Temperatury we Wrocławiu

Julia Wołk-Łaniewska, Szymon Stano

6 lutego 2024

1 Wstęp

W niniejszym projekcie zajmiemy się analizą danych dotyczących wysokości temperatury we Wrocławiu w zależności od czasu przez 1462 dni, tj. 4 lata. Dane zostały zaczerpnięte ze strony: https://meteostat.net/en/place/pl/wrocaw?s=12424&t=2020-01-01/2024-01-01.

Pojedynczy wiersz zawiera:

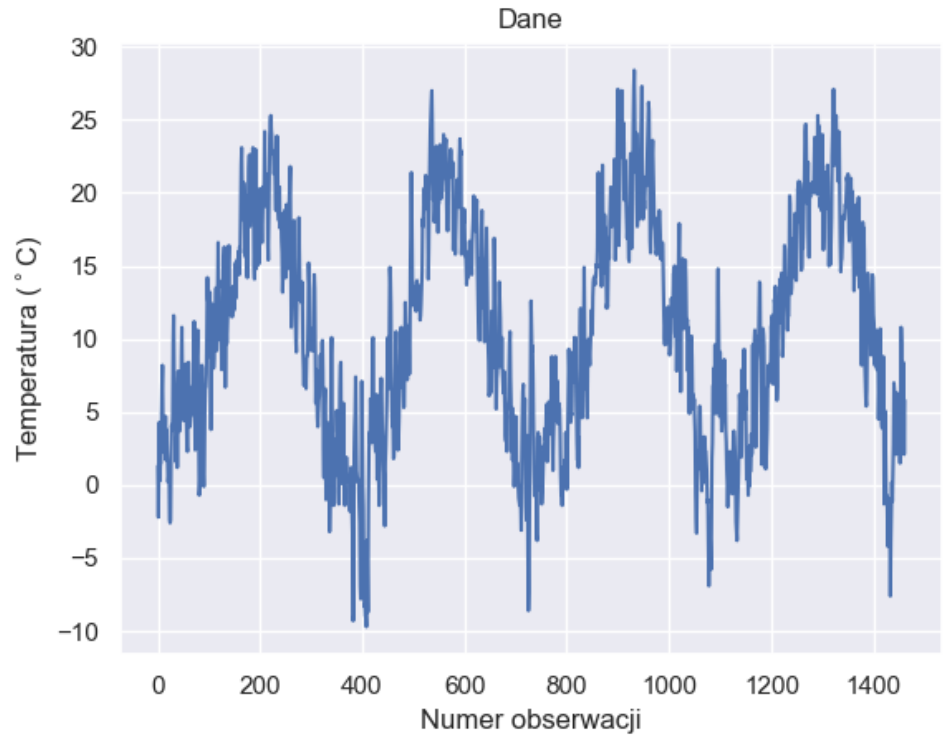
- datę,
- średnią temperaturę $[^\circ C]$,
- minimalną temperaturę $[^\circ C]$,
- maksymalną temperaturę $[^\circ C]$,
- całkowite opady [mm],
- głębokość pokrywy śnieżnej [mm],
- kierunek wiatru,
- prędkość wiatru $[\frac{km}{h}]$,
- prędkość najsilniejszego podmuchu wiatru $[\frac{km}{h}]$,
- ciśnienie atmosferyczne $[^\circ C]$,
- czas nasłonecznienia [h].

Poniżej zajmiemy się analizą wartości średniej temperatury, dopasowaniem modelu ARMA, oceną jakości dopasowania oraz analizą residuów. Wszystkie obliczenia oraz wykresy zostały wykonane w języku *Python*.

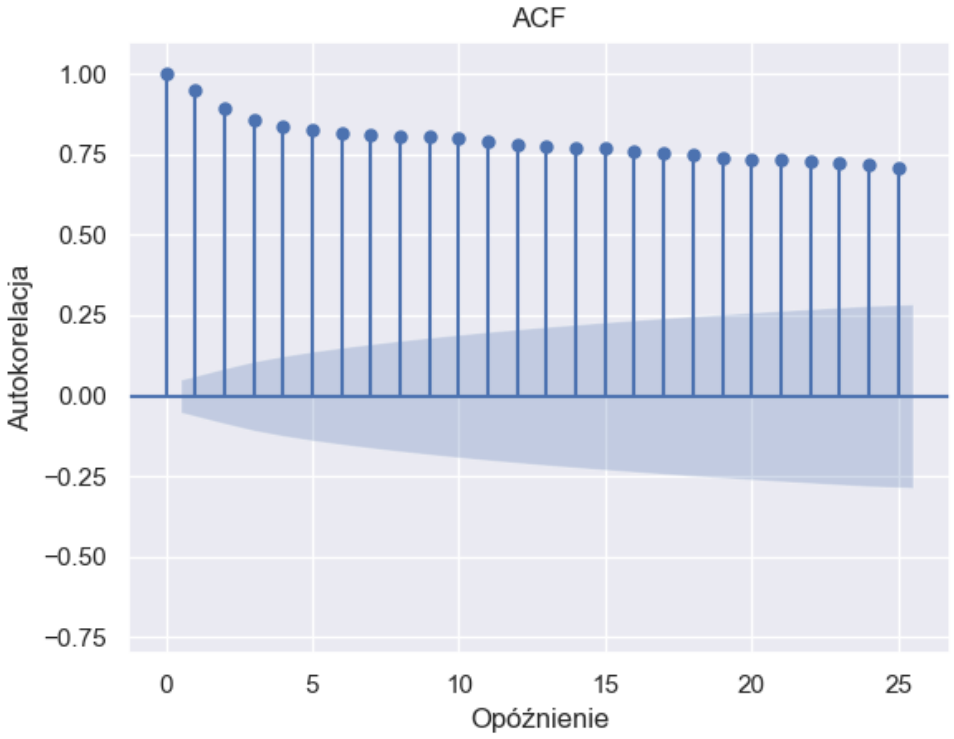
2 Przygotowanie danych do analizy

2.1 Dane surowe

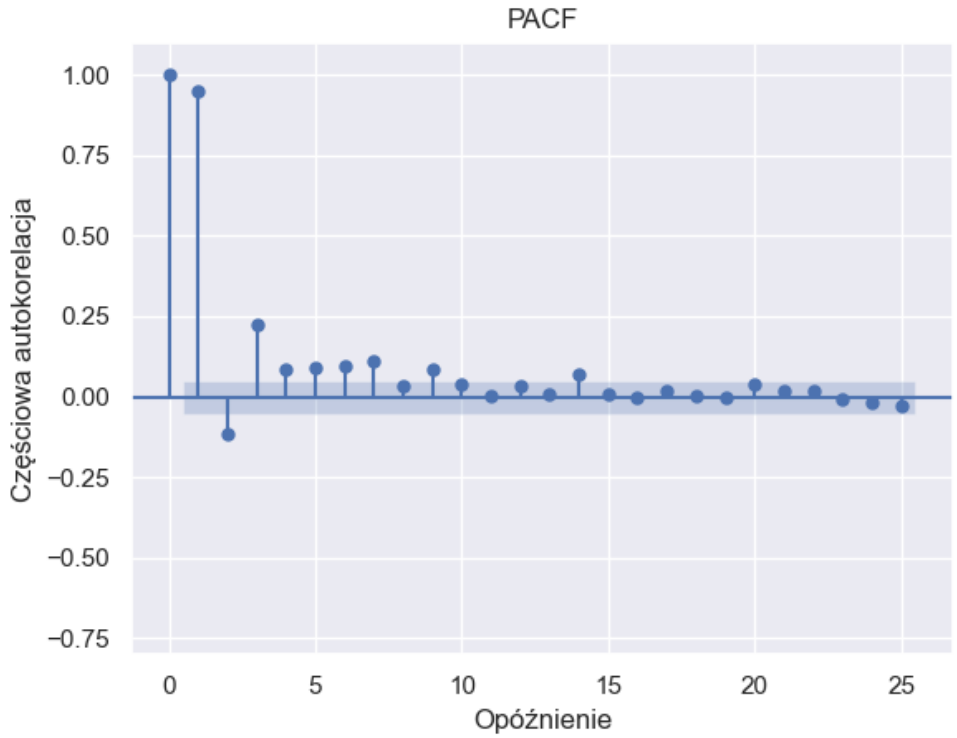
Analizę rozpoczęliśmy od narysowania wykresu danych w zależności od czasu (Wykres 1) oraz stworzenia dla nich wykresów funkcji ACF oraz PACF (Wykres 2 i 3).



Rysunek 1: Wykres zależności średniej temperatury od czasu



Rysunek 2: Wykres funkcji ACF dla surowych danych



Rysunek 3: Wykres funkcji PACF dla surowych danych

2.2 Dane po dekompozycji

Jak możemy zobaczyć na powyższym wykresie (Wykres 1) w danych występuje sezonowość. W celu usunięcia funkcji deterministycznych wykorzystaliśmy dekompozycję Wold'a.

Stosując tę dekompozycję zakładamy, że początkowo dane są określone przez:

$$X_t^* = X_t + s(t) + m(t),$$

gdzie:

- X_t^* - dane,
- X_t - szereg czasowy stacjonarny w słabym sensie,
- $s(t)$ - sezonowa funkcja determinityczna,
- $m(t)$ - liniowa funkcja deterministyczna.

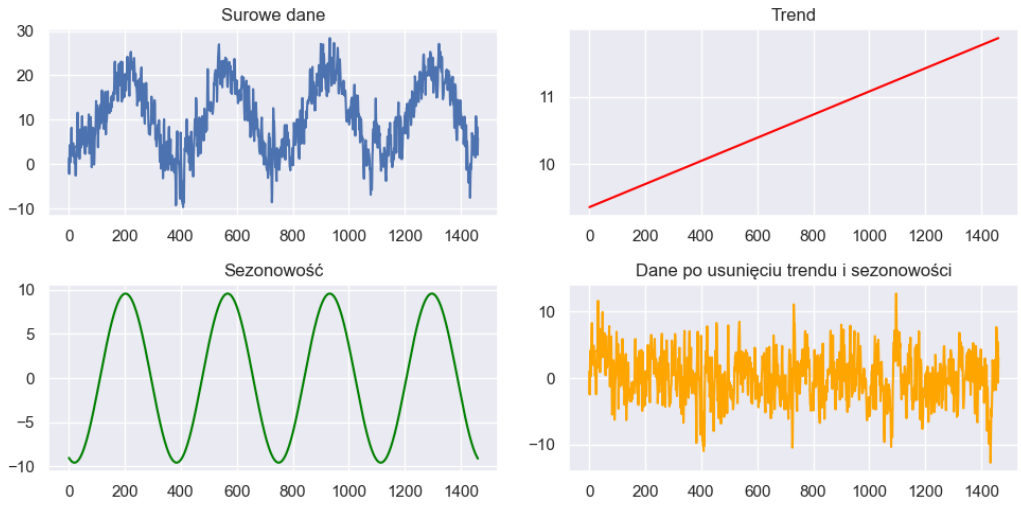
Przeprowadzając tę dekompozycję dopasowujemy najpierw sezonową funkcję do danych, następnie odejmujemy jej wartości w poszczególnych momentach w czasie otrzymując:

$$X_t^{**} = X_t + m(t).$$

Następnie w analogiczny sposób postępujemy z liniową funkcją deterministyczną otrzymując stacjonarny szereg czasowy X_t.

My przeprowadziliśmy dekompozycję Wold'a dopasowując:

- $s(t) = 0.0017t + 9.36126019563272$,
- $m(t) = -9.59\cos(0.017(t - 19.65))$.



Rysunek 4: Dekompozycja

Przyglądając się funkcjom dopasowanym do danych (Wykres 4) możemy zauważyć, że w ciągu rozważanego przez nas okresu czasu temperatury we Wrocławiu nieznacznie wzrosły. Oczywiście można było się tego spodziewać, ze względu na postępujące globalne ocieplenie klimatu. Analizując wygląd trendu liniowego widzimy, że w przeciągu 4 lat temperatury wzrosły średnio o $2.51°C$, co daje średnio $0.628°C$ rocznie. Dodatkowo widzimy, że minima lokalne przypadają w wartościach odpowiadających 20 stycznia każdego roku, co może wskazywać na średnio najchłodniejszy dzień w roku, z kolei lokalizacja maksimum lokalnych wskazuje na 22 lipca jako średnio najcieplejszy dzień w roku. Przyglądając się dalej funkcji okresowej możemy stwierdzić, że średnie temperatury oscylują wokół wartości $10°C$ i wachają się między $25°C$, a $-7°C$, co daje ogólne pojęcie o zachowaniu średnich temperatur we Wrocławiu na przestrzeni ostatnich lat. Dodatkowo okres funkcji wynosi dokładnie $\omega = 0.01722173762194247 = \frac{2\pi}{265}$, co wskazuje na zgodność z faktycznym zachowaniem średnich temperatur w czasie.

Następnie stworzyliśmy wykres danych po dekompozycji w zależności od czasu (Wykres 5) oraz wykresy funkcji ACF oraz PACF (Wykres 6 i 7).

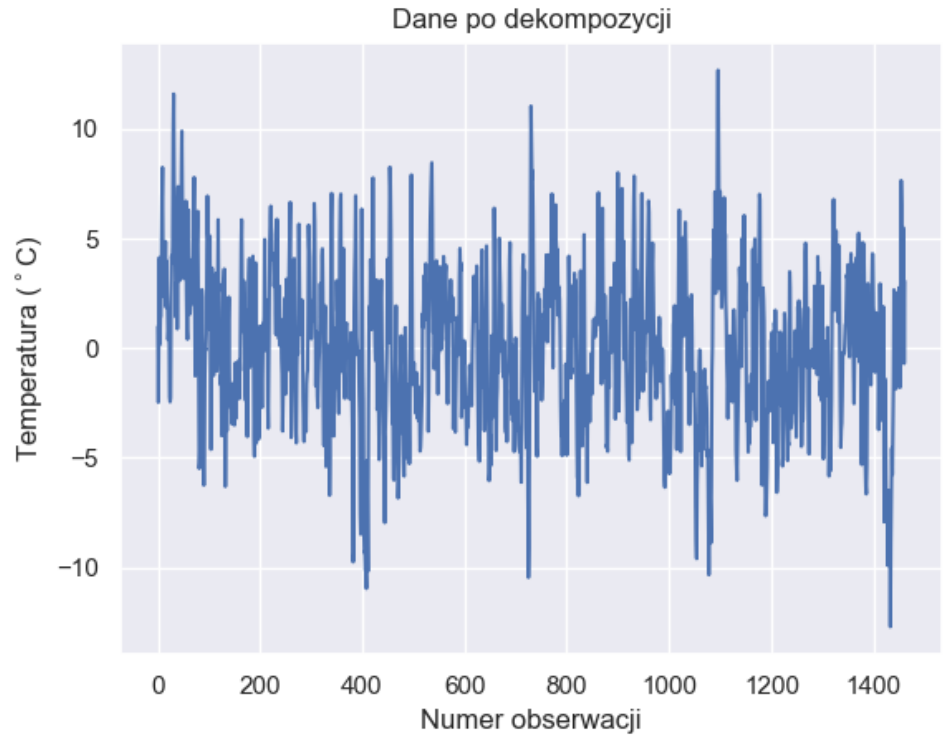
Widzimy, że po dekompozycji dane zachowują się w sposób typowy dla modelu ARMA(p,q), chociaż na podstawie wykresu zależności średniej temperatury od czasu (Wykres 5) możemy zauważyć, że występuje też spora grupa danych odstających. Natomiast jeśli chodzi o zachowanie funkcji ACF oraz PACF dla danych po dekompozycji (Wykres 6) zanikają one względnie jednostajnie do 0, przy czym wykres funkcji PACF zanika zdecydowanie szybciej.

Dodatkowo przeprowadziliśmy wzmocniony test Dickey'a-Fuller'a, który weryfikuje hipotezę zerową, która mówi, że szereg czasowy nie jest stacjonarny, dla danych przed i po dekompozycji. Rezultaty umieściliśmy w tabeli poniżej (Tabela 1).

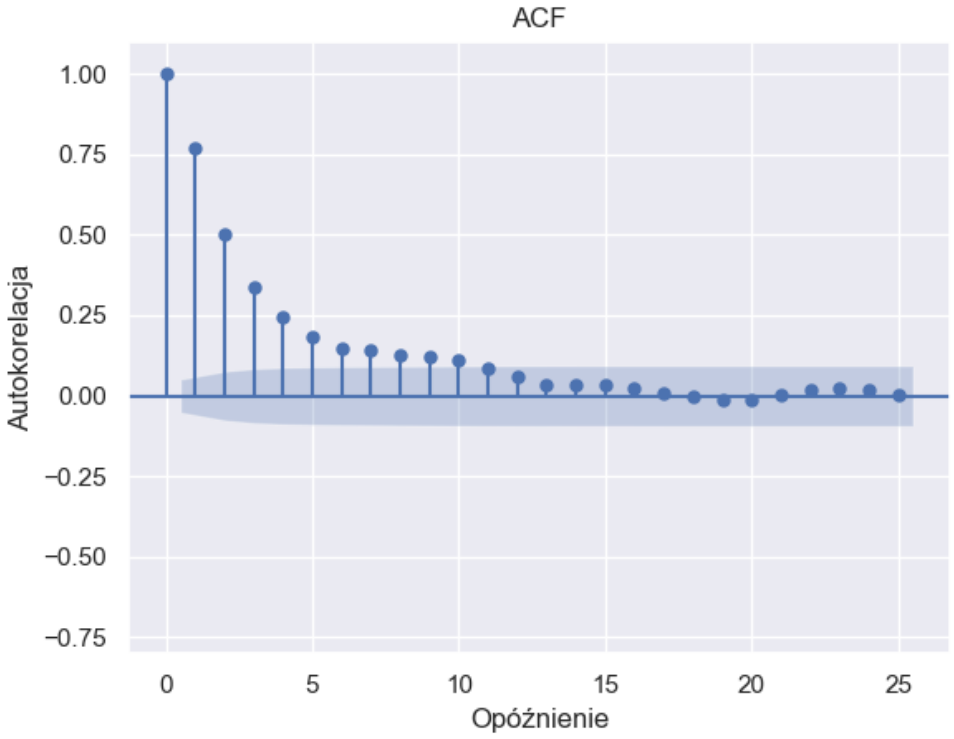
Dane	Statystyka testowa	p-wartość
surowe	-2.597	0.094
po dekompozycji	-13.418	0.0

Tabela 1: Wzmocniony test Dickey'a-Fuller'a

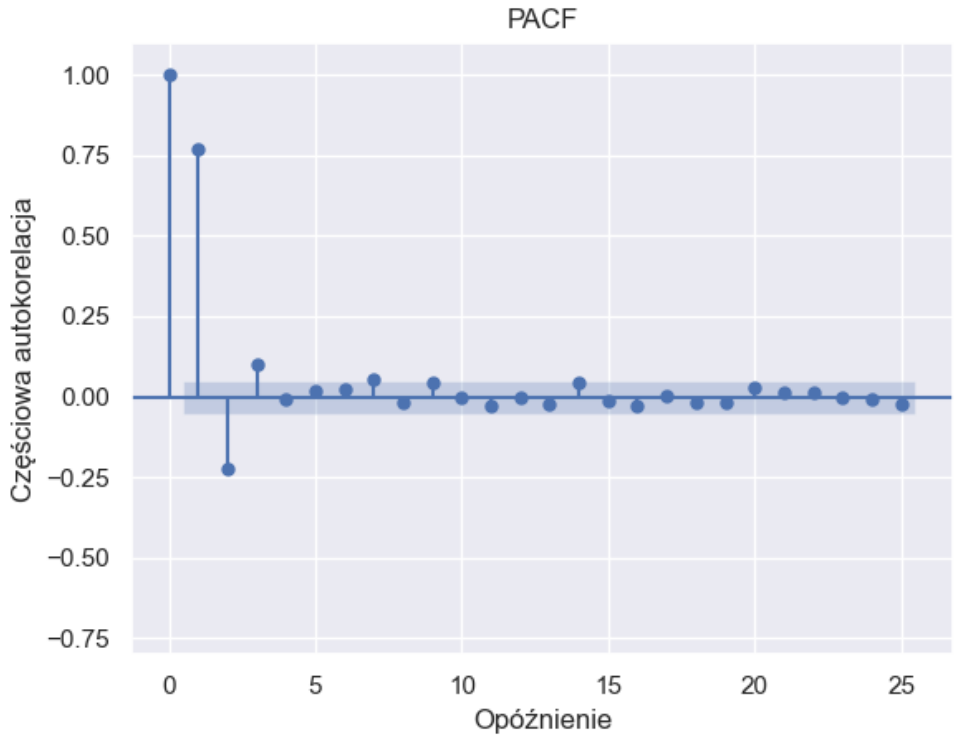
Jak widzimy dla danych po dekompozycjy p-wartość, wskazuje na odrzucenie hipotezy zerowej, zatem dane po dekompozycji są stacjonarne. Z kolei dla danych przed dekompozycją uzyskujemy przeciwny rezultat: p-wartość jest duża, zatem nie ma podstaw do odrzucenia hipotezy zerowej.



Rysunek 5: Dane po dekompozycji



Rysunek 6: Wykres funkcji ACF dla danych po dekompozycji



Rysunek 7: Wykres funkcji PACF dla danych po dekompozycji

3 Modelowanie danych przy pomocy ARMA(p,q)

3.1 Model ARMA(p,q)

Jak wyżej wspominaliśmy, dane będziemy chcieli dopasowywać do modelu ARMA(p,q), który opisany jest poniższym twierdzeniem (1).

Twierdzenie 1 *Szereg czasowy* $\{X_t\}_{t\in Z}$ *jest modelem ARMA(p,q), jeśli jest stacjonarny w słabym sensie i spełnia następujące równanie:*

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

gdzie $\{Z_t\}_{t\in Z} \sim WN(0, \sigma^2)$ *oraz wielomiany:*

- *autoregresji* - $\Phi(z) = 1 - \phi_1 z - ... - \phi_p z^p$,
- *średniej ruchomej* - $\Theta(z) = 1 + \theta_1 z + ... + \theta_q z^q$,

nie mają wspólnych pierwiastków.

3.2 Dobranie rzędu modelu

Dopasowanie odpowiedniego modelu ARMA(p,q) rozpoczęliśmy od znalezienia rzędów modelu - p i q. W tym celu wykorzystaliśmy kryteria informacyjne. Ich działanie polega na znalezieniu parametrów p i q, które minimalizują wartość statystyki stosowanej w odpowiednim kryterium. Wykorzystaliśmy kryteria, zdefiniowane w opisany sposób w wykorzystanej przez nas funkcjach wbudowanych języka Python:

- Akkaiko Information Criterion $AIC = -2ln(\hat{L}) + 2k$,
- Bayesion Information Criterion $BIC = -2ln(\hat{L}) + kln(N)$,
- Hannan and Quinn Information Criterion $HQIC = -2ln(\hat{L}) + 2kln(ln(N))$,

gdzie:

- $\hat{L}$ - wyestymowana funkcja największej wiarygodności,
- k - liczba estymowanych parametrów,
- N - długość próby.

Otrzymane wyniki umieściliśmy w tabeli poniżej (Tabela 2):

Kryterium	Wartość p	Wartość q
AIC	1	1
BIC	1	1
HQIC	1	1

Tabela 2: Rząd modelu

Jak możemy zauważyć rezultaty otrzymane dzięki wszystkim kryteriom pokrywają się, zatem przyjęliśmy następujące wartości p i q:

- p = 1,
- q = 1.

Dalej analizować będziemy zgodność danych z modelem ARMA(1,1):

$$X_t - \phi X_{t-1} = Z_t + \theta Z_{t-1}.$$

3.3 Estymacja parametrów modelu

Następnie estymowaliśmy parametry modelu, w tym celu wykorzystliśmy metody:

- Hannan'a-Rissanen'a,
- innowacji,
- przestrzeni stanu.

Otrzymane wartości parametrów umieściliśmy w tabeli poniżej (Tabela 3).

Metoda	ϕ	θ
Hannan'a-Rissanen'a	0.6420264993548227	0.32868760691053983
innowacji	0.641244743485922	0.3291874541327416
przestrzeni stanu	0.6412499071378781	0.32918410912208174

Tabela 3: Wartości parametrów

Jak widzimy wszystkie metody zwróciły podobne rezultaty. Szczególnie metody innowacji oraz przestrzeni stanu dały wyniki zgodne do 5 cyfr znaczących. Zatem opierając się na otrzymanych wynikach (Tabela 3) przyjęliśmy następujące wartości parametrów:

- $\phi = 0.6412$,
- $\theta = 0.3291$.

Zatem ostatecznie dopasowaliśmy do danych następujący model.

$$X_t - 0.6412 X_{t-1} = Z_t + 0.3291 Z_{t-1}.$$

Poniżej sprawdziliśmy jakoś dopasowania dobranego przez nas modelu.

4 Ocena dopasowania modelu

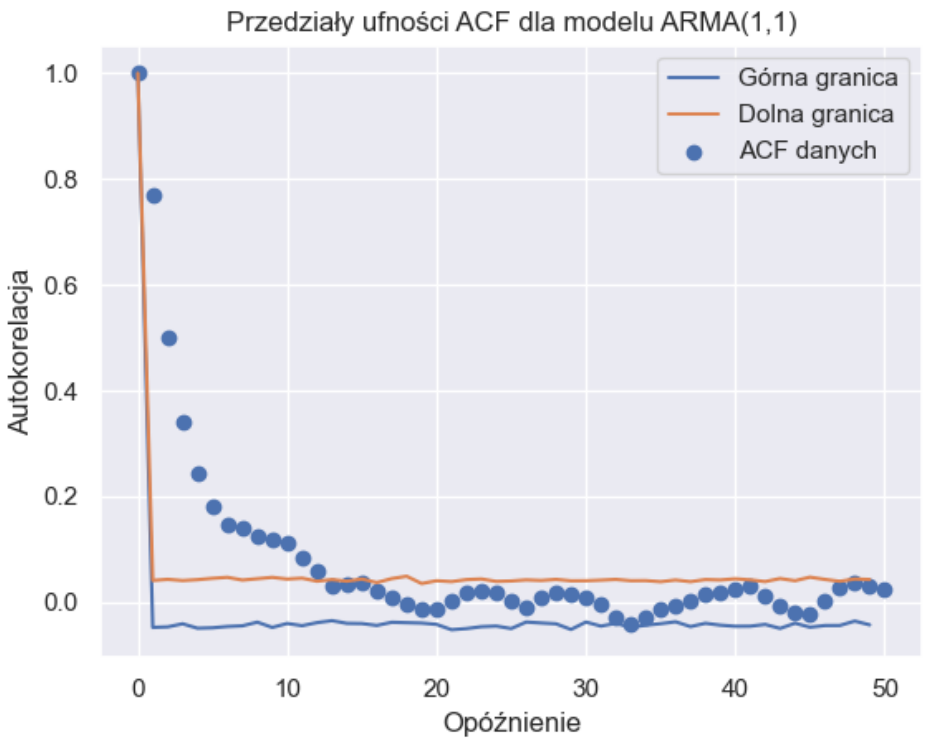
4.1 Przedziały ufności dla ACF i PACF

Na początku za pomocą metody Monte Carlo wyznaczyliśmy przedziały ufności dla funkcji ACF oraz PACF. W tym celu wysymulowaliśmy $M = 200$ trajektorii, dla których dla każdego przesunięcia $h = 1, 2, ..., 50$ wyznaczyliśmy przedziały ufności na poziomie ufności α. W tabeli poniżej (Tabela 4) przedstawiliśmy uzyskane przez nas wyniki dla różnych poziomów ufności.

poziom ufności	ilość wartości odstających dla ACF [%]	ilość wartości odstających dla PACF [%]
0.1	24%	16%
0.05	24%	16%
0.01	16%	14%

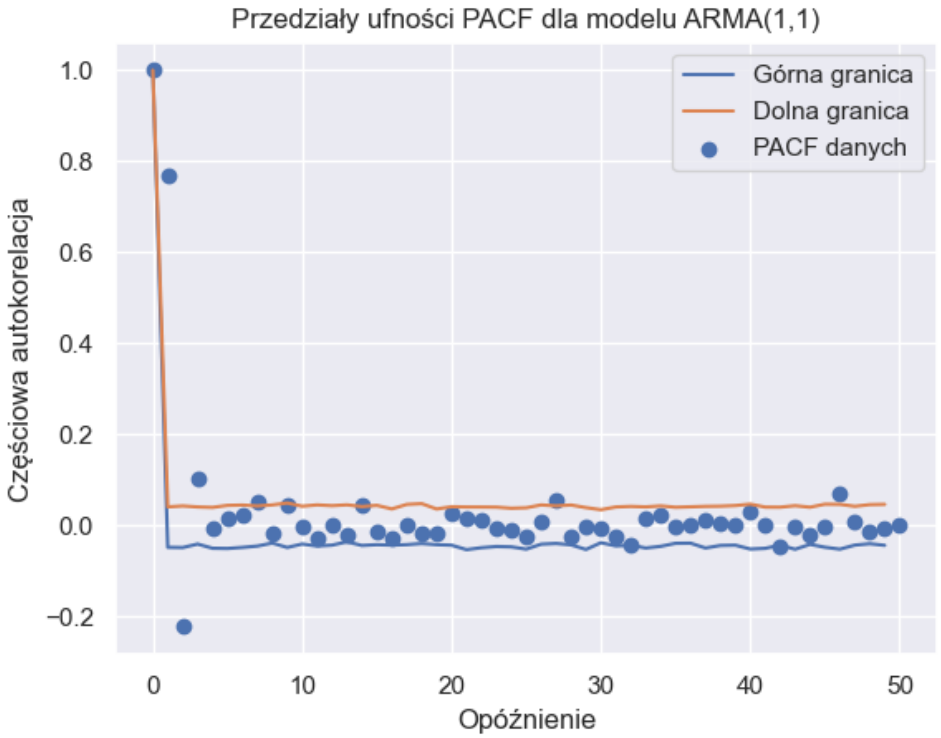
Tabela 4: Przedziały ufności dla ACF i PACF

Dodatkowo stworzyliśmy wykresy przedstawiające przedziały ufności na poziomie ufności $\alpha = 0.05$ dla funkcji ACF oraz PACF. Umieściliśmy je poniżej (Wykres 8 oraz 9).



Rysunek 8: Przedziały ufności na poziomie ufności $\alpha = 0.05$ dla funkcji ACF

Jak widzimy odstaje więcej wartości, niż byśmy się tego spodziewali. Już przy sporządzaniu wykresu zależności średniej temperatury od czasu zauważyliśmy, że dane mają tendencję do większego odstawania. Zatem jest to zgodne z naszymi wcześniejszymi rozważaniami. Widzimy dodatkowo, że dla funkcji ACF odstaje więcej danych, niż dla funkcji PACF. Wartym zauważenia jest również fakt, że dla poziomu ufności $\alpha = 0.1, 0.05$ odstaje porównywalna ilość obserwacji. Wynika to prawdopodobnie z faktu, że funkcje ACF oraz PACF stosunkowo wolno zanikają do zera, szczególnie funkcja autokorelacji, z kolei dla większych wartości opóźnienia obie funkcje chowają się w przedziałach ufności.



Rysunek 9: Przedziały ufności na poziomie ufności $\alpha = 0.05$ dla funkcji PACF

4.2 Porównanie linii kwantylowych z trajektorią

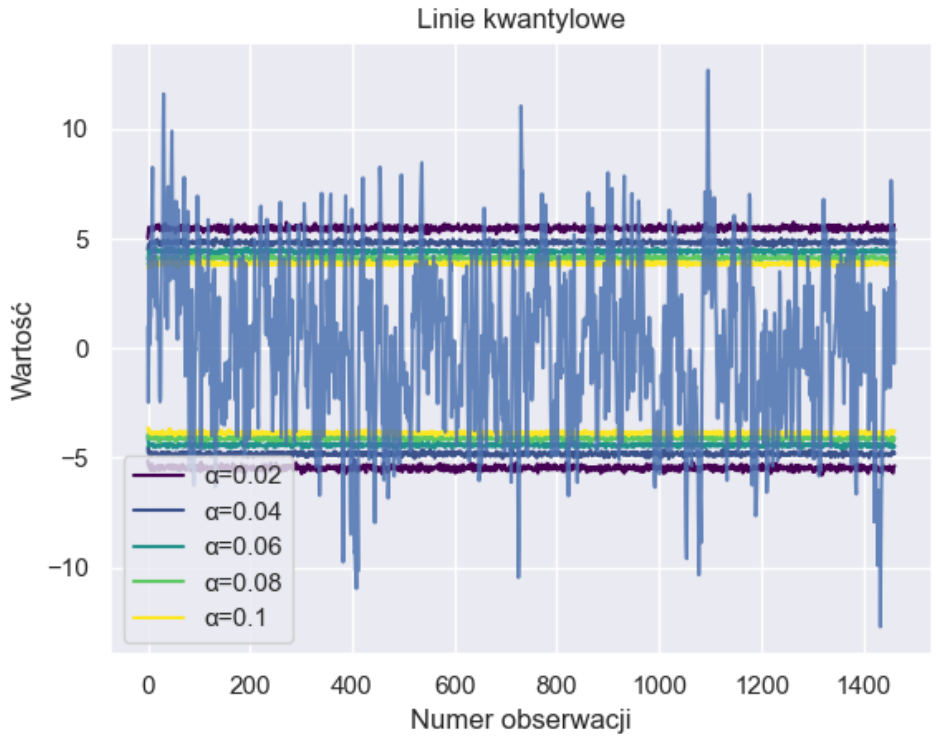
Następnie wyznaczyliśmy linie kwantylowe. W tym celu wykorzystaliśmy metodę Monte Carlo. Wysymulowaliśmy M = 10000 trajektorii, dla których wyznaczyliśmy kwantyle w każdym momencie t=1, ..., 1462. Następnie z uzyskanych kwantylii stworzyliśmy linie kwantylowe. Wykres z porównaniem linii kwantylowych z trajektorią umieściliśmy poniżej (Wykres 10).

Dodatkowo sporządziliśmy tabelę, w któej umieściliśmy ilość odstających wartości dla danych lini kwantylowych. Umieściliśmy ją poniżej (Tabela 5).

poziom ufności	ilość danych odstających [%]
0.1	26.94%
0.08	23.53%
0.06	19.90%
0.04	15.39%
0.02	10.94%

Tabela 5: Linie kwantylowe

Jak widzimy spora grupa danych odstaje, czego spodziewaliśmy się po sporządzeniu przedziałów ufności funkcji ACF i PACF. Widzimy jednak, że dla wszystkich sprawdzanych przez nas wartości są to odstępstwa rzędu kilku procent.



Rysunek 10: Wykres linii kwantylowych

5 Analiza residuów

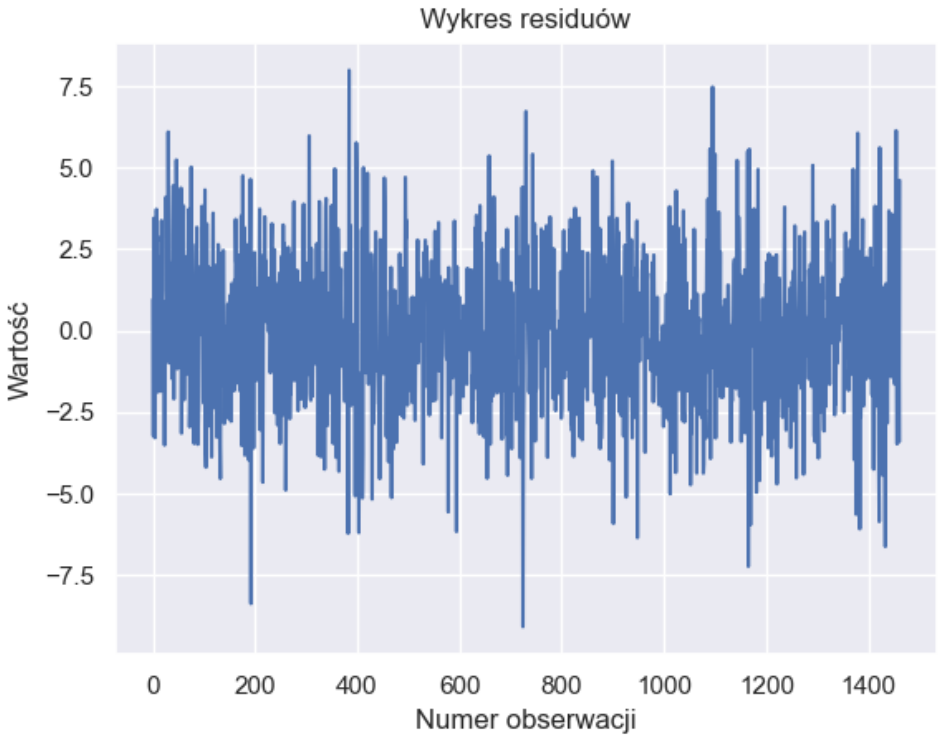
W tej części zajmiemy się analizą residuów Z_t. Będziemy sprawdzali, czy spełniają one poniższe kryteria:

1. $\forall_{t=1,..,n}\ EZ_i = 0$
2. $\forall_{t=1,..,n}\ VarZ_i = \sigma^2$
3. $\epsilon_1, ..., Z_n$ - zmienne losowe, nieskorelowane,
4. $\forall_{t=1,..,n}\ Z_i \sim WN(0, \sigma^2)$.

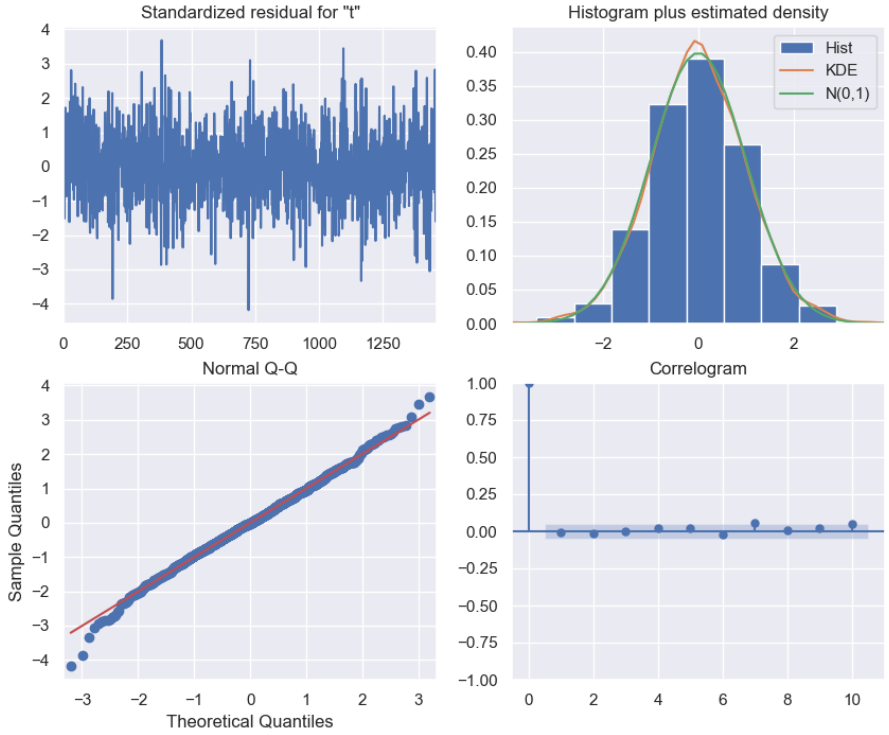
Analizę rozpoczeliśmy od narysowania wykresu zależności residuów od czasu (Wykres 11), funkcji ACF i PACF (Wykres 13 i 14) oraz wykresów badających podstawowe własności residuów (Wykres 12).

Przyglądając się wykresowi zależności wartości residuów od czasu możemy zauważyć, że oscylują one wokół wartości 0, co pozwala wstępnie przypuszczać, że $EZ_t = 0\ \forall_t$. Dodatkowo widzimy, że wahają się one o podobne wartości dla wszystkich rozważanych przez nas t, jednak występują pojedyncze odstające wartości.

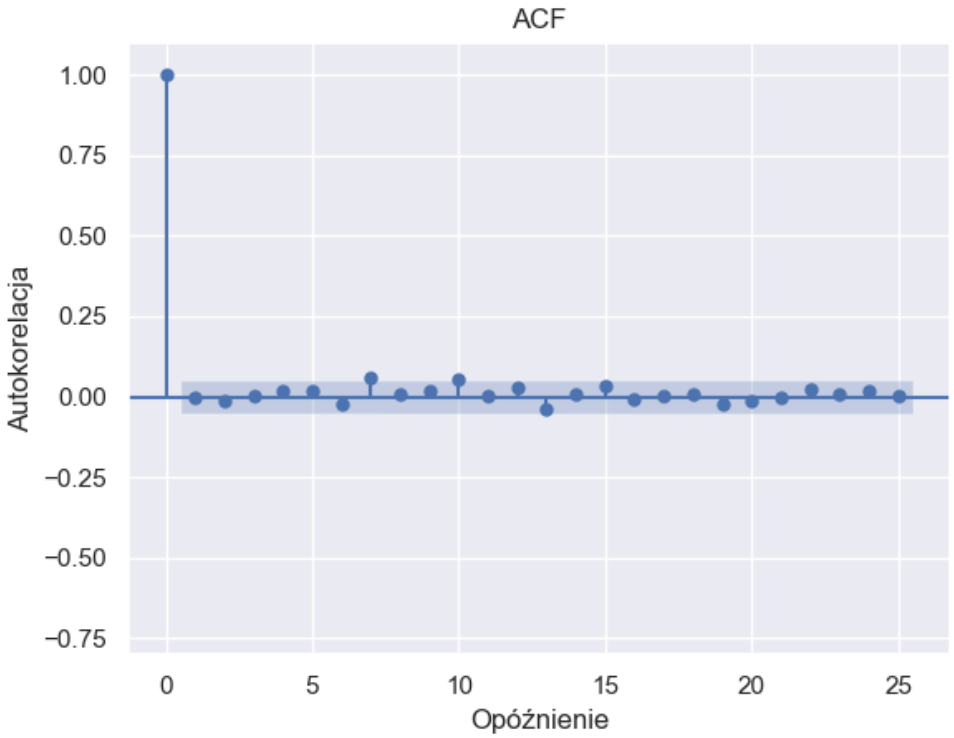
Analizując z kolei wykres kwantylowy residuów (Wykres 12) możemy zauważyć, że wykazuje on standardowe zachowanie, jednak dla krańcowych wartości nieco odstaje od kwantyli rozkładu normalnego. Może być to spowodowane obecnością grupy danych, które odstają, co wyżej zauważyliśmy.



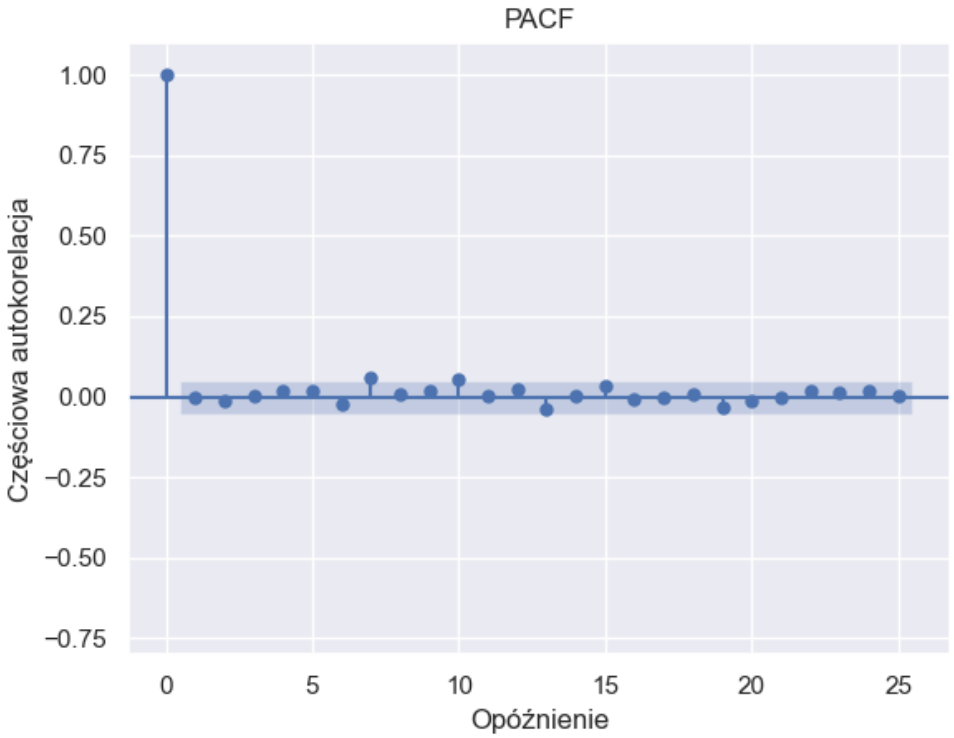
Rysunek 11: Residua



Rysunek 12: Residua - podstawowe wykresy



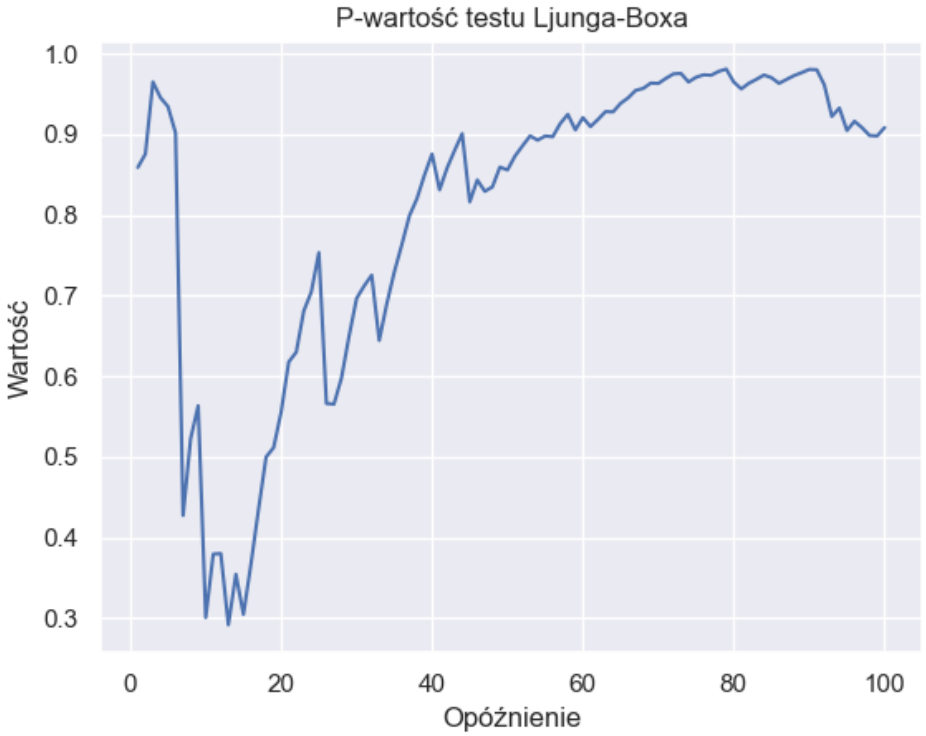
Rysunek 13: Wykres funkcji ACF dla residuów



Rysunek 14: Wykres funkcji PACF dla residuów

Z kolei analizując wykresy funkcji ACF i PACF dla residuów, widzimy, że już od przesunięcia $h = 1$ wartości tych funkcji oscylują wokół 0. Zatem z własności funkcji autokorelacji i częściowej funkcji

autokorelacji możemy wywnioskować, że residua są nieskorelowane. Dodatkowo w celu weryfikacji założenia o nieskorelowaniu residuów przeprowadziliśmy test Ljung'a-Box'a, którego wyniki umieściliśmy na wykresach poniżej (Wykres 16 - wartości staystyki oraz 15 - p-wartości).



Rysunek 15: P-wartości testu Ljung'a-Box'a

Jak widzimy na podstawie wykresu p-wartości testa Ljung'a-Box'a (Wykres **??**) p-wartości osiągają nie spadają poniżej 0.3, jednak dla większości lagów wynoszą ponad 0.5. Dodatkowo możemy zauważyć, że od pewnego momentu p-wartości stabilizują się na poziomie około 0.9. Zatem nie mamy podstaw do odrzucenia hipotezy zerowej, która mówi o niezależności badanych danych. Pokrywa się to z przeprowadzoną przez nas wcześniej analizą funkcji ACF oraz PACF. Pozwala to zatem przypuszczać, że residua są nieskorelowane.

Następnie zajęliśmy się sprawdzeniem założenia dotyczącego wartości średniej. W tym celu przeprowadziliśmy T - testy, którego wyniki umieściliśmy poniżej (Tabela 6) oraz policzyliśmy wartość oczekiwaną dla residuów, która wyniosła $EZ_t = -0.000789$

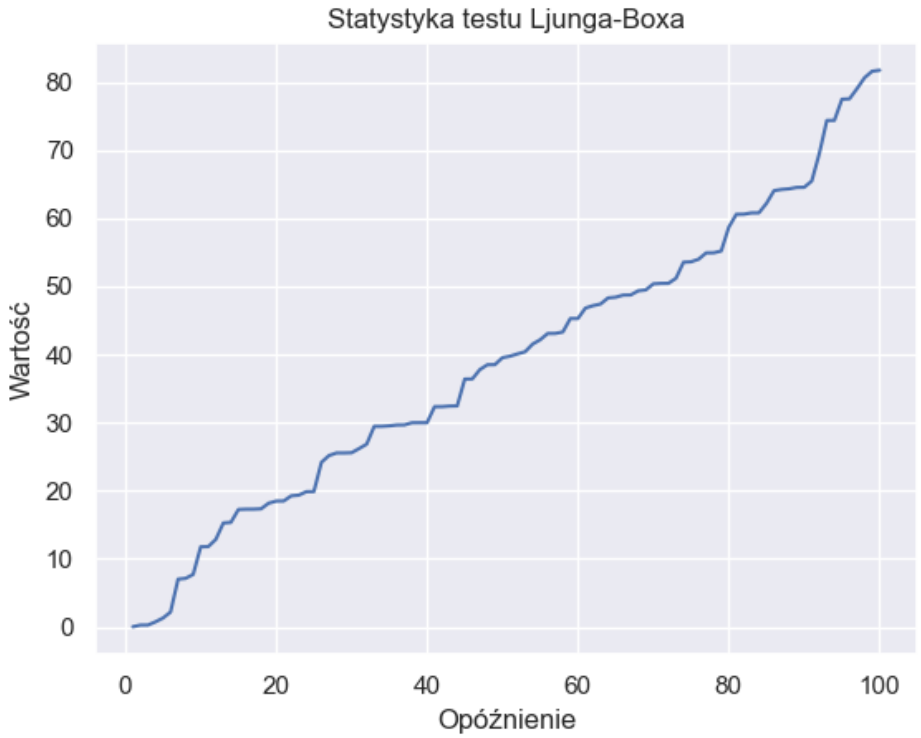
Test	Statystyka testowa	p-wartość
T - test	-0.013860050259897831	0.9889435262806625

Tabela 6: T - test

Jak widzimy po wynikach T-testu nie ma podstaw do odrzucenia hipotezy zerowej, która mówi, że wartość oczekiwana próbki wynosi 0 dla każdej obserwacji. Dodatkowo obliczona przez nas wartość oczekiwana jest bliska zeru. Zatem możemy przypuszczać, że również to założenie jest spełnione.

Dalej sprawdziliśmy założenie odnoszące się do wariancji. W tym celu przeprowadziliśmy testy statystyczne, które razem z ich wynikami umieściliśmy w tabeli poniżej (Tabela 7).

Tym razem p-wartości również nie dają powodów, aby odrzucić hipotezę zerową, która mówi, że wariacja próbki jest stała. Jednak p-wartości nie są aż tak wysokie, może być to spowodowane obecnością dość dużej grupy odstających wartości residuów.



Rysunek 16: Statystyki testu Ljung'a-Box'a

Test	Statystyka testowa	p-wartość
White'a	3.182759506128903	0.20364443876574165
Goldfeld'a-Quandt'a	0.9718529195679072	0.650089274971263

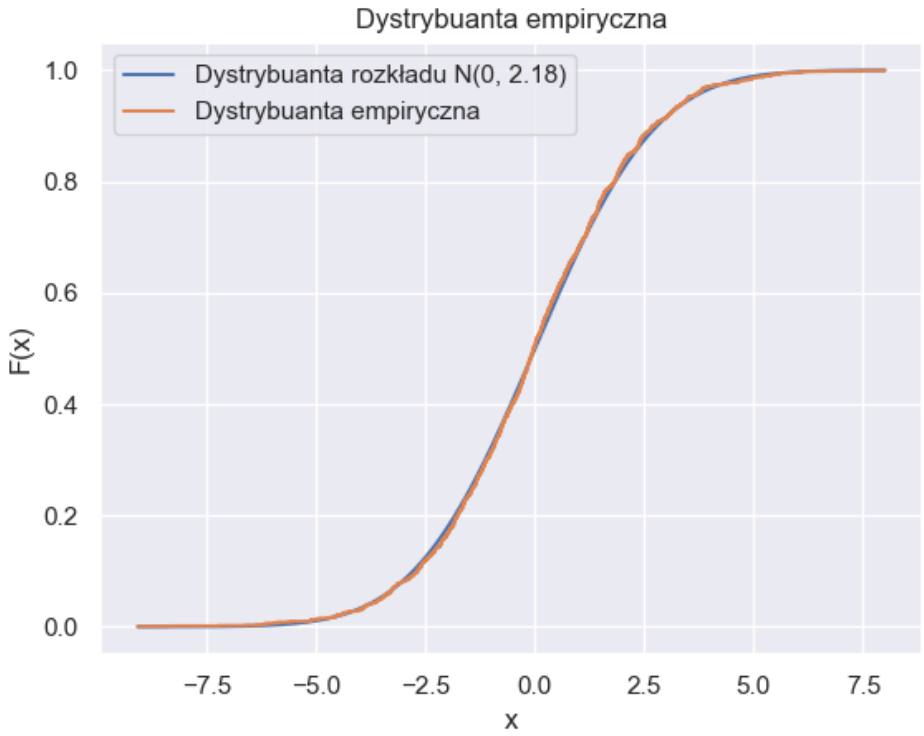
Tabela 7: Testy sprawdzające stałość wariancji

Jako ostatnie sprawdzaliśmy założenie dotyczące rozkładu residuów. W tym celu sporządziliśmy histogram i wykres gęstości residuów, które porównaliśmy z teorytycznymi odpowiednikami dla rozkłądu normalnego $N(0, \sigma^2)$, gdzie $\sigma^2 = 2.18$ - wariancja wyestymowana z danych. Wykresy umieściliśmy poniżej (Wykres 17 oraz 18) Dodatkowo przeprowadziliśmy odpowiednie testy statystyczne, których rezultaty zamieściliśmy w poniżej tabeli (Tabela 8).

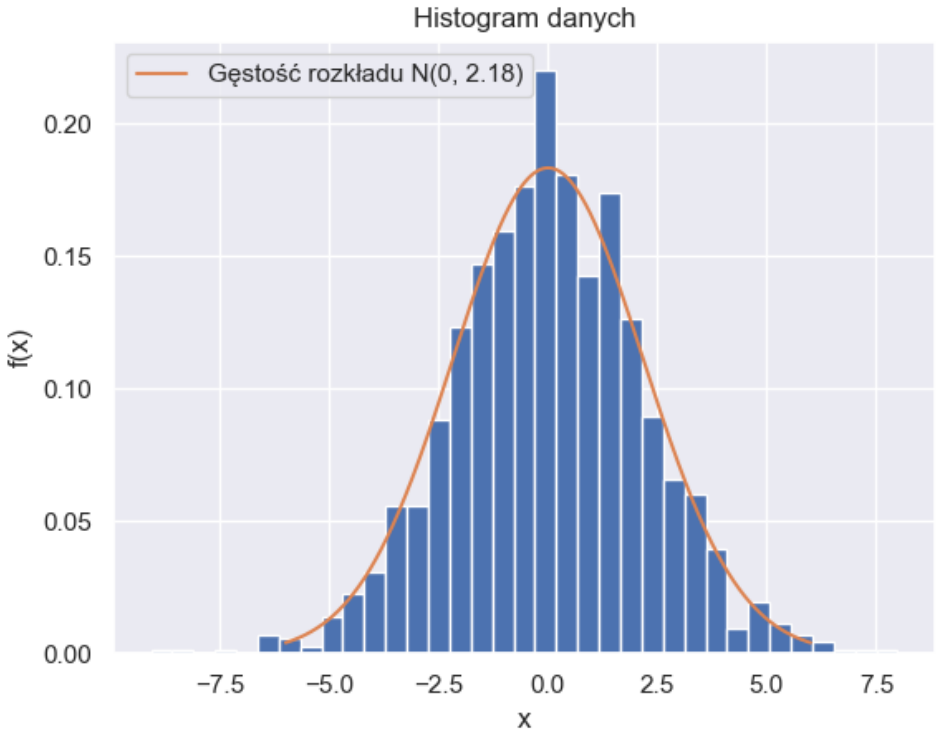
Test	Statystyka testowa	p-wartość
Shapiro-Wilka	0.9976593852043152	0.03220255300402641
Jarque-Bera	14.174994045106164	0.0008354859531705159

Tabela 8: Testy sprawdzające normalność residuów

Jak możemy zauważyć zaróno histogram, jak i dystrybunata wskazują na rozkład normalny residuów. Jednak rezultaty przeprowadzonych przez nas testów statystycznych nie potwierdzają tego. Może to wynikać z faktu, że powyższe testy są bardzo precyzyjne, a widzieliśmy wcześniej, że residua przyjmują momentami stosunkowo większe wartości, dlatego testy mogą odrzucać hipotezę o normalności dla rozważanych danych. Jednak korzystjąc z porównania gęstości teorytycznycznej i emiprycznej oraz porównania dla funkcji dystrybuanty możemy przypuszczać, że residua mają rozkład normalny.



Rysunek 17: Wykres dystrybuanty residuów



Rysunek 18: Histogram residuów

6 Podsumowanie

Zatem podsumowując udało nam się dopasować model ARMA(1,1) do wybranych przez nas danych pogodonowych. Dzięki przeprowadzonej analizie możemy stwierdzić, że jakość dopasowania jest wysoka. Możemy tak wnioskować na podstawie przedziałów ufności funkcji ACF oraz PACF, a także wykresu linii kwantylowych dla danych. Zauważyliśmy, że wykresy te zachowywały się porządanie, jednak w obu przypadkach występowała grupa wartości odstających. Do podobnych wniosków prowadzi nas analiza zachowania residuów, które spełniły wszystkie założenia występujące w modelu, jednak wykres kwantylowy, jak i wyniki testu Ljunga-Boxa oraz nawet sam wykres wartości residuów w zależności od czasu pokazują, że część obserwacji odstaje.

Prowadzi nas to do konkluzji, że pomimo dobrego dopasowania modelu dla danych pogodowych część obserwacji będzie wykazywać nieprzewidywalne zachowanie. Jest to wynik, którego mogliśmy się spodziewać, ponieważ odnosząc to do życia codziennego pomimo, że modele przewidujące pogodę są wysokiej jakości nie rzadko zdarza się, że nie pokrywają się lub też w dużym stopniu odstają od stanu rzeczywistego.

Mniej pozytywne wnioski możemy sformułować po analizie dopasowanych przez nas przy dekompozycji funkcji. O ile zachowanie funkcji okresowej świetnie odwzorowuje rzeczywistość, co jest niewątpliwym pozytywem, to zachowanie funkcji liniowej, która mówi nam o wzroście średnich temperatur we Wrocławiu jest niepokojące. Pokazuje to nam jak bliskie nam są problemy wydające się tak odległe, jak globalne ocieplenie klimatu oraz jak tak proste narzędzia matematyczne, jakim jest model ARMA świetnie je opisują.