

Analiza danych rzeczywistych przy pomocy modelu ARMA

Oliwia Makuch
Arkadiusz Kijaczek

Komputerowa Analiza
Szeregów Czasowych

2024/2025

ROZDZIAŁ 1

WSTĘP

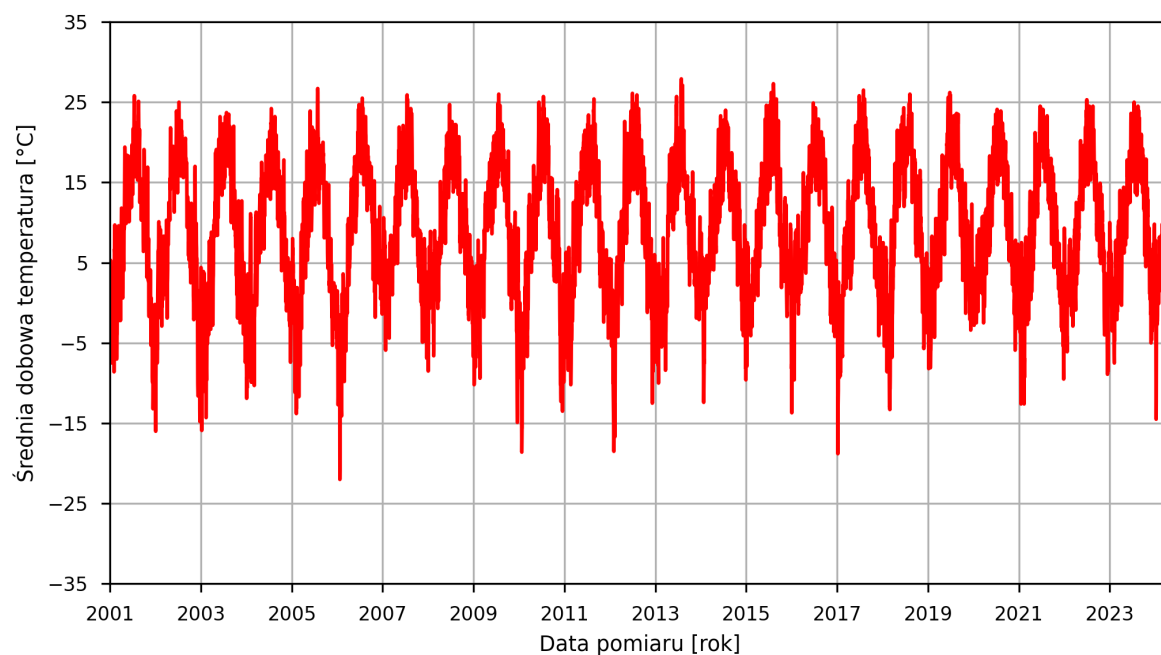
Celem pracy jest analiza danych rzeczywistych przy pomocy modelu ARMA.

Analiza przeprowadzona zostanie na zbiorze danych pogodowych pozyskanych z Instytutu Meteorologii i Gospodarki Wodnej w latach 2001–2023. Dla każdego dnia z tego okresu pozyskana została średnia dobową temperatura dla stacji Pszczyna. Dane użyte do analizy można uzyskać pod adresem: dane.imgw.pl. Ich wizualizacja znajduje się na [Wykresie 1.1](#).

Dla uzyskanych danych, zostanie przeprowadzona analiza jakościowa, a na podstawie oczyszczonych danych wyodrębniony zostanie zbiór testowy. Dane zostaną poddane dekompozycji na składowe trendu, sezonowości i szumu. Dla uzyskanego szeregu czasowego zostanie zbadana stacjonarność, autokorelacja oraz autokorelacja cząstkowa.

Na podstawie uzyskanych wyników przy pomocy kryteriów informacyjnych zostanie dobrany model ARMA, dla którego zostaną oszacowane wartości wielomianów auto–regresji i średniej ruchomej.

W kolejnym etapie zostanie przeprowadzona ocena dopasowania modelu do danych oraz weryfikacja założeń dotyczących szumu.



Wykres 1.1: Średnia dzienna temperatura powietrza w latach 2001–2023 zmierzona w stacji IMiGW w Pszczynie.

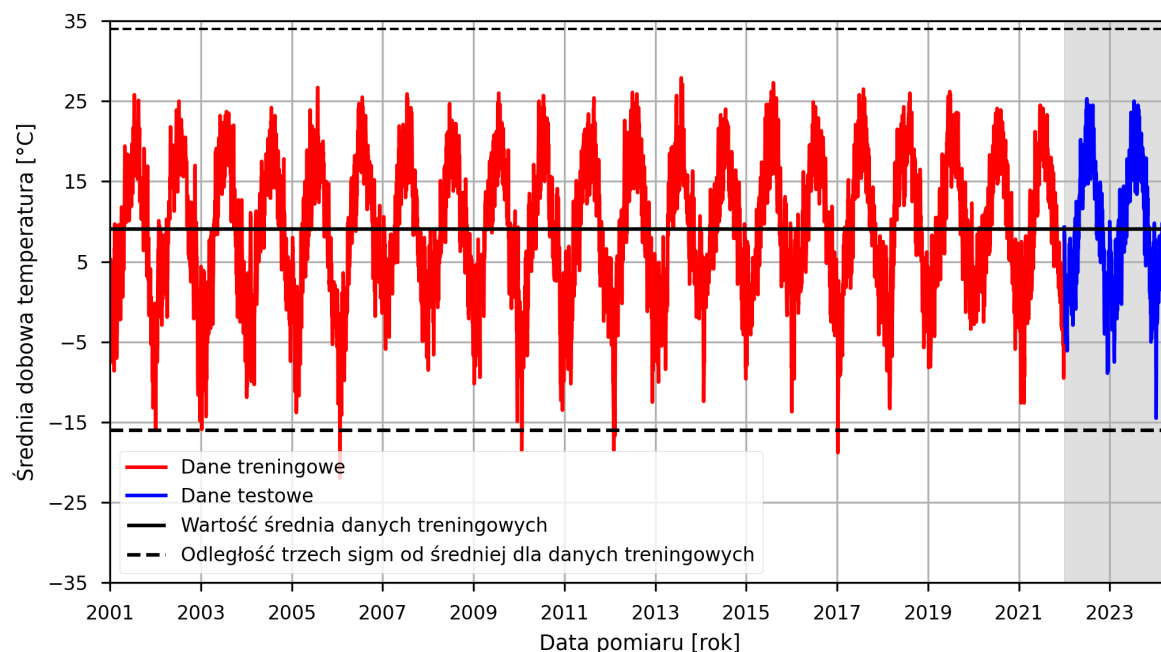
ROZDZIAŁ 2

PRZYGOTOWANIE DANYCH DO ANALIZY

Dane pozyskane z IMiGW dla stacji Pszczyna w latach 2001–2023 są danymi jakościowymi, ponieważ nie zawierają brakujących wartości, a liczba dni pomiarowych odpowiada różnicy między pierwszym a ostatnim dniem w zbiorze danych. Dodatkowo każda z wartości znajduje się w przedziale od -22 do 28 stopni Celsjusza, co odpowiada spodziewanym wartościom temperatury w okolicy miejscowości Pszczyna. Średnia wartość temperatury zbioru danych wynosi około 9 stopni Celsjusza, a jej odchylenie standardowe wynosi około 8,4 stopni Celsjusza. Temperatura poniżej -16 stopni Celsjusza występuje w zbiorze danych 10 razy, a temperatura powyżej 36 stopni Celsjusza występuje 0 razy. Stąd do analizy zostanie wykorzystany cały zbiór danych liczący 8400 obserwacji.

Na potrzeby analizy każdej dacie pomiaru przypisana zostanie wartość liczbowa odpowiadająca liczbie dni, które upłynęły od 1 stycznia 2001 roku – najwcześniejszej daty w zbiorze. W ten sposób uzyskany zostanie szereg czasowy, który posłuży jako podstawa dalszej analizy.

Jako zbiór testowy zostanie wyodrębniony okres od 1 stycznia 2020 roku do 31 grudnia 2023 roku, co odpowiada 1461 obserwacjom.



Wykres 2.1: Średnia dzienna temperatura powietrza w latach 2001–2023 zmierzona w stacji IMiGW w Pszczynie z wyodrębnionym zbiorem testowym i oznaczoną wartością średnią oraz zakresem trzech odchyleń standardowych.

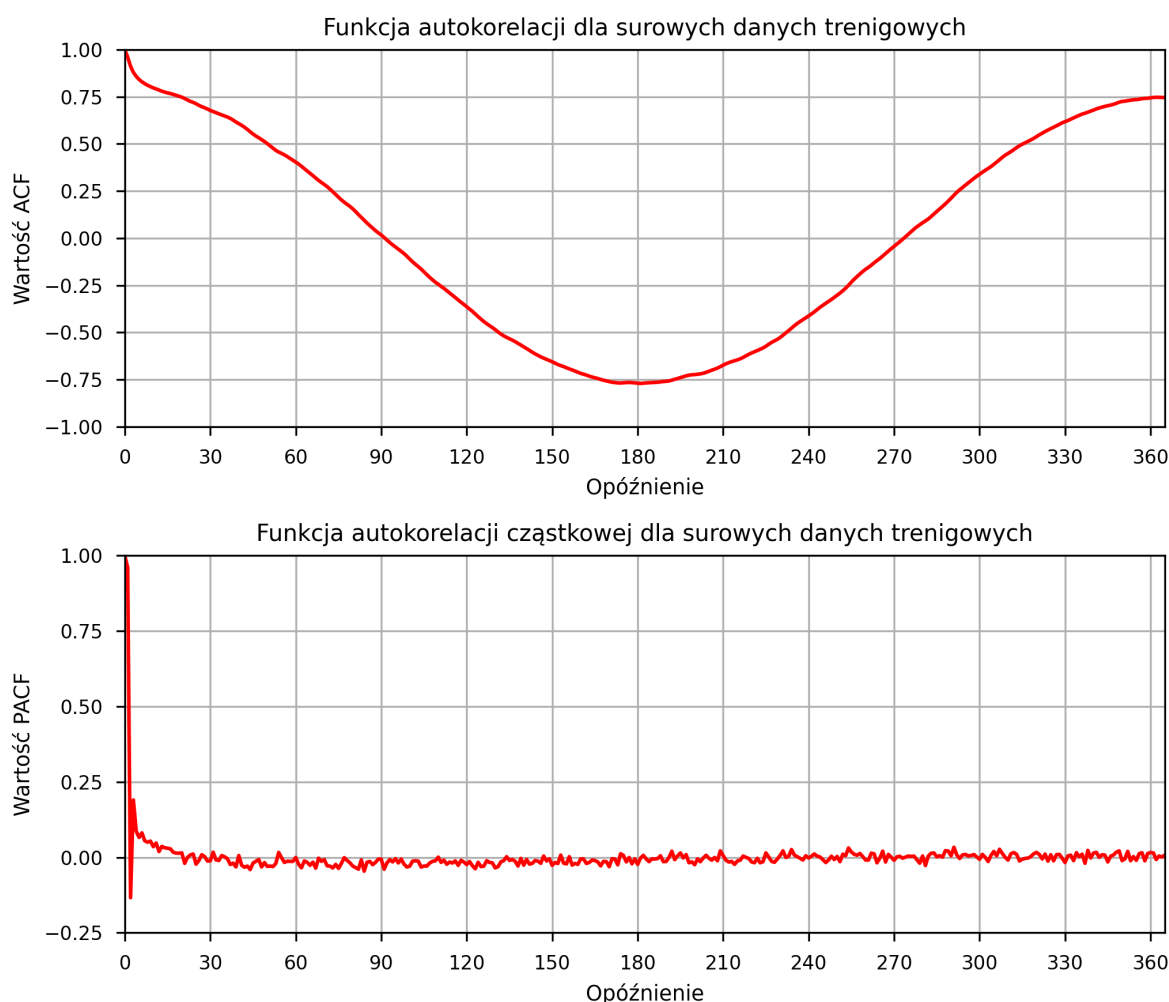
Dane z wyodrębnionym zbiorem testowym zostały przedstawione na [Wykresie 2.1](#).

Na [Wykresie 2.2](#) przedstawiono funkcję autokorelacji (ACF) oraz funkcję autokorelacji cząstkowej (PACF) dla surowych danych treningowych.

Górny wykres ukazuje wartości autokorelacji dla różnych opóźnień, co pozwala na identyfikację zależności sezonowych trendów w danych. Wartości ACF wykazują silną dodatnią korelację dla małych opóźnień, która stopniowo maleje, osiągając minimum około opóźnienia 180 dni, a następnie ponownie rośnie. Taki kształt wskazuje na roczny cykl sezonowy, co jest zgodne z oczekiwaniami dla szeregu czasowego temperatury powietrza.

Dolny wykres przedstawia funkcję autokorelacji cząstkowej (PACF), która pokazuje zależności między danymi po wyeliminowaniu wpływu wcześniejszych opóźnień. Widać, że wartości PACF szybko spadają po pierwszych kilku opóźnieniach, co sugeruje, że szereg czasowy można modelować przy użyciu procesu autoregresyjnego o niewielkiej liczbie opóźnień.

Analiza obu wykresów wskazuje na silne zależności sezonowe oraz sugeruje konieczność uwzględnienia sezonowej składowej w dalszym modelowaniu danych.



Wykres 2.2: Wartości funkcji autokorelacji oraz autokorelacji cząstkowej w zależności od opóźnienia dla surowych danych treningowych.

Wartości autokowariancji zostały obliczone na podstawie wzoru na autokowariancję i autoko-

relację empiryczną dla szeregu czasowego. Autokowariancja empiryczna z próby dla opóźnienia h została obliczona ze wzoru:

$$\gamma_h = \frac{1}{n} \sum_{i=1}^{n-h} (x_i - \bar{x})(x_{i+h} - \bar{x}),$$

gdzie:

- x_i - wartość obserwacji w czasie i ,
- n - liczba punktów danych w szeregu czasowym,
- h - opóźnienie.

Funkcja autokorelacji została następnie wyznaczona jako normalizacja autokowariancji względem wartości dla opóźnienia zerowego:

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

Funkcja cząstkowej autokorelacji została obliczona na podstawie danych treningowych z użyciem metody Yule-Walkera. PACF wyraża związek pomiędzy wartością x_t i x_{t+h} , po uwzględnieniu wpływu pośrednich zmiennych (obserwacji z pośrednich okresów). Funkcja cząstkowej autokorelacji dla opóźnienia h została obliczona jako:

$$\text{PACF}(h) = \hat{\alpha}_h,$$

gdzie $\hat{\alpha}_h$ to współczynnik autoregresyjny dla opóźnienia h wyznaczony metodą Yule-Walkera.

Funkcję PACF dla różnych opóźnień obliczono za pomocą funkcji `pacf` z pakietu `statsmodels`, która stosuje równania Yule-Walkera do oszacowania współczynników autoregresyjnych.

Przeprowadzono także test ADF (Augmented Dickey-Fuller), który sprawdza, czy szereg czasowy jest stacjonarny, tzn. czy nie zawiera pierwiastka jednostkowego, który sugerowałby obecność trendu lub niestacjonarności. Wyniki testu dla surowych danych zostały przedstawione w Tabeli 2.1.

Wielkość	Wartość
Statystyka ADF	-5.7414
Wartość krytyczna (5%)	-2.8619
p-wartość	$6.26 \cdot 10^{-7}$

Tabela 2.1: Wyniki testu ADF dla surowych danych

Na podstawie uzyskanej p-wartości (mniejszej niż 0.05) test odrzucił hipotezę zerową, która zakładała obecność pierwiastka jednostkowego. Wniosek z testu jest taki, że dane są stacjonarne.

Należy jednak zauważyć, że w przypadku danych z niewielkim trendem test ADF może błędnie wskazać stacjonarność, mimo że dane zawierają sezonowość oraz pewien słaby trend. Takie dane mogą zostać błędnie uznane za stacjonarne, ponieważ test ADF nie zawsze wychwytuje nieznaczące trendy.

DEKOMPOZYCJA SZEREGU CZASOWEGO

Trend szeregu czasowego zidentyfikowano przy użyciu metody regresji liniowej z modelu:

$$y(t) = a \cdot t + b,$$

gdzie współczynniki a i b wyznaczono za pomocą metody najmniejszych kwadratów.

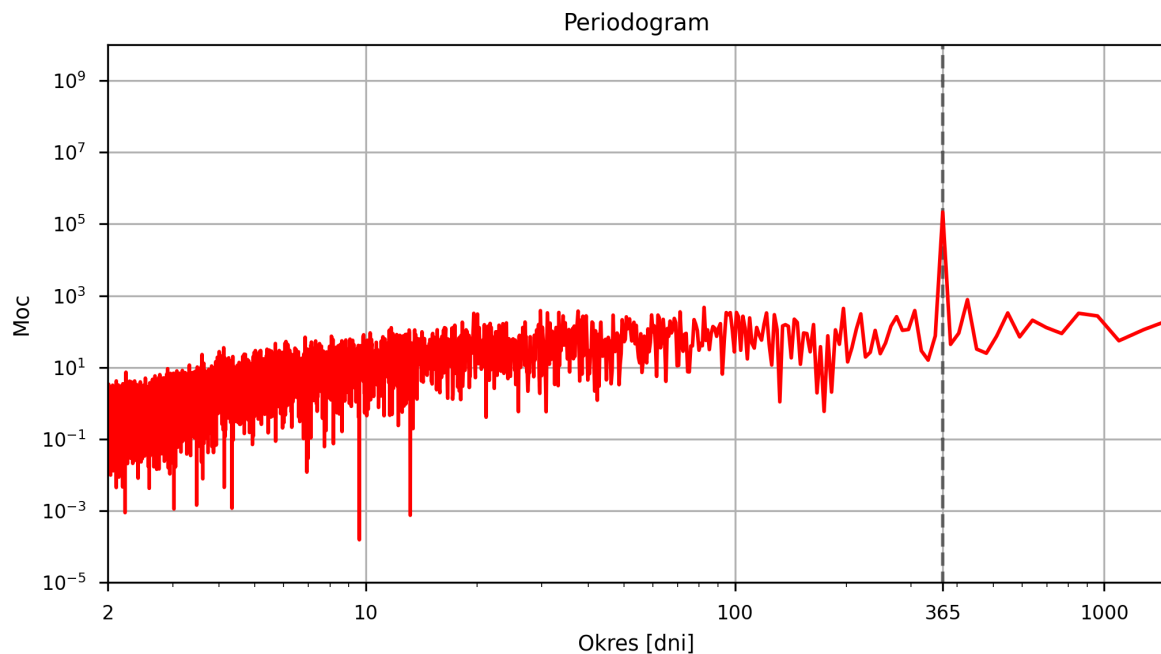
$$a = \frac{\sum_{t=1}^n (t - \bar{t})(X_t - \bar{X})}{\sum_{t=1}^n (t - \bar{t})^2},$$
$$b = \bar{X} - a \cdot \bar{t},$$

gdzie:

- X_t – obserwacja zmiennej losowej w chwili t ,
- n – liczba obserwacji w szeregu czasowym.

Dopasowana funkcja trendu regresji liniowej ma postać $y(t) = 0.000199 \cdot t + 8.244$.

Wyodrębniony trend usunięto z danych, pozostawiając wartości opisane jedynie przez komponenty sezonowe i szum.



Wykres 2.3: Periodogram dla szeregu czasowego średniej dziennej temperatury powietrza w latach 2001–2023 zmierzonej w stacji IMiGW w Pszczynie.

W celu identyfikacji sezonowości w szeregach czasowych, sporządzono periodogram, który umożliwia określenie dominujących częstotliwości w danych poprzez zastosowanie szybkiej

transformacji Fouriera. Periodogram przedstawia wartości funkcji

$$f(x) = \frac{|\text{FFT}(X_t)|^2}{n},$$

gdzie X_t to wartości szeregu czasowego po usunięciu trendu, a n to liczba obserwacji. Periodogram został przedstawiony na [Wykresie 2.3](#).

Na podstawie periodogramu wyznaczono okres T , równy w zaokrągleniu do liczby całkowitej 365 dni, dla którego wartość funkcji $f(x)$ była największa, co wskazuje na obecność rocznej sezonowości w danych. Otrzymane wyniki są zgodne z oczekiwaniami w przypadku szeregów czasowych temperatury powietrza.

Następnie w celu odwzorowania sezonowości dopasowano sinusoidalny model regresyjny o funkcji postaci:

$$g(t) = A \sin\left(\frac{2\pi}{T}(t + B)\right), \text{ gdzie}$$

- A – amplituda sinusoidy, określająca zakres zmienności funkcji,
- T – okres, reprezentujący długość jednego cyklu sinusoidalnego,
- B – przesunięcie fazowe.

Dopasowanie parametrów A , B oraz T przeprowadzono z użyciem metody najmniejszych kwadratów, wykorzystując funkcję `curve_fit` z biblioteki `scipy`. Jako punkt początkowy estymacji p_0 przyjęto $p_0 = [-10, 0, T]$, gdzie T to wartość okresu wyznaczonego wcześniej na podstawie periodogramu. Z dopasowanej funkcji sinusoidalnej wyznaczono wartość sezonowych składowych. Aby wyeliminować wpływ sezonowości, składową sinusoidalną odjęto od szeregu czasowego bez trendu.

Dopasowana funkcja sinusoidalna opisująca sezonowość szeregu czasowego z podstawionymi przybliżonymi parametrami ma postać:

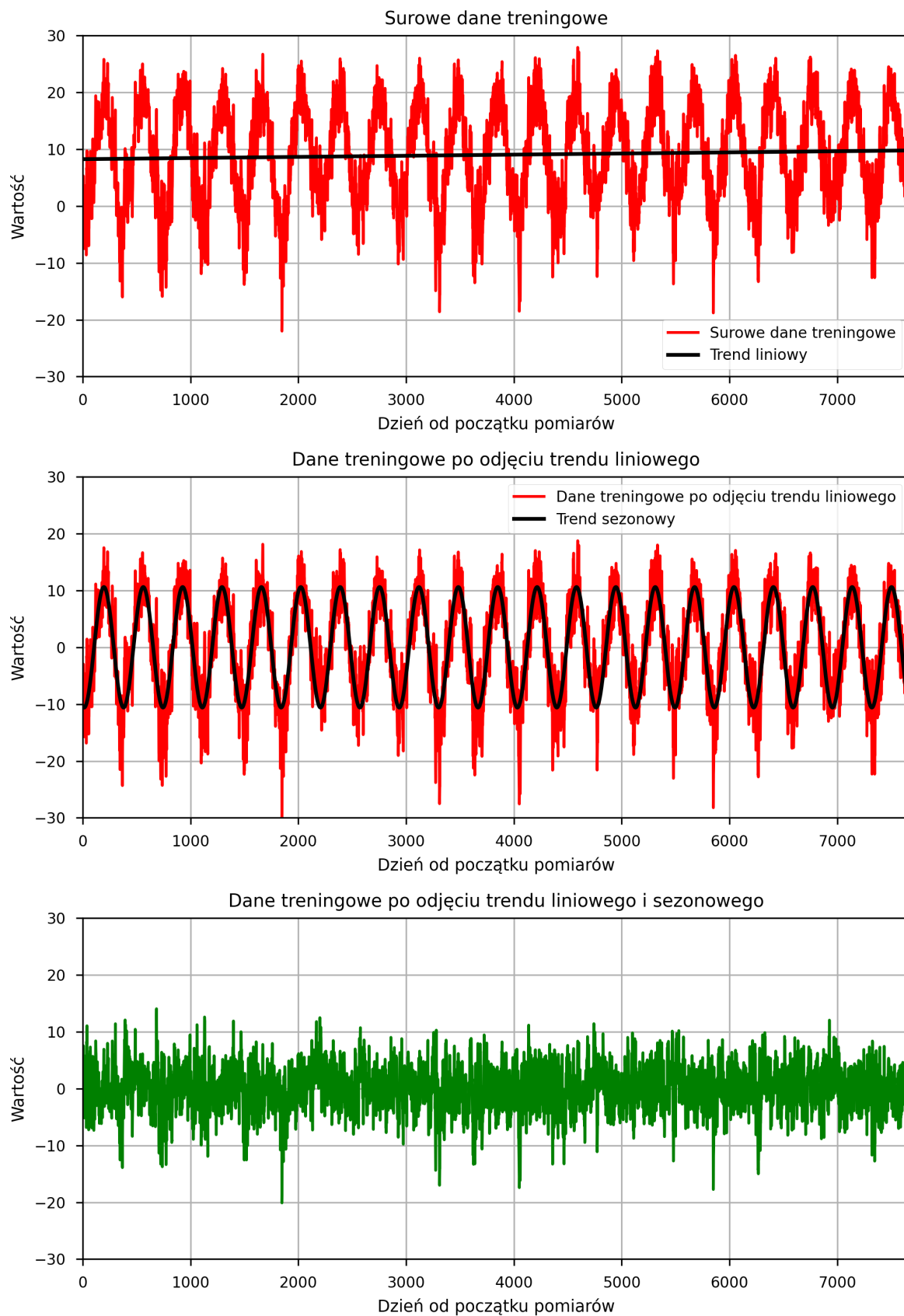
$$g(t) = -10.603 \cdot \sin\left(\frac{2\pi}{365.373} \cdot (t - 652.810)\right)$$

Proces usuwania zależności liniowej oraz okresowej z danych treningowych został przedstawiony na [Wykresie 2.4](#).

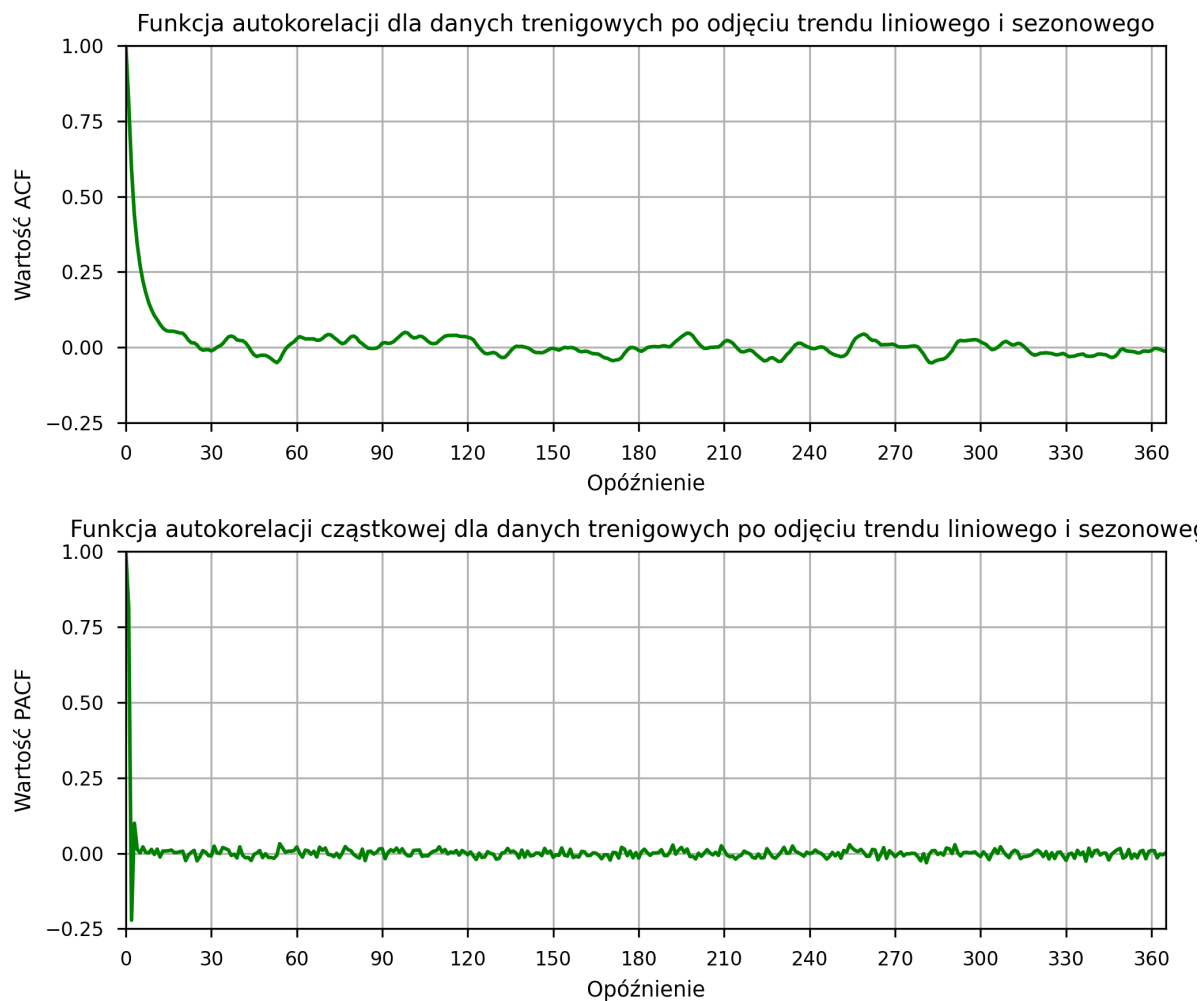
Wartości przedstawione na dolnym panelu wyglądają na względnie losowe i niezależne od czasu, co może sugerować brak wyraźnych struktur w danych resztowych. Dokładne potwierdzenie stacjonarności wymaga jednak dalszych testów statystycznych.

[Wykres 2.5](#) przedstawia wartości funkcji autokorelacji (ACF) oraz autokorelacji cząstkowej (PACF) dla szeregu czasowego uzyskanego po odjęciu trendu i sezonowości z danych treningowych.

Górny panel wykresu prezentuje funkcję autokorelacji, gdzie widoczne są wyraźne spadki w miarę wzrostu opóźnienia, co sugeruje ograniczenie długoterminowych zależności w danych po odjęciu dominujących składowych. Na dolnym panelu zaprezentowano funkcję autokorelacji cząstkowej, która wskazuje na znaczne osłabienie zależności po kilku pierwszych opóźnieniach. Taka struktura sugeruje, że pozostały szereg czasowy może być dobrze opisany za pomocą procesu ARMA o niskich rzędach autoregresji i średniej ruchomej.



Wykres 2.4: Dekompozycja szeregu czasowego średniej dziennej temperatury powietrza w latach 2001–2023 zmierzonej w stacji IMiGW w Pszczynie.



Wykres 2.5: Wartości funkcji autokorelacji oraz autokorelacji cząstkowej w zależności od opóźnienia dla szeregu uzyskanego po odjęciu wyznaczonego trendu oraz sezonowości z danych treningowych.

Przeprowadzono także test ADF (Augmented Dickey-Fuller) dla szeregu czasowego uzyskanego po odjęciu trendu i sezonowości. Test sprawdza, czy szereg czasowy jest stacjonarny, tzn. czy nie zawiera pierwiastka jednostkowego, który sugerowałby obecność trendu lub niestacjonarności. Wyniki testu zostały przedstawione w Tabeli 2.2. Na podstawie uzyskanej p-wartości (mniejszej od 0,05) test odrzucił hipotezę zerową, która zakładała obecność pierwiastka jednostkowego. Wniosek z testu jest taki, że otrzymany szereg jest stacjonarny.

Wielkość	Wartość
Statystyka ADF	−28.1782
Wartość krytyczna (5%)	−2.8619
p-wartość	0.0

Tabela 2.2: Wyniki testu ADF dla szeregu po odjęciu trendu i sezonowości

ROZDZIAŁ 3

MODELOWANIE DANYCH PRZY POMOCY ARMA

W celu dobrania rzędu modelu $\text{ARMA}(p, q)$, wykorzystano kryteria informacyjne: AIC (Akaike Information Criterion), BIC (Bayesian Information Criterion) oraz HQIC (Hannan-Quinn Information Criterion). Rozważono maksymalne wartości parametrów $p = 10$ oraz $q = 3$, a modele dopasowano trzema różnymi metodami: `statespace`, `innovations_mle` oraz `hannan_rissanen`. Poniżej przedstawiono główne kroki przeprowadzonej analizy:

- Iterowano po wartościach p i q , dopasowując model ARMA do szeregu czasowego otrzymanego po usunięciu sezonowości i trendu z surowych danych.
- Dla każdego modelu zapisano wartości kryteriów AIC, BIC oraz HQIC, jak również metodę estymacji zastosowaną do dopasowania.
- Uzyskane wyniki uporządkowano w tabeli i wybrano model o najmniejszych wartościach poszczególnych kryteriów.

Wyniki analizy wskazują, że optymalnym modelem według wszystkich kryteriów informacyjnych jest model $\text{ARMA}(3, 0)$ dopasowany przy użyciu metody `innovations_mle`. Szczegółowe wyniki przedstawiono w Tabeli 3.1.

p	q	AIC	BIC	HQIC	Metoda
3	0	33785.88	33820.61	33797.79	innovations_mle

Tabela 3.1: Optymalny model $\text{ARMA}(p, q)$ według kryteriów informacyjnych

Na podstawie powyższych wyników wybrano model $\text{ARMA}(3, 0)$ jako najlepiej dopasowany do otrzymanego szeregu.

Po wyborze modelu $\text{ARMA}(3, 0)$ jako optymalnego, przeprowadzono estymację jego parametrów przy użyciu metody `innovations_mle`.

Na podstawie uzyskanych wyników, dopasowany model $\text{ARMA}(3, 0)$ można zapisać w postaci:

$$X_t = 1.019 \cdot X_{t-1} - 0.322 \cdot X_{t-2} + 0.100 \cdot X_{t-3} + \varepsilon_t,$$

gdzie ε_t oznacza składnik losowy o wariancji $\sigma^2 = 4.786$.

Szczegółowa ocena dopasowania modelu została przedstawiona w kolejnym rozdziale.

ROZDZIAŁ 4

OCENA DOPASOWANIA MODELU

W celu oceny dopasowania modelu do oczyszczonych danych znaleziono przedziały ufności dla funkcji autokowariancji oraz autokorelacji szeregu czasowego AR(3).

Przedziały ufności zostały wyznaczone poprzez wielokrotne generowanie próbek o rozmiarze $n = 1000$ z wykorzystaniem procesu ARMA(3,0) z parametrami dopasowanymi do danych. Dla każdej próbki obliczono wartości autokowariancji oraz autokorelacji dla $h = 0, 1, \dots, 50$. Następnie, dla każdego opóźnienia h , określono kwantyle rozkładu empirycznego autokowariancji i autokorelacji, odpowiadające poziomom istotności $\alpha = 0.05$. Kwantylny dolny $\alpha/2$ i kwantylny górny $1 - \alpha/2$ wyznaczyły granice przedziałów ufności.

Wyniki analizy przedstawiono na [Wykresie 4.1](#). Górny panel obrazuje wartości autokowariancji empirycznej oraz teoretycznej, z naniesionymi przedziałami ufności. Dolny panel przedstawia analogiczne wyniki dla autokorelacji. Szare obszary oznaczają zakresy przedziałów ufności, które wizualnie wskazują na rozrzut oczekiwanych wartości.

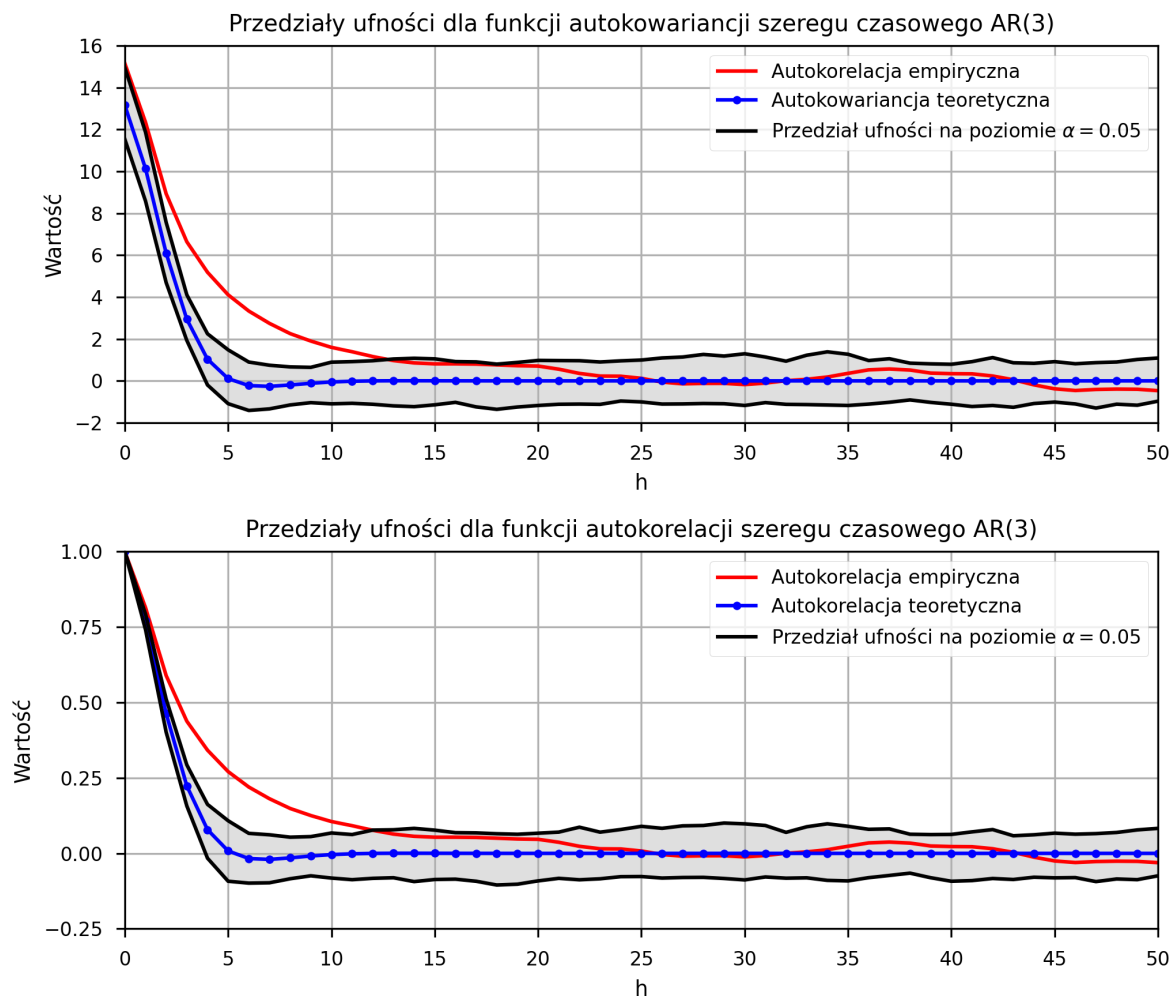
Z analizy wykresów wynika, że dla opóźnień $h > 13$ zarówno wartości autokowariancji, jak i autokorelacji dla otrzymanego szeregu mieszczą się w przedziałach ufności. Jednak dla mniejszych opóźnień (tj. $h \leq 13$) wartości te przekraczają górne granice przedziałów. Może to świadczyć o tym, że składnik szumu nie pochodzi z rozkładu normalnego, co wprowadza dodatkowe odstępstwa od teoretycznych założeń modelu ARMA(3,0).

Pomimo tych odstępstw, model ARMA(3,0) można uznać za stosunkowo dobrze dopasowany, ponieważ w większych opóźnieniach, gdzie dominują istotne zależności dla danych szeregowych, wartości mieszczą się w oczekiwanych granicach. Sugeruje to, że model skutecznie odwzorowuje strukturę danych i może być wykorzystywany do analizy predykcyjnej, choć potencjalna nienormalność szumu wymagałaby dalszych badań.

Podsumowując, wyniki analizy potwierdzają, że model ARMA(3,0) w dużej mierze odwzorowuje charakterystyki oczyszczonego szeregu czasowego, co czyni go użytecznym narzędziem w badaniach na danych meteorologicznych. Niemniej jednak, identyfikacja szumu wskazuje na możliwość udoskonalenia modelu poprzez uwzględnienie alternatywnych rozkładów lub dodatkowych komponentów.

Oprócz analizy autokowariancji i autokorelacji, przeprowadzono również ocenę funkcji autokorelacji cząstkowej (PACF) dla szeregu czasowego AR(3).

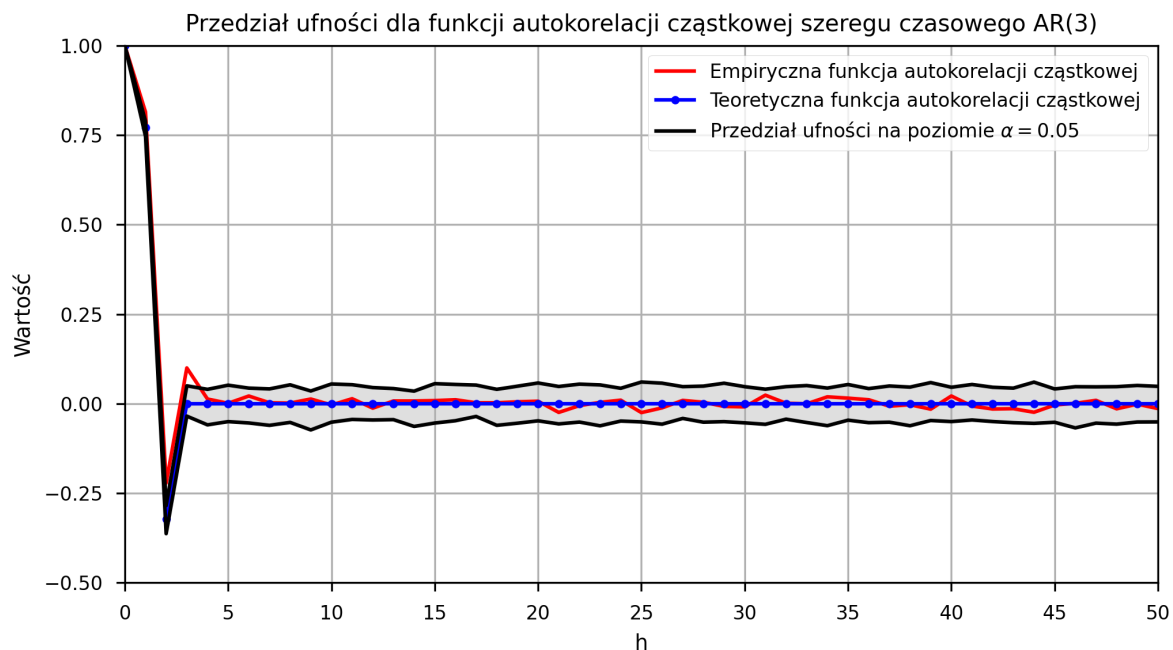
Przedziały ufności wyznaczono na podstawie metody Monte Carlo, analogicznie jak dla funkcji ACVF i ACF. Wartości funkcji autokorelacji cząstkowej zostały obliczone przy użyciu równań Yule-Walkera.



Wykres 4.1: Porównanie funkcji autokowariancji i autokorelacji otrzymanego szeregu z teoretycznymi wartościami funkcji autokowariancji i autokowariancji szeregu AR(3) oraz empirycznie wyznaczonych przedziałów ufności dla AR(3) na poziomie $\alpha = 0.05$.

Wyniki analizy przedstawiono na [Wykresie 4.2](#). Z analizy wynika, że empiryczne wartości funkcji PACF w zdecydowanej większości mieszczą się w wyznaczonych przedziałach ufności. Dla mniejszych opóźnień (tj. $h \leq 6$) wartości funkcji PACF wykazują odstępstwa od przedziałów, co ponownie może sugerować, że składnik szumu w otrzymanym szeregu nie pochodzi z rozkładu normalnego. Analiza funkcji PACF dostarcza dodatkowych informacji o strukturze zależności w szeregach czasowych, a uzyskane wyniki potwierdzają, że model ARMA(3,0) stanowi dobrą, choć nie idealną, aproksymację dla oczyszczonych danych.

W celu oceny dopasowania modelu AR(3) do oczyszczonych danych przeprowadzono także porównanie linii kwantylowych szeregu czasowego wygenerowanego na podstawie modelu z rzeczywistym szeregiem oczyszczonych danych treningowych. Linie kwantylowe zostały wyznaczone dla rzędów 10%, 20%, ..., 90%, przy czym dla każdego kwantyla obliczono empiryczne wartości na podstawie 10^4 próbek o rozmiarze równym długości szeregu rzeczywistego. Wyniki przedstawiono na [Wykresie 4.3](#).



Wykres 4.2: Porównanie funkcji autokorelacji cząstkowej otrzymanego szeregu z teoretycznymi wartościami funkcji autokorelacji cząstkowej szeregu AR(3) oraz empirycznie wyznaczonych przedziałów ufności dla AR(3) na poziomie $\alpha = 0.05$.

Z analizy wynika, że rzeczywisty szereg czasowy w większości opóźnień mieści się w wyznaczonych przedziałach kwantylowych, co świadczy o dobrym dopasowaniu modelu AR(3). Niemniej jednak, dokładne oszacowanie pokazuje, że niektóre obserwacje wychodzą poza granice wyznaczone przez linie kwantylowe, co może być związane z nienormalnością rozkładu szumu lub obecnością efektów nieliniowych w danych.

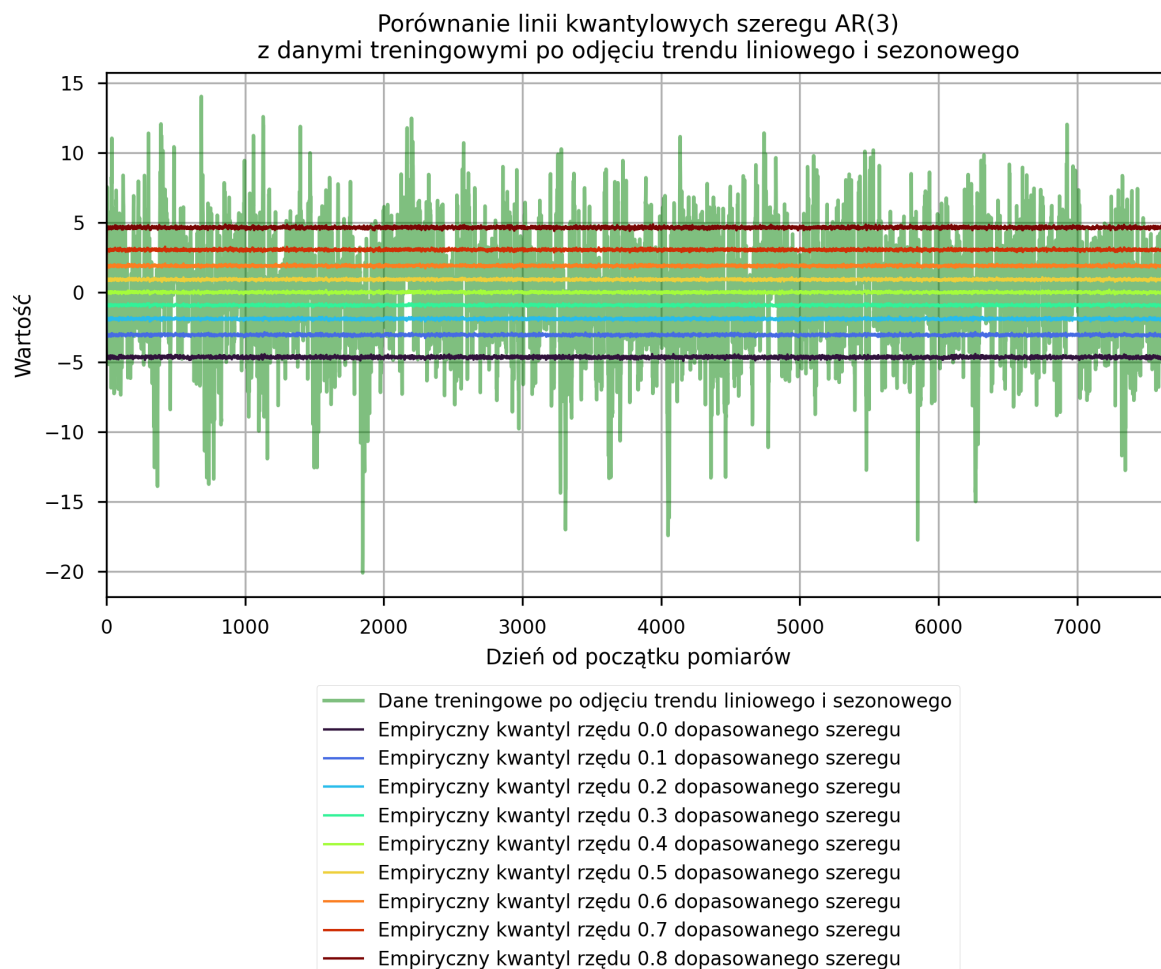
Szczegółowe statystyki rozkładu obserwacji między przedziałami kwantylowymi, przedstawione w Tabeli 4.1, wskazują, że udział danych w każdym przedziale jest bliski teoretycznie oczekiwanej wartości (10%), z niewielkimi odchyleniami.

Tabela 4.1: Rozkład obserwacji w przedziałach kwantylowych dla modelu AR(3).

Przedział ($Q_n - Q_{n+1}$)	Procent obserwacji (%)
$Q_{0.0} - Q_{0.1}$	9.70
$Q_{0.1} - Q_{0.2}$	9.17
$Q_{0.2} - Q_{0.3}$	9.35
$Q_{0.3} - Q_{0.4}$	9.58
$Q_{0.4} - Q_{0.5}$	9.14
$Q_{0.5} - Q_{0.6}$	10.47
$Q_{0.6} - Q_{0.7}$	10.25
$Q_{0.7} - Q_{0.8}$	10.89

Procent obserwacji w poszczególnych przedziałach kwantylowych waha się od 9.14% do 10.89%, co świadczy o braku wyraźnych odchyleń i równomiernym rozkładzie danych. Choć dopasowanie modelu jest ogólnie zadowalające, różnice między skrajnymi wartościami mogą sugerować,

że model nie jest w pełni idealny. Tego rodzaju odchylenia mogą wskazywać, że rozkład szumu w modelowanym szeregu nie jest normalny i może być asymetryczny.

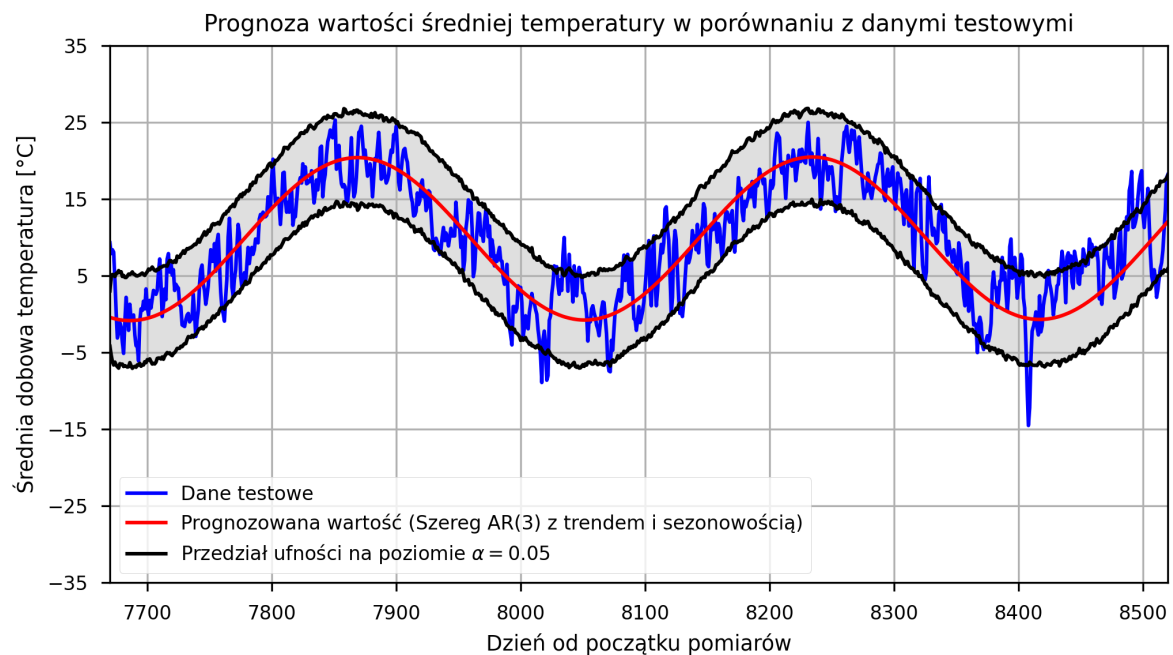


Wykres 4.3: Porównanie empirycznie wyznaczonych linii kwantylowych szeregu czasowego AR(3) z danymi treningowymi po odjęciu trendu liniowego i sezonowości.

Aby sprawdzić dopasowanie modelu do danych, porównano również prognozę z rzeczywistym zbiorem danych testowych. Prognozy zostały uzyskane za pomocą modelu AR(3), trendu i sezonowości. Na podstawie wygenerowanych symulacji z modelu AR(3) obliczono przedziały ufności na poziomie 5% dla przyszłych wartości.

Wizualizacja wyników, przedstawiona na [Wykresie 4.4](#) pokazuje porównanie rzeczywistych danych testowych z prognozowanymi wartościami oraz przedziały ufności. Na wykresie widać, jak prognozowane wartości (uwzględniające zarówno komponenty losowe, jak i deterministyczne) pokrywają się z rzeczywistymi danymi, a także jak przedziały ufności uchwytują zakres możliwych wyników w przyszłości. Około 88% wartości ze zbioru testowego mieści się w przedziale między 5% a 95% kwantylami prognozowanych wartości.

Wyniki wskazują, że model AR(3) dobrze odwzorowuje dane testowe i może służyć do prognozowania wartości średniej dziennej temperatury powietrza w Pszczynie w przyszłości.



Wykres 4.4: Porównanie rzeczywistych danych testowych z prognozowanymi wartościami średniej dziennej temperatury powietrza w Pszczynie. Około 88% wartości ze zbioru testowego mieści się w przedziale między 5% a 95% kwantylami prognozowanych wartości.

ROZDZIAŁ 5

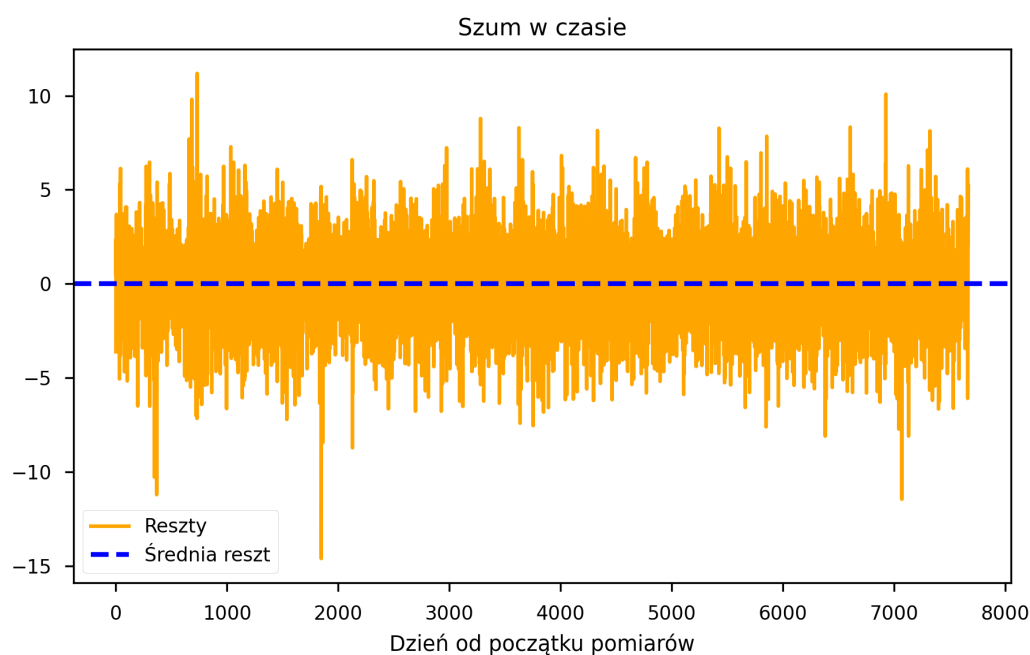
WERYFIKACJA ZAŁOŻEŃ DOTYCZĄCYCH SZUMU

Zakłada się, że residua po dopasowaniu modelu AR(3) powinny spełniać warunki białego szumu. Oznacza to, że:

- Średnia reszt powinna wynosić zero dla wszystkich t , tj. $\mathbb{E}[e_t] = 0$,
- Wariancja reszt powinna być stała i skończona, tj. $\text{Var}(e_t) = \sigma^2$ dla każdego t ,
- Autokorelacja reszt powinna wynosić 1 dla opóźnienia $h = 0$, a dla innych opóźnień $h > 0$ autokorelacja powinna być zerowa.

W celu przeprowadzenia analizy szumu, zbadano residua uzyskane po porównaniu modelu AR(3) z szeregiem uzyskanym po usunięciu trendu i sezonowości z danych. W tym celu wykorzystano atrybut `resid` obiektu `model` z biblioteki `statsmodels`. Analizę rozpoczęto od wizualnej oceny szumu, przedstawiając dane na [Wykresie 5.1](#). Wartości reszt nie wykazują widocznych zależności czasowych, co sugeruje brak struktury w danych resztowych.

Dalsza analiza opiera się na dokładniejszych metodach statystycznych, które pozwalają zweryfikować losowość i zgodność reszt z założeniami modelu.



Wykres 5.1: Wykres wartości resztowych szeregu otrzymanego po usunięciu trendu i sezonowości z danych.

W celu zbadania, czy średnia reszt wynosi zero, przeprowadzono test t-Studenta dla otrzymanego szumu. Hipoteza zerowa zakłada, że średnia reszt wynosi zero, a hipoteza alternatywna, że średnia jest różna od zera. Wyniki testu t-Studenta dla reszt zostały przedstawione w Tabeli 5.1. Uzyskano t-statystykę równą 0.0120 oraz p-wartość 0.9904. Ponieważ p-wartość jest znacznie większa niż przyjęty poziom istotności $\alpha = 0.05$, nie ma podstaw do odrzucenia hipotezy zerowej. Oznacza to, że możemy przyjąć, że średnia reszt wynosi zero.

Wielkość	Wartość
t-statystyka	0.0120
p-wartość	0.9904

Tabela 5.1: Wyniki testu t-Studenta dla średniej reszt

W celu zbadania, czy wariancja reszt wykazuje autokorelację, przeprowadzono test ARCH. Test ten ma na celu sprawdzenie, czy wariancja reszt zmienia się w czasie, co wskazywałoby na obecność heteroskedastyczności (zależność wariancji od czasu). Hipoteza zerowa testu ARCH zakłada, że wariancja reszt jest stała, a hipoteza alternatywna mówi, że wariancja jest zmienna. Wyniki testu ARCH zostały przedstawione w Tabeli 5.2. Statystyka testu wynosi 138.3902, a p-wartość jest równa 9.006×10^{-25} . Ponieważ p-wartość jest znacznie mniejsza niż poziom istotności $\alpha = 0.05$, odrzucamy hipotezę zerową, co sugeruje, iż wariancja reszt wykazuje autokorelację i jest zmienna w czasie.

Wielkość	Wartość
Statystyka ARCH	138.3902
p-wartość	9.006×10^{-25}

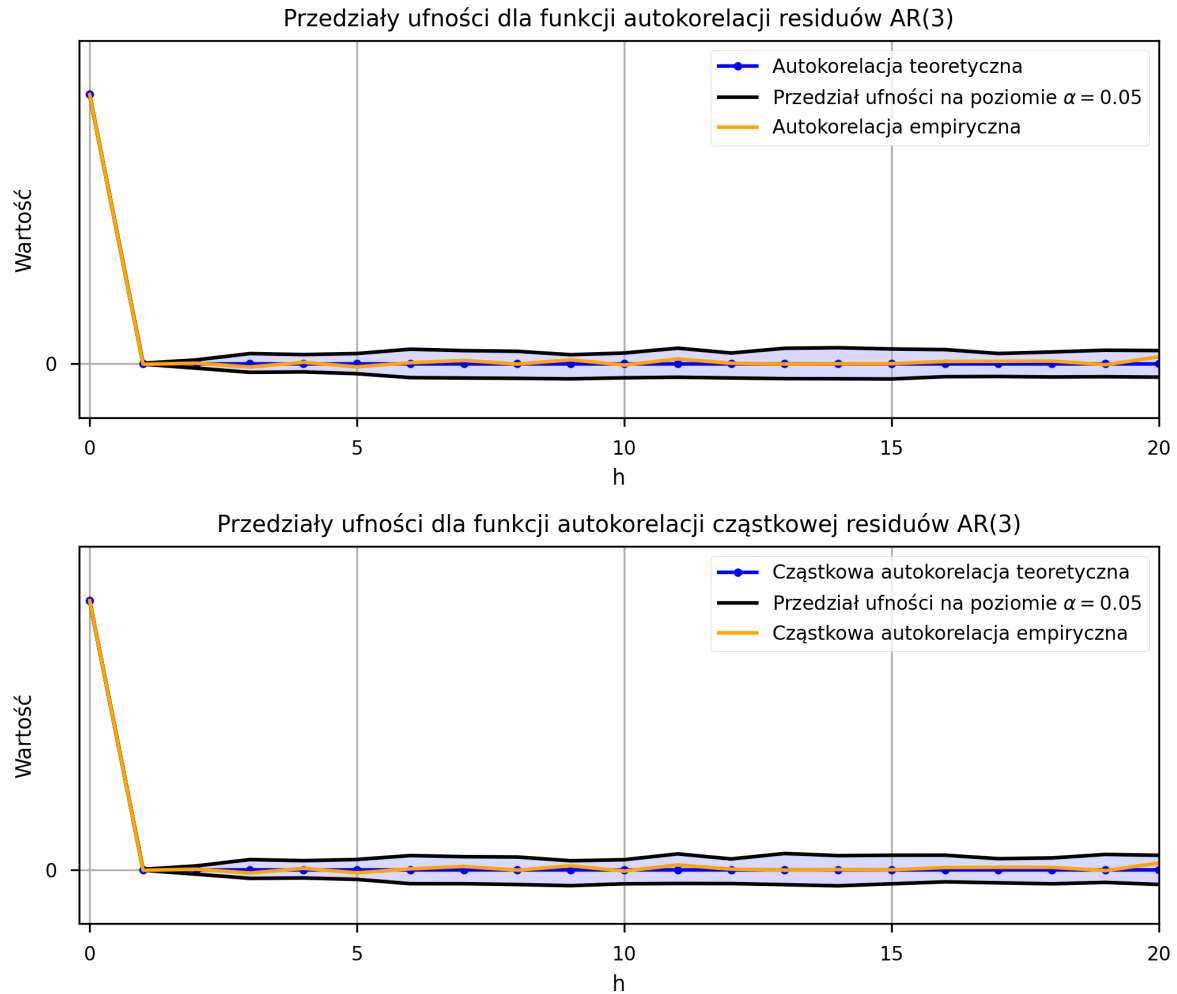
Tabela 5.2: Wyniki testu ARCH dla wariancji reszt

Dodatkowo, dla dalszej analizy wariancji reszt przeprowadzono test Levene'a, który służy do porównania równości wariancji w różnych grupach. W tym przypadku, reszty zostały podzielone na 2, 5 i 10 grup, a dla każdej z tych konfiguracji przeprowadzono test. Hipoteza zerowa testu Levene'a zakłada, że wariancje są równe (stałe), a hipoteza alternatywna mówi, że wariancje są różne. Wyniki testu przedstawiono w Tabeli 5.3.

Dla podziału na 2 grupy nie ma podstaw do odrzucenia hipotezy zerowej, co sugeruje, że wariancja reszt jest stała. Natomiast dla podziału na 5 i 10 grup p-wartości były mniejsze niż 0.05, co pozwala na odrzucenie hipotezy zerowej i wskazuje na możliwość zmiennej wariancji reszt w tych przypadkach. Wnioski te sugerują pewną zmienność wariancji w danych, co wskazuje na obecność heteroskedastyczności.

Liczba grup	Statystyka Levene'a	p-wartość
2	1.0355	0.3164
5	2.5510	0.0354
10	2.5202	0.0296

Tabela 5.3: Wyniki testu Levene'a dla wariancji reszt w zależności od liczby grup



Wykres 5.2: Porównanie funkcji autokorelacji i autokorelacji cząstkowej dopasowanego modelu z teoretycznymi wartościami funkcji autokowariancji i autokowariancji oraz empirycznie wyznaczonych przedziałów ufności dla residuów AR(3) na poziomie $\alpha = 0.05$.

Wnioski z przeprowadzonych testów sugerują, że wariancja reszt jest zmienna w czasie, co wskazuje na obecność heteroskedastyczności, a tym samym na naruszenie założeń białego szumu, które zakładałyby stałą wariancję reszt.

Analiza funkcji autokorelacji (ACF) oraz autokorelacji cząstkowej (PACF) reszt została przedstawiona na [Wykresie 5.2](#).

Dla reszt z dopasowanego modelu AR(3) obliczono funkcje autokorelacji (ACF) oraz cząstkowej autokorelacji (PACF) na podstawie dostępnych danych. Obliczone wartości ACF i PACF uzyskano przy użyciu odpowiednich funkcji z biblioteki `statsmodels`, z uwzględnieniem opóźnień do $h_{\max} = 20$. Teoretyczne wartości ACF zostały określone przy założeniu, że reszty modelu powinny mieć postać białego szumu.

Aby ocenić zachowanie funkcji ACF i PACF w kontekście losowego szumu, przeprowadzono symulację prób z procesów AR(3) z zadanymi współczynnikami autoregresji oraz wariancją szumu. Zrealizowano $N = 100$ prób o długości $n = 1000$, z których każda została dopasowana do modelu ARIMA(3,0,0). Dla residuów każdej z prób obliczono wartości ACF i PACF. Na podstawie wyników symulacji dla każdej z funkcji ACF i PACF obliczono empiryczne przedziały ufności na poziomie $\alpha = 0.05$.

Wyniki analizy wykazały, że dla reszt z modelu AR(3) wartości funkcji ACF i PACF leżą w granicach wyznaczonych przedziałów ufności, co wskazuje na brak autokorelacji w resztach. Jest to silna przesłanka, że reszty mogą stanowić biały szum, choć należy pamiętać, że brak autokorelacji nie wyklucza całkowicie możliwości istnienia bardziej złożonych, nieliniowych zależności.

W celu dokładniejszego zbadania autokorelacji reszt, przeprowadzono test Ljunga-Boxa dla różnych opóźnień (1, 2, 5 oraz 10). Test ten służy do weryfikacji, czy w szeregu czasowym występuje autokorelacja w podanym zakresie opóźnień. Hipoteza zerowa testu Ljunga-Boxa zakłada brak autokorelacji reszt (są one niezależne), natomiast hipoteza alternatywna wskazuje na obecność autokorelacji.

Wyniki testu dla wybranych opóźnień przedstawiono w Tabeli 5.4. P-wartości dla wszystkich opóźnień (1, 2, 5, 10) są znacznie większe od przyjętego poziomu istotności $\alpha = 0.05$. W związku z tym brak jest podstaw do odrzucenia hipotezy zerowej, co oznacza, że reszty nie wykazują autokorelacji w badanych zakresach. Wyniki te wspierają wniosek, że reszty nie są skorelowane.

Opóźnienia	Statystyka Ljunga-Boxa	p-wartość
1	0.0143	0.9048
2	0.0691	0.9660
5	2.2799	0.8092
10	5.5791	0.8493

Tabela 5.4: Wyniki testu Ljunga-Boxa dla różnych opóźnień (1, 2, 5, 10)

Sprawdzono także normalność reszt, ponieważ wiele testów statystycznych, takich jak testy na autokorelację czy heteroskedastyczność, zakłada, że błędy modelu mają rozkład normalny. Ponadto, model ARMA z biblioteki statsmodels generuje próbki z szumem pochodzącym z rozkładu normalnego.

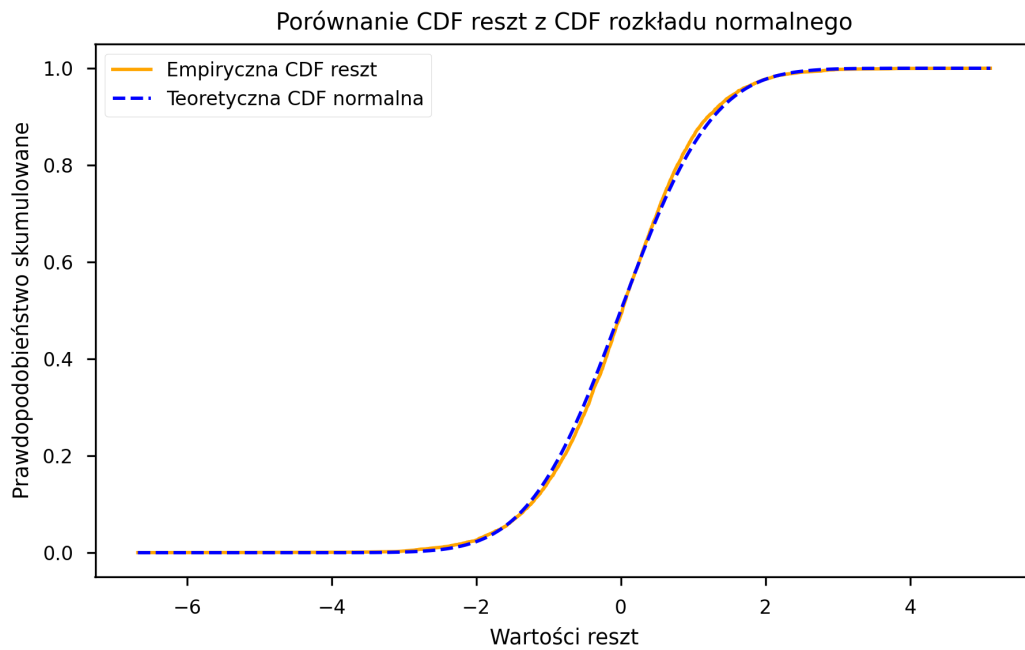
Reszty e_t zostały znormalizowane według wzoru:

$$e_t^* = \frac{e_t - \bar{e}_t}{\sigma_{e_t}}$$

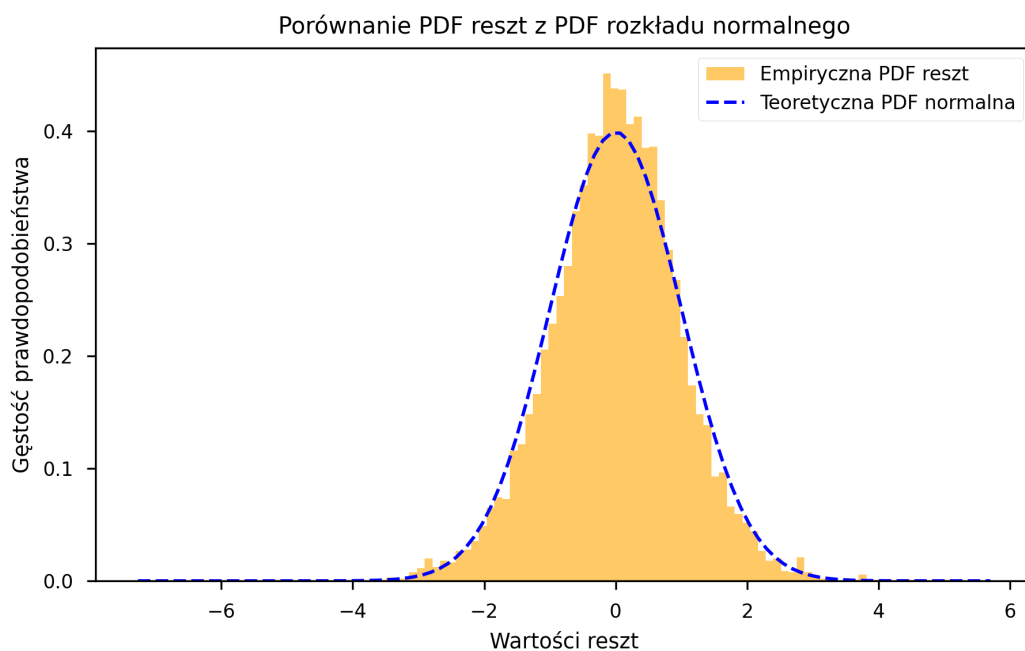
gdzie \bar{e}_t to średnia reszt, a σ_{e_t} to ich odchylenie standardowe. Normalizacja umożliwia porównanie rozkładu reszt z rozkładem $\mathcal{N}(0, 1)$.

Na Wykresie 5.3 przedstawiono porównanie empirycznej dystrybuanty znormalizowanych reszt z teoretycznym rozkładem normalnym. Na pierwszy rzut oka można stwierdzić, że rozkład reszt wygląda na zgodny z założeniem normalności, ponieważ linie empirycznej i teoretycznej dystrybuanty są bardzo zbliżone. Minimalne różnice pojawiają się lokalnie, lecz ich wielkość nie wskazuje na wyraźne odstępstwa od normalności.

Na Wykresie 5.4 przedstawiono porównanie gęstości prawdopodobieństwa dla znormalizowanych reszt modelu z gęstością teoretycznego rozkładu normalnego. Empiryczną gęstość wyznaczono za pomocą histogramu z funkcji hist z biblioteki Matplotlib, który przedstawia rozkład znormalizowanych reszt. Wysokość słupków histogramu została znormalizowana tak, aby cała pod wykresem wynosiła 1, zgodnie z definicją gęstości prawdopodobieństwa. Analiza wykresu wskazuje, że empiryczny rozkład reszt wykazuje pewne odchylenia od teoretycznego rozkładu normalnego, szczególnie w centralnej części.



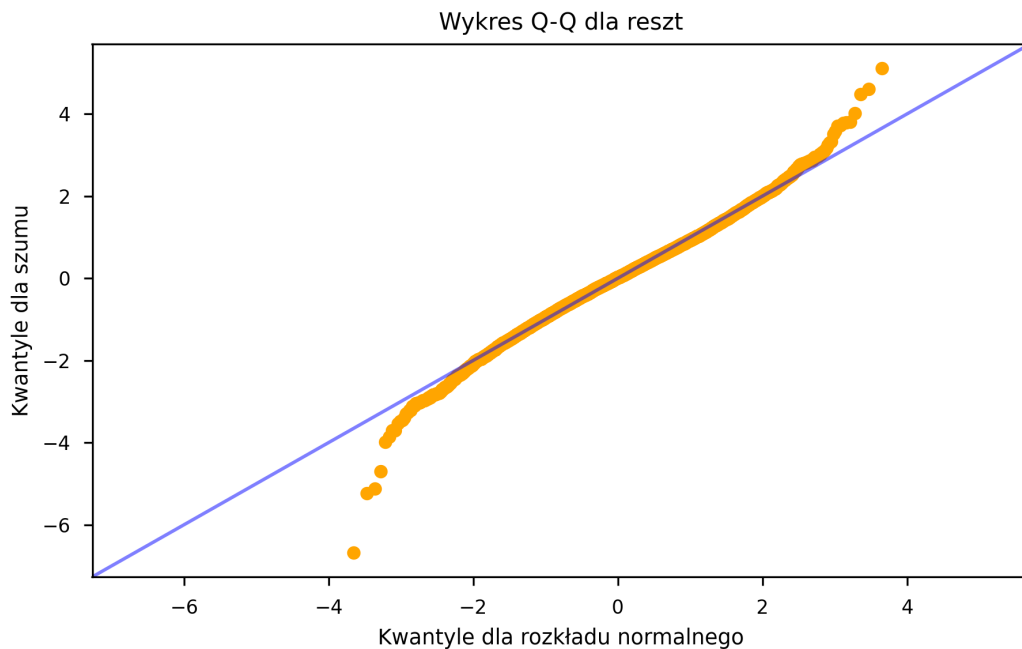
Wykres 5.3: Porównanie dystrybuanty rozkładu $\mathcal{N}(0, 1)$ z dystrybuantą empiryczną znormalizowanych residuów.



Wykres 5.4: Porównanie gęstości rozkładu $\mathcal{N}(0, 1)$ z dystrybuantą empiryczną znormalizowanych residuów.

Na [Wykresie 5.5](#) przedstawiono wykres Q-Q dla znormalizowanych reszt modelu, który pozwala na ocenę zgodności rozkładu reszt z rozkładem normalnym. Został on wygenerowany za pomocą funkcji `qqplot` z biblioteki `statsmodels`, gdzie punkty reprezentują kwantyle empiryczne reszt w odniesieniu do kwantyli teoretycznych rozkładu normalnego. Analiza wykresu

wskazuje, że reszty odbiegają od linii prostej na krańcach rozkładu, co sugeruje pewne odchylenia od normalności, choć w centralnej części rozkładu punkty są stosunkowo blisko linii odniesienia.



Wykres 5.5

W celu jednoznacznego ustalenia zgodności rozkładu reszt z rozkładem normalnym przeprowadzono trzy testy statystyczne: test Shapiro-Wilka, test Kołmogorowa-Smirnowa oraz test D'Agostino-Pearsona. Każdy z przeprowadzonych testów zakładał hipotezę zerową (H_0), że rozkład jest normalny. Wyniki przedstawiono w Tabeli 5.5. P-wartości we wszystkich testach są znacznie mniejsze od przyjętego poziomu istotności $\alpha = 0.05$, co prowadzi do odrzucenia hipotezy zerowej o normalności rozkładu. Tym samym można stwierdzić, że rozkład reszt istotnie odbiega od normalnego.

Test	Statystyka	p-wartość	Wniosek
Shapiro-Wilk	0.9936	4.504×10^{-18}	Odrzucamy H_0
Kołmogorow-Smirnow	0.0219	0.0012	Odrzucamy H_0
D'Agostino-Pearson	182.9455	1.879×10^{-40}	Odrzucamy H_0

Tabela 5.5: Wyniki testów normalności dla znormalizowanych reszt. Każdy z przeprowadzonych testów zakładał hipotezę zerową (H_0), że rozkład jest normalny.

ROZDZIAŁ 6

ZAKOŃCZENIE

W niniejszej pracy dokonano analizy danych rzeczywistych dotyczących średnich dziennych temperatur w Pszczynie przy wykorzystaniu modeli autoregresyjnych. Proces analizy obejmował wstępne przygotowanie danych, ich dekompozycję na składowe trendu, sezonowości oraz szumu, a następnie modelowanie przy pomocy ARMA.

W pierwszym etapie pracy zweryfikowano stacjonarność szeregu czasowego oraz określono dominujące struktury zależności czasowych. Analiza autokorelacji i autokorelacji cząstkowej wykazała obecność silnych zależności sezonowych, co potwierdziło konieczność uwzględnienia sezonowości w modelu. Następnie dokonano eliminacji komponentów trendu i sezonowości.

Na podstawie kryteriów informacyjnych AIC, BIC oraz HQIC wybrano optymalny model autoregresyjny, którym okazał się model AR(3). Jego parametry zostały oszacowane przy użyciu metody największej wiarygodności, a uzyskane wyniki wskazały na dobrą zgodność modelu z danymi treningowymi. Weryfikacja dopasowania modelu poprzez analizę autokorelacji reszt wykazała, że dla dłuższych opóźnień wartości mieszczą się w przedziałach ufności, co potwierdza poprawność przyjętego modelu. Niemniej jednak, dla krótszych opóźnień zaobserwowano niewielkie odchylenia, które mogły wskazywać na obecność nienormalności w składniku losowym. Szczególnie widoczne było to w analizie funkcji autokorelacji cząstkowej, gdzie wartości dla kilku pierwszych opóźnień nieznacznie przekraczały teoretyczne przedziały ufności.

Przeprowadzone testy statystyczne wykazały, że składnik szumu modelu nie spełnia założenia normalności. Testy Shapiro-Wilka, Kołmogorowa-Smirnowa oraz D'Agostino-Pearsona jednoznacznie odrzuciły hipotezę o normalnym rozkładzie reszt. Oznacza to, że w przyszłych badaniach warto rozważyć zastosowanie innych rozkładów dla składnika losowego, które mogłyby lepiej odzwierciedlać rzeczywiste właściwości danych.

Analiza szumu wykazała jednak, że większość kluczowych założeń modelu została spełniona. Średnia reszt okazała się bliska zeru, co sugeruje brak systematycznych błędów w modelu. Ponadto, test Ljunga-Boxa nie wykazał istotnej autokorelacji reszt, a funkcja cząstkowej autokorelacji nie wskazała na obecność zależności, które mogłyby sugerować niedostateczne dopasowanie modelu. Głównym wyzwaniem pozostaje zmienność wariancji reszt, na co wskazały wyniki testu ARCH. W kolejnej próbie warto zwrócić uwagę na zmienność wariancji reszt i spróbować dopasować modele, które ją uwzględniają.

Podsumowując, model AR(3) okazał się skutecznym narzędziem do analizy i prognozowania analizowanego szeregu czasowego. Pomimo pewnych ograniczeń związanych z rozkładem szumu, model ten dobrze odwzorowuje strukturę zależności w danych meteorologicznych. Dalsza praca mogłaby skupić się na eksploracji alternatywnych modeli probabilistycznych dla składnika losowego.