Komputerowa analiza szeregów czasowych

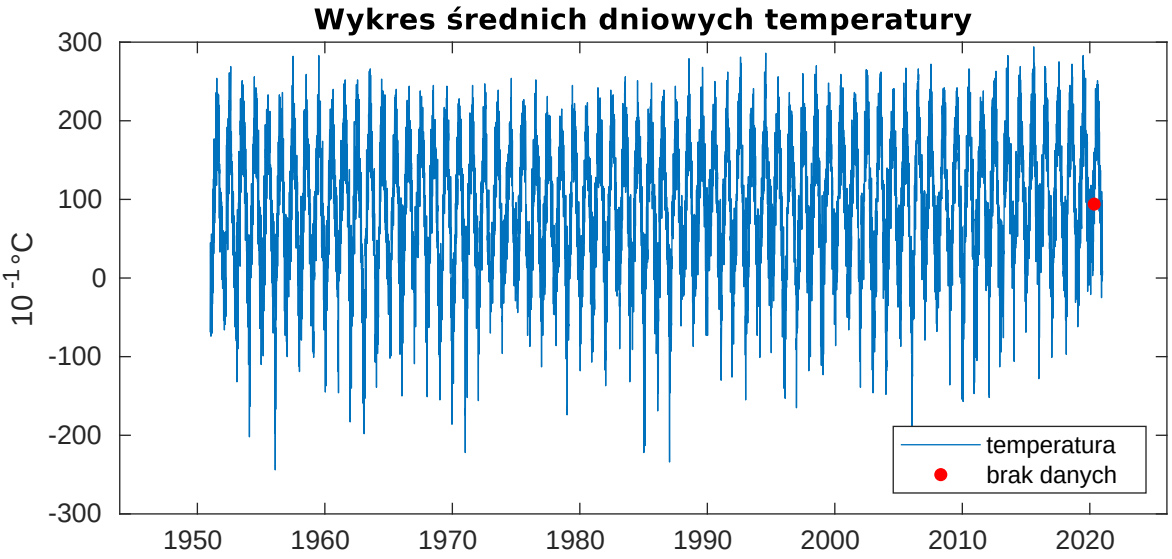
Raport z analizy danych

Bogna Jaszczak
Andrzej Puć

1 sierpnia 2021

1 Przedstawienie danych i cel analizy

Dane pochodzą ze strony **European Climate Assessment and Dataset** (`https://www.ecad.eu`) i liczą 25567 obserwacji. Opisują one średnią dobową temperaturę zarejestrowaną we Wrocławiu między 1.01.1951 a 31.12.2020 . Wartości temperatur wyrażone są w 0.1 stopnia Celsjusza. Pobrany szereg czasowy przedstawiony jest na Rysunku 1.



Rysunek 1: Wykres przedstawiający całe zebrane dane dniowe od roku 1951. Zawierają jeden brak pomiaru oznaczony w oryginalnym pliku numerycznie przez -9999.

Celem raportu będzie odtworzenie podejścia Box'a - Jenkins'a modelowania procesów stochastycznych opisanego w rozdziale 20.2.5 Econometric Analysis Williama H. Greene [1]. Dane poddamy więc dekompozycji Wolda a następnie spróbujemy dopasować do nich model ARMA z odpowiednimi parametrami. Jako główne kryterium dopasowania modelu przyjęte zostanie podobieństwo residuów do białego szumu.

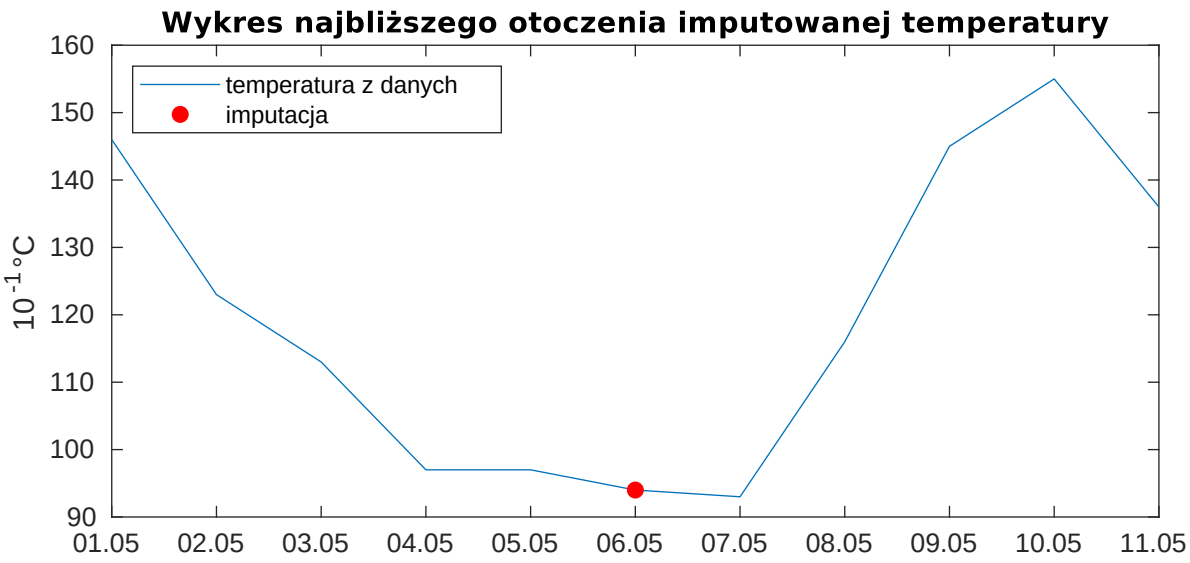
2 Przygotowanie danych do analizy

W tym rozdziale skupimy się głównie na wybraniu odpowiedniej wartości w miejsce brakującego pomiaru. Zostanie także przeprowadzona wstępna graficzna analiza danych.

2.1 Imputacja danych

Zakładamy, że dane pochodzące z nieznacznie odległych stacji pomiarowych nie różnią się znacząco. Zakładamy też, że przyczyna wystąpienia braku danych jest całkowicie losowa. Na podstawie takich założeń można, ze świadomością popełnienia pewnego błędu, użyć pomiaru z pobliskiej stacji w miejscu braku danych. Na podstawie indeksacji danych można zatem

określić dokładną datę w której występuje brak (6 maj 2020) i podstawić w to miejsce wartość pomiaru wynoszącą 9.4 stopni Celcjusza [2].



Rysunek 2: Imputowana wartość temperatury wydaje się dobrze oddawać charakter danych.

Wykres na Rysunku 2 wskazuje na możliwość popełnienia dużego błędu podczas imputowania na podstawie pomiarów najbliższych. Dużo mniejszy błąd zostałby prawdopodobnie popełniony przy zastosowaniu wstawiania ostatniej znanej wartości (Last value carried forward, [3]). Podejście oparte na wykorzystaniu pomiaru z innej stacji może okazać się jednak najlepsze. W daleszej analizie przyjmujemy więc już dane po takiej imputacji.

2.2 Podstawowe statystyki

W celu lepszego zilustrowania danych zostaną one w tym rozdziale podzielone na przedziały długości jednego roku. W taki sposób łatwiej będzie zobrazować zarówno trend jak i okresowość temperatury. Należy podkreślić, że podział został wykonany na podstawie długości roku przestępnego (366 dni). Taki podział wymusza obliczenia średniej (4) i odchylenia standardowego (5) temperatury dnia 29.02 na podstawie osiemnastu pomiarów z lat przestępnych.

Na wykresie przedstawionym na Rysunku 3 zauważalny jest trend wzrostowy. Co ciekawe staje się on wyraźniejszy po roku 1990. Można wnioskować, że średnia roczna temperatura rośnie. Obliczając współczynniki kierunkowy prostej regresji dla wszystkich analizowanych danych otrzymamy 0.3631. Ten sam współczynnik obliczony na podstawie danych od roku 1990 do 2020 wynosi 0.6984.

Wykres średniej temperatury dla każdego dnia roku (Rysunek 4) pokazuje spodziewany kształt. Maksimum jest osiągane w miesiącach letnich. Dla analizowanych danych dzień z maksymalną średnią temperaturą to pierwszy sierpnia roku przestępnego czyli drugi sierpnia roku nieprzestępnego.

Odchylenie standardowe widoczne na Rysunku 5 jest znacznie większe dla miesięcy zimowych i wczesno wiosennych niż dla letnich. Może to wynikać z wpływu prądów niosących gwałtowne spadki lub wzrosty temperatury. Warto zauważyć, że takie różnice w odchyleniu standardowym mogą negatywnie wpływać na dopasowanie modeli liniowych.

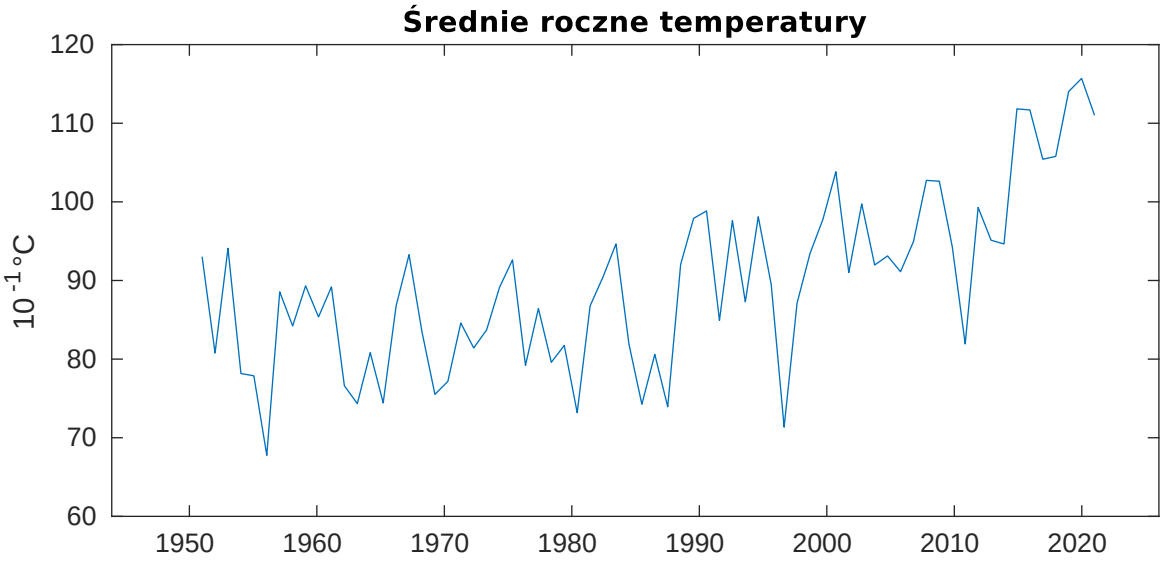
2.3 Dekompozycja Wolda

Zakładamy, że nasze dane opisane są modelem:

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

gdzie:

- $m(t)$ to zauważalny w danych trend nieokresowy,
- $s(t)$ składowa deterministyczna okresowa,
- $\{X_t\}_{t\in\mathbb{Z}}$ to szereg czasowy stacjonarny w słabym sensie.

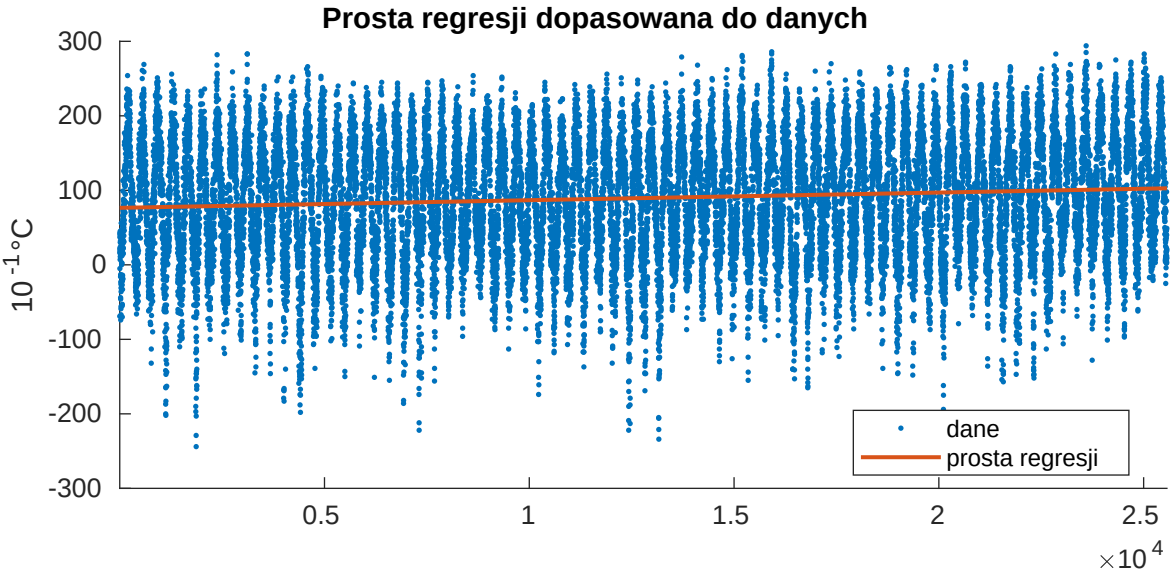


Rysunek 3: Wykres średnich arytmetycznych temperatury rocznej obliczanych na podstawie średnich dniowych temperatur.

Przed przystąpieniem do analizy szeregu czasowego niezbędnym krokiem jest dokonanie **dekompozycji Wolda**, czyli usunięcia składowych deterministycznych $m(t)$ oraz $s(t)$.
W danych zauważalny jest niewielki trend liniowy. Aby go usunąć, wyznaczymy najpierw prostą regresji metodą najmniejszych kwadratów. Otrzymujemy tym sposobem, że

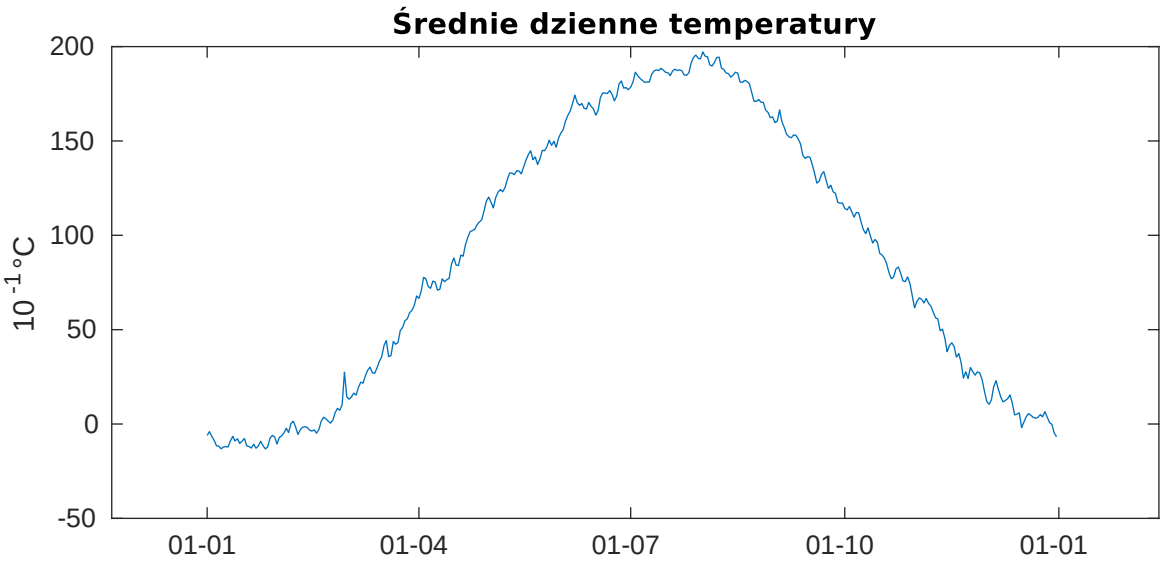
$$\hat{m}(t) = 76.5044 + 0.001t.$$

Dopasowaną do danych prostą regresji zaprezentowano na wykresie (Rysunek 6).

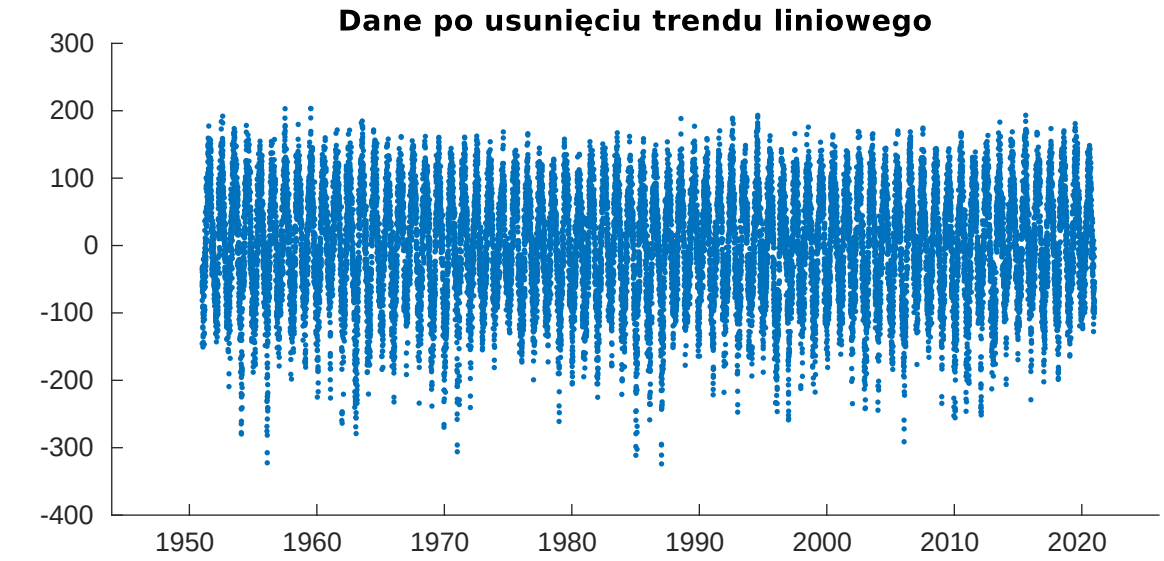


Rysunek 6: Po nałożeniu prostej regresji na dane bardziej widoczne stały się niskie wartości temperatur przed 1990 rokiem.

Zauważmy, że trend ten jest rosnący. Wiąże się to bezpośrednio z ocieplaniem klimatu - średnie temperatury w kolejnych latach systematycznie rosną (porównaj z wykresem 3). Dane po usunięciu trendu liniowego zaprezentowano na wykresie (Rysunek 7).



Rysunek 4: Średnie dobowe temperatury obliczone na podstawie siedemdziesięciu lat.



Rysunek 7: Dane po usunięciu trendu liniowego.

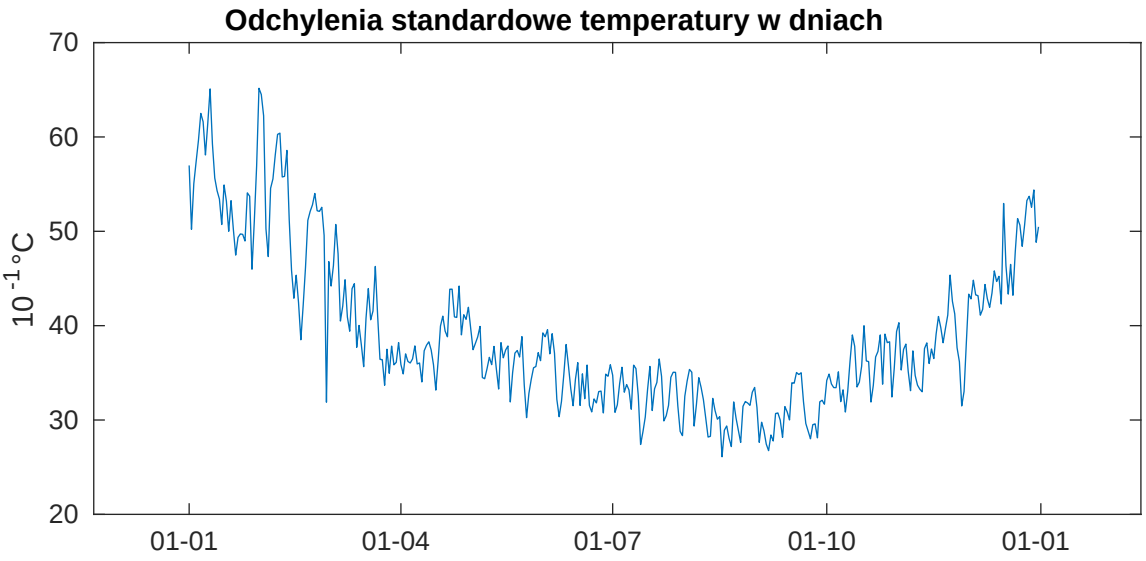
W danych bardzo wyraźnie widoczna jest okresowość. Wynika ona oczywiście z faktu, że Wrocław znajduje się w strefie klimatu umiarkowanego przejściowego. Średnie dobowe temperatury wyraźnie wzrastają w miesiącach letnich oraz obniżają się w miesiącach zimowych (porównaj z wykresem średnich dniowych 4). Do usunięcia okresowości wykorzystamy fukcję `fit`. Umożliwia ona znalezienie parametrów a, b i c modelu:

$$s(t) = a \cdot \sin(b \cdot t + c).$$

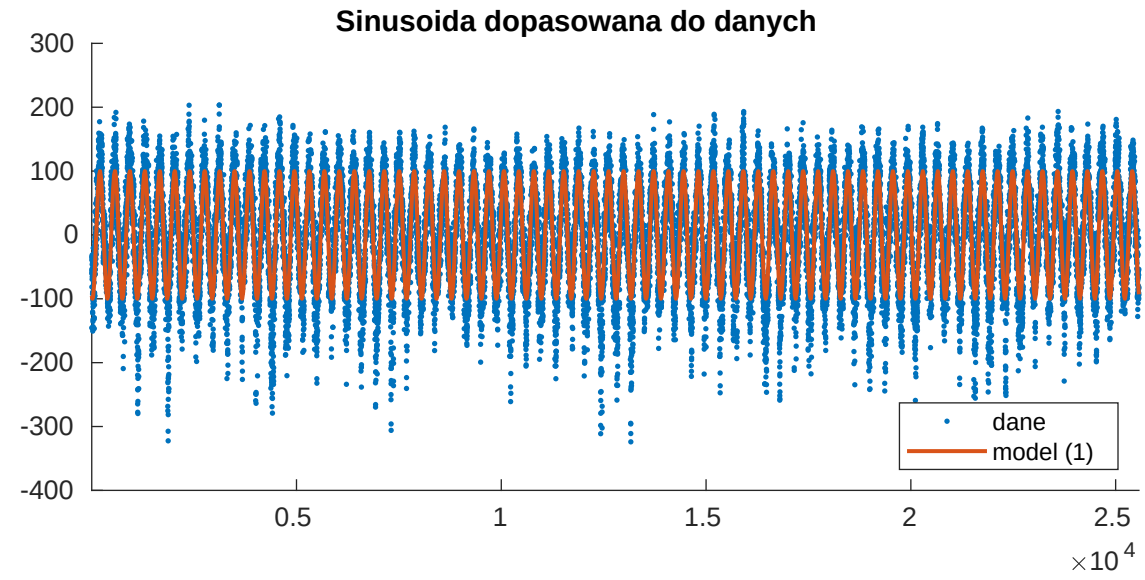
Po wywołaniu funkcji dla naszych danych otrzymujemy, że

$$\hat{s}(t) = 100 \cdot \sin(0.0172 \cdot t - 1.893). \tag{1}$$

Dopasowaną do danych sinusoidę przedstawiono na wykresie (Rysunek 8).

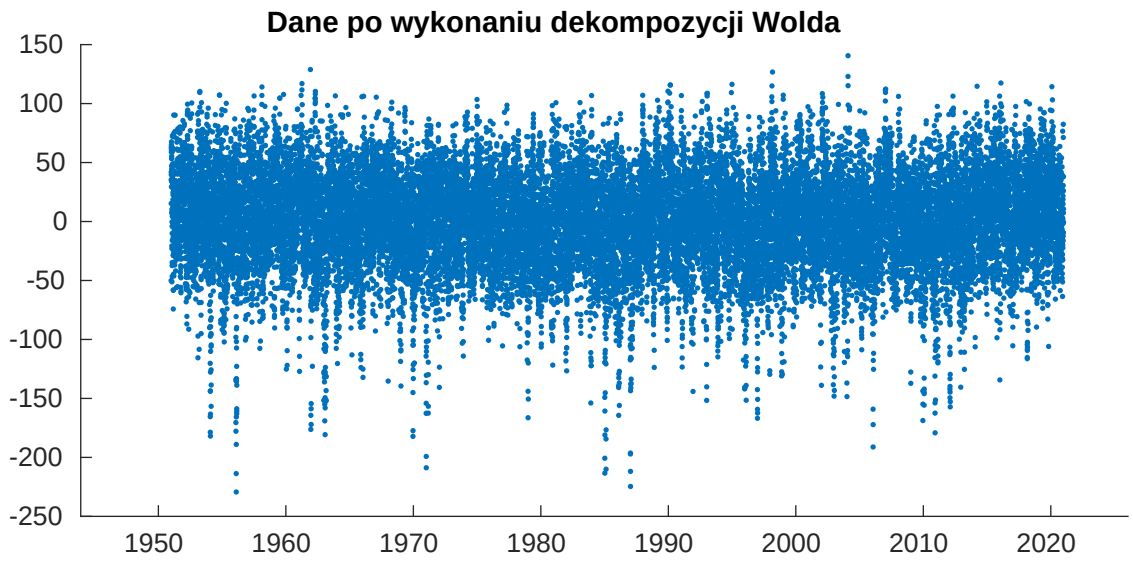


Rysunek 5: Odchylenie standardowe obliczone na takich samych zasadach jak średnia z wykresu 4.



Rysunek 8: Krzywa opisana modelem (1) wyraźnie gorzej opisuje dane z lat 1951-1990 gdzie występowało o wiele więcej skrajnie niskich temperatur.

Ostateczenie, po dokonaniu dekompozycji Wolda, dane prezentują się jak na Rysunku 9. Odjęcie modelu (1) wyraźnie uwydatniło skrajne wartości temperatur. Co więcej, wartości skrajne pojawiają się dość regularnie po czym można wnioskować, że w procesie dekompozycji nie została usunięta cała okresowość danych temperatury lub że pewna cykliczność pozostała w danych.



Rysunek 9: Odjęcie od danych modelu okresowości (1) uwydatniło temperatury ujemne z lat 1951 - 1990.

2.4 ACF oraz PACF

W kolejnym kroku porównamy ACF (funkcję autokorelacji) oraz PACF (funkcję częściowej autokorelacji) dla danych przed dokonaniem dekopozycji Wolda oraz po niej.

Definicja 1 ***Funkcja autokorelacji***
Niech $\{X_t\}_{t\in\mathbb{Z}}$ *będzie szeregiem czasowym o skończonej wartości oczkiwanej i skończonym drugim momencie. Wtedy funkcja autokorelacji dana jest wzorem*

$$\rho_X(t,s) = \frac{\gamma_X(t,s)}{\sqrt{\gamma_X(t,t)\cdot\gamma_X(s,s)}} = \frac{cov(X_t,X_s)}{\sqrt{Var[X_t]\cdot Var[X_s]}},$$

gdzie γ *to funkcja autokowariancji zdefiniowana dla szeregu czasowego drugiego rzędu jako*

$$\gamma(t,s) = cov(X_t,X_s).$$

Definicja 2 ***Funkcja częściowej autokorelacji***
Niech $\{X_t\}_{t\in\mathbb{Z}}$ *będzie stacjonarnym szeregiem czasowym postaci*

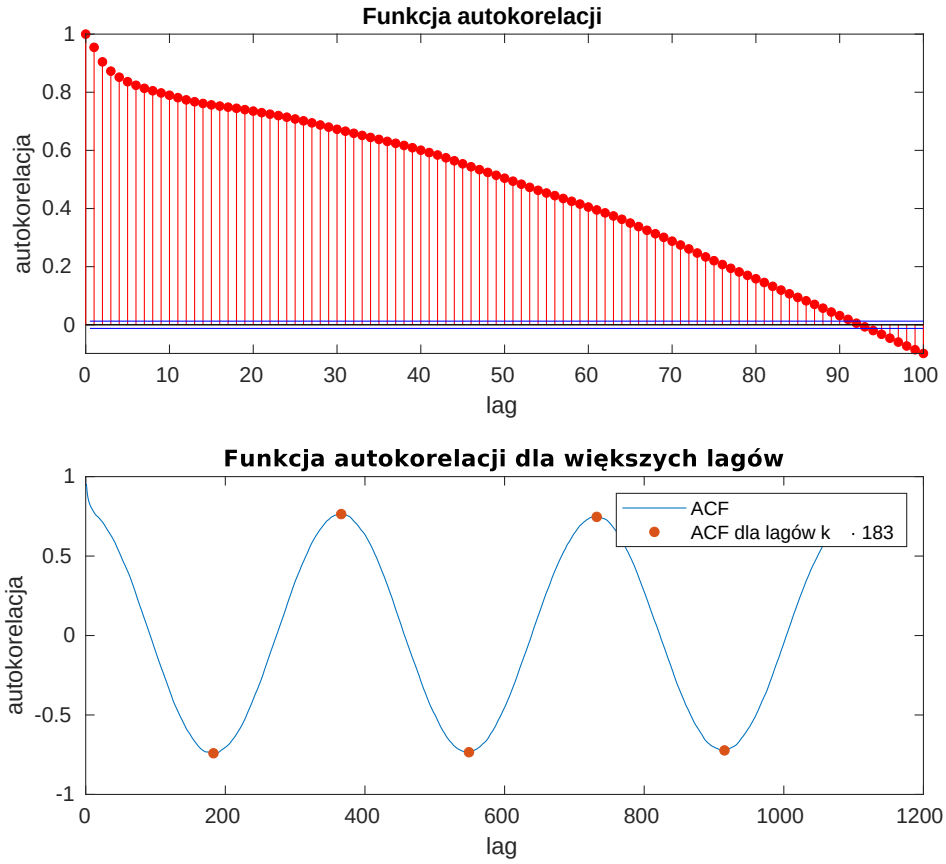
$$X_t = \phi_0 + \phi_{k1}X_{t-1} + ... + \phi_{kk}X_{t-k} + \xi_t,$$

gdzie ξ_t *to ciąg nieskorelowanych zmiennych losowych o średniej zero. Wtedy*

$$PACF(k) = \phi_{kk}.$$

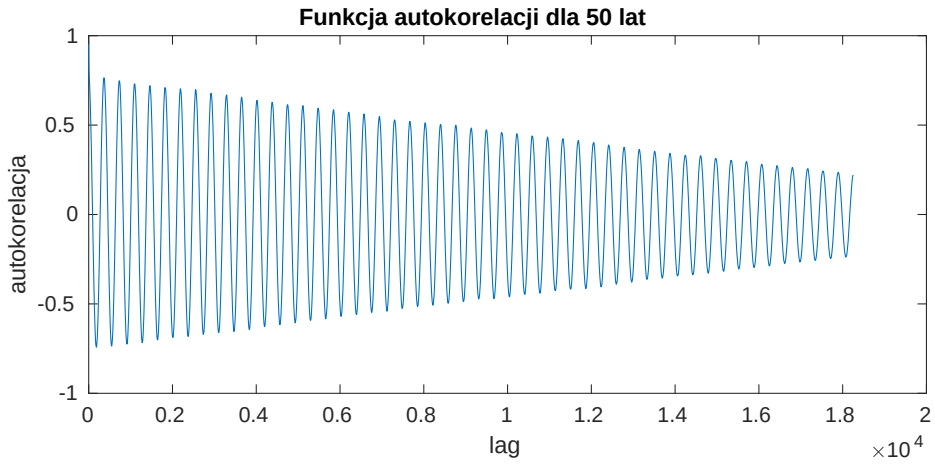
2.5 ACF

Funkcja autokorelacji zdefiniowana jak w 1 informuje nas o tym, na ile dane obserwacje są istotnie związane z obserwacjami zaobserwowanymi wcześniej (o stałym przesunięciu czasowym). Gdy ACF przyjmuje wartości bliskie 0 oznacza to, że korelacja między danymi jest bardzo słaba.

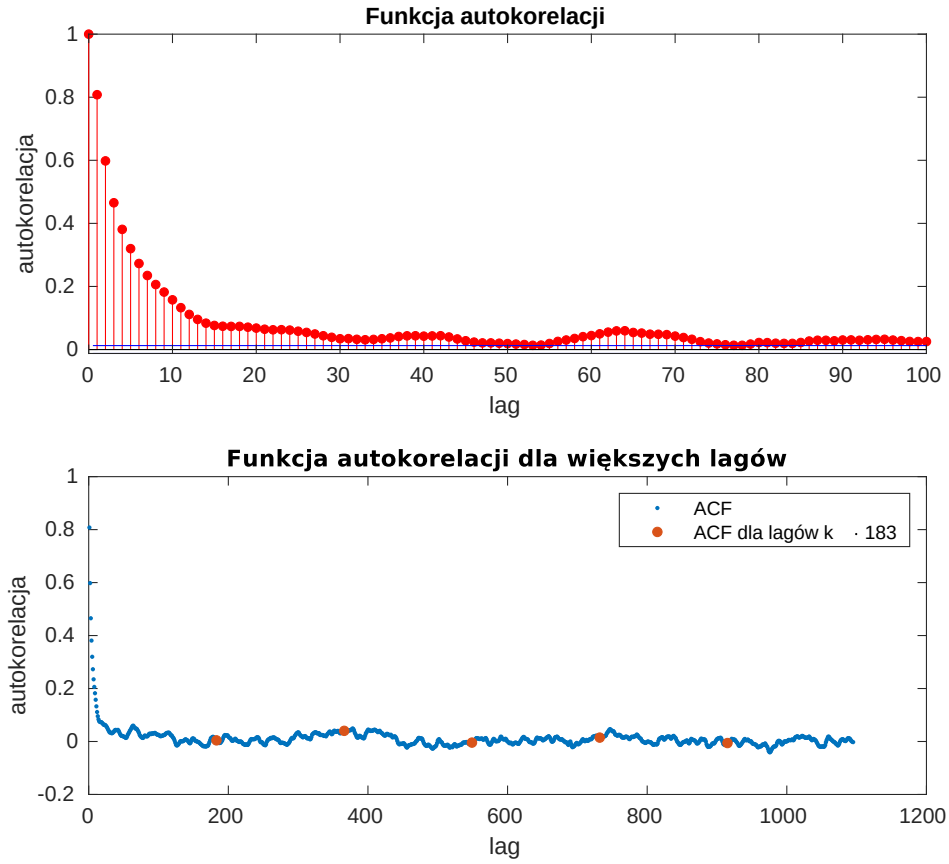


Rysunek 10: ACF dla danych przed dokonaniem dekompozycji Wolda. Poziome niebieskie linie wyznaczają przedział ufności na poziomie istotności 0.05. Drugi wykres pokazuje jak silna jest zależność pomiędzy temperaturami w dniach oddalonych czaswo o pełne lata.

Zgodnie ze wstępnymi wnioskami dotyczącymi okresowości temperatury w poszczególnych latach są od siebie zależne co bardzo dobrze obrazuje wykres na Rysunku 10. Dla lagów będących w przybliżeniu wielokrotnościami półrocza, na wykresie $k = \{1, 2, 3, 4, 5\}$, wartość funkcji autokorelacji to około 0.75. Można jednak zauważyć, że zależność ta jest silniejsza dla lat mniej oddalonych i monotonicznie maleje. Dla lagów wynoszących około 50 lat ACF przyjmuje wartości mniejsze od 0.3 (Rysunek11). Na wykresach 10 i 12 przedstawiono wartości ACF dla lagów od 0 do 100. Widzimy, że po dekompozycji zdecydowanie zmniejszyły się wartości funkcji autokorelacji. Zaniknęła również bardzo widoczna wcześniej okresowość. Wartości ACF stopiowo maleją, jednak nie osiągają wartości 0. Wskazuje to na fakt, że najprawdopodobniej nie mamy do czynienia z szeregiem czasowym MA. Widoczna jest też pewna cykliczność spadków i wzrostów wartości funkcji.



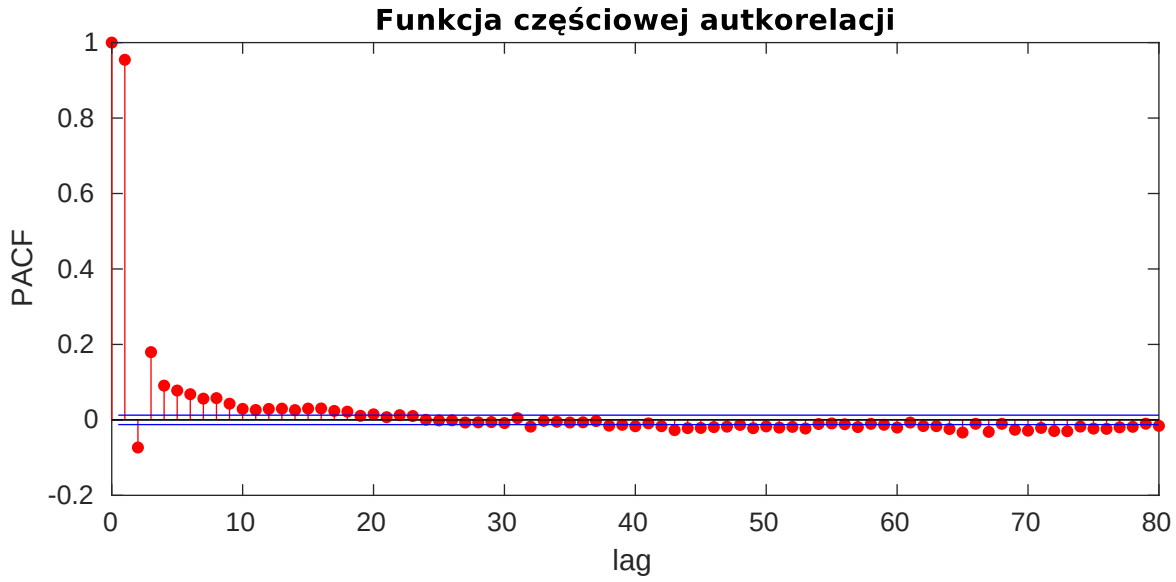
Rysunek 11: Monotonicznie malejące zależności pomiędzy temperaturami dla rosnących różnic w czasie.



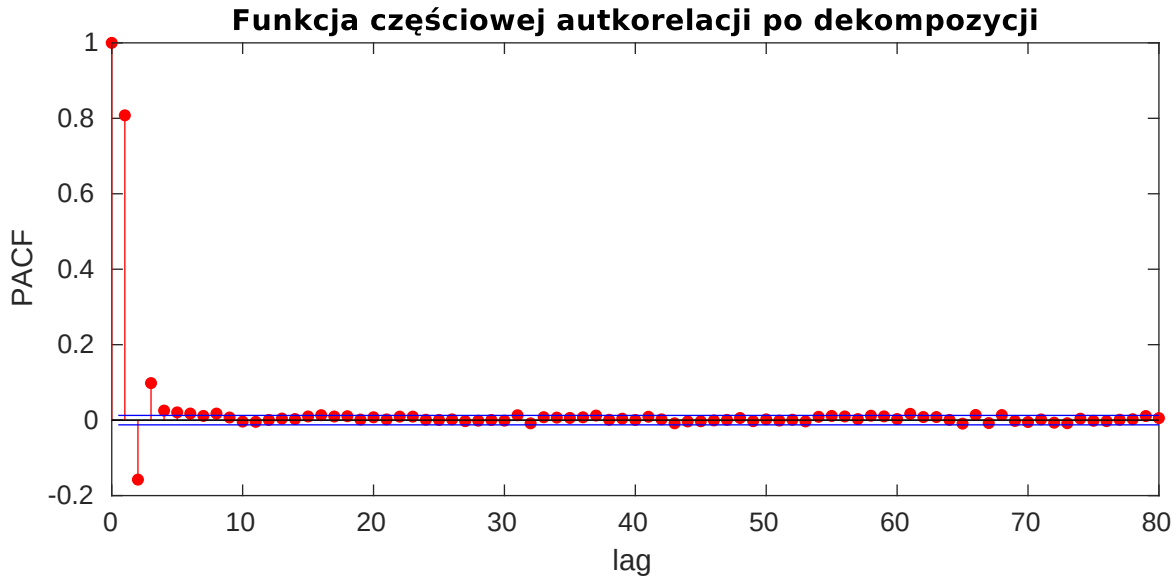
Rysunek 12: ACF dla danych po dokonaniu dekompozycji Wolda.

2.6 PACF

Funkcja częściowej autokorelacji zdefiniowana jak w 2 zwraca wartość autokorelacji między dwoma obserwacjami, jednak bez uwzględniania zależności między obserwacjami znajdującymi się pomiędzy nimi. Na przykład wartość PACF dla lagu k to wartość autokorelacji między obserwacjami z_t, z_{t+k} bez uwzględniania wpływu lagów rzędu od 1 do $k-1$.



Rysunek 13: PACF dla danych przed dokonaniem dekompozycji Wolda. Poziome niebieskie linie wyznaczają przedział ufności na poziomie istotności 0.05.



Rysunek 14: PACF dla danych po dokonaniu dekompozycji Wolda. Poziome niebieskie linie wyznaczają przedział ufności na poziomie istotności 0.05.

Na wykresach 13 i 14 przedstawiono wartości PACF dla lagów od 0 do 100. Funkcja częściowej autokorelacji przyjmuje dla danych przed dekompozycją Wolda wartości istotnie większe od zera nawet dla znacznych lagów. Dla danych po dekompozycji od lagu p = 6 PACF przyjmuje wartości nieistotnie większe od zera. Może to wskazywać na fakt, że mamy do czynienia z modelem AR(6).

3 Modelowanie przy pomocy ARMA

Do znalezienia odpowiedniego rzędu modelu ARMA(p, q) wykorzystamy narzędzie ITSM. Po analizie ACF oraz PACF spodziewamy się, że $0 \leqslant p, q \leqslant 6$. Wykonując symulacje sprawdzamy wszystkie kombinacje wartości p i q z tych przedziałów. Po zastosowaniu wbudowanej metody największej wiarygodności otrzymaliśmy, że najlepiej dopasowanym do naszych danych

modelem jest ARMA(2, 6) postaci:

$$X_t = 1.792X_{t-1} - 0.7972X_{t-2} + Z_t - 0.8470Z_{t-1} - 0.2459Z_{t-2} + 0.05077Z_{t-3} + 0.04668Z_{t-4} + 0.01548Z_{t-5} + 0.01308Z_{t-6}. \quad (2)$$

Parametry modelu są estymowane przy założeniu znanego rozkładu szumu. Ze względu na ograniczony dostęp do dokumentacji metod stosowanych przez ITSM sprawdzimy podczas analizy residuów, czy założony w nim rozkład jest normalny czy t-Studenta. Wybór modelu jest ściśle związany z wartościami kryteriów informacyjnych. W naszej pracy będą rozważane dwa z nich:

- **AIC**, czyli **kryterium informacyjne Akaikego.** Wartość AIC wyznacza się ze wzoru:

  $$AIC = -2\sum_j \ln(\hat{\pi}_j) + 2q,$$

  gdzie $\hat{\pi}_j$ to estymowane prawdopodobieństwo (przy założeniach danego modelu) uzyskania takiej właśnie wartości obserwacji j jaka była naprawdę uzyskana, a q to liczba parametrów modelu.
  W narzędziu ITSM wykorzystywana jest poprawiona wersja AIC, czyli **AICc.** Zdefiniowana jest ona wzorem:

  $$AICc = AIC + \frac{2k^2 + 2k}{n - k - 1},$$

  gdzie n to liczba obserwacji, a k to liczba parametrów modelu.
  Zarówno w przypadku AIC, jak i AICc im mniejsza wartość kryterium informacyjnego, tym lepsze dopasowanie modelu do danych.

- **BIC**, czyli **Bayesowskie kryterium informacyjne Schwartza.** Wartość BIC wyznacza się ze wzoru:

  $$BIC = k\ln(n) - 2\ln(\hat{L}),$$

  gdzie $\hat{L}$ to zmaksymalizowana funkcja wiarygodności, n to liczba obserwacji, a k to liczba parametrów modelu.
  Podobnie jak w przypadku AIC oraz AICc preferowany jest model o najniższej wartości BIC.

w przypadku dobranego modelu ARMA(2, 6) wartości kryteriów informacyjnych to:

- **AICC**: 232570,
- **BIC** : 232622.

Rezultaty podane przez ITSM przedstawiono na Rysunku 15.

```
========================================
ITSM::(Maximum likelihood estimates)
========================================

Method: Maximum Likelihood


ARMA Model:
X(t) = 1.792 X(t-1) - .7972 X(t-2)
     + Z(t) - .8470 Z(t-1) - .2459 Z(t-2) + .05077 Z(t-3)
     + .04668 Z(t-4) + .01548 Z(t-5) + .01308 Z(t-6)

WN Variance = .521783E+03

AR Coefficients
      1.791988       -.797221

Standard Error of AR Coefficients
       .040378        .036941

MA Coefficients
      -.847041       -.245869        .050770        .046676
       .015477        .013083

Standard Error of MA Coefficients
       .040895        .008383        .011911        .009760
       .008831        .008159

(Residual SS)/N = .521783E+03

AICC = .232570E+06
BIC  = .232622E+06

-2Log(Likelihood) = .232552E+06

Accuracy parameter = .100000E-08

Number of iterations = 1

Number of function evaluations = 57727

Uncertain minimum.
```

Rysunek 15: Model ARMA dopasowany do danych przez ITSM metodą największej wiarogodności.

Wybór takiego modelu może wydawać się zaskakujący. Zgodnie z Tabelą 1, stopniowo malejąca funkcja autokorelacji oraz brak wystąpienia wartości istotnie większych od zera dla lagów większych od 6 w przypadku funkcji cząstkowej autokorelacji wskazywałyby na model AR(6) [4]. Może to wynikać z faktu, że zależności w naszym modelu ARMA(2, 6) mogą być dobrze przybliżane szeregiem czasowym AR(6).

	$AR(p)$	$MA(q)$	$ARMA(p,q)$
ACF	maleje	istotnie większa od zera tylko dla lagów $< q$.	maleje
PACF	istotnie większa od zera tylko dla lagów $< p$.	maleje	maleje

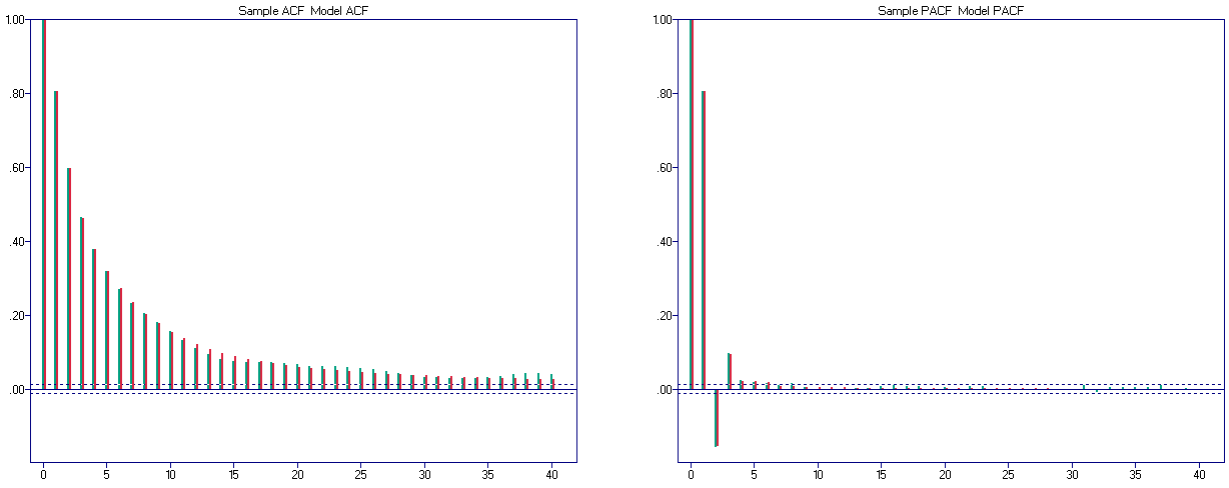
Tabela 1: Tabela wskazująca interpretacje różnych kształtów funkcji ACF i PACF.

Sprawdźmy jeszcze jakie wartości przyjmują kryteria informacyjne dla modelu AR(6):

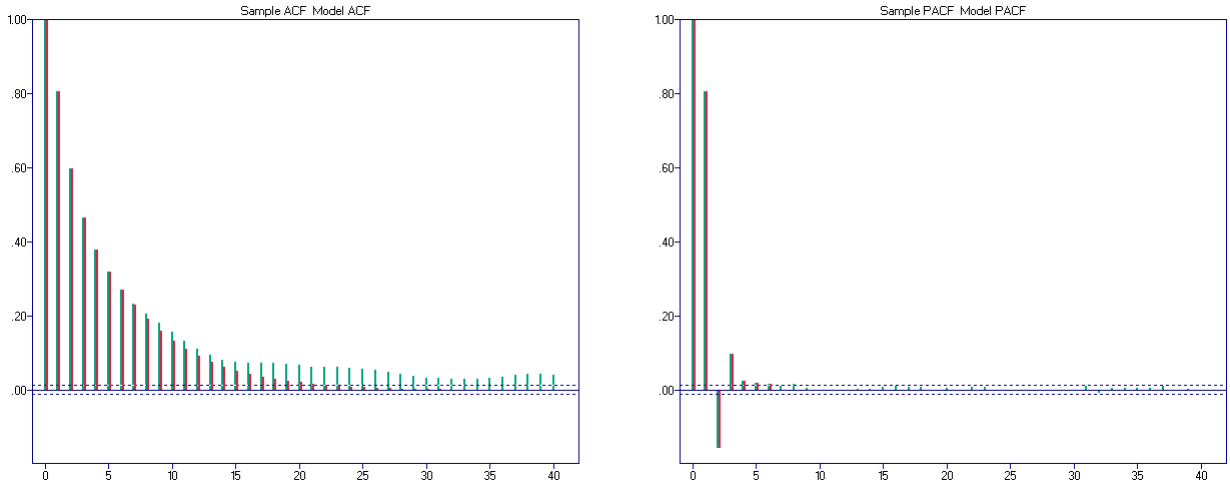
- **AICC**: 232589,
- **BIC** : 232634.

Możemy zaobserwować, że obie te wartości są większe od odpowiadających kryteriów informacyjnych dla modelu ARMA (2, 6), czyli model jest rzeczywiście gorzej dopasowany do danych.

W celu ostatecznego rozstrzygnięcia, że model ARMA(2, 6) jest lepiej dopasowany do danych niż model AR(6), wykonamy porównanie funkcji ACF i PACF naszych danych z odpowiadającymi funkcjami rozważanych modeli. Do graficznego porównania użyte zostaną wykresy generowane przez ITSM.



Rysunek 16: Porównanie ACF i PACF danych (kolor zielony) z ACF i PACF modelu ARMA(2, 6) (kolor czerwony).



Rysunek 17: Porównanie ACF i PACF danych (kolor zielony) z ACF i PACF modelu AR(6) (kolor czerwony).

Jak możemy zaobserwować na wykresach ACF i PACF modelu ARMA(2, 6) (patrz 16) niemal dokładnie pokrywają się z odpowiadającymi funkcjami wyznaczoymi dla danych. W przypadku modelu AR(6) (patrz 17) wartości PACF dla danych i modelu są bardzo zbliżone, natomiast w przypadku funkcji ACF od lagu piętnastego zaczynają występować wyraźne rozbieżności.
Rozważania te pokazują, że choć analiza wykresów funkcji ACF oraz PACF może być bardzo pomocna w wyznaczaniu odpowiedniego modelu ARMA, to nie może ona stanowić jedynego kryterium wyboru.

4 Weryfikacja poprawności modelu

4.1 Analiza residuów

Przypomnijmy, że dobraliśmy do danych model ARMA(2, 6), który wyraża się wzorem:

$$X_t = 1.792X_{t-1} - 0.7972X_{t-2} + Z_t - 0.8470Z_{t-1} - 0.2459Z_{t-2} + 0.05077Z_{t-3} + 0.04668Z_{t-4} + 0.01548Z_{t-5} + 0.01308Z_{t-6},$$

gdzie Z_t to nieskorelowane zmienne losowe o średniej 0 i wariancji $\sigma^2 = 521.783$. W celu weryfikacji poprawności naszego modelu zweryfikujemy, czy residua są nieskorelowanymi zmiennymi losowymi o średniej $\mu = 0$ i wariancji $\sigma^2 = 521.783$. Wykonamy również serię testów mających na celu sprawdzenie, czy pochodzą one z rozkładu normalnego lub t-Studenta - parametry modelu ARMA były bowiem estymowane metodą największej wiarygodności.

4.1.1 Wartości średniej oraz wariancji

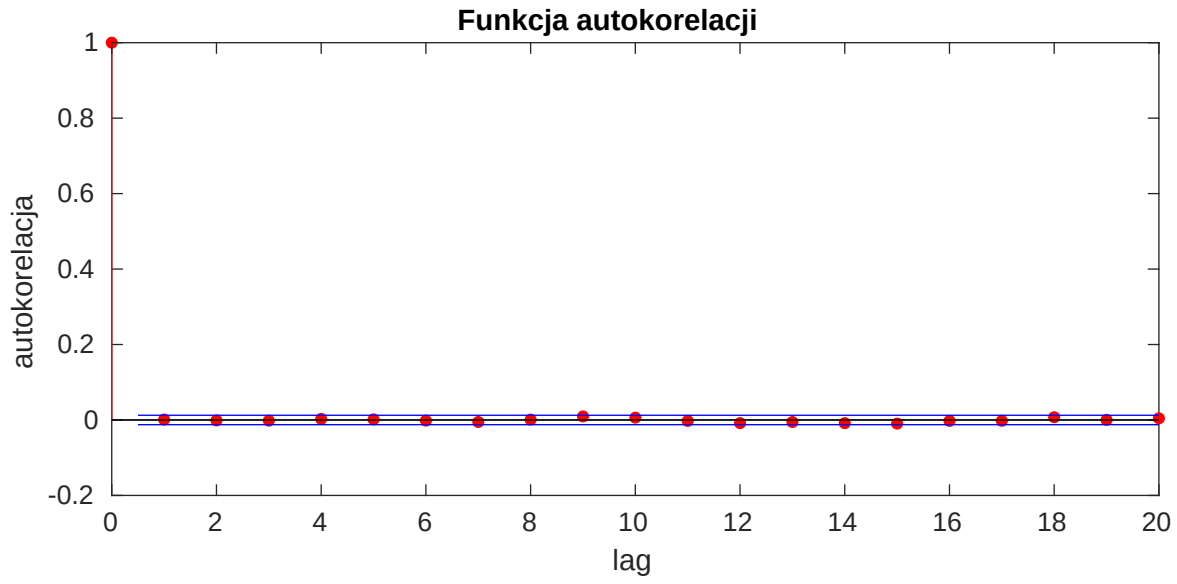
W wyniku symulacji otrzymaliśmy następujące wartości estymatorów paramterów:

- $\hat{\mu} = -0.00037669$,
- $\hat{\sigma}^2 = 521.8029$.

Empiryczna wartość μ jest bliska zeru, stąd można wnioskować, że warunek o średniej równej 0 jest spełniony. Wartość $\hat{\sigma}^2$ jest bardzo zbliżona do wartości wariancji Z_t, wynoszącej 521.783.

4.1.2 Funkcja autokorelacji

Aby zweryfikować, czy residua są nieskorelowane wyznaczyliśmy ACF. Odpowiedni wykres przedstawiono na Rysunku 18. Jak widzimy, funkcja autokorelacji przyjmuje wartość 1 dla lagu równego 0, zaś w pozostałych przypadkach osiąga wartości nieistotnie różne od 0. Wnioskujemy, że zgodnie z założeniami **residua są nieskorelowane**.



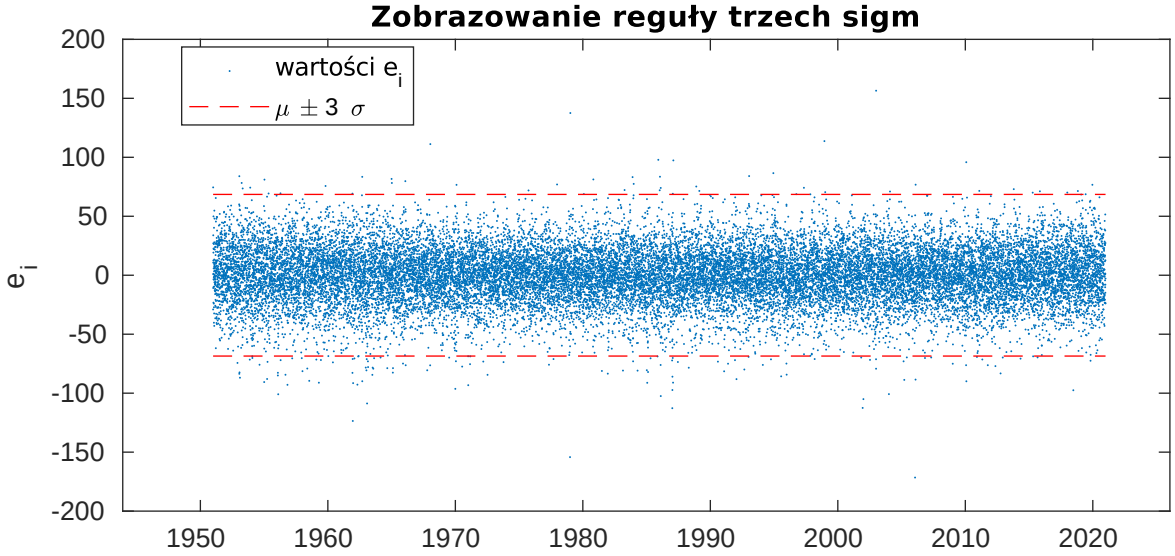
Rysunek 18: Funkcja autokorelacji wyznaczona dla residuów. Liniami poziomymi oznaczony jest przedział ufności na poziomie istotności $\alpha = 0.05$.

4.1.3 Normalność residuów

Na koniec sprawdzimy, czy residua pochodzą z rozkładu normalnego o parametrach μ i σ. W tym celu skorzystamy z wykresu obrazującego regułę trzech sigm, porównania dystrybuanty teoretycznej z dystrybuantą empiryczną, porównania gęstości teoretycznej z jądrowym estymatorem gęstości, analizy wykresu kwantylowego oraz kilku testów statystycznych.

- **Zobrazowanie reguły trzech sigm**

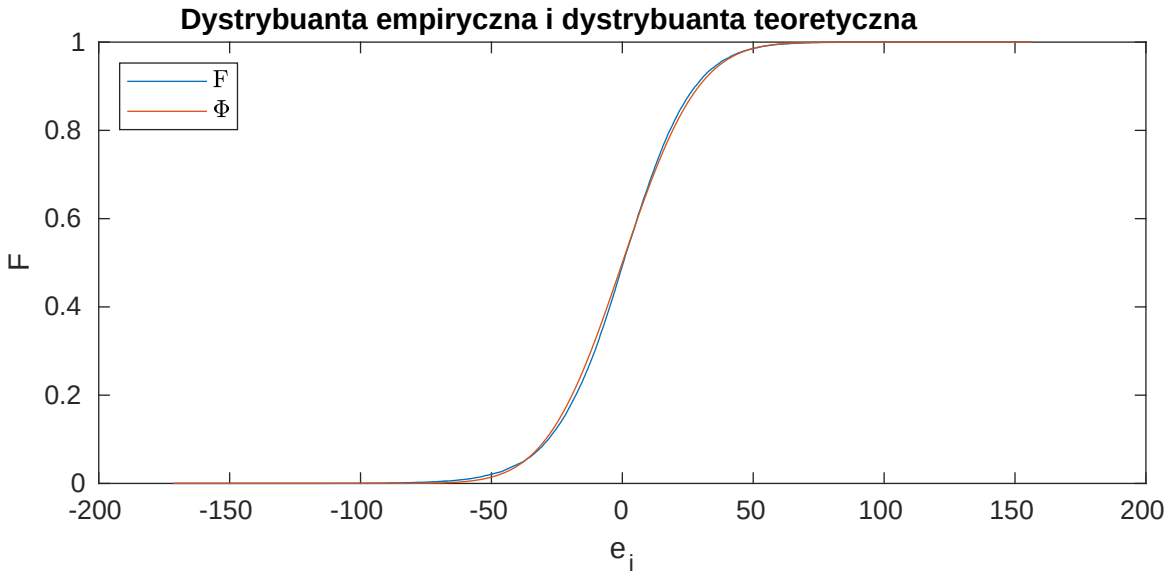
  Wykres 19 przedstawia graficznie regułę trzech sigm. Parametry użyte do konstrukcji przedziału są parametrami teoretycznymi, tj. $\mu = 0,\ \sigma = \sqrt{521.783}$. Do takiego przedziału wpada 99.3% residuów. Jest to wynik bliski teoretycznemu, który wynosi 99.7% co może wskazywać na duże podobieństwo do rozkładu normalnego. Jednocześnie taka różnica może wskazywać też na rozkład o cięższych ogonach.



Rysunek 19:

- **Porównanie dystrybuant**

  Jak możemy zauważyć na wykresie przedstawionym na Rysunku 20 dystrybuanta empiryczna niemal dokładnie pokrywa się z dystrybuantą rozkładu normalnego. Może to wskazywać na fakt, że residua pochodzą z tego rozkładu.

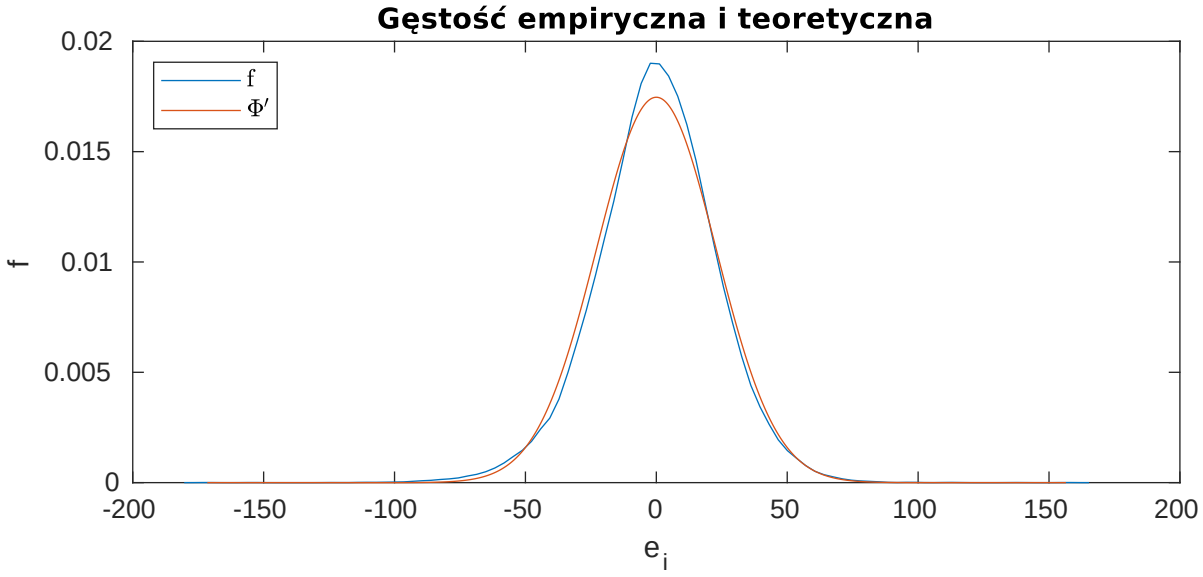


Rysunek 20: Porównanie dystrybuanty empirycznej z dystrybuantą teoretyczną.

- **Porównanie gęstości teoretycznej z jądrowym estymatorem gęstości**

  Na wykresie przedstawionym na Rysunku 21 przedstawiono porównanie teoretycznej gęstości rozkładu normalnego z jądrowym estymatorem gęstości. Choć funkcje te są do siebie zbliżone kształtem, jądrowy estymator gęstości przyjmuje

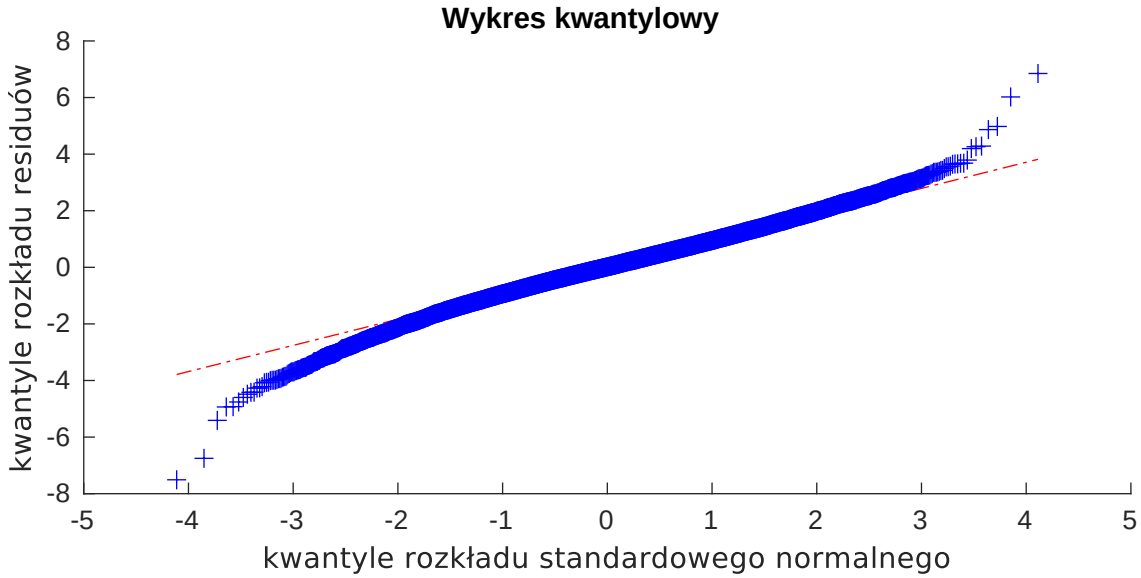
wyraźnie większe wartości dla argumentów bliskich zera. Może to wskazywać na fakt, że rozkład residuów ma cięższe ogony od rozkładu normalnego.



Rysunek 21: Porównanie gęstości teoretycznej z jądrowym estymatorem gęstości.

- **Wykres kwantylowy**

  Analizując wykres kwantylowy 22 residuów ponownie możemy wysnuć wniosek, że rozkład residuów ma cięższe ogony niż rozkład normalny. Wskazują na to wyraźne odchylenia od linii prostej na obu końcach. Ponownie budzi to wątpliwości dotyczące normalności rozkładu residuów.



Rysunek 22: Wykres kwantylowy dla residuów.

- **Testy statystyczne** W celu ostatecznej weryfikacji hipotezy dotyczącej normalności residuów, wykonanych zostało kilka testów statystycznych na poziomie istotności $\alpha = 0.05$ opisanych poniżej.
  - **Test Kołmogorowa-Smirnowa** - test nieparametryczny, który opiera się na odległości supremum pomiędzy dystrybu-antą empiryczną a dystrybuantą rozkładu referencyjnego (w naszym przypadku dystrybuantą rozkładu normalnego).

– **Test Jarque-Bera** - test weryfikujący, czy dane z próby mają skośność i kurtozę pasujące do rozkładu normalnego.
– **Test Andersona-Darlinga**- test jest oparty o ważoną odległość Cramera von Misesa pomiędzy dystrybuantami empiryczną i teoretyczną z wagami odpowiadającymi odwrotności wariancji dystrybuanty empirycznej.

Wyniki wykonania powyższych testów w Matlabie zaprezentowano w poniższej tabeli (Tabela 2).

Tabela 2: Wyniki testów statystycznych wykonanych dla residuów

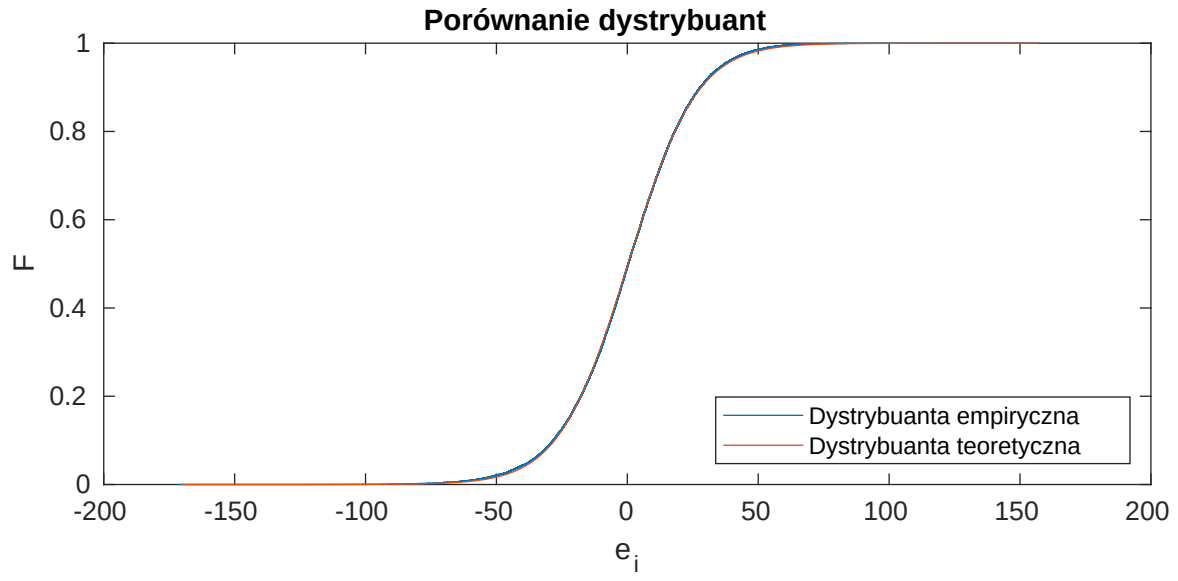
test	**wartość funkcji testowej**	**p-wartość**
test Kołmogorowa-Smirnowa	1	$9.0765 \cdot 10^{-13}$
test Jarque-Bera	1	< 0.001
test Andersona-Darlinga	1	< 0.0005

Każdy z powyższych testów odrzucił hipotezę zerową czyli hipotezę, że dane pochodzą z rozkładu normalnego. P-wartość w każdym z trzech przypadków jest mniejsza od ustalonego wcześniej poziomu istotności $\alpha = 0.05$. Wszystko to świadczy o tym, że dane mogą nie pochodzić z rozkładu normalnego.
Ostatecznie sprawdziliśmy, że residua spełniają warunki białego szumu: mają średnią bardzo zbliżoną do 0, skończoną wariancję oraz są nieskorelowane. Nie pochodzą one jednak z rozkładu normalnego, ale rozkładu o cięższych ogonach.

4.1.4 Pochodzenie residuów z rozkładu t-Studenta

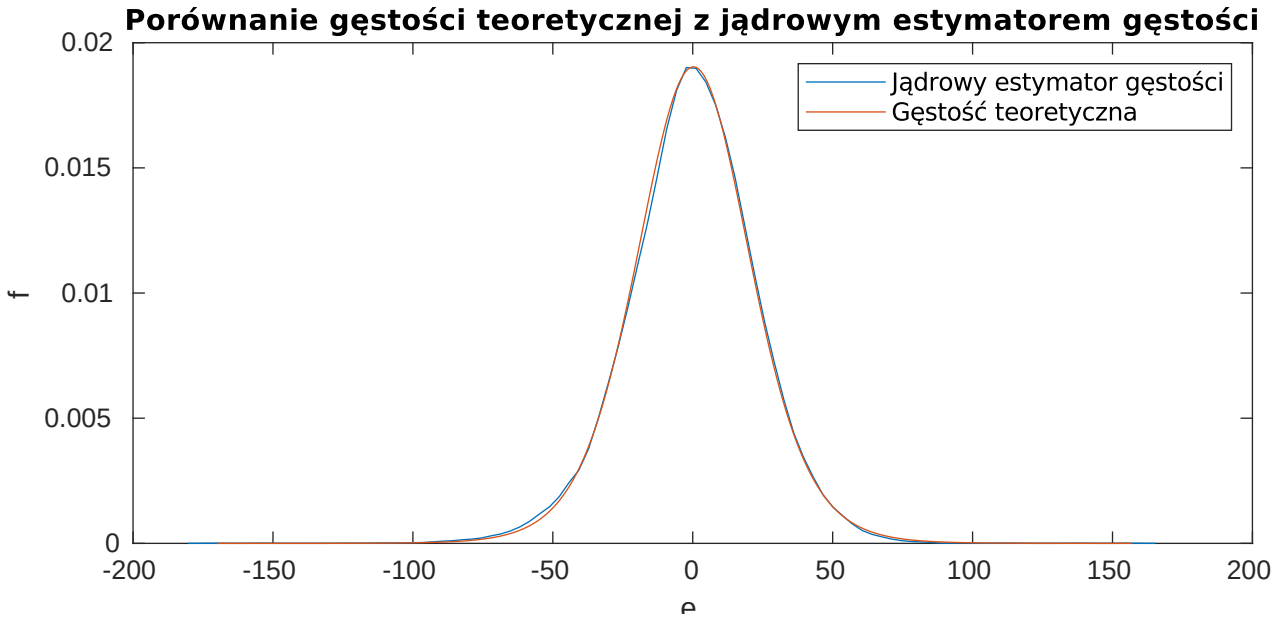
Obserwacje dotyczące ogonów rozkładu szumu, które bardzo dobrze obrazuje wykres 22, wskazują na rozkład t-Studenta. Aby potwierdzić tę hipotezę przeprowadzimy podobną analizę jak dla rozkładu normalnego.

- **Porównanie dystrybuant**
  Jak możemy Rysunku 23 dystrybuanta empiryczna niemal dokładnie pokrywa się z dystrybuantą przeskalowanego rozkładu t-Studenta. Może to świadczyć o tym, że residua rzeczywiście pochodzą z tego rozkładu.



Rysunek 23: Porównanie dystrybuanty empirycznej residuów z dystrybuantą przeskalowanego rozkładu t-Studenta.

- **Porównanie gęstości**
  Na Rysunku 24 przedstawiono porównanie jądrowego estymatora gęstości residuów z teoretyczną gęstością przeskalowanego rozkładu t-Studenta. W przeciwieństwie do rozważanego wcześniej rozkładu normalnego, tym razem funkcje pokrywają się niemal dokładnie. Świadczy to o tym, że najprawdopodobniej residua nie pochodzą z rozkładu normalnego, a z rozkładu t-Studenta.



Rysunek 24: Porównanie jądrowego estymatora gęstości residuów z teoretyczną gęstością przeskalowanego rozkładu t-Studenta.

- **Test Kołmogorowa-Smirnowa**
  Dla potwierdzenia wcześniejszych obserwacji wykonaliśmy test Kolmogorowa-Smirnowa (3) na poziomie istotności $\alpha = 0.05$. Zwrócił on wartość 0, co oznacza brak podstaw do odrzucenia hipotezy, że residua pochodzą z rozkładu t-Studenta. Możemy więc przyjąć, że **residua pochodzą z rozkładu t-Studenta.**

	wartość funkcji testowej	**p-wartość**
test Kołmogorowa-Smirnowa	0	0.2861

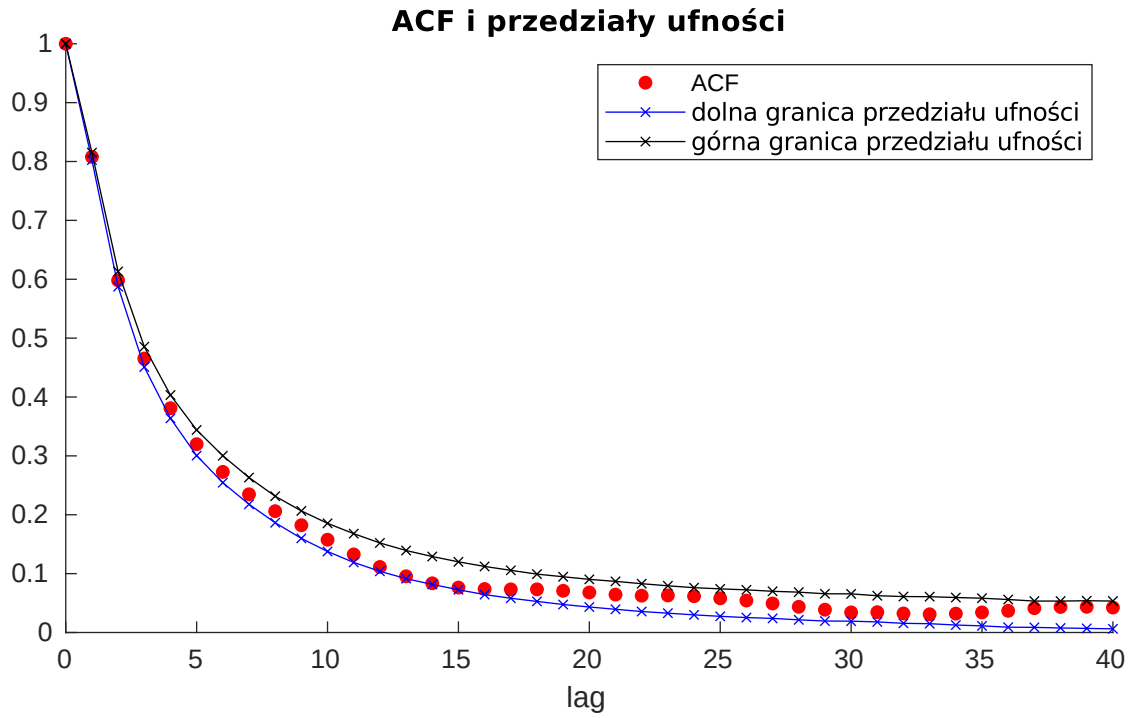
Tabela 3: Wyniki testu Kołmogorowa-Smirnowa.

Zarówno ten wniosek jak i fakt, że `MATLAB` sugeruje bliźniaczo podobne wyniki dla szumu o rozkładzie t-Studenta, skłania nas ku założeniu, że ITSM estymuje parametry właśnie przy takim założeniu.

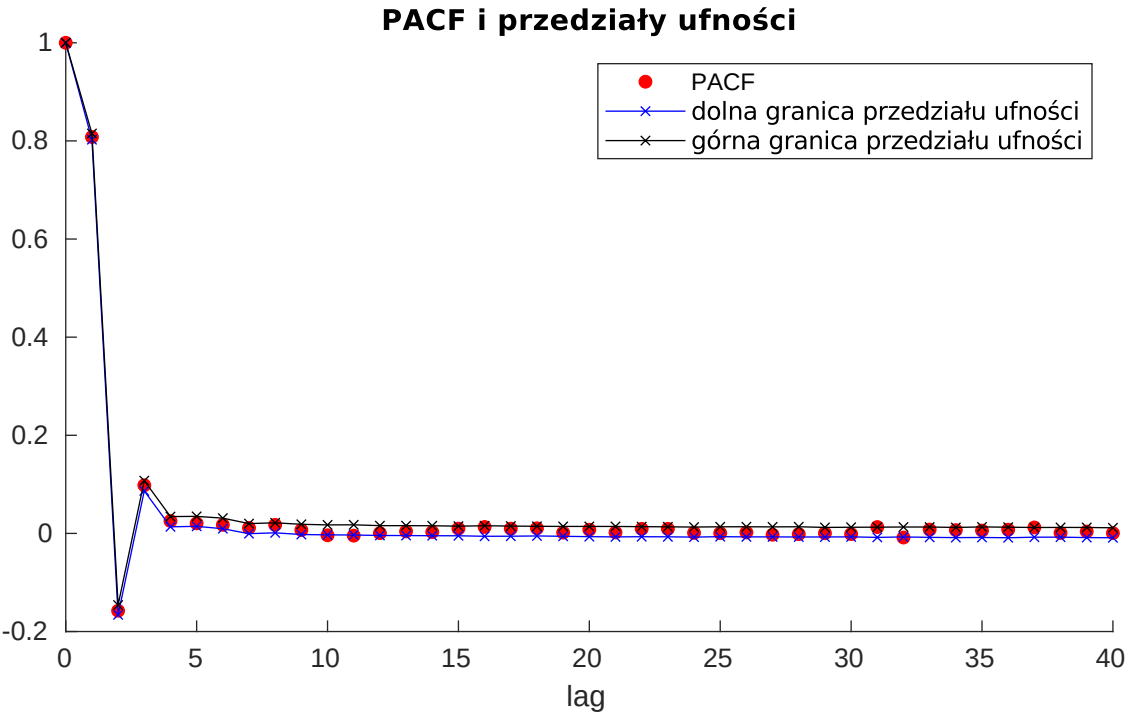
4.2 Porównanie ACF i PACF z przedziałami ufności wyznaczonymi symulacyjnie

W celu lepszego określenia poprawności dopasowania modelu ARMA(2, 6) do średnich dniowych temperatur z lat 1951-2020 wyznaczymy symulacyjnie przedziały ufności dla ACF i PACF modelu (2) i wyrysujemy na ich tle funkcje dla danych. Należy zwrócić uwagę, że pomimo wyznaczenia rozkładu t-Studenta jako rozkładu szumu podejmujemy poniżej konsekwentnie próbę porównania tego założenia z założeniem o normalności.

4.2.1 Przedziały ufności symulowane dla szumu z rozkładu normalnego

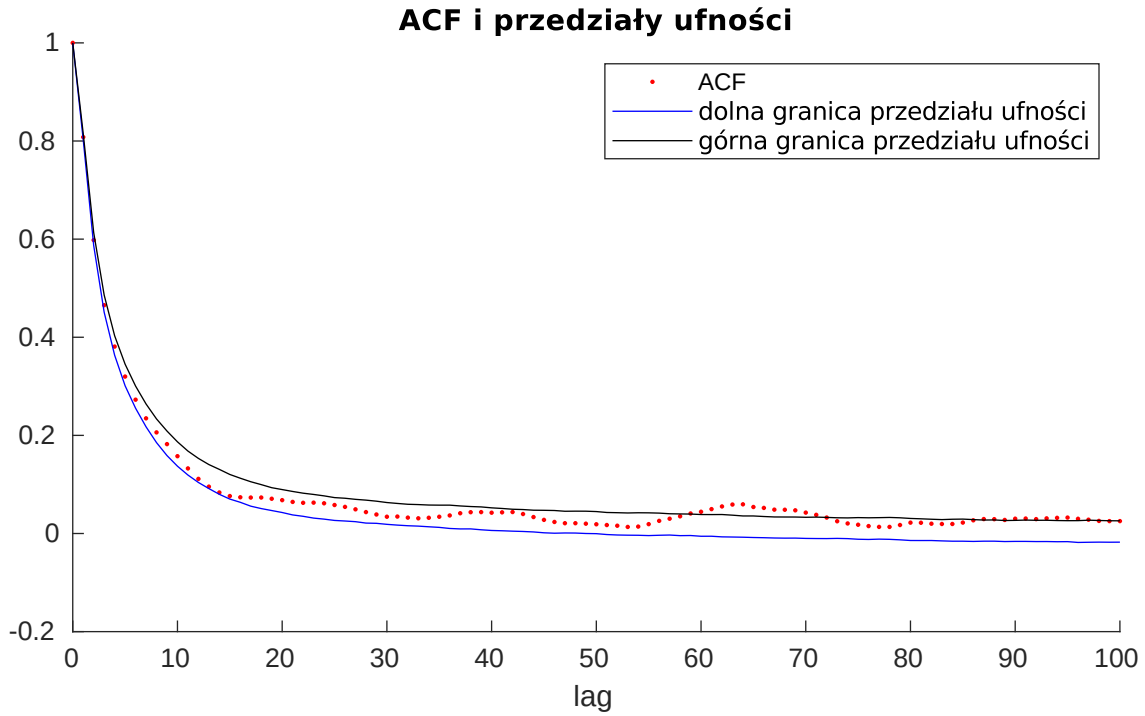


Rysunek 25: Zobraznowanie ACF w symulacyjnym przedziale ufności dla poziomu istotności $\alpha = 0.05$.

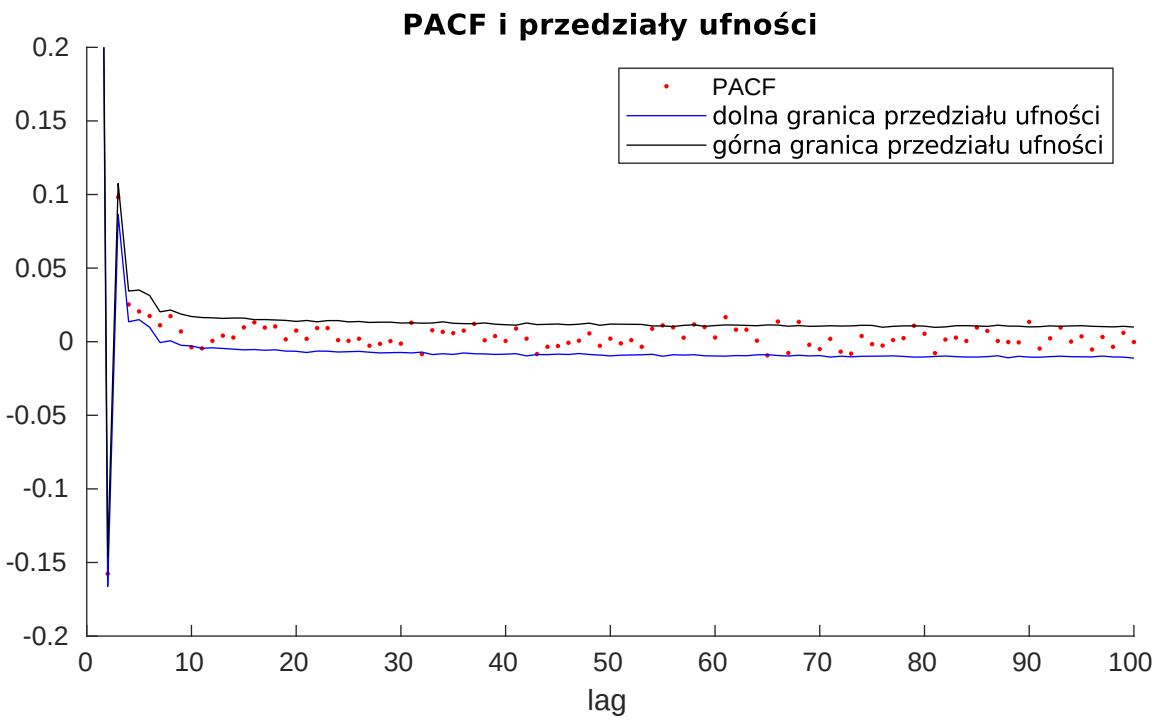


Rysunek 26: Zobraznowanie PACF w symulacyjnym przedziale ufności dla poziomu istotności $\alpha = 0.05$.

Wykres 25 jest oparty o przedziały ufności wyznaczone na podstawie empirycznych kwantyli próbki tysiąca trajektorii funkcji autokorelacji wyliczonych na podstawie tysiąca trajektorii modelu (2) o długości odpowiadającej długości oryginalnej próbki. Górna granica przedziału ufności jest zatem wyznaczona przez kwantyl rzędu 0.95 a dolna przez kwantyl rzędu 0.05. Analogicznie przebiegała syulacja 26 dla PACF. Na obu wykresach widoczne jest spodziewane dla modelu 2 zachowanie, tj. wartości funkcji mieszczą się w przedziałach ufności. Niepokojące może być natomiast zauważalne, naprzemienne stykanie się wartości ACF z dolną i górną granicą przedziału. Aby sprawdzić, czy to tylko miejscowa anomalia wykonamy wykresy dla lagów od 1 do 100.



Rysunek 27: Wykres ACF dla większych lagów w celu zobrazowania cyklicznych zachowań funkcji.

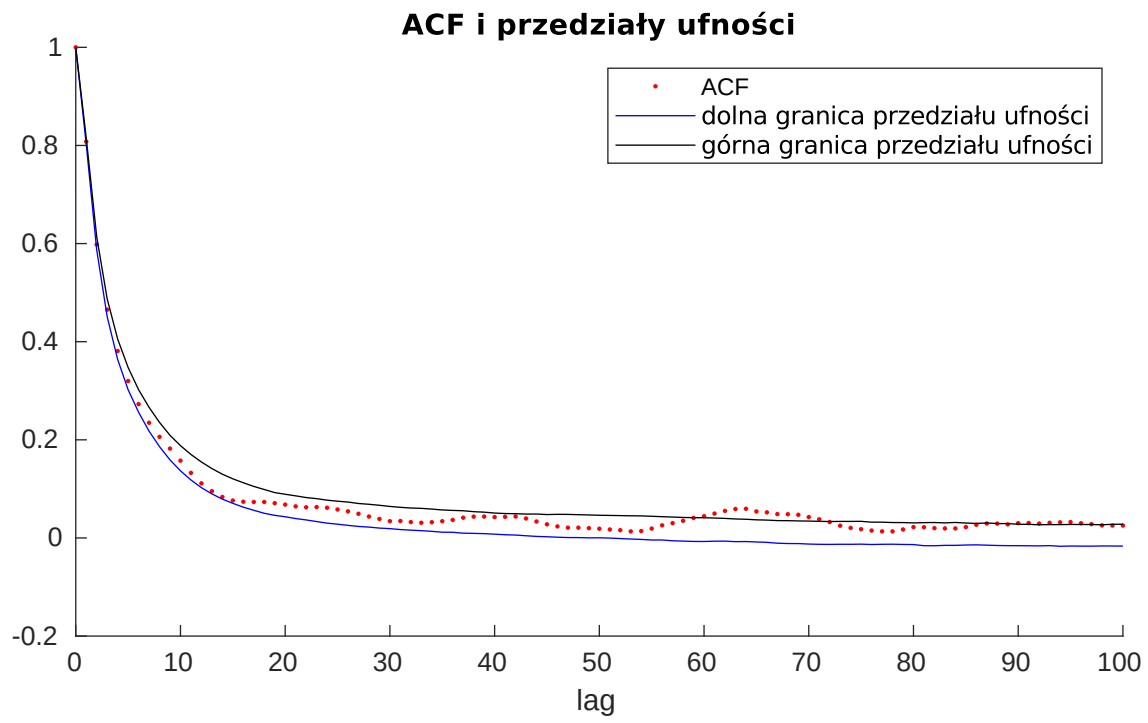


Rysunek 28: Wykres z uciętą wartością w zerze dla lepszego zobrazowania PACF na dłuższym przedziale.

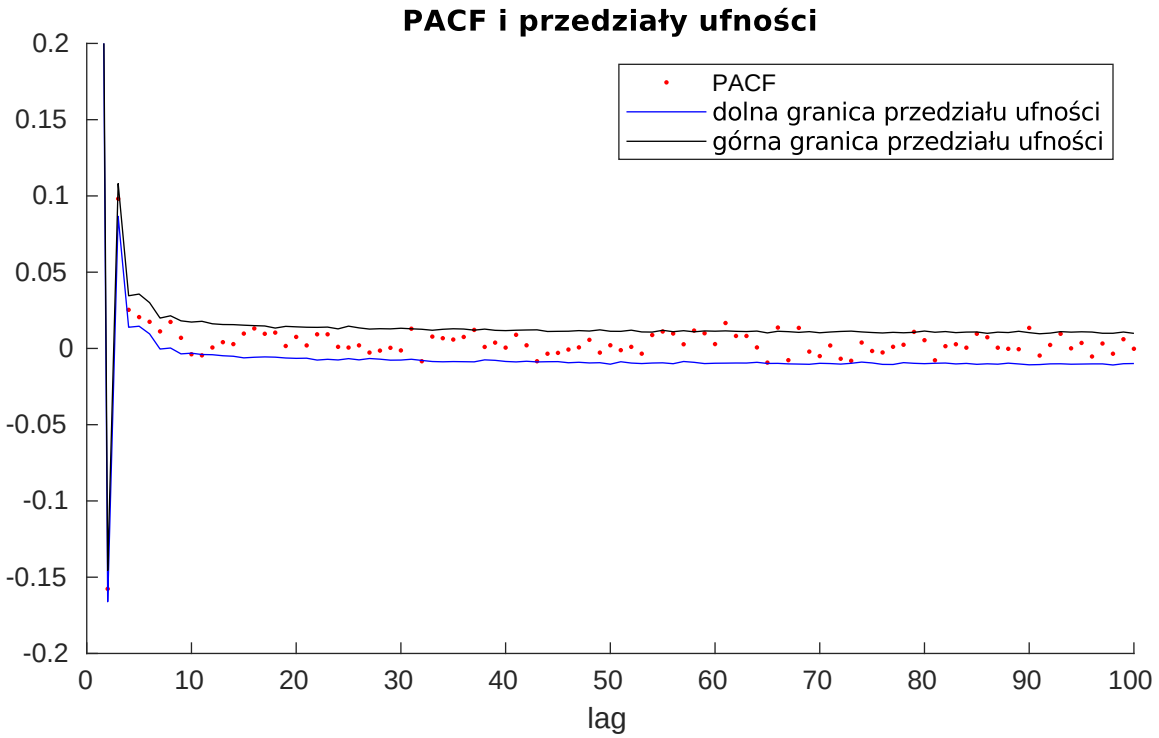
Wyniki symulacji dla większych lagów dały w przypadku PACF (wykres 28) spodziewane rezultaty. Poza przedziałem ufności znajduje się około pięć wartości funkcji co jest zgodne z teoretycznym $\alpha \cdot 100$. Wartości ACF (wykres 27) wykazują natomiast cykliczne zachowanie i powyżej lagu 60 wychodzą poza przedział ufności. Na podstawie symulacji można zatem stwierdzić, że na poziomie istotności $\alpha = 0.05$ analizowana funcja autokorelacji nie jest funkcją autokorelacji modelu ARMA(2,6) wyrażonego przez (2) z założeniem normalnści szumu.

4.2.2 Przedziały ufności symulowane dla szumu z rozkładu t-Studenta

Analogicznie przedziały ufności dla wartości ACF i PACF zostały wysymulowane dla założenia o pochodzeniu szumu z rozkładu t-Studenta. Łatwo zauważyć, że wyniki symulacji z wykresów 29 i 30 są wizualnie nierozróżnialne od tych dla rozkładu normalnego (27, 28). W tym przypadku symulacja także odrzuca na poziomie istotności $\alpha = 0.05$ hipotezę, że ACF danych to ACF modelu (2) oraz także nie odrzuca założenia, że PACF danych jest PACF modelu (2).



Rysunek 29: Wykres ACF dla większych lagów w celu zobrazowania zachowania funkcji. Przedziały ufności symulowane przy założeniu rozkładu szumu t-Studenta.



Rysunek 30: Wykres PACF dla założenia o rozkładzie szumu t-Studenta.

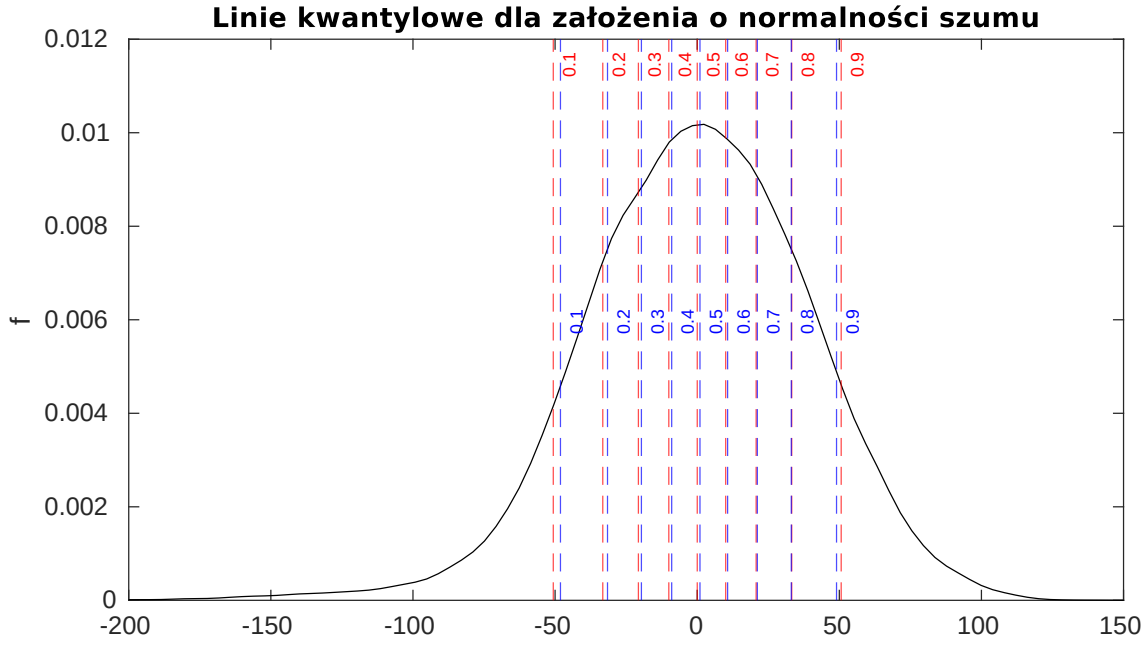
4.3 Linie kwantylowe

Kolejnym sposobem na weryfikację poprawności dobranego modelu jest analiza kwantyli. Aby tego dokonać, przeprowadziliśmy symulację Monte Carlo. Wygenerowaliśmy N = 1000 trajektorii teoretycznego modelu ARMA(2, 6) - wpierw przyjmując, że szum pochodzi z rozkładu normalnego, a następnie przyjmując, że pochodzi z przeskalowanego rozkładu t-Studenta. Dla każdej z trajektorii wyznaczyliśmy kwantyle odpowiednich rzędów. Następnie wyliczyliśmy średnią dla kwantyli danego rzędu i sprawdziliśmy, jaka część danych jest mniejsza lub równa od wyznaczonej tym sposobem wartości kwantyla.
Wyniki dla szumu z rozkładu normalnego przedstawiono w Tabeli 4. Możemy zaobserwować, że odsetek danych, które są mniejsze lub równe od wartości danego kwantyla są zbliżone do jego rzędu. Oznacza to, że model dość dobrze opisuje badane dane.

Tabela 4: Wartości kwantyli uzyskane w wyniku symulacji - założenie normalnego rozkładu szumu

rząd kwantyla	**wartość kwantyla**	**dane $\leqslant$ kwantyl**	**dane $>$ kwantyl**
0.1	-50.6289	0.0876	0.9124
0.2	-33.2373	0.1865	0.8135
0.3	-20.7125	0.2902	0.7098
0.4	-10.0054	0.3883	0.6117
0.5	-0.0033	0.4892	0.5109
0.6	9.9933	0.5922	0.4078
0.7	20.6915	0.6942	0.3058
0.8	33.2031	0.7994	0.2006
0.9	50.5680	0.9062	0.0938

Aby przedstawić, jak wyznaczone za pomocą symulacji Monte Carlo kwantyle mają się do empirycznych kwantyli naszych danych, przedstawiliśmy ich porównanie na Rysunku 31. W celu zwiększenia czytelności zdecydowaliśmy się na naniesienie linii odpowiadających danym kwantylom na wykres jądrowego estymatora gęstości naszych danych.
Możemy zaobserwować, że kwantyle wyznaczone symulacyjnie są zbliżone do kwantyli wyznaczonych dla danych. Ponownie oznacza to, że model ten dobrze estymuje badane przez nas dane.



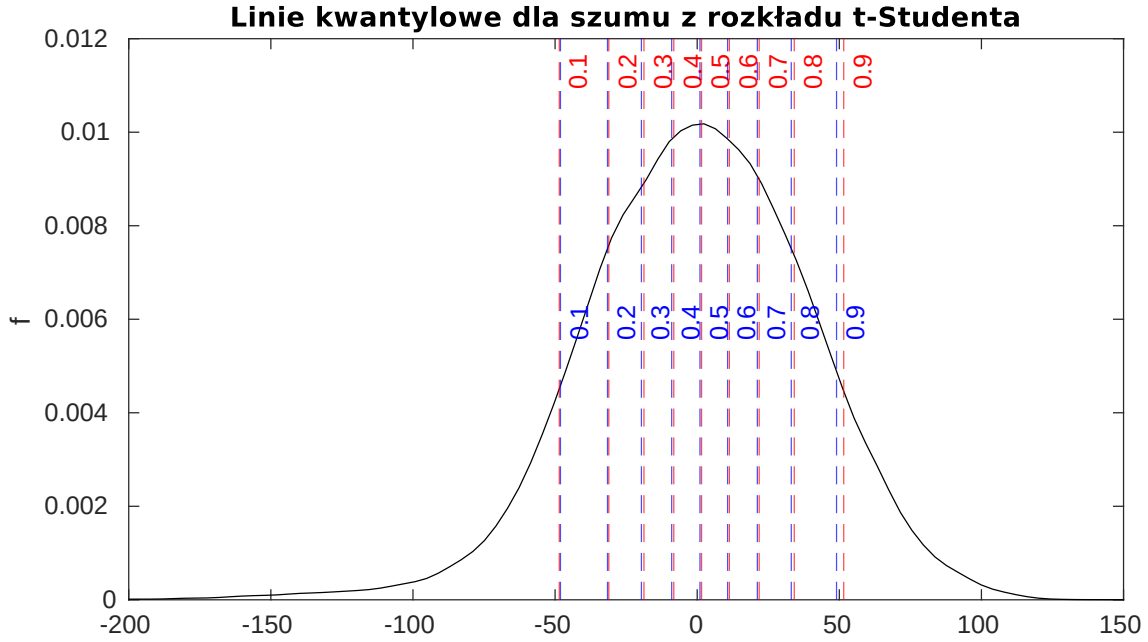
Rysunek 31: Porównanie kwantyli wyznaczonych symulacyjnie (kolor czerwony) z kwantylami wyznaczonymi dla danych (kolor niebieski).

Wyniki dla szumu z przeskalowanego rozkładu t-Studenta przedstawiono w Tabeli 5. Możemy zaobserwować, że odsetek danych, które sąmniejsze lub równe od wartości danego kwantyla ponownie są zbliżone do jego rzędu. Oznacza to, że model dobrze opisuje badane dane.

Tabela 5: Wartości kwantyli uzyskane w wyniku symulacji - założenie, że szum pochodzi z przeskalowanego rokzładu t - Studenta

rząd kwantyla	**wartość kwantyla**	**dane $\leqslant$ kwantyl**	**dane $>$ kwantyl**
0.1	-48.5458	0.0979	0.9021
0.2	-31.1293	0.2031	0.7969
0.3	-18.7736	0.3070	0.6930
0.4	-8.2834	0.4068	0.5932
0.5	1.4987	0.5052	0.4948
0.6	11.2812	0.6050	0.3950
0.7	21.7677	0.7054	0.2946
0.8	34.1373	0.8072	0.1928
0.9	51.5483	0.9114	0.0886

Dla porównania kwantyli empirycznych wyznaczonych dla danych z wyznaczonymi symulacyjnie kwantylami modelu teoretycznego przy założeniu, że szum pochodzi z przeskalowanego rozkładu t-Studenta, zdecydowaliśmy się przedstawić je na analogicznym do poprzedniego przypadku wykresie (Rysunek 32). Ponownie możemy zaobserwować, że wartości poszczególnych kwantyli są do siebie bardzo zbliżone, na ogół bardziej niż przy założeniu normalności szumu. Oznacza to, że model opierający się na szumie z rozkładu t-Studenta bardzo dobrze opsiuje badany zbiór danych.



Rysunek 32: Porównanie kwantyli wyznaczonych symulacyjnie (kolor czerwony) z kwantylami wyznaczonymi dla danych (kolor niebieski) dla założenia, że szum ma rozkład t-Studenta.

5 Podsumowanie raportu

Przed przystąpieniem do zebrania wyników raportu należy przypomnieć z jakimi błędami godziliśmy się podczas jego przygotowywania. Pierwszym, choć dla tak dużej ilości danych niewielkim, jest imputacja pomiaru z innej stacji w miejsce brakującego pomiaru temperatury (wykonana w 2.1). Kolejną niedogodnością wpływającą negatywnie na dopasowanie modeli liniowych do danych może być pozostawiona w nich cykliczność, która wynika z powtarzalnych ekstremalnych temperatur (patrz wykres 9). Takie wartości nie mogły zostać usunięte poprzez odjęcie prostej funkcji okresowej jak sinusoida stąd w dalszych badaniach należałoby wziąć pod uwagę inne metody usuwania okresowości i zastosować metody usuwania cykliczności. Same wyniki raportu wskazują, zgodnie z kryterium opartym na analizie residuów, że modelem ARMA dobrze opisującym dane jest model ARMA(2, 6) opisany wzorem (2) i wyznaczony metodą największej wiarygodności przy założeniu, że rozkładem szumu jest rozkład t-Studenta. Fakt, że empiryczne ACF na poziomie istotności 0.05 nie jest funkcją autokorelacji wybranego modelu może wynikać ze wspomnianej wcześniej cykliczności pozostawionej w danych.

Literatura

[1] W. H. Greene, *Econometric Analysis*, 5th ed. Prentice Hall, 2003.

[2] [Online]. Available: http://meteomanz.com/

[3] A. Gelman and J. Hill, *Missing-data imputation*, ser. Analytical Methods for Social Research. Cambridge University Press, 2006, p. 529–544.

[4] A. Aue, "3.3: The pacf of a causal arma process," Aug 2020. [Online]. Available: https://stats.libretexts.org/Bookshelves/Book:˙Time˙Series˙Analysis˙(Aue)/3:˙ARMA˙Processes/3.3:˙The˙PACF˙of˙a˙Causal˙ARMA˙Process?fbclid=IwAR1hMsjj1Ygf502VN3L76QPobJ7qXezSkbGB7TteVKZn-EuommQJlLVFmhw