Raport
Komputerowa analiza szeregów czasowych
Analiza temperatury w Pakistanie
od stycznia 1901 r. do grudnia 1950 r.

Paweł Stępień, 276 038
Jan Zawadzki, 267 878

31 stycznia 2025

1 Wstęp

1.1 Cel raportu

Celem niniejszego raportu jest analiza zmian temperatury w Pakistanie w okresie od stycznia 1901 roku do grudnia 1950 roku [1] przy użyciu modelu ARMA (Autoregressive Moving Average). Pakistan to kraj położony w Azji Południowej, który charakteryzuje się zróżnicowanym klimatem. Ze względu na swoje położenie geograficzne, kraj ten doświadcza zarówno ekstremalnych upałów, jak i znaczących wahań temperatury w skali roku, co czyni go interesującym obszarem do analizy. W obliczu rosnącego zainteresowania tematyką klimatyczną i jej wpływem na życie codzienne, analiza pogody staje się nie tylko aktualna, ale również istotna z perspektywy naukowej i społecznej. Raport ma na celu zbadanie struktury szeregu czasowego, identyfikację istotnych trendów i sezonowości oraz ocenę dopasowania modelu do danych rzeczywistych.

1.2 Informacje o danych

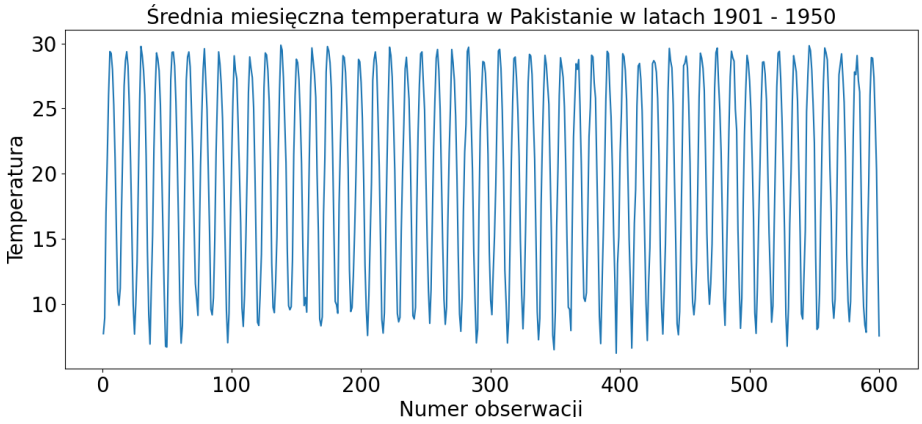
Dane wykorzystane w niniejszym raporcie pochodzą z platformy Kaggle, która jest popularnym źródłem zbiorów danych do analiz statystycznych i modelowania. Zbiór danych obejmuje informacje dotyczące temperatury w Pakistanie w okresie od stycznia 1901 roku do grudnia 1950 roku. Jest to łącznie 50 lat, co daje 600 miesięcy.
Dane zostały zebrane przez Zeeshan-ul-hassan Usmani, amerykańskiego naukowca. Jest niezależnym konsultantem ds. sztucznej inteligencji i data science. Współpracował z Organizacją Narodów Zjednoczonych, Farmer's Insurance, Wal-Mart, Best Buy, 1-800-Flowers, Planned Parenthood, Vicrtoria's Secret, MetLife, SAKS Analytics, Departamentem Zdrowia Karoliny Północnej i kilkoma uniwersytetami w USA, Pakistanie, Kanadzie, Wielkiej Brytanii, Litwie, Chinach, Bangladeszu, Irlandii, Sri Lance i na Bliskim Wschodzie. Udzielił również

dwóch wywiadów dla stacji CNN. Portal ocenia go jako bardzo wiarygodne źródło informacji i zajmuje on bardzo wysokie miejsca w rankingach platformy.

2 Przygotowanie danych do analizy

2.1 Wizualizacja danych

Na sam początek tworzymy wykres prezentujący dane. Widzimy, że dane tworzą swego rodzaju „prostokąt", czyli przez 50 lat w Pakistanie średnia temperatura w danym miesiącu była mniej więcej stała. Może to sugerować, że szereg, z którym mamy tu do czynienia, jest stacjonarny - na razie nie jest to jednak udowodnione.



Rysunek 1: Wykres przedstawiający średnią miesięczną temperaturę w Pakistanie

2.2 Wykres ACF i PACF dla danych surowych

Definicja. *Autokorelacja (ACF)*
narzędzie matematyczne często używane w przetwarzaniu sygnałów do analizowania funkcji lub serii wartości. Mniej formalnie jest to statystyka opisująca, w jakim stopniu dany wyraz szeregu zależy od wyrazów poprzednich w szeregu czasowym. Autokorelacja jest funkcją, która argumentowi naturalnemu k przypisuje wartość współczynnika korelacji Pearsona pomiędzy szeregiem czasowym a tym samym szeregiem cofniętym o k jednostek czasu. [2] Funkcja autokorelacji dla szeregu czasowego $\{X_t\}_{t\in\mathbb{Z}}$ *wyraża się wzorem*

$$\rho(t,s) = \frac{\gamma(t,s)}{\sqrt{\gamma(t,t)\gamma(s,s)}}$$

gdzie

$$\gamma(t,s) = \mathbb{E}[X_t X_s] - \mathbb{E}[X_t] \cdot \mathbb{E}[X_s],\ t,s \in \mathbb{Z}$$

Wzór na autokorelację próbkową (empiryczną):

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

gdzie

$$\hat{\gamma}(h) = \frac{1}{n} \sum_{t=1}^{n-|h|} (x_{t+h} - \bar{x})(x_t - \bar{x}),\ \textit{gdzie } \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Definicja. *Funkcja autokorelacji cząstkowej (ang. partial autocorrelation function, PACF)*
stosowana w analizie szeregów czasowych miara korelacji cząstkowej stacjonarnego szeregu czasowego z jego własnymi opóźnionymi wartościami. Autokorelacja cząstkowa dla danego rzędu opóźnienia jest wyznaczona z wyłączeniem wpływu korelacji dla wszystkich krótszych opóźnień, pod tym względem różni się od zwykłej funkcji autokorelacji, która nie kontroluje pozostałych opóźnień. [3]

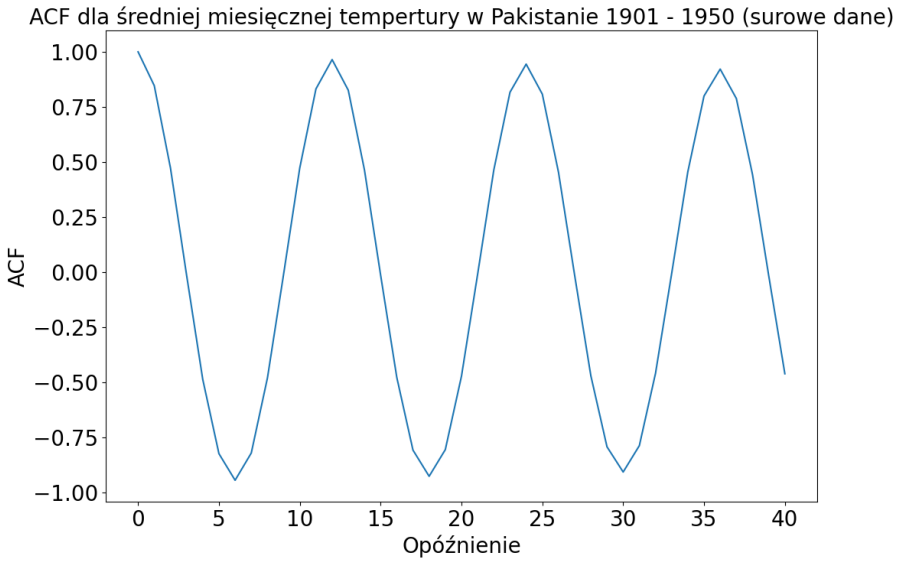
Definicja. *Przedział ufności*
podstawowe narzędzie estymacji przedziałowej. Niech cecha X ma rozkład w populacji z nieznanym parametrem θ. Z populacji wybieramy próbę losową $(X_1, X_2, \ldots, X_n)$. Przedziałem ufności o współczynniku ufności $1-\alpha$ nazywamy taki przedział (θ_1, θ_2), który spełnia warunek:

$$P(\theta_1 < \theta < \theta_2) = 1 - \alpha,$$

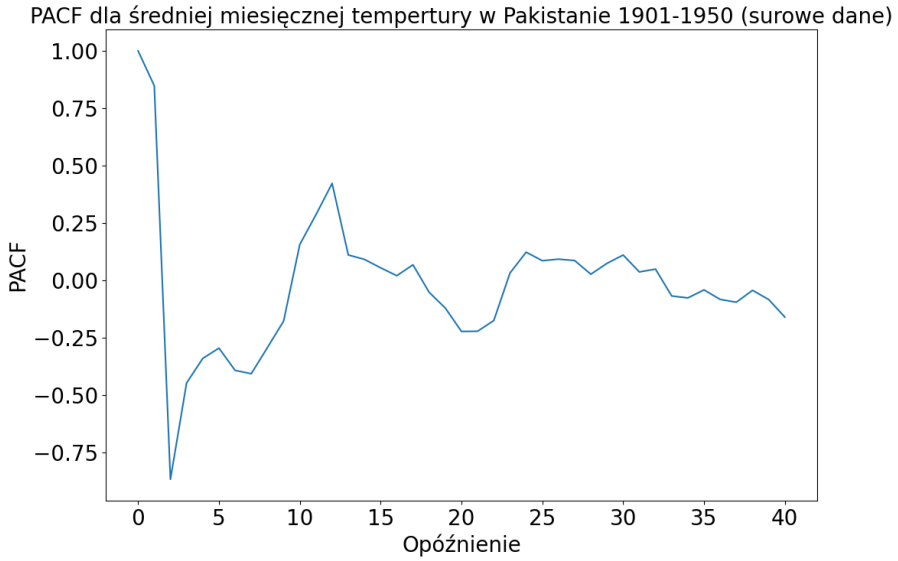
gdzie θ_1 i θ_2 są funkcjami wyznaczonymi na podstawie próby losowej [4].

Pierwszym etapem badania szeregu czasowego jest znalezienie funkcji autokorelacji i funkcji częściowej autokorelacji dla surowych danych.

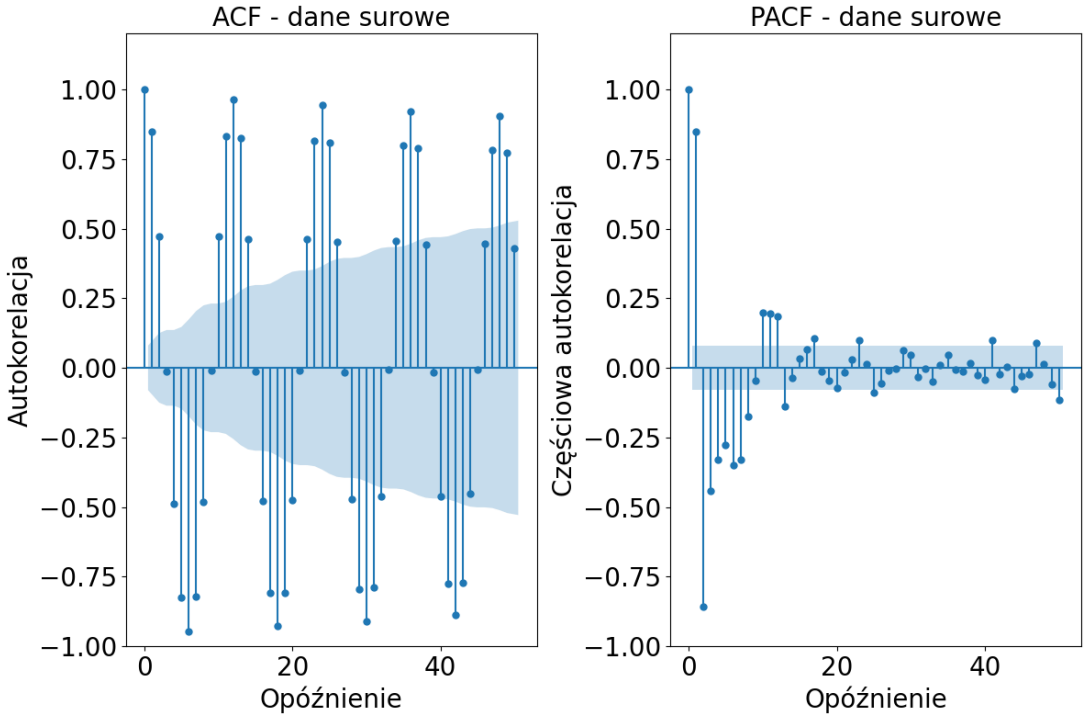
Widzimy, że dla surowych danych (Rys. 2,3,4) funkcje ACF i PACF wyglądają „brzydko" - funkcja autokorelacji przypomina funkcję sinusoidalną, natomiast funkcja częściowej autokorelacji ma bliżej nieokreślony kształt. Pamiętamy, że dla ACF i dla PACF od pewnego momentu w czasie wartości powinny mieścić się w przedziale ufności, czyli być statystycznie równe zeru. Na razie taka sytuacja nie ma miejsca.



Rysunek 2: Wykres przedstawiający ACF dla surowych danych



Rysunek 3: Wykres przedstawiający PACF dla surowych danych



Rysunek 4: Wykres przedstawiający ACF i PACF dla surowych danych z przedziałami ufności

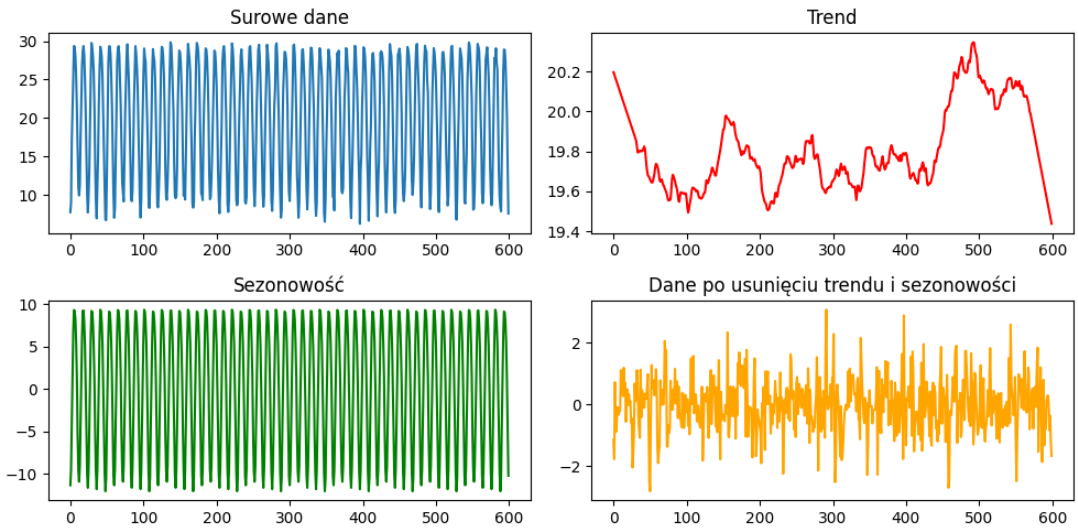
2.3 Dekompozycja szeregu czasowego

Dokonujemy zatem dekompozycji szeregu czasowego - usuwamy trend oraz sezonowość, by w efekcie otrzymać „oczyszczone" dane (Rys. 5).

Sama dekompozycja szeregu czasowego jest, w teorii, dość łatwą czynnością - zakładamy, że nasz szereg czasowy X_t jest w postaci

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

gdzie $u(t)$ jest trendem, $s(t)$ jest komponentem sezonowym ($u(t)$, $s(t)$ są funkcjami deterministycznymi), a $\{Y_t\}_{t\in\mathbb{Z}}$ jest szeregiem czasowym liniowym; w celu uzyskania „czystego" szeregu czasowego odejmujemy od X_t trend $u(t)$ i komponent sezonowy $s(t)$ - jest to tak zwana *dekompozycja Walda*. W praktyce trudne jest samo znalezienie funkcji $u(t)$, $s(t)$, dlatego wykorzystuje się do tego na przykład funkcje wbudowane w Pythonie.



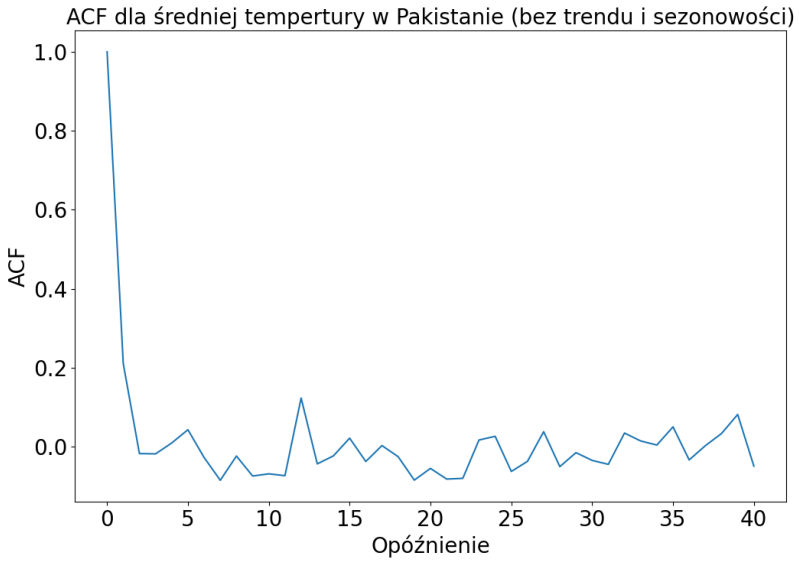
Rysunek 5: Wykres przedstawiający surowe dane, trend, sezonowość i oczyszczone dane.

2.4 Wykres ACF i PACF dla danych oczyszczonych

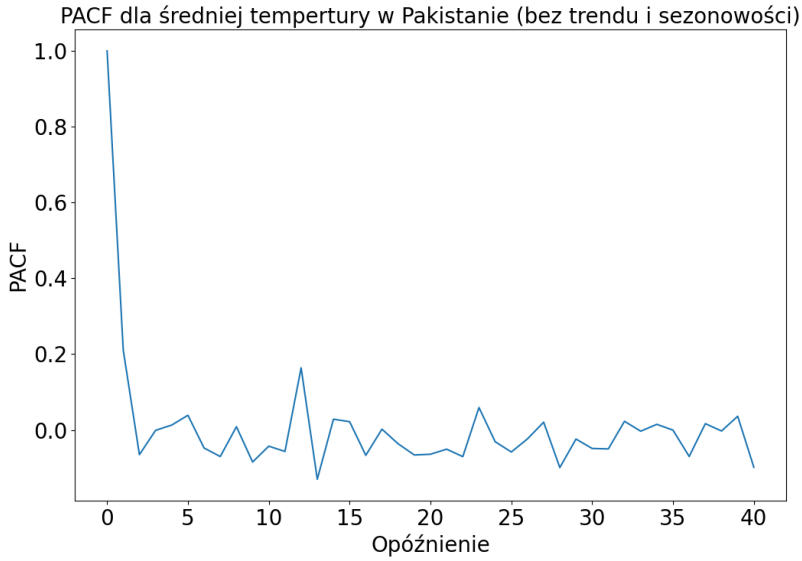
Mamy „oczyszczony" szereg czasowy $\{Y_t\}_{t\in\mathbb{Z}}$, możemy zatem obliczyć jego funkcję autokorelacji i funkcję częściowej autokorelacji.

Widzimy na wykresach (Rys. 6,7), że od pewnego momentu wartości funkcji ACF i PACF oscylują wokół zera; nakładamy na wykresy ACF i PACF przedział ufności, żeby sprawdzić, czy wartości te są statystycznie równe zeru.

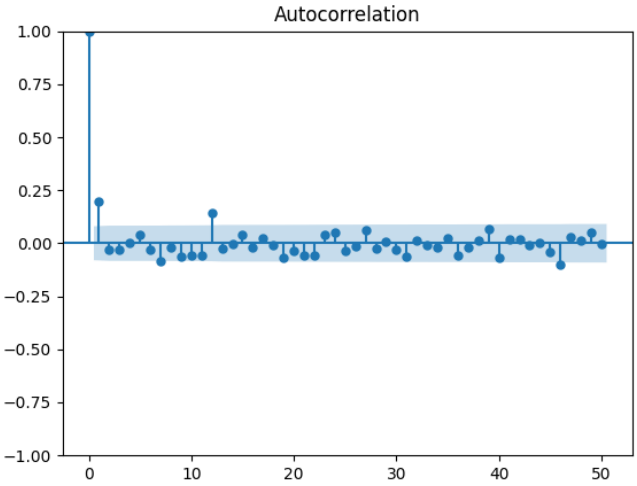
Choć pewne „kropki" wychodzą poza przedział ufności (Rys. 8,9), co jest nieuniknione dla danych empirycznych, to zasadniczo mieszczą się w tym przedziale.



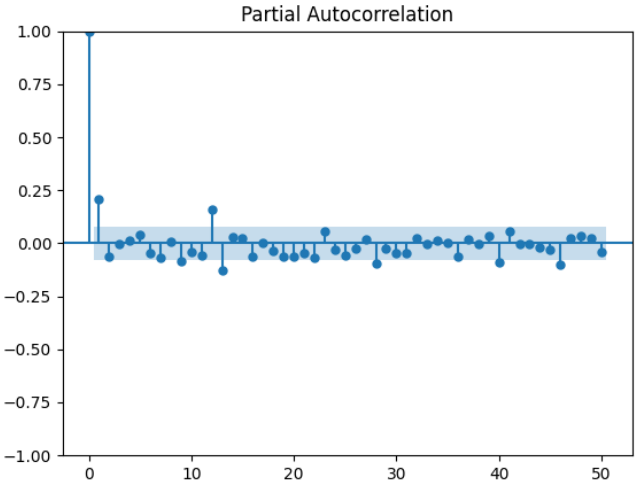
Rysunek 6: Wykres przedstawiający ACF oczyszczone dane



Rysunek 7: Wykres przedstawiający PACF oczyszczone dane



Rysunek 8: Wykres przedstawiający ACF oczyszczone dane



Rysunek 9: Wykres przedstawiający PACF oczyszczone dane

2.5 Test ADF - stacjonarność szeregu

Definicja. *Test Augmented Dickey–Fuller (ADF)*
jest statystycznym testem wykorzystywanym do sprawdzania, czy seria czasowa

jest jednostajna (stationary), tzn. czy jej wartości nie mają długozasięgowych trendów (czy nie zawierają pierwiastka jednostkowego). Test jest rozszerzeniem klasycznego testu Dickey'ego-Fullera, który uwzględnia dodatkowe opóźnienia (lag) w modelu, aby lepiej radzić sobie z autokorelacją reszt [5].
Hipotezy testowe są następujące:

- *Hipoteza zerowa (H_0): Seria czasowa ma pierwiastek jednostkowy, tzn. $\beta = 0$, co oznacza, że seria jest niestacjonarna.*
- *Hipoteza alternatywna (H_1): Seria jest stacjonarna, tzn. $\beta < 0$.*

Definicja. *P-wartość*
prawdopodobieństwo uzyskania wyników testu co najmniej tak samo skrajnych, jak te zaobserwowane w rzeczywistości (w próbie losowej z populacji), obliczone przy założeniu, że hipoteza zerowa jest prawdziwa. Może być interpretowana jako miara niezgodności danych z założonym modelem, wyrażonym w hipotezie zerowej [5].

Na koniec wykonamy jeszcze test ADF (*Augmented Dickey-Fuller Test*) w celu sprawdzenia stacjonarności szeregu $\{Y_t\}$; wyniki prezentują się następująco:

- Statystyka ADF: -11.280429
- p-wartość: 0.00

Skoro p-wartość jest mniejsze od 0,05, to możemy odrzucić hipotezę zerową o niestacjonarności szeregu $\{Y_t\}$.

3 Modelowanie danych przy pomocy ARMA

Definicja. *Model ARMA(p,q) (Autoregressive Moving Average)*
jest modelem szeregów czasowych, który łączy dwa składniki: autoregresję (AR) oraz średnią ruchomą (MA). Jest używany do modelowania i prognozowania wartości szeregów czasowych, które są zależne zarówno od przeszłych wartości zmiennej (autoregresja), jak i od przeszłych błędów prognozy (średnia ruchoma). Model ARMA jest definiowany jako:

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

gdzie $\{Z_t\} \sim WN(0, \sigma^2)$ oraz wielomiany

$$\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p$$

$$\theta(z) = 1 + \theta_1 z + \cdots + \theta_q z^q$$

nie mają wspólnych pierwiastków [6].

3.1 Dobranie rzędu modelu

Sprawdziliśmy już, że szereg czasowy $\{Y_t\}$ jest stacjonarny - to jeden z warunków, by szereg czasowy można było modelować przy użyciu modelu ARMA(p;q). Zakładamy teraz, że szereg $\{Y_t\}$ można modelować używając modelu ARMA(p;q) i próbujemy wyznaczyć rząd modelu.

W celu wyznaczenia rzędu modelu ARMA(p;q) używa się tak zwanych *kryteriów informacyjnych* (*information criterion*). Szukaliśmy rzędu modelu ARMA przy pomocy trzech kryteriów informacyjnych: AIC (*Akaike information criterion*), BIC (*Bayesian information criterion*) oraz HQIC (*Hannan–Quinn information criterion*). Wyniki przeprowadzonych obliczeń prezentowane są w Tabeli 1.

p	q	AIC	BIC	HQIC
2	2	1351.676965	1378.058542	1361.946782
2	2	1351.676965	1378.058542	1361.946782
2	2	1351.676965	1378.058542	1361.946782

Tabela 1: Współczynniki p, q oraz wartości kryteriów informacyjnych

Wszystkie użyte kryteria informacyjne dały ten sam wynik, wobec czego przyjęliśmy w naszym modelu $p = 2$, $q = 2$. Nasz model ma zatem następującą postać

$$Y_1 - \Phi_1 Y_{t-1} - \Phi_2 Y_{t-2} = Z_t + \theta_1 Z_{t-1} + \theta_2 Z_{t-2}. \tag{1}$$

3.2 Estymacja parametrów modelu

Szukamy ϕ_1,ϕ_2, θ_1, θ_2. W tym celu użyliśmy funkcji wbudowanych w Pythonie. Korzystając z dwóch metod: `statespace` oraz `innovations_mle`. Wyniki estymacji parametrów zaprezentowane są na grafice (Rys. 10). Dokładne wyniki obliczeń prezentują tabele (Tabela 2 i 3). Ponieważ p-wartość dla stałej jest w obu wypadkach bardzo wysokie, nie wprowadzamy stałej do wielomianów w równaniu 1. Widzimy, że w estymowanych wartościach parametrów ϕ_1,ϕ_2, θ_1, θ_2 dwiema metodami są różnice. Decydujemy się wybrać parametry wyznaczone metodą `innovations_mle`, ponieważ błąd standardowy dla każdego wyestymowanego parametru jest mniejszy niż w wypadku metody `statespace`. Wobec tego mamy $\phi_1 = 1,5791$,$\phi_2 = -0,7162$, $\theta_1 = -1,9071$, $\theta_2 = 0,9075$, a równanie 1 przyjmuje postać

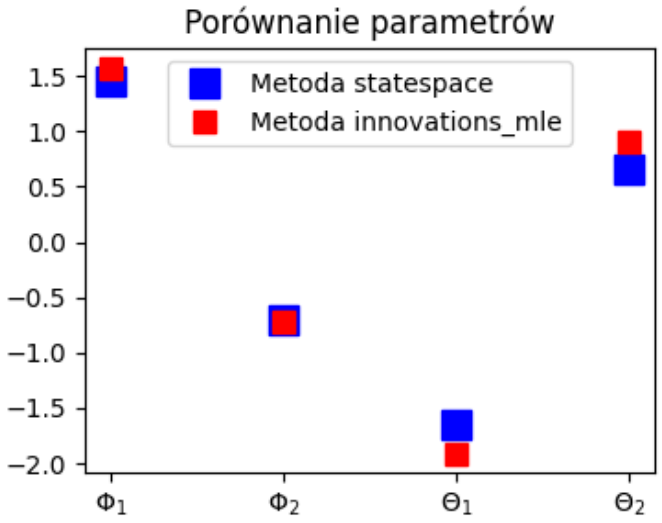
$$Y_1 - 1,5791 \cdot Y_{t-1} + 0,7162 \cdot Y_{t-2} = Z_t - 1,9071 \cdot Z_{t-1} + 0,9075 \cdot Z_{t-2}$$

	Wartość	Błąd standardowy	p-wartość
stała	0,0004	0,001	0,644
ϕ_1	1,4464	0,055	0,00
ϕ_2	-0,6992	0,036	0,00
θ_1	-1,6570	0,061	0,00
θ_2	0,6635	0,061	0,00

Tabela 2: Wyniki dla metody `statespace`

	Wartość	Błąd standardowy	p-wartość
stała	$3,628 \cdot 10^{-05}$	0,00	0,835
ϕ_1	1,5791	0,033	0,00
ϕ_2	-0,7162	0,029	0,00
θ_1	-1,9071	0,027	0,00
θ_2	0,9075	0,027	0,00

Tabela 3: Wyniki dla metody `innovations_mle`



Rysunek 10: Wykres przedstawiający porównanie wyestymowanych parametrów.

3.3 Pierwiastki wielomianów modelu ARMA

Wyliczamy pierwiastki wielomianów AR i MA:

- Pierwiastki wielomianu AR: $z_1 = 1,102 - 0,425i$, $z_2 = 1.102 + 0.425i$
- Pierwiastki wielomianu MA: $x_1 = 1,004$, $x_2 = 1,097$

Wielomiany AR i MA nie mają wspólnych pierwiastków, co oznacza, że dobrze dobraliśmy parametry modelu.

3.4 Przyczynowość i odwracalność modelu

Definicja. *Model ARMA(p;q)* $\{X_t\}$ *jest przyczynowy (jest przyczynową funkcją* $\{Z_t\}_{t\in\mathbb{Z}}$*), jeśli istnieją stałe* $\{\Psi_j\}_{j=0}^{\infty}$ *takie, że*

$$\sum_{j=0}^{\infty} |\{\Psi_j\}| < +\infty$$

oraz

$$X_t = \sum_{j=0}^{\infty} \Psi_j Z_{t-j}$$

Warunek gwarantujący przyczynowość modelu ARMA(p;q) jest następujący:

$$\Phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p \neq 0 \ \forall |z| \leq 1$$

Definicja. *Model ARMA(p;q)* $\{X_t\}$ *jest odwracalny, jeśli istnieją stałe* $\{\Pi_j\}_{j=0}^{\infty}$ *takie, że*

$$\sum_{j=0}^{\infty} |\{\Pi_j\}| < +\infty$$

oraz

$$Z_t = \sum_{j=0}^{\infty} \Pi_j X_{t-j}$$

Warunek gwarantujący odwracalność modelu ARMA(p;q) jest następujący:

$$\Theta(z) = 1 + \theta_1 z + \cdots + \theta_q z^q \neq 0 \ \forall |z| \leq 1$$

Otrzymany szereg jest szeregiem przyczynowym i odwracalnym, ponieważ wartość bezwzględna z pierwiastków wielomianu $\Phi(z)$ jest równa $|z_{1,2}| = 1,181 > 1$ oraz wartości bezwzględne z pierwiastków wielomianu $\Theta(z)$ są równe $|z_1| = 1,004 > 1$, $|z_2| = 1,097 > 1$.

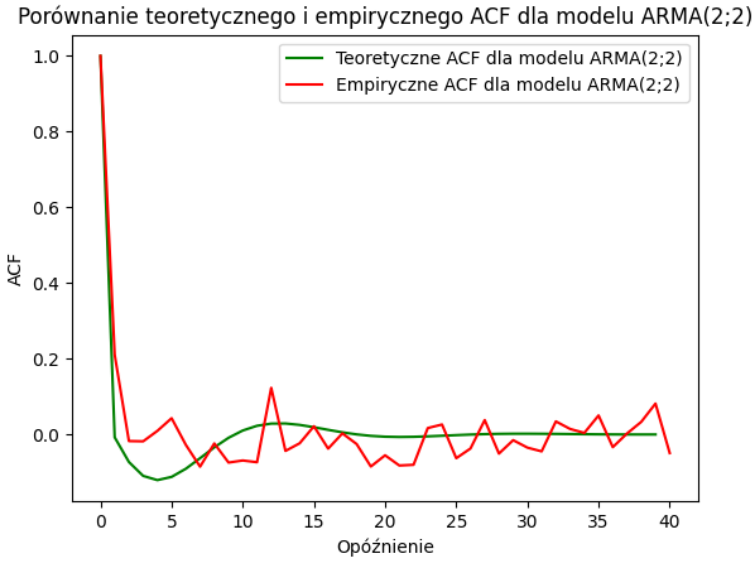
4 Ocena dopasowania modelu

4.1 Przedziały ufności dla ACF i PACF

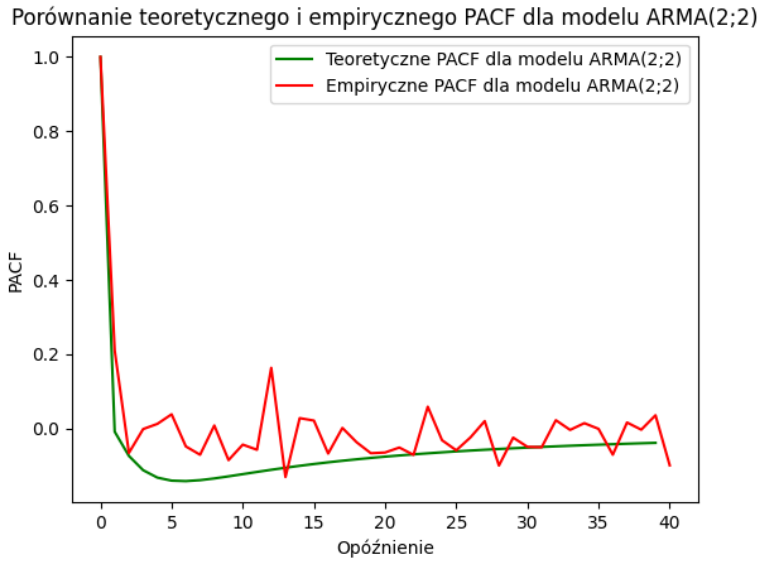
Znając parametry modelu sprawdzamy, czy jego teoretyczna funkcja autokorelacji i teoretyczna funkcja częściowej autokorelacji pokrywają się ze swoimi empirycznymi odpowiednikami (Rys. 11, 12).

Widzimy, że wartości teoretyczne i empiryczne w dużej mierze się pokrywają. To kolejne potwierdzenie, że nasz model jest dobrze dobrany. Co więcej,

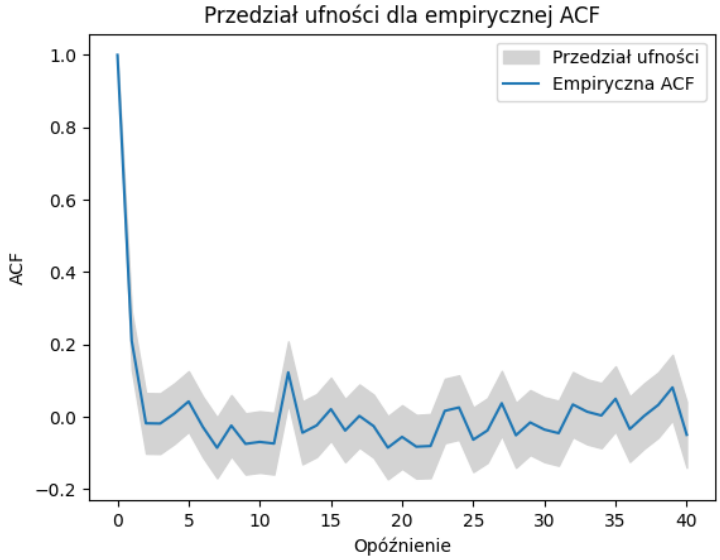
skoro empiryczna funkcja autokorelacji i empiryczna funkcja częściowej autokorelacji od pewnego momentu przyjmują wartości oscylujące w okolicach zera, to potwierdzamy tym samym brak korelacji między danymi. Oczywiście empiryczna funkcja autokowariancji i empiryczna funkcja częściowej autokorelacji w pewnym procencie pokrywają się ze swoimi teoretycznymi odpowiednikami. W celu zilustrowania tego faktu konstruujemy przedział ufności (w tym wypadku na poziomie 95%) i nakładamy go na wykres ACF i PACF (Rys. 13,14).



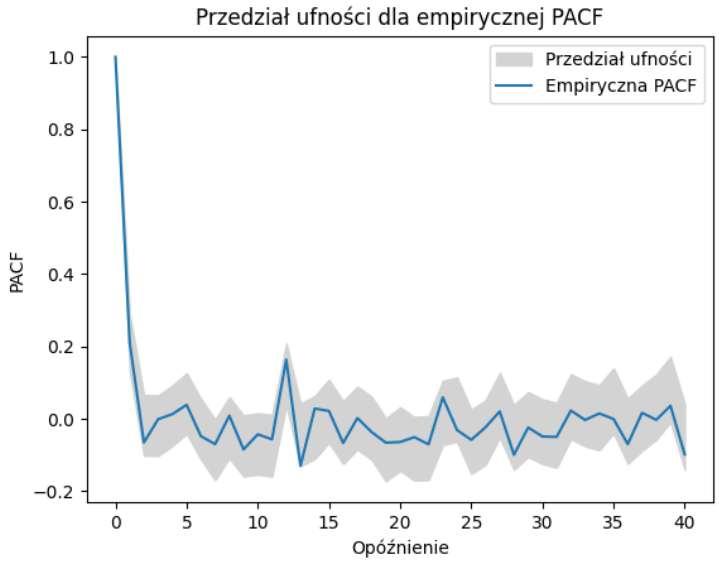
Rysunek 11: Wykres przedstawiający empiryczne i teoretyczne ACF dla oczyszczonych danych.



Rysunek 12: Wykres przedstawiający empiryczne i teoretyczne PACF dla oczyszczonych danych.



Rysunek 13: Wykres przedstawiający dla empirycznej ACF dla oczyszczonych danych.



Rysunek 14: Wykres przedstawiający dla empirycznego PACF dla oczyszczonych danych.

4.2 Linie kwantylowe

Definicja. *Kwantyl [7]*
Kwantylem rzędu p, gdzie $0 \leq p \leq 1$, w rozkładzie empirycznym P_X zmiennej losowej X nazywamy każdą liczbę x_p, dla której spełnione są nierówności

$$P_X([-\infty, x_p]) \geq p$$

oraz

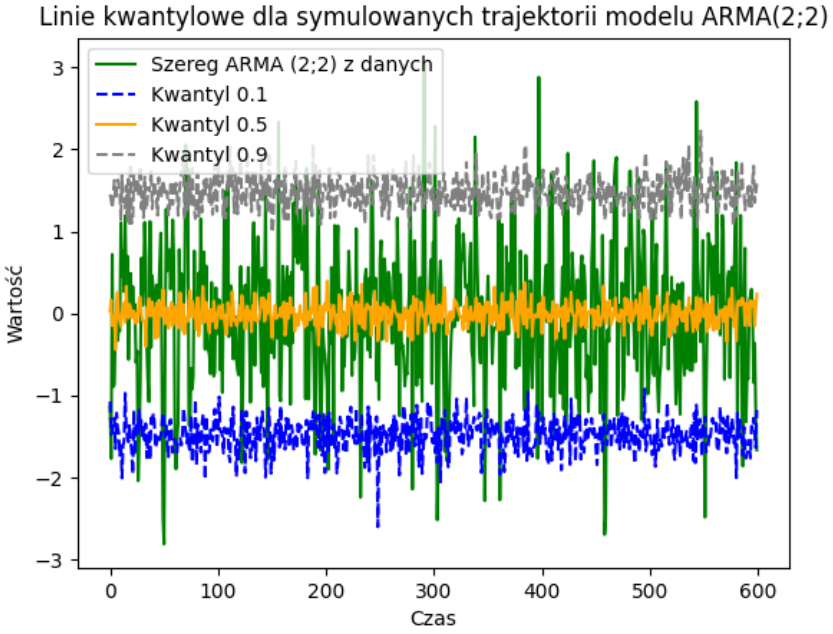
$$P_X([x_p, infty]) \geq 1 - p$$

W szczególności, kwantylem rzędu p jest taka wartość x_p zmiennej losowej, że wartości mniejsze lub równe od x_p są przyjmowane z prawdopodobieństwem co najmniej p, zaś wartości większe lub równe od x_p są przyjmowane z prawdopodobieństwem co najmniej $1 - p$.

Analizując dane tworzymy też wykres linii kwantylowych (powstałych z obliczania kwantyli symulowanych trajektorii ARMA(2;2)), który nakładamy na wykres szeregu $\{Y_t\}$.

Grafika klarownie przedstawia nasze dane (Rys. 15); widzimy, że duża część obserwacji znajduje się między liniami kwantylowymi 0,1 i 0,9, a linie kwantylowe są rozmieszczone w mniej więcej równych odstępach - możemy stąd wywnioskować, że dane są rozłożone symetrycznie. Warto też zwrócić uwagę na fakt, że wartości ekstremalne występują dość rzadko (obserwacja powyżej kwantyla 0,9

lub poniżej kwantyla 0,1). Ponadto linia kwantylowa 0,5 (czyli średnia) oscyluje w okolicach zera. A skoro dane są rozłożone symetrycznie, wartości ekstremalne występują rzadko, a średnia rozkładu jest w przybliżeniu równa zeru, to możemy przypuszczać, że dane pochodzą z rozkładu normalnego. Faktycznie, po wykonaniu testu Kołmogorowa-Smirnowa dla naszych danych otrzymujemy p-wartość równą $0.487 > 0{,}05$, czyli możemy odrzucić hipotezę o braku normalności rozkładu.



Rysunek 15: Wykres przedstawiający linie kwantylowe i oczyszczone dane

5 Szum

5.1 Wykres wartości resztowych

Definicja. *Wartości resztowe*
W statystyce residuum (reszta) to różnica między obserwowaną wartością a wartością przewidywaną przez model statystyczny. W kontekście regresji liniowej, dla każdej obserwacji i, residuum e_i oblicza się jako:

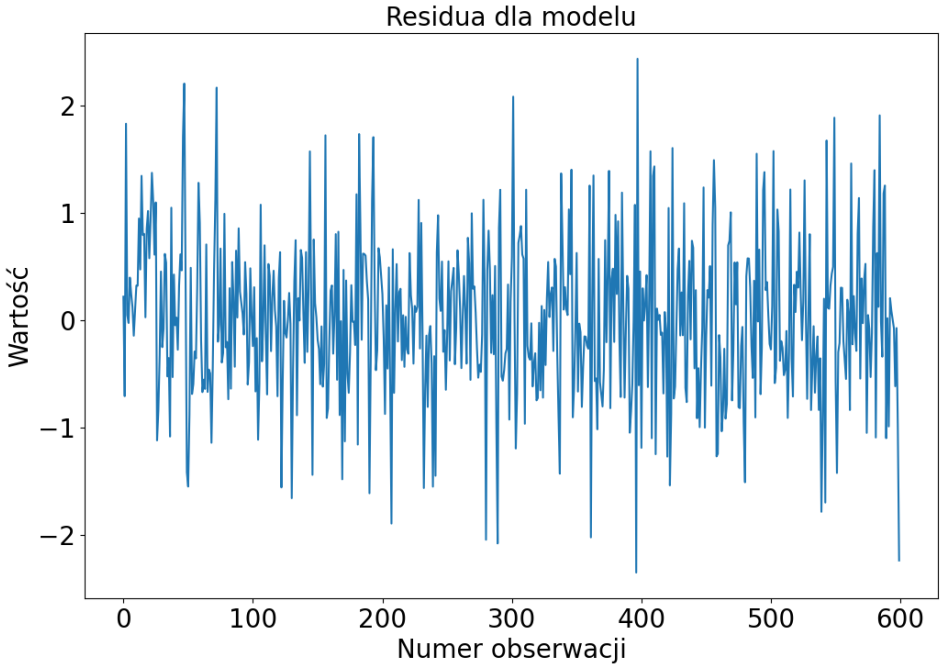
$$e_i = y_i - \hat{y}_i$$

gdzie:

- y_i *to wartość obserwowana dla i-tej jednostki,*

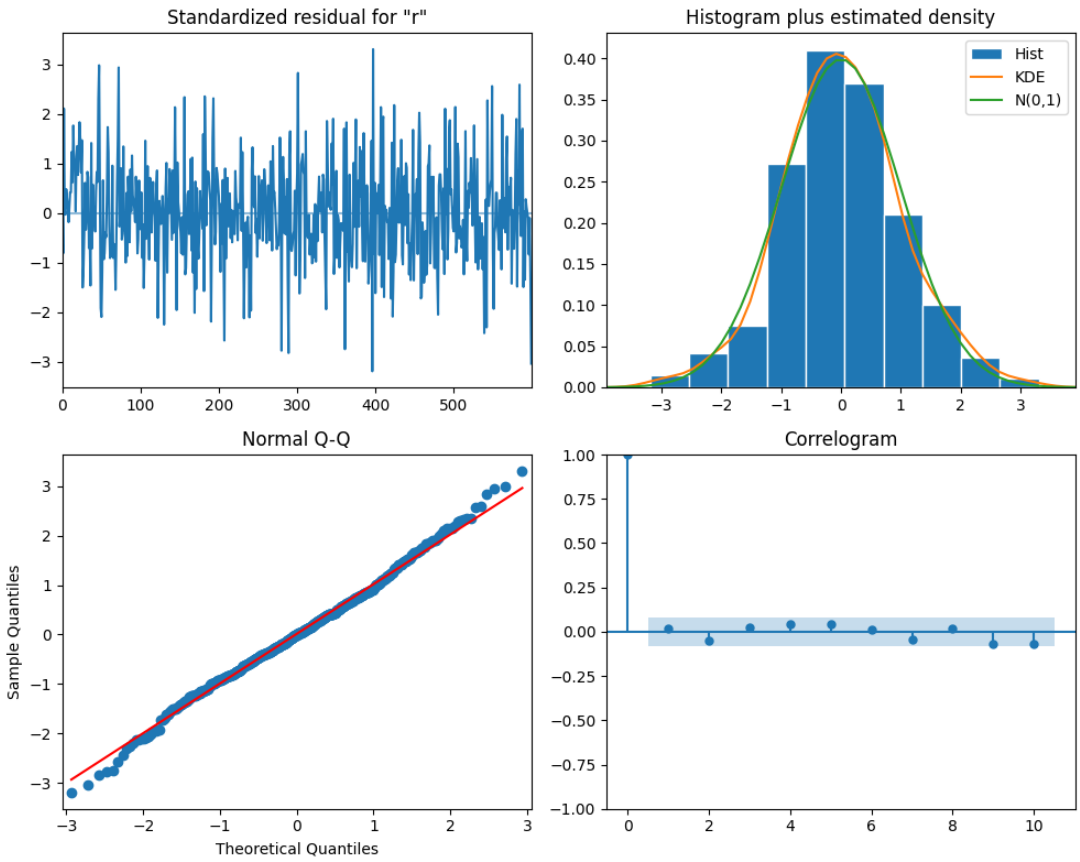
- $\hat{y}_i$ *to wartość przewidywana przez model dla tej samej jednostki [8].*

Istotnym czynnikiem podczas badania szeregów czasowych jest badanie residuów szeregu. W naszym szeregu $\{Y_t\}$ residua wyodrębniliśmy za pomocą funkcji wbudowanej w Pythonie, w wyniku czego otrzymaliśmy wykres wartości resztowych (Rys. 16).



Rysunek 16: Wykres przedstawiający residua dla modelu

Za pomocą funkcji `plot_diagnostics` tworzymy wykresu, które pomagają graficznie ocenić, czy residua pochodzą z rozkładu normalnego, czy z rozkładu t-Studenta (Rys. 17).



Rysunek 17: Wykres przedstawiający porównanie rozkładów na podstawie residuów

5.2 Dystrybuanta i gęstość szumu

Definicja. *Dystrybuanta*
funkcja rzeczywista jednoznacznie wyznaczająca rozkład prawdopodobieństwa, a więc zawierająca wszystkie informacje o tym rozkładzie. [9]
Niech $\mathbb{P}$ będzie rozkładem prawdopodobieństwa na prostej. Funkcję $F\colon \mathbb{R} \to \mathbb{R}$ daną wzorem

$$F(t) = \mathbb{P}((-\infty, t]),$$

nazywamy dystrybuantą rozkładu $\mathbb{P}$.

Definicja. *Gęstość prawdopodobieństwa*
nieujemna funkcja rzeczywista, określona dla rozkładu prawdopodobieństwa, taka

że całka z tej funkcji, obliczona w odpowiednich granicach, jest równa prawdopodobieństwu wystąpienia danego zdarzenia losowego. Funkcję gęstości definiuje się dla rozkładów prawdopodobieństwa jednowymiarowych i wielowymiarowych. Rozkłady mające gęstość nazywane są rozkładami ciągłymi. [10]
Niech P będzie rozkładem prawdopodobieństwa w przestrzeni $\mathbb{R}^N$ (w szczególności na prostej rzeczywistej $\mathbb{R}$).

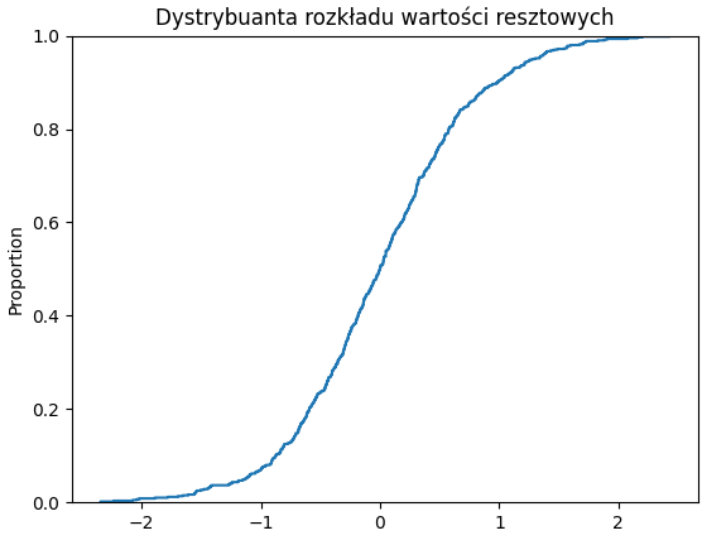
Gęstością rozkładu prawdopodobieństwa P nazywa się nieujemną funkcję borelowską $f\colon \mathbb{R}^N \to \mathbb{R}_+ \cup \{0\}$, taką, że dla każdego zbioru borelowskiego $B \subseteq \mathbb{R}^N$ zachodzi równość:

$$P(B) = \int_B f(x)\,dx,$$

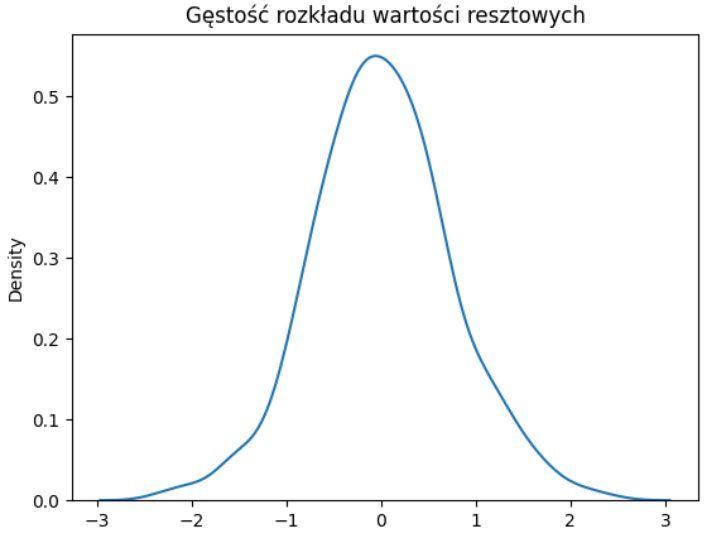
tzn. całka z funkcji f obliczona na zbiorze B jest równa prawdopodobieństwu $P(B)$ przypisanemu zbiorowi B.

Definicja. *Histogram prawdopodobieństwa*
(znany również jako histogram gęstości) jest to szczególny przypadek histogramu, w którym wysokości słupków odpowiadają gęstości prawdopodobieństwa. Jest to narzędzie służące do estymacji rozkładu prawdopodobieństwa zmiennej losowej na podstawie próbki danych [12].

Szczególnie istotny jest wykres Normal Q-Q. Wiemy, że rozkład t-Studenta ma ciężkie ogony, co na wykresie *Q-Q plot* charakteryzuje się „rozjazdem" wartości obserwacji od linii reprezentujące wartości z rozkłądu normalnego. W tym wypadku odchylenia wartości obserwacji od teoretycznych wartości rozkładu normalnego są niewielkie, co oznacza, że szum nie pochodzi z rozkładu t-Studenta. Natomiast na histogramie widzimy, że gęstość rozkładu naszych danych w znacznej mierze pokrywa się z gęstością rozkładu normalnego. Jako ostanie „graficzne potwierdzenie" normalności rozkładu wartości resztowych mogą posłużyć wykresy dystrybuanty i gęstości (Rys. 18,19).



Rysunek 18: Wykres przedstawiający dystrybuantę szumu



Rysunek 19: Wykres przedstawiający gęstość szumu

5.3 Testy statystyczne

Oczywiście „graficzne potwierdzenie" nie jest dowodem na normalność rozkładu - wszak rozkład normalny i rozkład t-Studenta mają bardzo podobne wykresy gęstości i dystrybuanty. Z tego powodu należy też przeprowadzić testy statystyczne. Przeprowadziliśmy dwa testy statystyczne: test Kołmogorowa-Smirnowa oraz test Andersona-Darlinga. Wyniki obu testów są następujące:

1. Test Kołmogorowa-Smirnowa na normalność
   - Statystyka testu: 0,029
   - p-wartość: 0,678
2. Test Andersona-Darlina
   - Statystyka testu: 0,714
   - Wartości krytyczne: 0,572; 0,652; 0,782; 0,912; 1,085

Możemy zauważyć, że p-wartość w teście Kołmogorowa-Smirnowa na normalność jest większa niż 0,05, a zatem możemy odrzucić hipotezę o braku normalności rozkładu. Podobnie w teście Andersona-Darlinga wartość statystyki testu jest mniejsza od wartości krytycznej na poziomie istotności 0,05, co oznacza, że badany rozkład jest rozkładem normalnym. Potwierdzamy zatem nasze przypuszczenia o normalności, które, póki co, mogliśmy poprzeć jedynie „rysunkami".

W celu dalszej analizy szumu przeprowadziliśmy także inne testy statystyczne: T-Test dla zbadania średniej rozkładu, zmodyfikowany test Levene'a i ARCH test dla zbadania wariancji rozkładu oraz test Ljunga-Boxa dla zbadania niezależności.

1. T-Test
   - Statystyka T: 0,535
   - p-wartość: 0,593

   Wykonując T-Test porównaliśmy szum szeregu $\{Y_t\}$ z szumem wygenerowanego komputerowo modelu ARMA(2;2). Od razu widzimy, że p-wartość jest większa od 0,05 (odrzucamy hipotezę zerową móiącą, że średnie próbek są różne), a odpowiadająca jej statystyka T przyjmuje małą wartość - nie ma więc istotnego odstępstwa od średniej. Zatem jakiekolwiek zauważalne różnice są jedynie wypadkiem losowego szumu o średniej równej zeru.

2. Modified Levene Test i ARCH test

   2.1. Modified Levene Test (MLT)
   - Statystyka Levene'a: 46,828
   - p-wartość: $1,23 \cdot 10^{-11}$

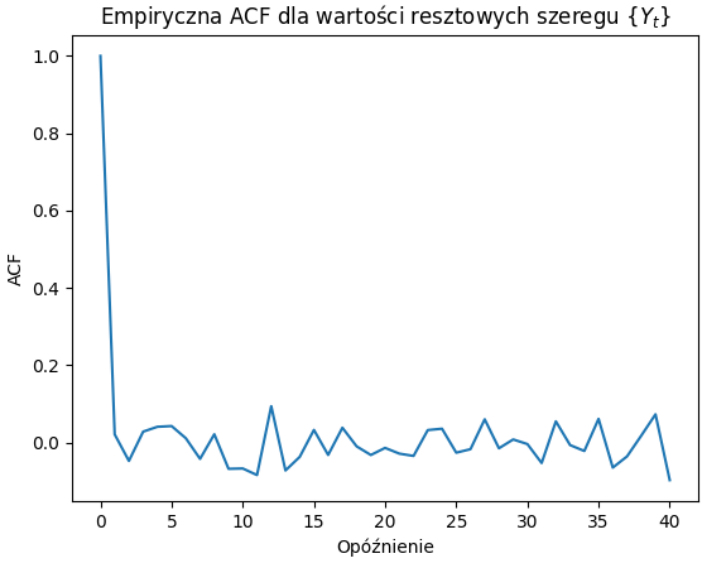
   2.2. ARCH test

- Statystyka F: 2,22
- : p-wartość: 0,001

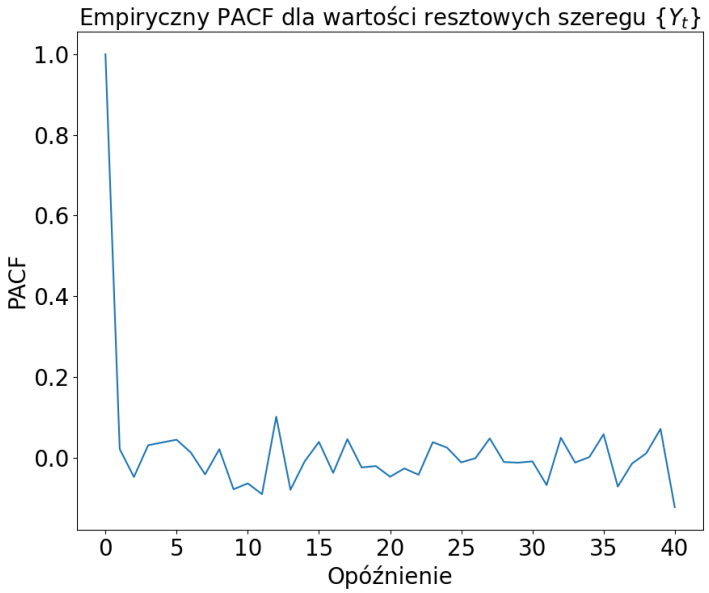
W wypadku MLT porównaliśmy szum z szeregu $\{Y_t\}$ z wygenerowanym szeregiem ARMA(2;2), o którym już wspominaliśmy. Widzimy, że p-wartość jest niemal równa zeru, czyli wariancje obu próbek znacząco się różnią. Wyniki otrzymane w teście ARCH wskazują zaś na to, że wariancja reszt jest stała.

3. Test Ljunga-Boxa

   W wypadku testu Ljunga-Boxa p-wartości rosły wraz z każdą kolejną obserwacją szumu - zaczynając od 0,038 a skończywszy na 0,311. Wysokie p-wartości w teście Ljunga-Boxa wskazują, że badane dane nie są autoskorelowane. W naszym wypadku jest to tylko potwierdzenie faktu, który widzimy wykresach ukazujących empiryczną funkcję autokorelacji i empiryczną funkcję częściowe autokorelacji dla wartości resztowych - od pewnego momentu wartości były statystycznie równe zeru, co oznacza, że obecne wartości szumu nie wpływają na jego przyszłe wartości (Rys. 20,21).



Rysunek 20: Wykres przedstawiający empiryczną ACF dla szumu



Rysunek 21: Wykres przedstawiający empiryczny PACF dla szumu

Oczywiście, niejako widzieliśmy te wykresy już wcześniej na korelogramie (Rys. 17). Zbierając zebrane przez nas informacje, wiemy, że szum szeregu $\{Y_t\}$ ma stałą w czasie średnią i wariancję oraz jego funkcja autokorelacji jest statystycznie różna od zera tylko dla opóźnienia równego zeru. Zatem szum szeregu $\{Y_t\}$ jest białym szumem.

6 Podsumowanie i wnioski

Badane przez nas dane dotyczyły średniej miesięcznej temperatury w Pakistanie w latach 1901 - 1950. W celu zbadania danych i modelowania ich za pomocą modelu ARMA dokonaliśmy dekompozycji tego szeregu, który, po „oczyszczeniu" z trendu i sezonowości, okazał się być szeregiem stacjonarnym. Następnym krokiem było znalezienie rzędu modelu i wyestymowanie parametrów - w naszym wypadku rząd modelu wynosił 2,2, a współczynniki wielomianów AR i MA przyjmowały wartości $\phi_1 = 1,5791$,$\phi_2 = -0,7162$, $\theta_1 = -1,9071$, $\theta_2 = 0,9075$. Wielomiany te nie miały wspólnych pierwiastków, wobec czego mogliśmy modelować nasze dane modelem ARMA. Badając szereg $\{Y_t\}$ doszliśmy do wniosku, że ma on rozkład symetryczny, a jego wartości resztowe, będące białym szumem, mają rozkład normalny. Co więcej, residua szeregu $\{Y_t\}$ są nieskorelowane, co wykazaliśmy wykonując test Ljunga-Boxa. Normalność szumu mogliśmy także zaobserwować na wykresach gęstości czy dystrybuanty rozkładu wartości resztowych.

Model ARMA(2;2), który uzyskaliśmy z danych, jest niemal podręcznikowym przykładem szeregu ARMA - jest stacjonarny, odwracalny, przyczynowy, pochodzi z rozkładu normalnego, wielomiany AR i MA nie mają wspólnych pierwiastków, jego wartości resztowe są białym szumem o rozkładzie normalnym. Pamiętajmy jednak, że zjawisko, które badaliśmy, dotyczyło zjawiska pogodowego, a dane pogodowe wręcz idealnie „wpasowują się" w model ARMA. W tej pracy przyglądaliśmy się bliżej wyłącznie średniej miesięcznej temperaturze w Pakistanie, a wysokość temperatury powietrza na jakimś poziomie jest zjawiskiem niejasko okresowym - każdej wiosny waha się ona między pewnymi wartościami, podobnie latem, jesienią czy zimą; w danym miesiącu również temperatura powietrza przyjmuje wartości z pewnego typowego dla danego okresu przedziału. Jest to więc szereg z natury stacjonarny, mający stałą w czasie średnią i wariancję. I choć wartości ekstremalne mogą się zdarzać, to należą one to rzadkości. Dane pogodowe są więc „wdzięcznym" kandydatem do modelowania z racji swojej powtarzalności - a na pewno było tak na początku XX wieku. W dzisiejszych czasach, w epoce zmian klimatu, takie modelowanie staje się coraz trudniejsze, ponieważ dużo częściej występują zjawiska ekstremalne - bardzo wysokie lub niskie temperatury, wyjątkowo dużo lub mało opadów atmosferycznych... Modelowanie pogody staje się coraz trudniejszym wyzwaniem, któremu będziemy musieli sprostać.

7 Bibliografia

Korzystaliśmy z języka programowania Python do pomocy w obliczeniach i wizualizacji wykresów.

Literatura

[1] Dane z Kaggle, `https://www.kaggle.com/datasets/zusmani/pakistan-temperature`,
dostęp: 31 stycznia 2025 r.

[2] Wikipedia ACF, `https://en.wikipedia.org/wiki/Autoregressive_moving-average_model`,
dostęp: 31 stycznia 2025 r.

[3] Wikipedia PACF, `https://pl.wikipedia.org/wiki/Funkcja_autokorelacji_cz%C4%85stkowej`,
dostęp: 31 stycznia 2025 r.

[4] Wikipedia przedziały ufności, `https://pl.wikipedia.org/wiki/Przedzia%C5%82_ufno%C5%9Bci`,
dostęp: 31 stycznia 2025 r.

[5] Wikipedia ADF, https://en.wikipedia.org/wiki/Augmented_Dickey%E2%80%93Fuller_test,
dostęp: 31 stycznia 2025 r.

[6] Wikipedia ARMA, https://pl.wikipedia.org/wiki/Autoregresja,
dostęp: 31 stycznia 2025 r.

[7] Wikipedia kwantyl, https://pl.wikipedia.org/wiki/Kwantyl,
dostęp: 31 stycznia 2025 r.

[8] Wartości resztowe, https://www.math.uni.wroc.pl/,
dostęp: 31 stycznia 2025 r.

[9] Wikipedia dystrybuanta, https://pl.wikipedia.org/wiki/Dystrybuanta,
dostęp: 31 stycznia 2025 r.

[10] Wikipedia gęstość, https://pl.wikipedia.org/wiki/Funkcja_g%C4%99sto%C5%9Bci_prawdopodobie%C5%84stwa,
dostęp: 31 stycznia 2025 r.

[11] Wikipedia histogram, https://pl.wikipedia.org/wiki/Histogram,
dostęp: 31 stycznia 2025 r.

[12] Wikipedia histogram, https://pl.wikipedia.org/wiki/Warto%C5%9B%C4%87_p,
dostęp: 31 stycznia 2025 r.