbiory (powyżej i poniżej jej wartości) o równej liczbie obserwacji wynosi  $10.61^{\circ}C$ . Oznacza to, że w połowie dni w badanym okresie temperatura była taka bądź wyższa.
- Otrzymane kwantyle pozwalają wysunąć wniosek, że 25% dni w Denver, to dni z temperaturą wyższą niż  $19.33^{\circ}C$ , natomiast przez 25% dni występuje temperatura  $2.89^{\circ}C$  bądź niższa.
- Wysoki współczynnik zmienności, bo wynoszący aż 0.97 świadczy o dużej zmienności temperatur. Lata w Denver są gorące, natomiast zimy mroźne.
- Współczynnik skośności wyniósł -0.24, co oznacza, że otrzymano niewielką lewostronną skośność. Sugeruje to, częstsze występowanie cieplejszych dni w Denver niż zimniejszych.
- Kurtoza dla temperatury osiągnęła wartość -0.66, co świadczy o tym, że jest więcej skrajnych wartości odstających, niż w rozkładzie normalnym co powoduje spłaszczenie rozkładu. Ma to swoje uzasadnienie, ponieważ rozważane są wartości temperatur w ciągu wielu lat, które co roku są do siebie zbliżone.

## 2 Przygotowanie danych do analizy

### 2.1 Zbadanie jakości danych

Przygotowując dane do analizy, zweryfikowano, czy nie występują w nich wartości skrajnie odstające. Rozkład temperatury zaprezentowany na wykresie 1.1 świadczy o braku takich obserwacji. Potwierdza to też wykres pudełkowy 1.2. Kolejnym krokiem było zbadanie, czy w badanym zbiorze znajdują się braki. Po przeanalizowaniu otrzymano, że analizowane dane są poprawne. Nie wykryto również błędów w próbkowaniu, ponieważ każdy pomiar temperatury dotyczy innego dnia, zatem dane te są równooddalone w czasie.

### 2.2 Podział danych na zbiór testowy oraz treningowy

W celu dokładnej analizy danych utworzono zbiór testowy, który zawiera 20 ostatnich obserwacji. Dla zwiększenia czytelności raportu pozostałe wartości, tworzące zbiór treningowy, oznaczono jako  $Y_t$  i to na nich dokonywano kolejnych przekształceń.



Wykres 2.1: Podział zbioru danych na dane treningowe i testowe

Na wykresie 2.1 przedstawiono wartości dla  $t \in [2500, 3421]$ , dla lepszej widoczności danych testowych.

### 2.3 Dekompozycja szeregu czasowego

Dekompozycja szeregu czasowego polega na wyodrębnieniu wszystkich możliwych elementów szeregu czasowego, takich jak trend czy wahania okresowe tak, aby otrzymana zmienna losowa  $X_t$  okazała się stacjonarnym procesem losowym  $\{X_t\}$ .

W dekompozycji Walda zakłada się, że

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

gdzie

- $Y_t$  to dane pomiarowe,
- $\{X_t\}$  jest szeregiem stacjonarnym w słabym sensie,
- $w(t)$  jest komponentem odpowiedzialnym za trend deterministyczny,
- $s(t)$  jest periodycznie zmienną funkcją zwaną sezonowością.

**Twierdzenie 1** Szereg czasowy  $\{X_t\}$  jest stacjonarny w słabym sensie, jeśli spełnione są następujące warunki:

1.  $\mu_X(t) = \mathbf{E}X_t = const$ ,
2.  $cov(X_t, X_{t+h}) = \gamma(h)$ , co oznacza, że funkcja autokowariancji zależy tylko od  $h$ .

Analizę rozpoczęto od wizualizacji danych przy pomocy funkcji autokorelacji oraz funkcji częściowej autokorelacji.

**Definicja 1** Funkcja autokorelacji (ACF) informuje o tym, jak zmienia się korelacja między dwiema wartościami szeregu oddalonymi od siebie o  $h$ , nazywanym opóźnieniem.

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

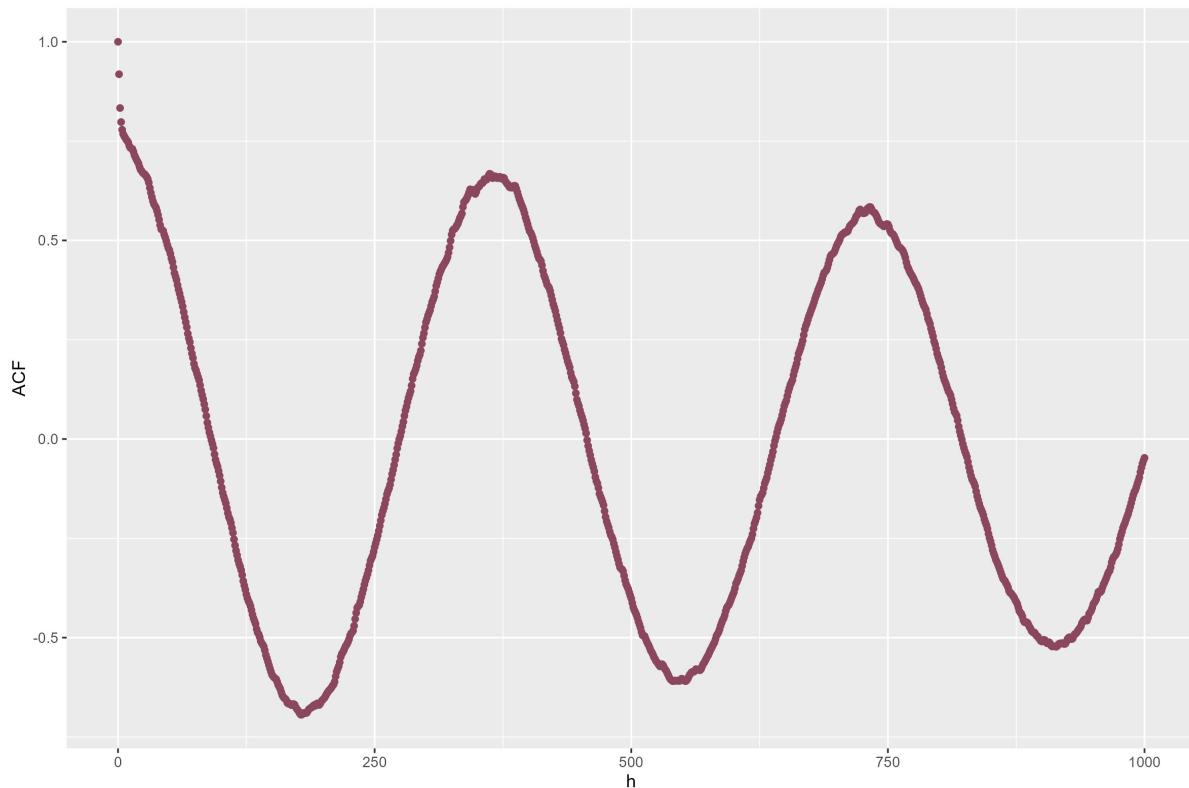
przy czym  $\hat{\gamma}(h)$  to funkcja autokowariancji zdefiniowana następującym wzorem:

$$\hat{\gamma}(h) = \frac{1}{n} \sum_{i=1}^{n-|h|} (e_{i+|h|} - \bar{e})(e_i - \bar{e}),$$

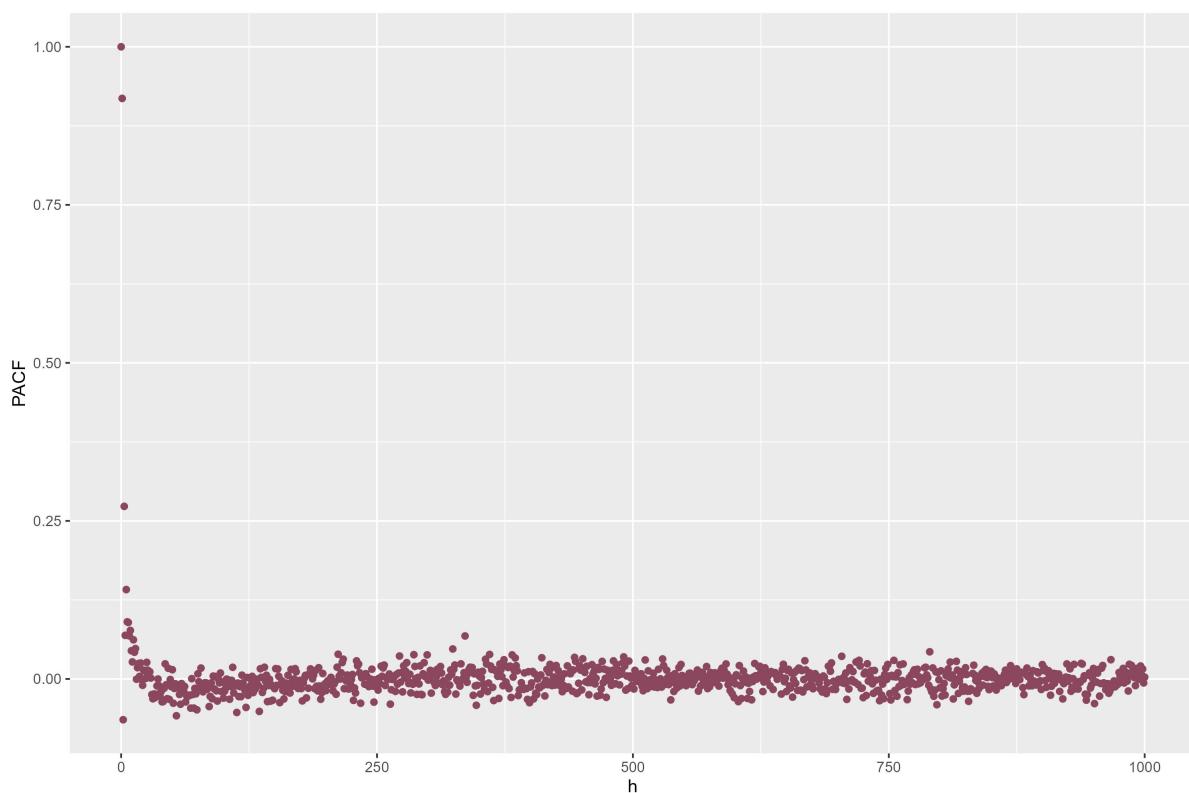
gdzie  $-n < h < n$ ,  $h \in \mathbb{Z}$ .

**Definicja 2** Funkcja częstekowej autokorelacji (PACF) dla danego opóźnienia jest częstekową korelacją szeregu czasowego z samym sobą, która uwzględnia wszystkie informacje, które pojawiają się między analizowanymi punktami w czasie.

Otrzymane ACF i PACF przedstawiono na wykresach 2.2 i 2.3.

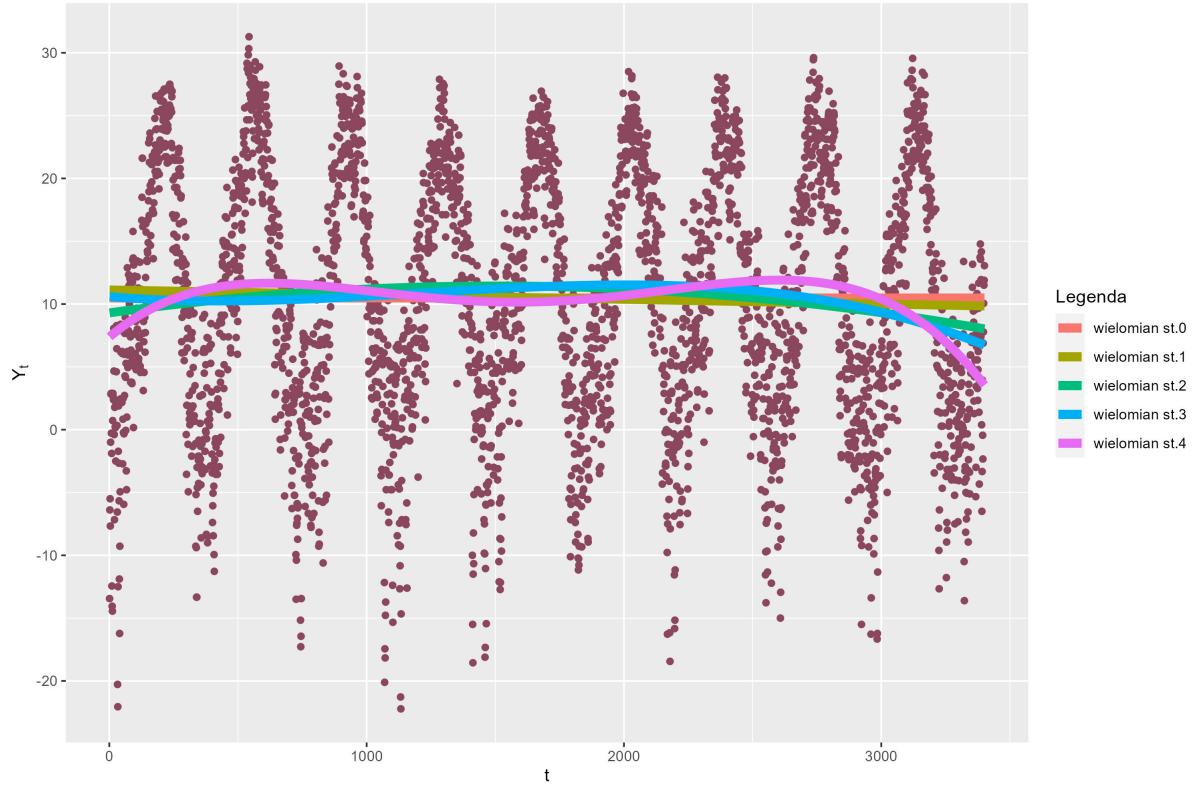


**Wykres 2.2:** Funkcja empirycznej autokorelacji dla  $Y_t$



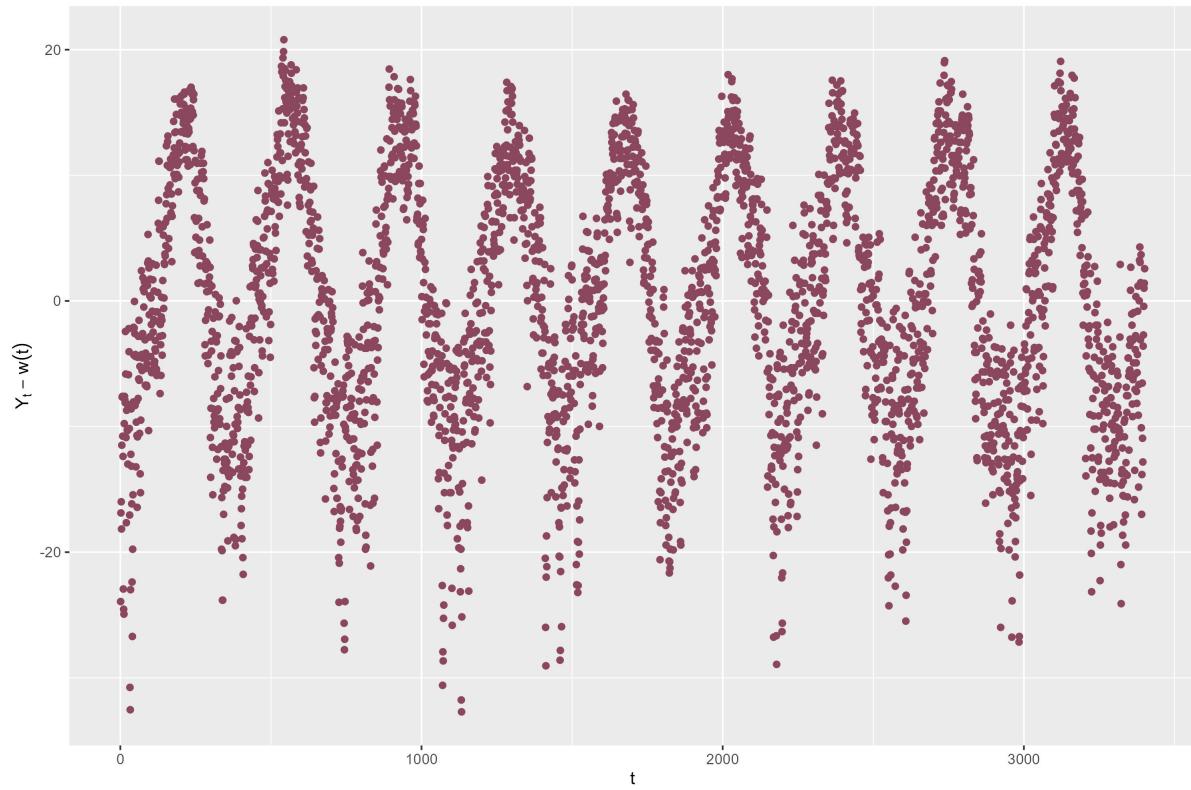
**Wykres 2.3:** Funkcja empirycznej autokorelacji cząstkowej dla  $Y_t$

W celu sprawdzenia niestacjonarności surowych danych przeprowadzono test ADF na poziomie istotności  $\alpha = 0.05$  dla  $Y_t$ , w którym otrzymano statystykę o wartości  $-4.0314$  oraz p-wartość równą  $0.01$ . Na tej podstawie odrzucono hipotezę zerową i przyjęto, że badany szereg jest stacjonarny w słabym sensie. Jednak na podstawie wykresu empirycznej autokorelacji 2.2 postanowiono wykonać dekompozycję Walda po to, by usunąć ewentualny trend i widoczną sezonowość. W tym celu do  $Y_t$  dopasowano wielomiany stopnia  $k \in \{0, 1, 2, 3, 4\}$  a wyniki zaprezentowano na wykresie 2.4.



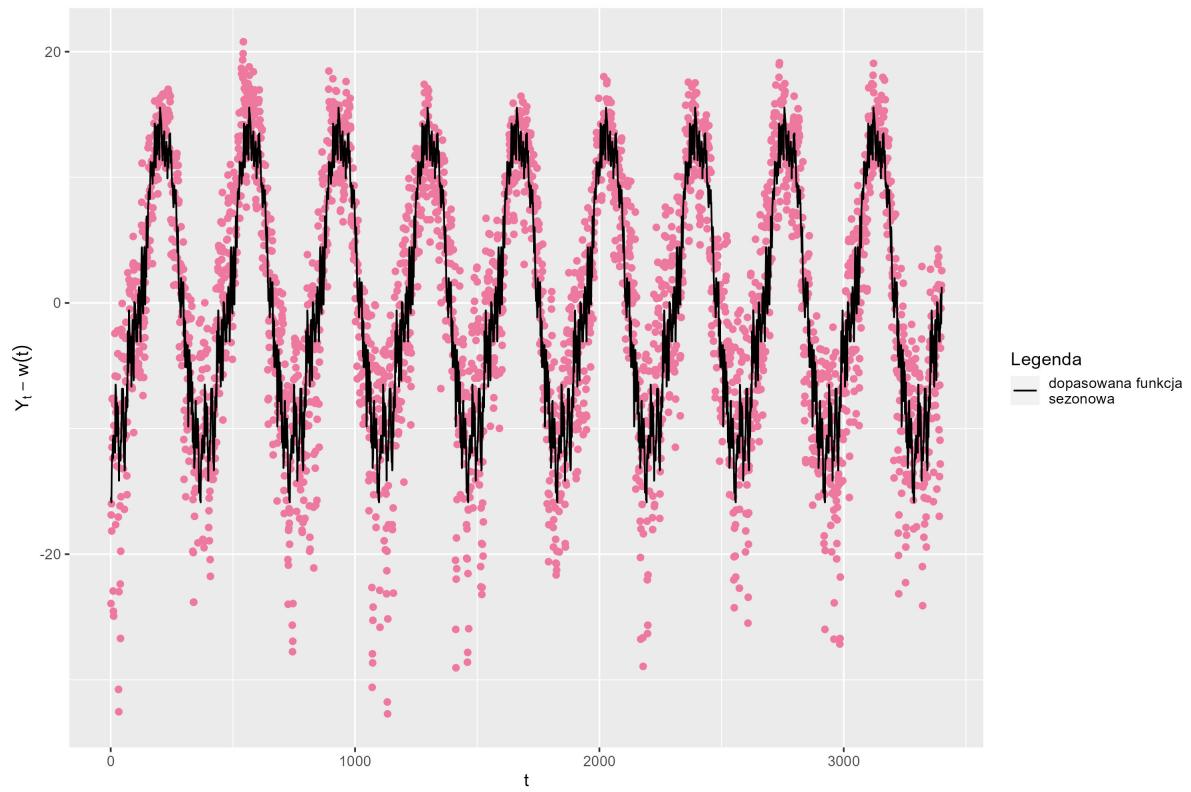
**Wykres 2.4:** Dopasowanie wielomianów do  $Y_t$

Zauważono, że w danych nie ma widocznego trendu. Zdecydowano się na usunięcie z danych wielomianu stopnia  $k = 0$  w celu ułatwienia dalszych przekształceń.

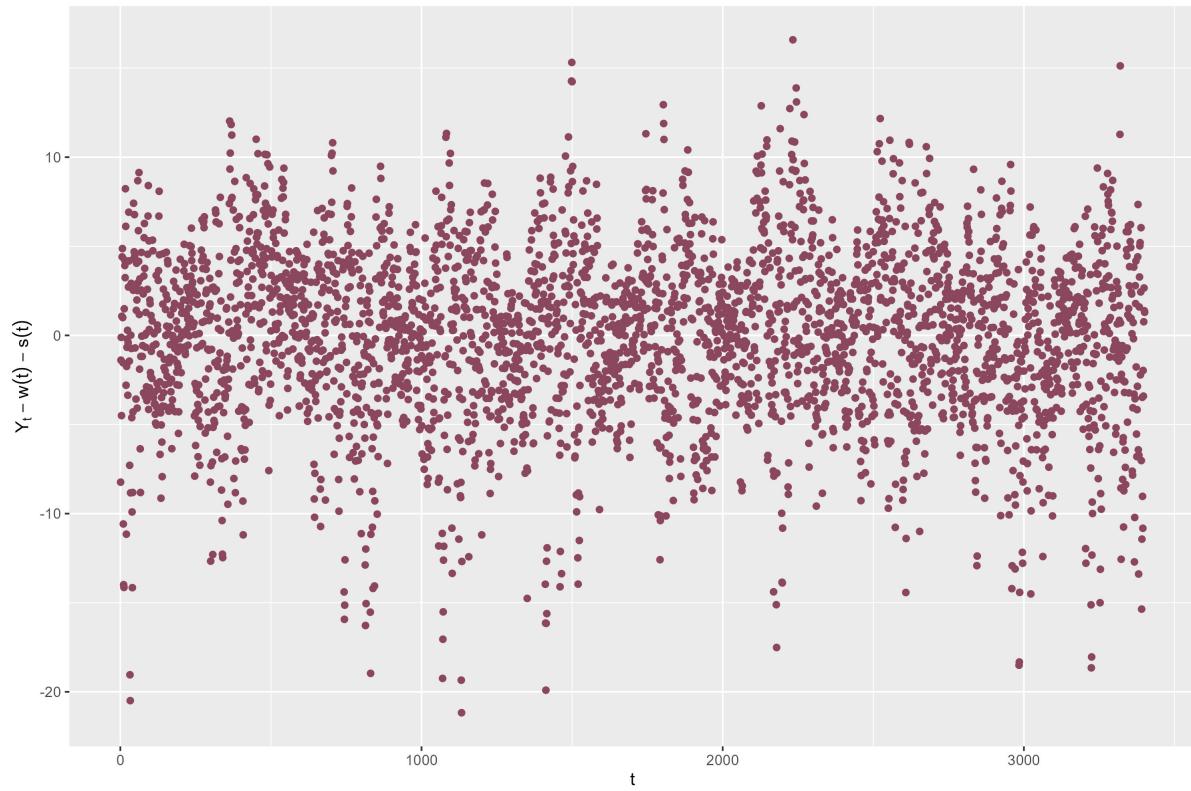


**Wykres 2.5:**  $Y_t$  z usuniętym odpowiednim wielomianem stopnia 0

Kolejnym etapem dekompozycji jest usunięcie sezonowości. W tym celu wykorzystano informację o cykliczności danych z rocznym okresem. Wyznaczono dla nich średnią co okres dla każdej wartości  $t$  z pojedynczego okresu. Tak otrzymaną średnią funkcję sezonową powielono tak, aby odpowiadała liczbie badanych obserwacji. Wyniki zaprezentowano na wykresie 2.6.

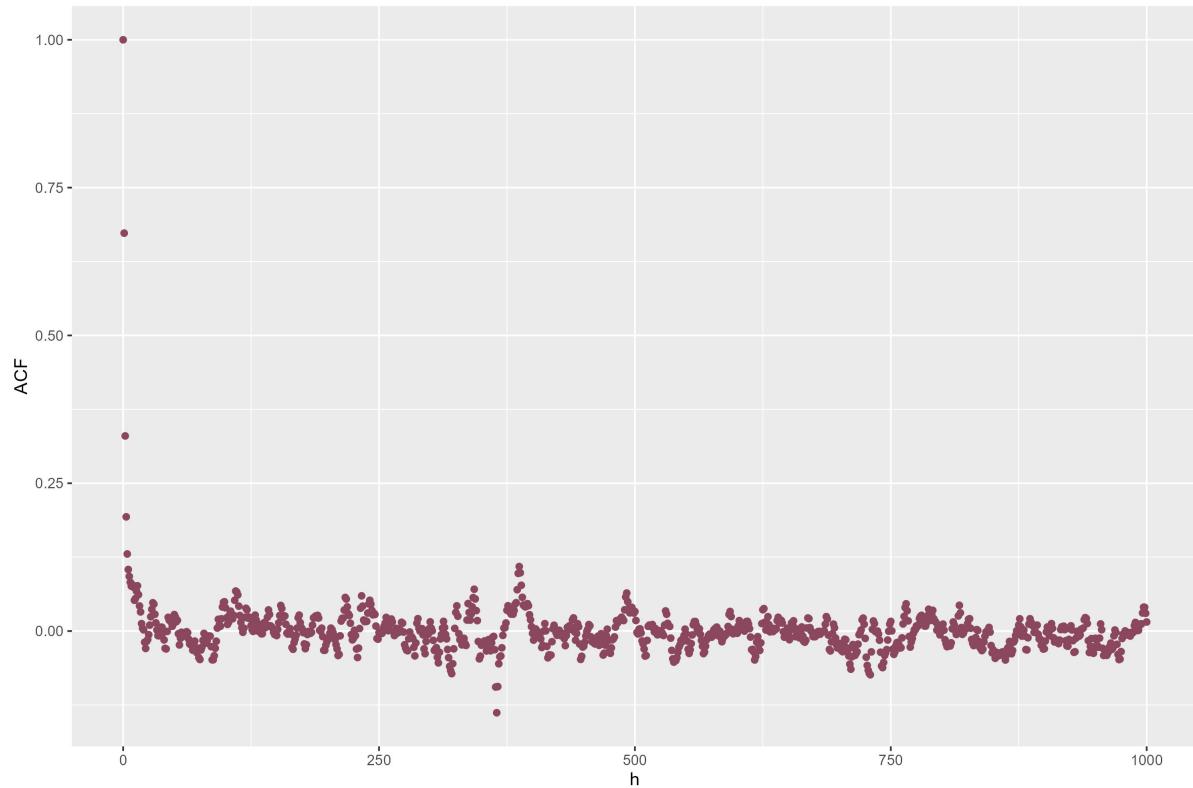


**Wykres 2.6:**  $Y_t - w(t)$  z nałożoną wyznaczoną sezonowością

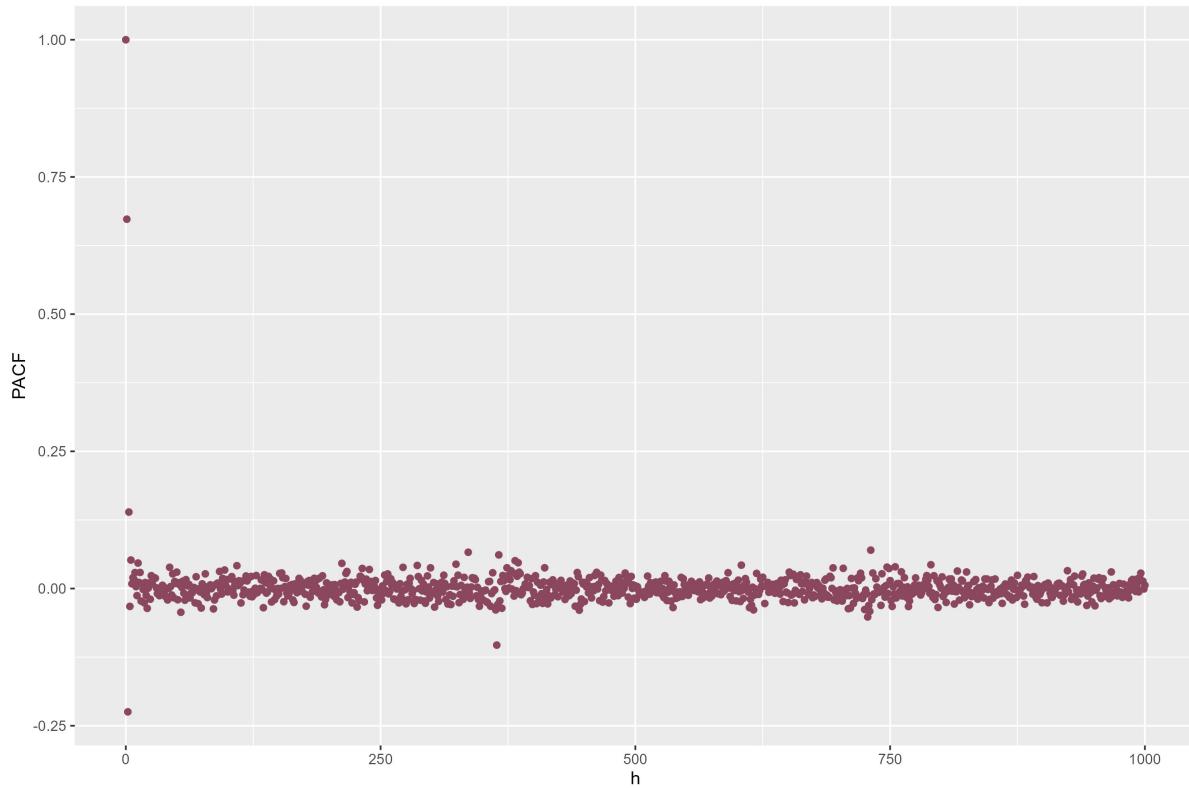


**Wykres 2.7:**  $Y_t - w(t)$  z usuniętą sezonowością

Po usunięciu sezonowości z danych, przekształcone wartości przedstawiono na wykresie 2.7. W końcowej fazie dekompozycji Walda zwizualizowano funkcje empirycznej auto-korelacji i autokorelacji częstkowej (wykresy 2.8 oraz 2.9), aby upewnić się, że dobrze usunięto sezonowość.



**Wykres 2.8:** Funkcja empirycznej autokorelacji dla  $Y_t - w(t) - s(t)$



**Wykres 2.9:** Funkcja empirycznej autokorelacji cząstkowej dla  $Y_t - w(t) - s(t)$

Na podstawie powyższych wykresów można stwierdzić, że dane są poprawne, ponieważ wykazują krótką pamięć i szybko zanikają do 0. W celu sprawdzenia, czy mimo przekształceń szereg pozostał stacjonarny, ponownie wykonano ADF test na poziomie istotności  $\alpha = 0.05$  tym razem na  $X_t = Y_t - w(t) - s(t)$  i otrzymano statystykę -11.787 oraz p-wartość równą 0.01. Zatem przyjęto, że szereg  $X_t$  jest stacjonarny w słabym sensie, co pozwala na przewidywanie przyszłego zachowania się szeregu.

### 3 Modelowanie danych przy pomocy ARMA

#### 3.1 Dobranie rzędu modelu

W celu dobrania rzędu modelu ARMA( $p, q$ ) wykorzystano dwa kryteria informacyjne.

- Bayesowskie kryterium informacyjne:

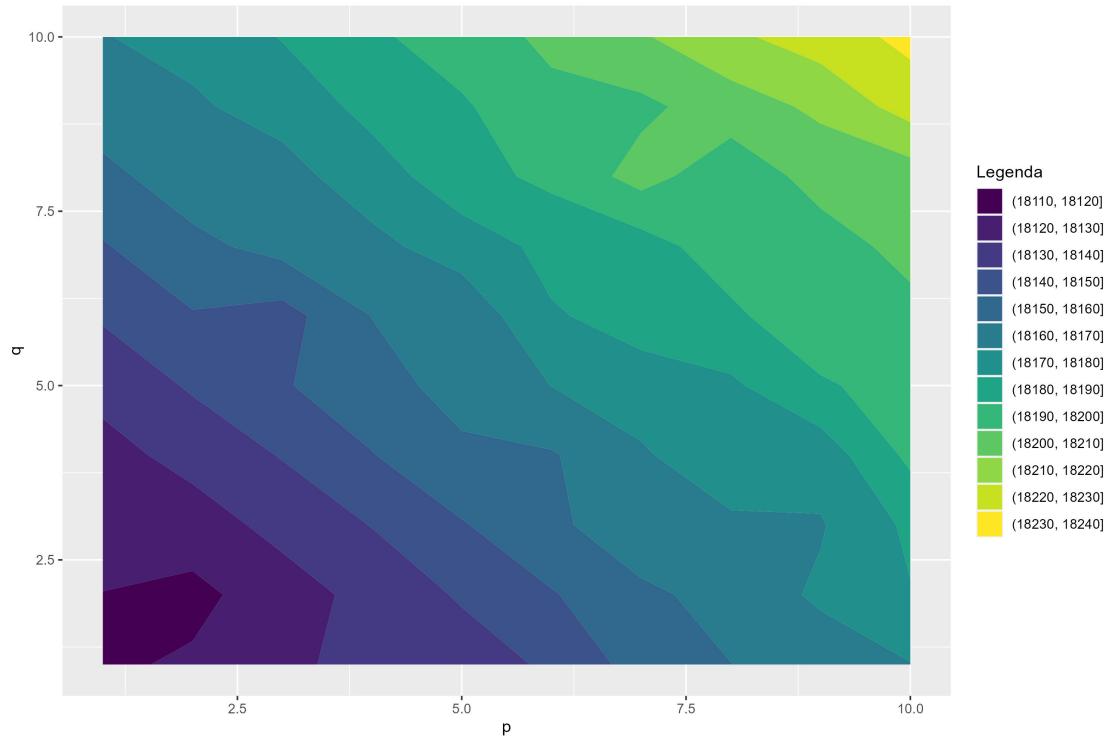
$$BIC = -2 \ln(L) + (p + q) \ln(n),$$

- Kryterium informacyjne Akaikiego:

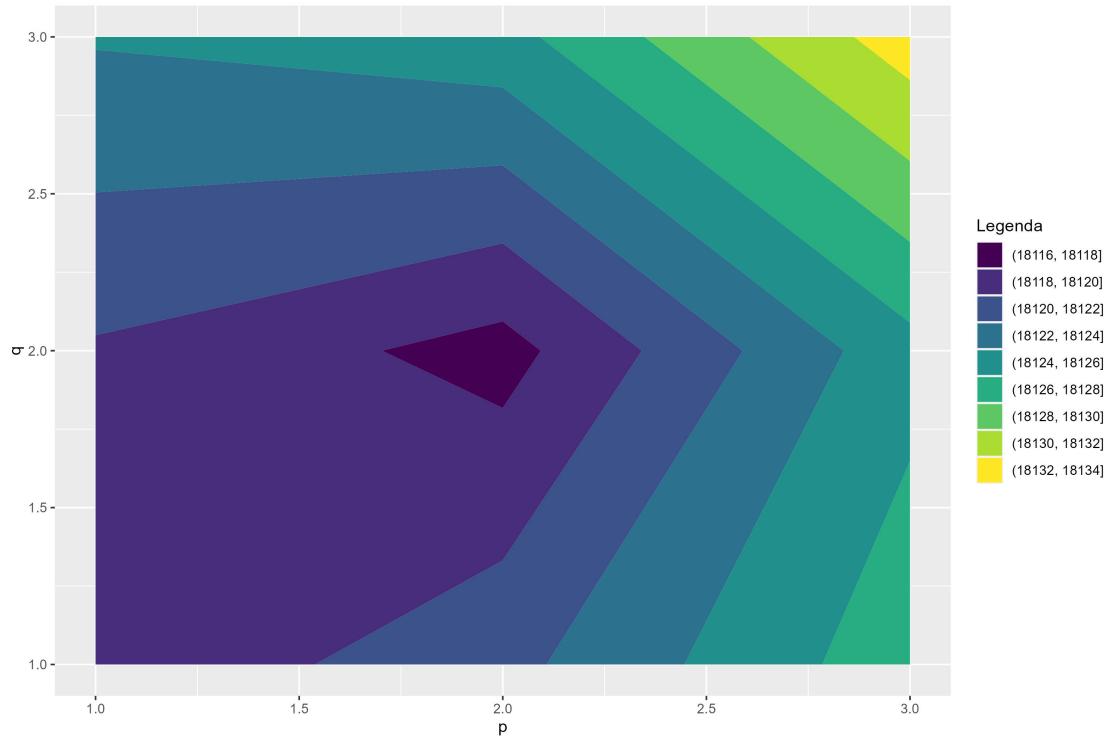
$$AIC = -2 \ln(L) + 2(p + q),$$

gdzie  $L$  to funkcja wiarygodności dla residuum, a  $n$  to liczba obserwacji.

Dla wybranego zakresu  $p \in (0, 10]$ ,  $q \in (0, 10]$  policzono wartości dla obu wyżej wymienionych kryteriów. Preferowane współczynniki modelu spośród wielu badanych par to te, dla których wartość kryterium będzie najmniejsza.

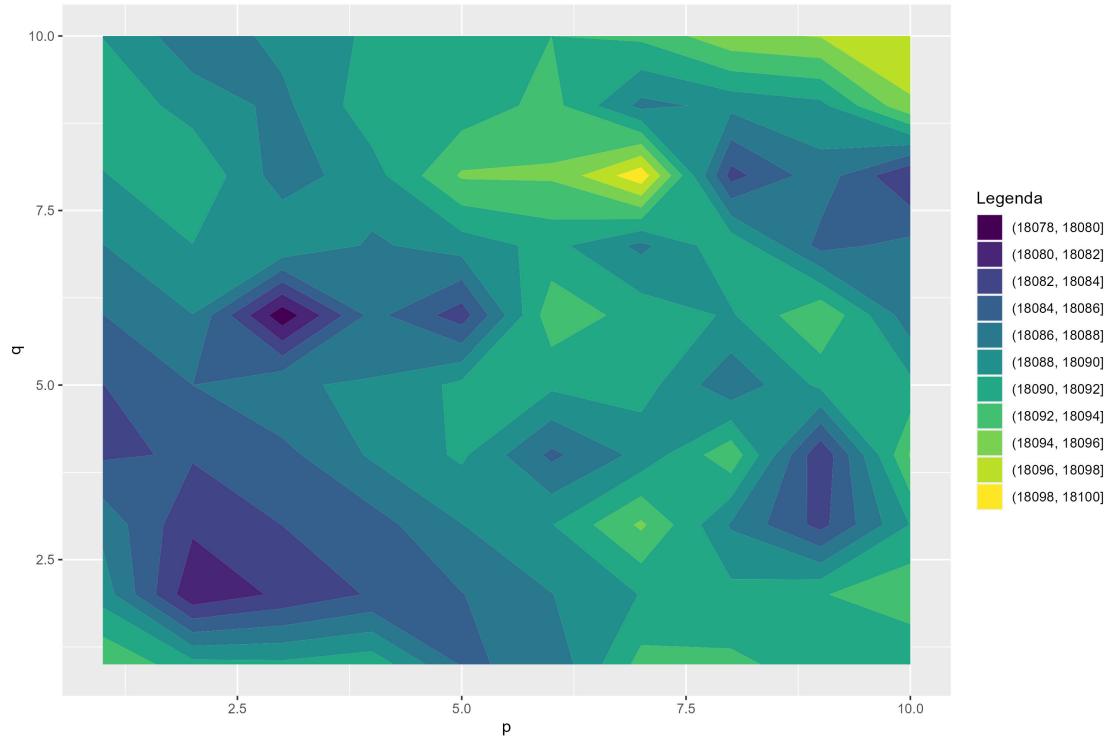


**Wykres 3.1:** Wartości BIC w zależności od  $p$  i  $q$

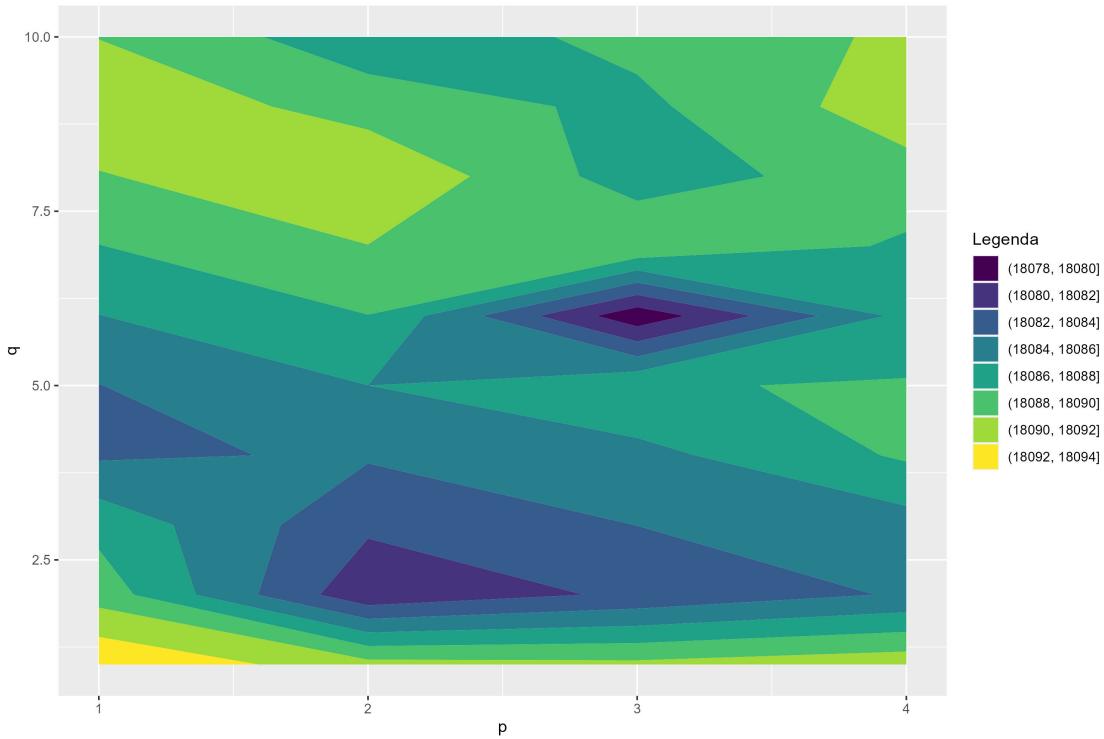


**Wykres 3.2:** Wartości BIC w zależności od  $p$  i  $q$  na przedziale  $(0, 3]$

Zauważono, że wartość BIC jest najmniejsza dla  $p$  i  $q$  z przedziału  $(0, 3]$ . Aby poznać dokładne wyniki, postanowiono wykonać wykres 3.2, który przedstawia Bayesowskie kryterium informacyjne, lecz ze zmniejszonym zakresem poszukiwanych wartości. Na jego podstawie odczytano optymalne wartości  $p = 2$  i  $q = 2$ .



**Wykres 3.3:** Wartości AIC w zależności od  $p$  i  $q$



**Wykres 3.4:** Wartości AIC w zależności od  $p \in (0, 4]$  i  $q \in (0, 10]$

Wykres 3.3 sugeruje zmniejszenie przedziału dla AIC do  $q \in (0, 10]$  i  $p \in (0, 4]$ . Na podstawie wykresu 3.4 można by stwierdzić, że optymalną parą współczynników jest  $p = 3$  i  $q = 7$ . Jednakże wykres 3.1 sugeruje dopasowanie zupełnie innej wartości. Ma to uzasadnienie w tym, że kryterium Akaikiego ma tendencję do tego, żeby wybierać model większego rzędu. Zwrócono zatem uwagę, że oba analizowane kryteria wykazały wartości najmniejsze dla  $p = 2$  i  $q = 2$  i to te współczynniki postanowiono przyjąć w analizowanym modelu ARMA( $p, q$ ).

### 3.2 Estymacja parametrów modelu wybraną metodą

Na podstawie uzyskanych w sekcji 3.1 współczynników modelu ARMA( $p, q$ ) wyestymowano parametry metodą największej wiarygodności. Przyjęto  $p = 2, q = 2$ .

$$X_t - \phi_1 X_{t-1} - \phi_2 X_{t-2} = Z_t + \theta_1 Z_{t-1} + \theta_2 Z_{t-2}.$$

Wykorzystano w tym celu funkcję arima z pakietu "stats" dla parametrów (2,0,2) i wybrano metodę "ML"<sup>1</sup>. Otrzymane w ten sposób wyniki umieszczone w tabeli 3.1.

**Tabela 3.1:** Parametry ARMA(2,2) dla badanych danych

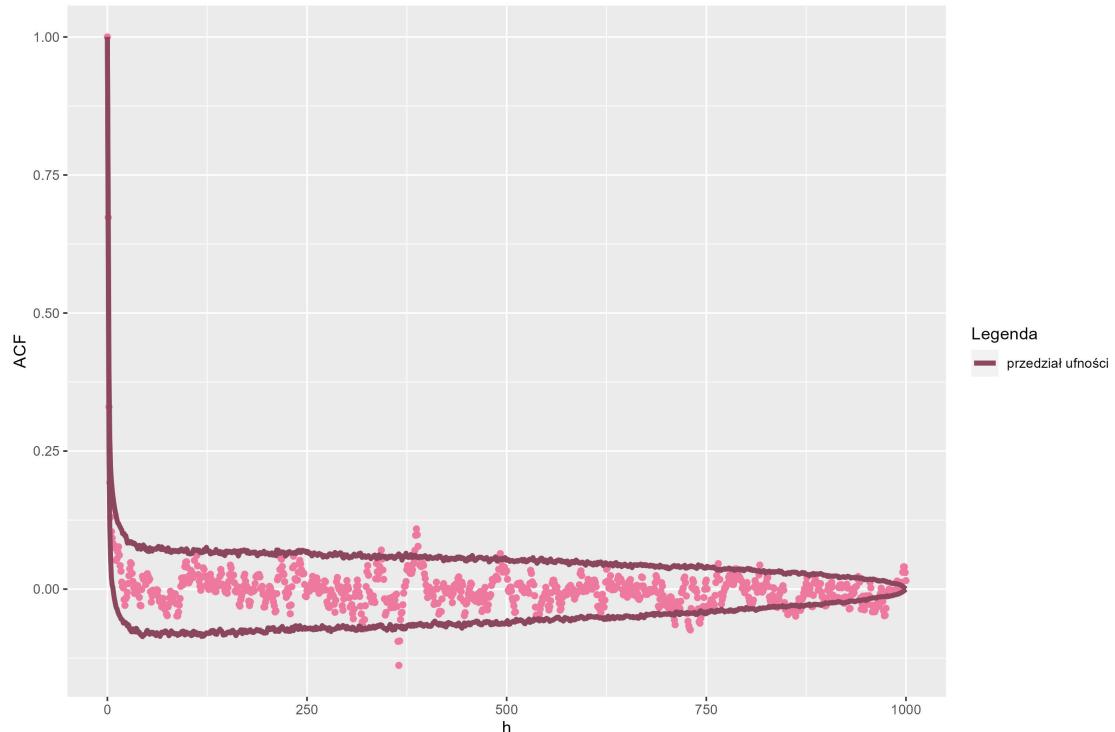
$\phi_1$	$\phi_2$	$\theta_1$	$\theta_2$	$\sigma^2$
1.279	-0.3365	-0.4173	-0.3864	11.88

<sup>1</sup>Maximum Likelihood - metoda największej wiarygodności.

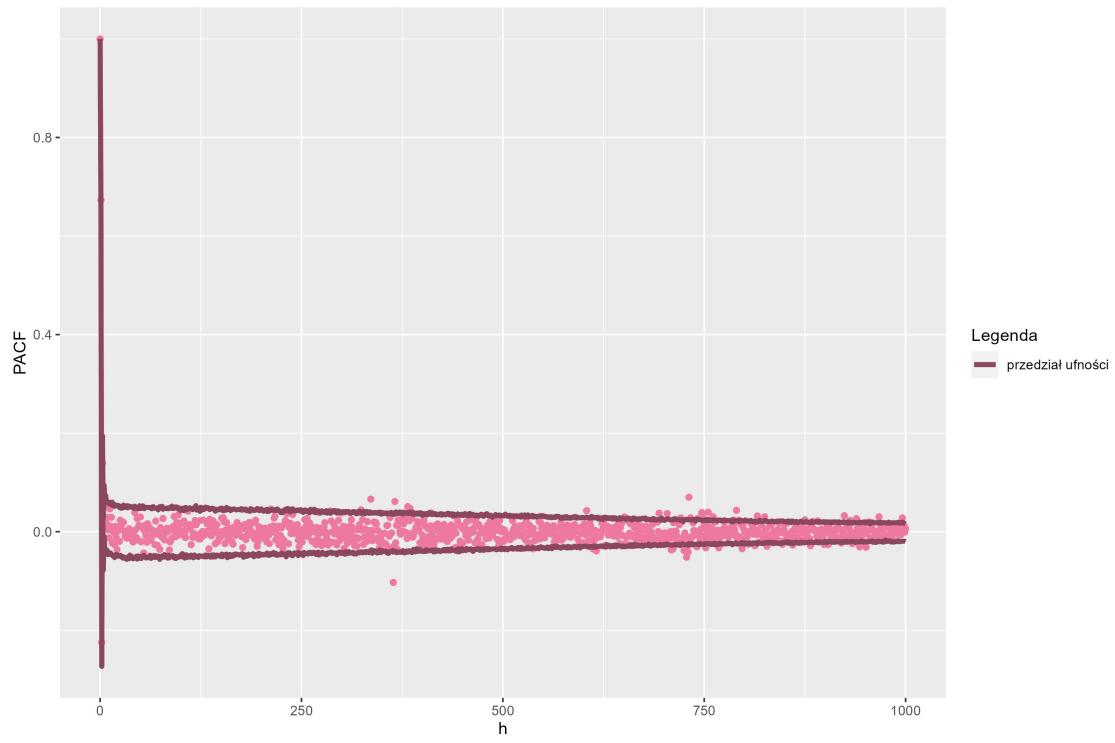
## 4 Ocena dopasowania modelu

### 4.1 Przedziały ufności dla PACF/ACF

W celu oceny dopasowania modelu wysymulowano 1000 przykładowych trajektorii z modelu ARMA(2, 2). Dla każdej z nich wyznaczono ACF i PACF dla lagów (opóźnień) z przedziału [0, 1000]. Następnie znaleziono kwantyle empirycznej autokorelacji i empirycznej autokorelacji cząstkowej na poziomach  $\alpha = 0.05$  i  $1 - \alpha = 0.95$ , które wykorzystano jako przedziały ufności ACF i PACF dla badanych danych.



**Wykres 4.1:** Wartości ACF z przedziałami ufności dla modelu ARMA(2, 2)

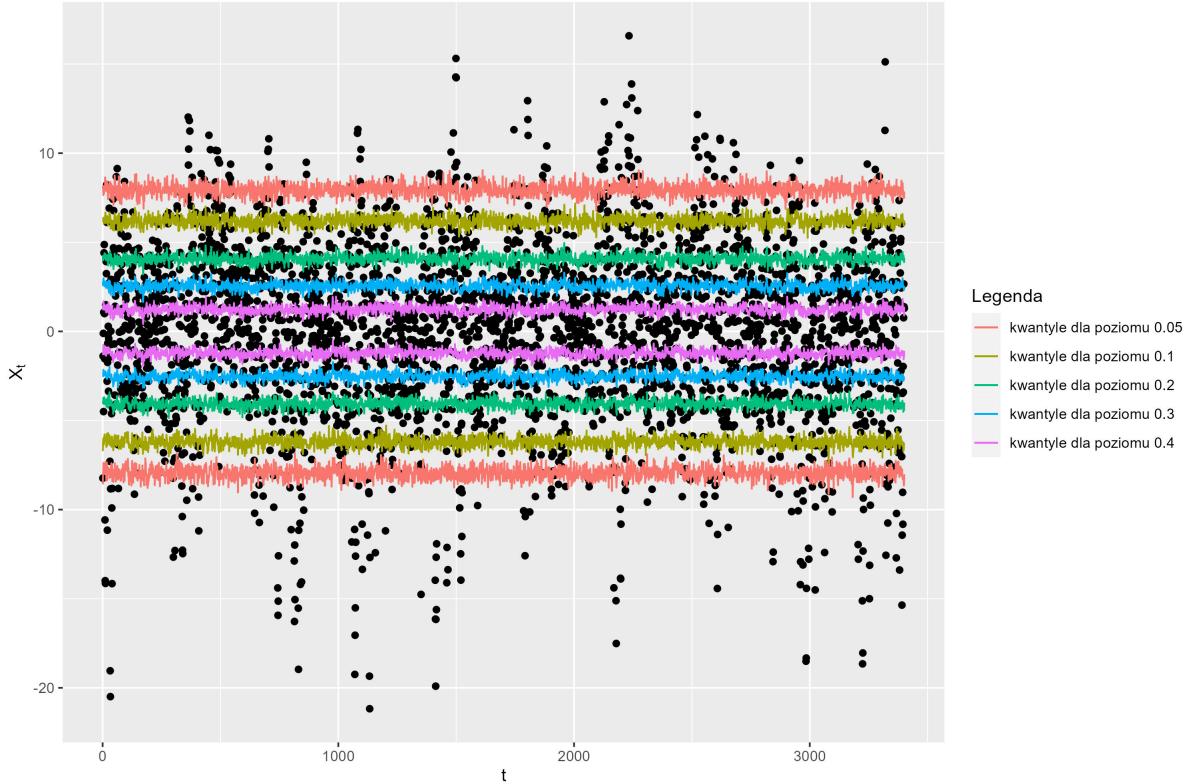


**Wykres 4.2:** Wartości PACF z przedziałami ufności dla modelu ARMA(2, 2)

Na podstawie wykresów 4.1 i 4.2 stwierdzono, że dopasowanie jest słuszne, ponieważ większość wartości zawiera się w odpowiednich przedziałach ufności.

## 4.2 Porównanie linii kwantylowych z trajektorią

W tym kroku wysymulowano 1000 przykładowych trajektorii, a następnie dla każdego  $t$  wyznaczono kwantyle dla wartości ze zbioru  $\alpha \in \{0.05, 0.1, 0.2, 0.3, 0.4\}$  i odpowiadających im  $1 - \alpha$ . Zaznaczono je na wykresie 4.3.



**Wykres 4.3:** Linie kwantylowe dla analizowanego szeregu czasowego

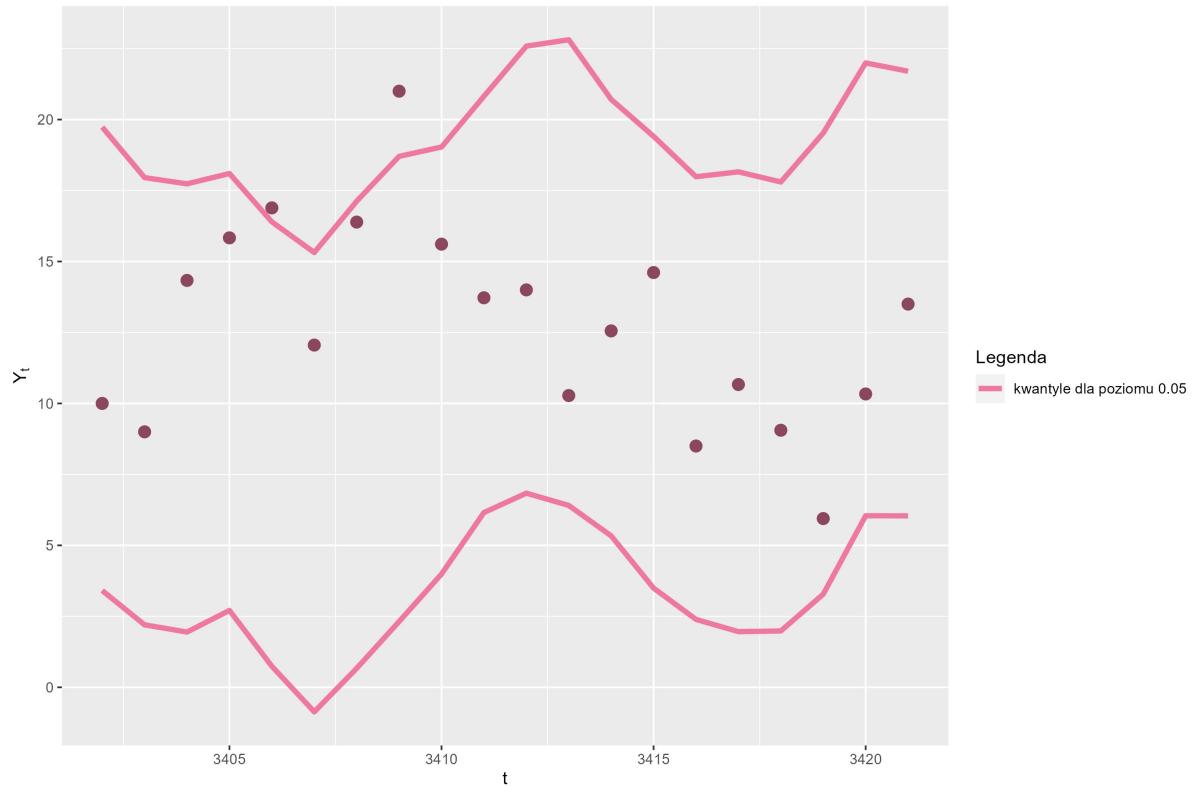
**Tabela 4.1:** Procent obserwacji znajdujących się pomiędzy odpowiednimi liniami kwantylowymi

0.05 i 0.95	0.1 i 0.9	0.2 i 0.8	0.3 i 0.7	0.4 i 0.6
90.532%	82.417%	63.423%	42.488%	21.317%

Wyliczone wartości zaprezentowane w tabeli 4.1 pozwalają stwierdzić, że dopasowanie jest poprawne, ponieważ dany procent wszystkich wartości znajdujących się między liniami kwantylowymi odpowiada wyznaczonemu przedziałowi ufności na badanym poziomie.

### 4.3 Prognoza dla przyszłych obserwacji

W celu prognozy wartości dla przyszłych obserwacji wyznaczono przedziały predykcji. Przyjęto poziom istotności  $\alpha = 0.05$ . Dla każdej wartości  $t$  z przedziału [3401, 3421] wyznaczono przedział ufności na podstawie kwantyli symulacji o tysiącu kroków Monte Carlo. Następnie dodano trend i sezonowość usunięte w wyniku dekompozycji Walda. Ostatecznie porównano wyniki z danymi testowymi na wykresie 4.4 i zauważono, że większość wartości mieści się w wyznaczonym przedziale predykcji.



Wykres 4.4: Dane testowe z przedziałami predykcyjnymi

### 4.4 Wnioski

- Otrzymany szereg stacjonarny poddano do analizy przy pomocy modelu ARMA( $p, q$ ). Wykorzystując wybrane kryteria informacyjne, z powodzeniem dobrano rząd  $p = 2$  oraz  $q = 2$ , a następnie wyestymowano parametry, które przedstawiono w tabeli 3.1. Dopasowanie, którego dokonano, zostało sprawdzone poprzez wyznaczenie przedziałów ufności dla funkcji autokowariancji oraz częściowej autokowariancji. Oba wykresy 4.1, 4.2 przedstawiają wartości, które mieszczą się w wyznaczonym przedziale ufności na poziomach  $\alpha = 0.05$  oraz  $1 - \alpha = 0.95$ , co pozwala stwierdzić poprawność modelu.
- Słuszne dopasowanie modelu potwierdza także wykonany wykres linii kwantylowych 4.3 oraz uzyskane dzięki niemu wartości w tabeli 4.1.
- Na skutek wydzielenia zbioru testowego, który nie był brany do analizy, możliwe było zweryfikowanie, czy będą się one znajdować w wyznaczonym przedziale ufności. Rezultaty uzyskane na wykresie 4.4 pozwalają uznać trafne dopasowanie modelu.

## 5 Weryfikacja założeń dotyczących szumu

Analiza wartości resztkowych polega na sprawdzeniu następujących warunków:

1.  $E\varepsilon_t = 0, \forall t = 1, 2, \dots, n,$
2.  $Var(\varepsilon_t) = \sigma^2, \forall t = 1, 2, \dots, n,$
3.  $\{\varepsilon_t\}_{t=1}^n$  są niezależnymi zmiennymi losowymi.

**Uwaga 1** Powyższe założenia są prawdziwe w klasycznym modelu ARMA. W tym raporcie rozważono jedno założenie (o rozkładzie  $\varepsilon_t$ ) ze względu na wykorzystanie metody największej wiarygodności do estymacji parametrów.

$$\varepsilon_t \sim \mathcal{N}(0, \sigma^2), \forall t = 1, 2, \dots, n.$$

Spełnienie tego warunku nie jest jednak wymagane w klasycznym modelu.

### 5.1 Założenie dotyczące średniej oraz wariancji

W celu zbadania założenia dotyczącego średniej wyznaczono residua badanego modelu, a następnie przedstawiono jego wartości na wykresie 5.1.



Wykres 5.1: Rozkład szumu

Z powyższego wykresu widać, że wartości oscylują wokół 0 przyjmując zarówno wartości ujemne oraz dodanie co może świadczyć o tym, że średnia jest równa 0. Podstawowe statystyki, które zostały wyliczone, znajdują się w tabeli 5.1.

**Tabela 5.1:** Średnia, wariancja i odchylenie standardowe wartości szumu

Średnia	Wariancja	Odchylenie standardowe
0.0023	11.88	3.44

Do sprawdzenia, czy wariancja residiów jest stała, wykorzystano ARCH test na poziomie istotności  $\alpha = 0.05$ . W wyniku przeprowadzenia testu otrzymano p-wartość mniejszą niż  $2.2 \cdot 10^{-16}$ . Na tej podstawie odrzucono hipotezę zerową, mówiącą o stałej wariancji, i przyjęto hipotezę alternatywną, stwierdzającą heteroskedastyczność residiów. Wynik testu oznacza, że założenie o homoskedastyczności szumu nie zostało spełnione. Sensowność dopasowania modelu i predykcje na jego podstawie mogą być wyjaśnione przez dużą liczbę obserwacji wykorzystanych do analizy. Jednak w celu otrzymania modelu lepiej opisującego dane można wybrać taki, który zakłada niestąłą wariancję (np. model GARCH).

## 5.2 Założenie dotyczące niezależności

W celu sprawdzenia niezależności reszt modelu szeregu czasowego wykonano test Ljunga-Boxa. Wykorzystuje on następujące hipotezy:

$$H_0 : \text{reszty są niezależne},$$

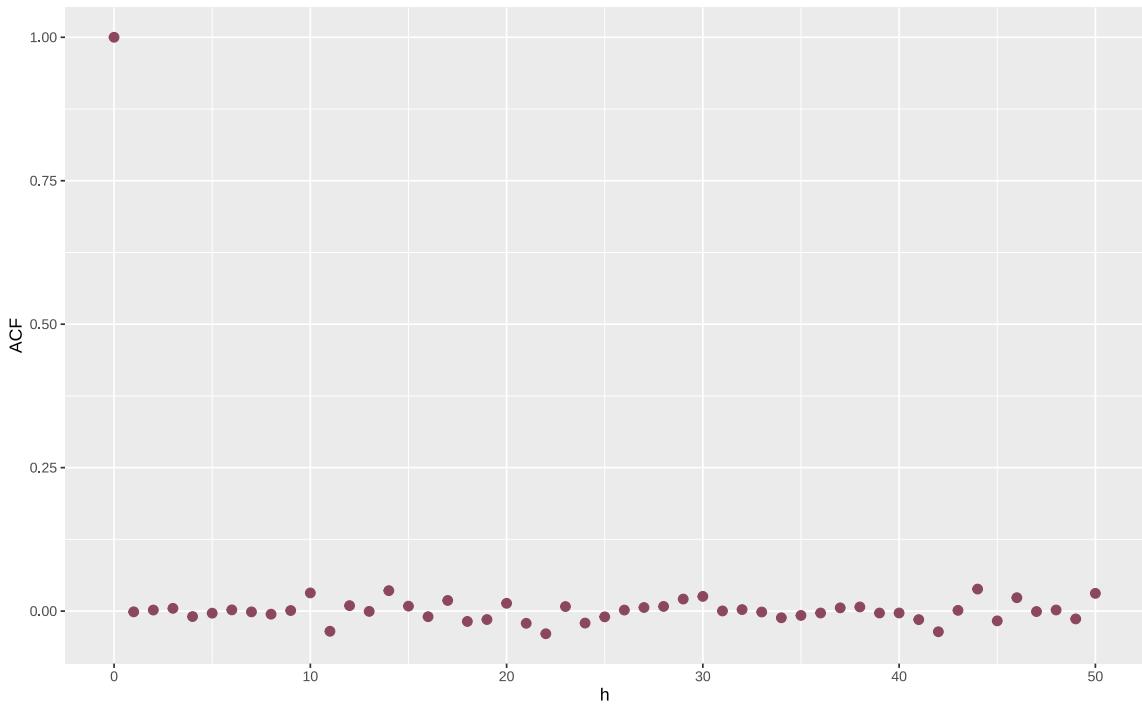
$$H_1 : \text{reszty nie są niezależnie rozłożone}.$$

Wartości, które otrzymano z testu, dla opóźnienia wynoszącego 50 przedstawiono w tabeli 5.2.

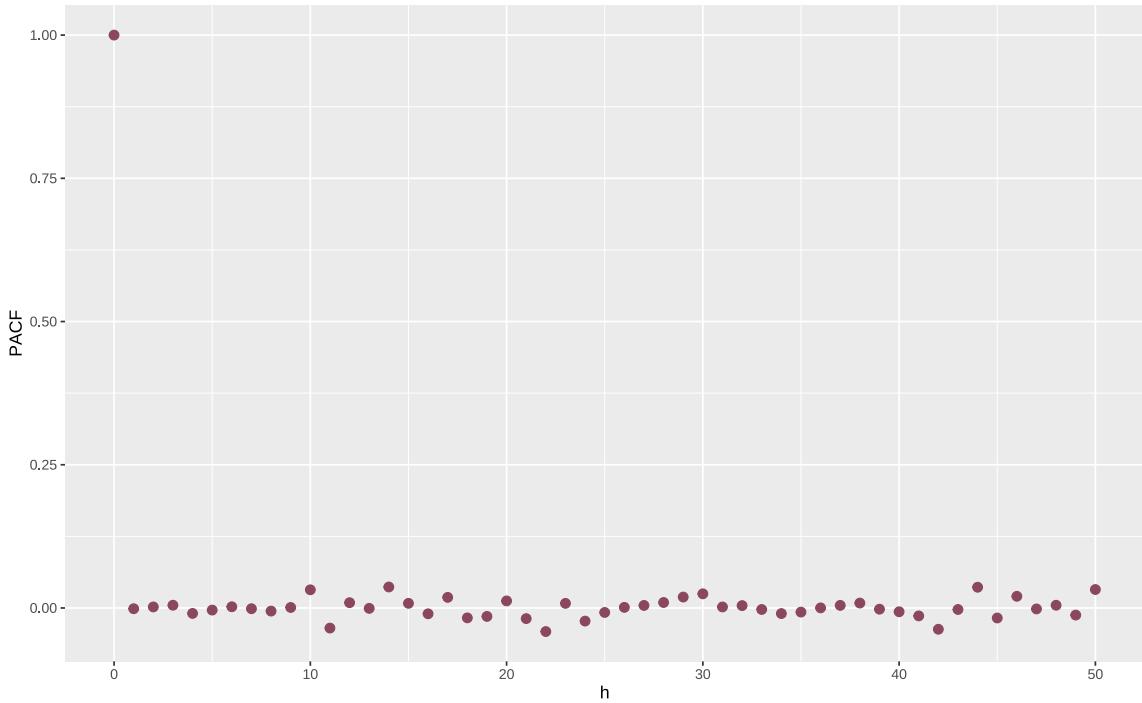
**Tabela 5.2:** Test normalności rozkładu

Test	Statystyka	P-wartość
Ljung-Box	48.431	0.5365

P-wartość jest większa od poziomu istotności  $\alpha = 0.05$ , co oznacza, że nie ma podstaw do odrzucenia hipotezy zerowej świadczącej o tym, że reszty dla badanego modelu szeregu czasowego są niezależne.



**Wykres 5.2:** Funkcja empirycznej autokorelacji dla wartości resztkowych w zależności od  $h$

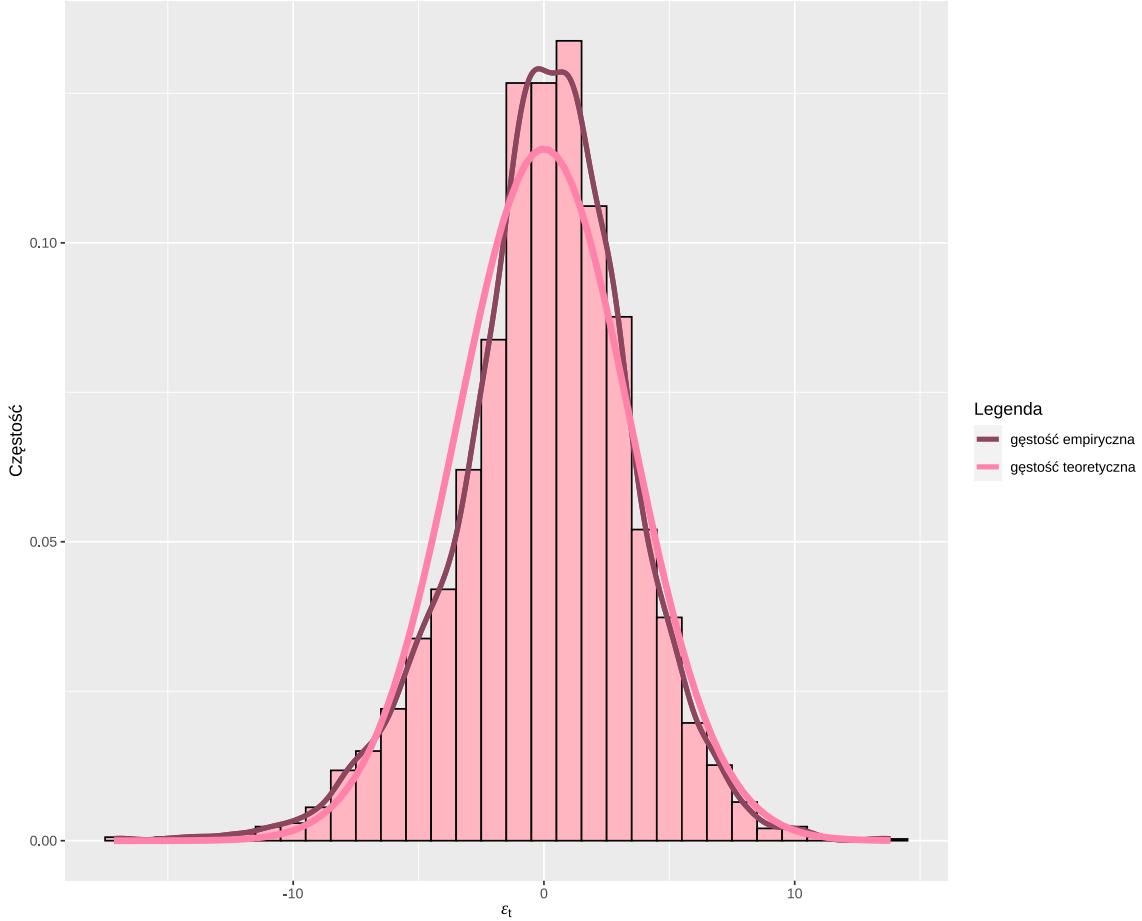


**Wykres 5.3:** Funkcja empirycznej autokorelacji częściowej dla wartości resztkowych w zależności od  $h$

Wniosek o niezależność wartości resztkowych potwierdzają również wykresy 5.2 i 5.3 wykonane dla  $h = 0, 1, 2, \dots, 50$ . Analizując je, widać, że dla  $h = 0$  oczywiście otrzymano 1, natomiast dla pozostałych  $h$  wartości empirycznej autokorelacji oraz częściowej autokorelacji są bliskie 0.

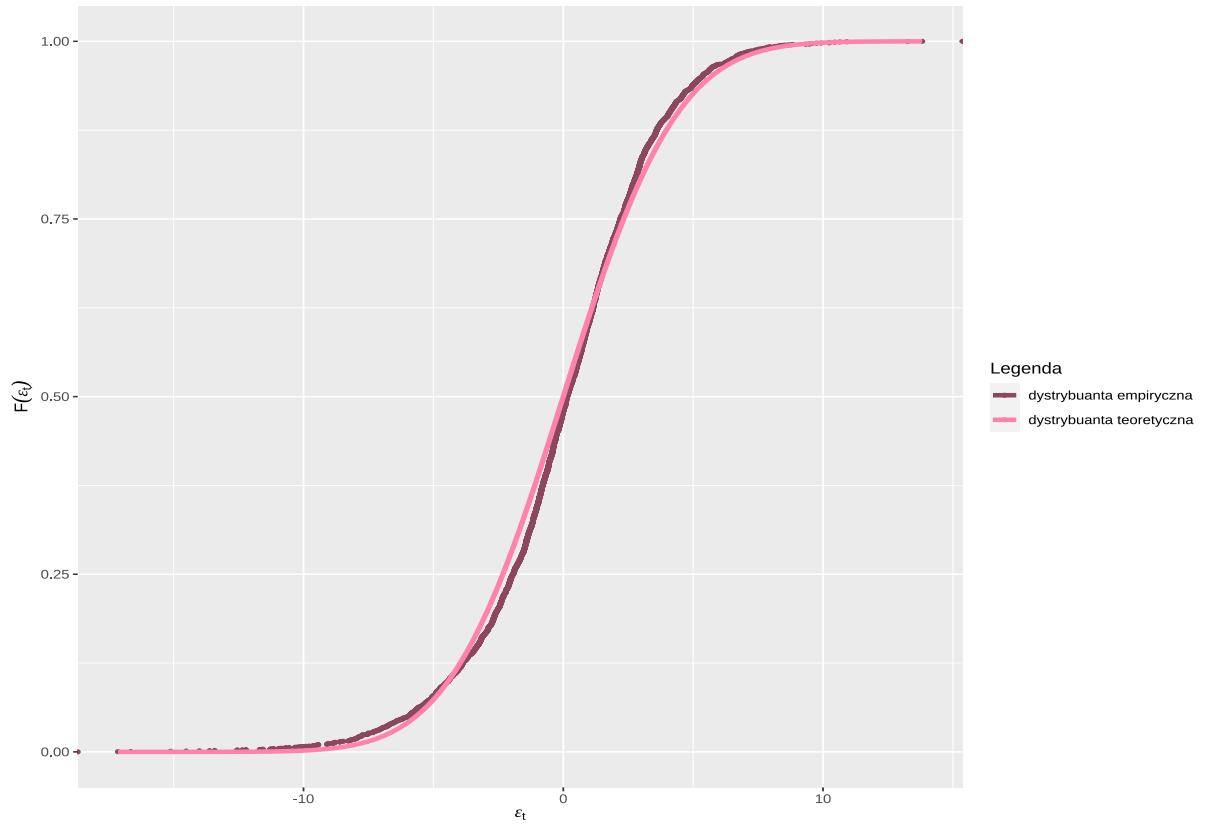
### 5.3 Założenie dotyczące normalności rozkładu

Ostatnim etapem było sprawdzenie, czy otrzymany szum faktycznie ma rozkład normalny. Z założeń modelu wiadomo, że  $\mu = 0$ . Dodatkowo w tabeli 5.1 otrzymano, że  $\sigma \approx \sqrt{S^2} \approx 3.44$ . Na wykresie 5.4 porównano gęstość empiryczną i teoretyczną.

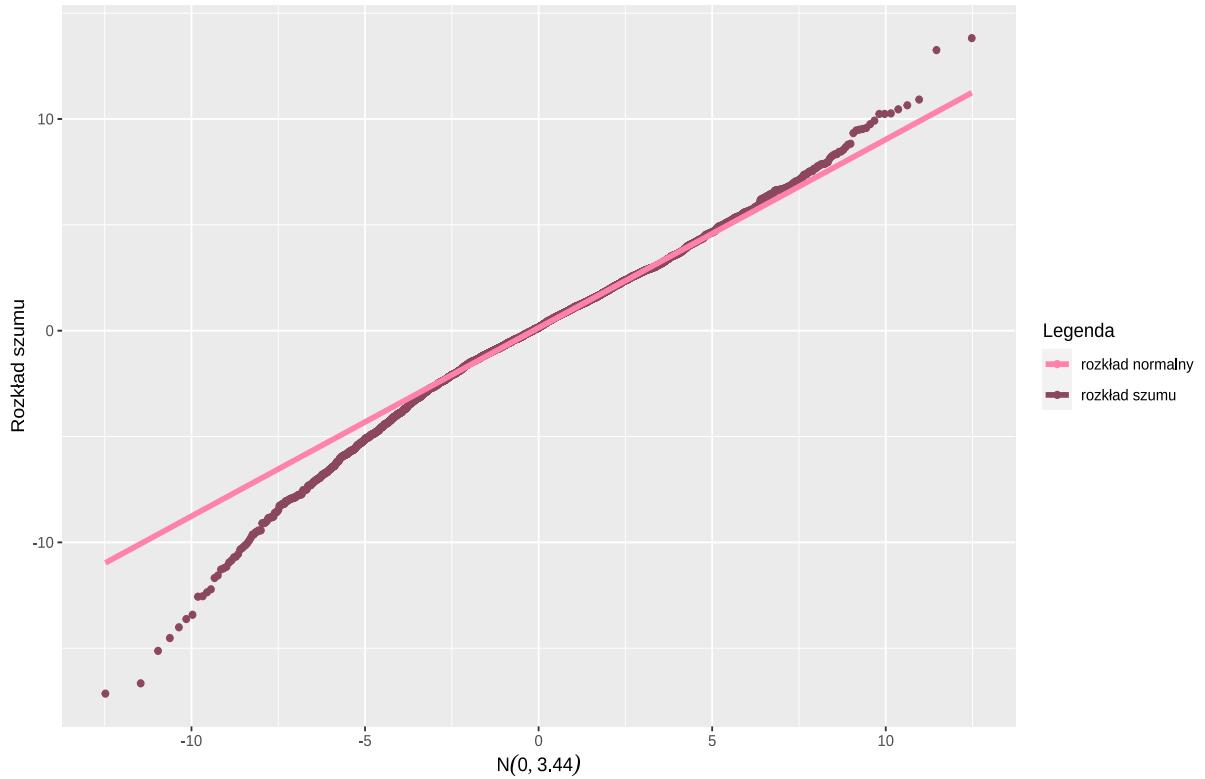


**Wykres 5.4:** Porównanie gęstości empirycznej  $\varepsilon_t$  i gęstości teoretycznej rozkładu  $\mathcal{N}(0, 3.44)$

Gęstości na powyższym wykresie w pewnych miejscach się pokrywają, co może sugerować dobre dopasowanie rozkładu. Sprawdzono również, jak prezentują się dystrybuanty oraz wartości resztowe na wykresie kwantylowym.



**Wykres 5.5:** Porównanie dystrybuanty empirycznej  $\varepsilon_t$  i dystrybuanty teoretycznej rozkładu  $\mathcal{N}(0, 3.44)$



**Wykres 5.6:** Porównanie kwantylów  $\varepsilon_t$  i teoretycznych kwantylów rozkładu  $\mathcal{N}(0, 3.44)$

Podobnie jak w przypadku gestości, porównując teoretyczną oraz empiryczną dystrybuantę zauważono, że badane dane mają rozkład bliski teoretycznemu. Z kolei wykres kwantylowy 5.6 przedstawia znaczne odchylenia punktów leżących na końcach krzywej, co może sugerować, że wartości nie mają rozkładu normalnego. Zdecydowano się zatem wykonać test normalności rozkładu.

Aby sprawdzić, czy  $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$ ,  $\forall t = 1, 2, \dots, n$  wykonano test Jarque-Bera, który bada normalność rozkładu. Przyjęto następujące hipotezy.

$$H_0 : \text{rozkład szumu jest rozkładem normalnym,}$$

$$H_1 : \text{rozkład szumu jest różny od rozkładu normalnego.}$$

Postawienie takiej hipotezy zerowej dokonano opierając się na graficznej reprezentacji danych na wykresach 5.4 oraz 5.5. Następnie wyznaczoną na podstawie statystyki testowej p-wartość porównano z poziomem istotności  $\alpha = 0.05$ . Wykonany test Jarque-Bera jednoznacznie wykazał p-wartość niemalże bliską 0, co potwierdziło, że badany szum nie pochodzi z rozkładu normalnego.

## 5.4 Wnioski

Na podstawie analizy wartości resztkowych sformułowano następujące wnioski.

- Średnia wartość badanych residuów jest bliska 0, zatem pierwsze z założeń jest zachowane.
- ARCH test jednoznacznie wskazuje na brak jednorodności wariancji residuów, co oznacza, że drugie z założeń nie jest spełnione. Jednak poprzez wykorzystanie do analizy dużej liczby obserwacji dopasowany model ARMA(2,2) jest wciąż skuteczny zarówno w opisie danych, jak i predykcji ich kolejnych wartości<sup>2</sup>.
- Wykonany test Ljunga-Boxa nie odrzucił badanej hipotezy zerowej o niezależności wartości resztkowych. Ponadto wykresy 5.2 i 5.3 potwierdzają, że analizowane zmienne są niezależne, ponieważ wartości funkcji empirycznej autokorelacji oraz cząstkowej autokorelacji dla  $h > 0$  są bliskie 0.
- Na podstawie wykresów 5.4 oraz 5.5 zauważono, że rozkład wartości resztkowych może być bliski rozkładowi normalnemu z parametrami  $\mu = 0, \sigma = 3.44$ . Te przyjęte zweryfikował wykres kwantylowy 5.6 oraz test Jarque-Bera, który wykazał, że badany szum nie pochodzi z rozkładu normalnego. Rozbieżność w tych wynikach może jedynie sugerować zbliżenie wartości analizowanych residuów do sprawdzanego rozkładu.

---

<sup>2</sup>Podczas dalszej analizy warto rozważyć dopasowanie modelu zakładającego niejednorodność wariancji, np. GARCH.

## **6 Zakończenie**

W niniejszej pracy wykorzystano poznane metody dotyczące analizy danych z wykorzystaniem modelu ARMA. Dla obserwacji dotyczących średnich temperatur w Denver na przestrzeni lat uzyskano dopasowanie w postaci ARMA(2, 2). Dokonano również oceny jego jakości, dzięki której stwierdzono, że jest on właściwie dobrany. Weryfikując założenia dotyczące szumu, nie potwierdzono wszystkich jego wytycznych. Może to sugerować, że przedstawiony w niniejszej pracy problem nie wyczerpuje w całości swoich możliwości. Pracę można byłoby rozszerzyć poprzez zbadanie innych modeli, które prawdopodobnie wykazywałby lepsze rezultaty.