Sprawozdanie

Arkadiusz Urbaniak (276034), Hubert Zawerbny (276001)

31 stycznia 2025

Spis treści

1 Wstęp

1.1 Cel raportu

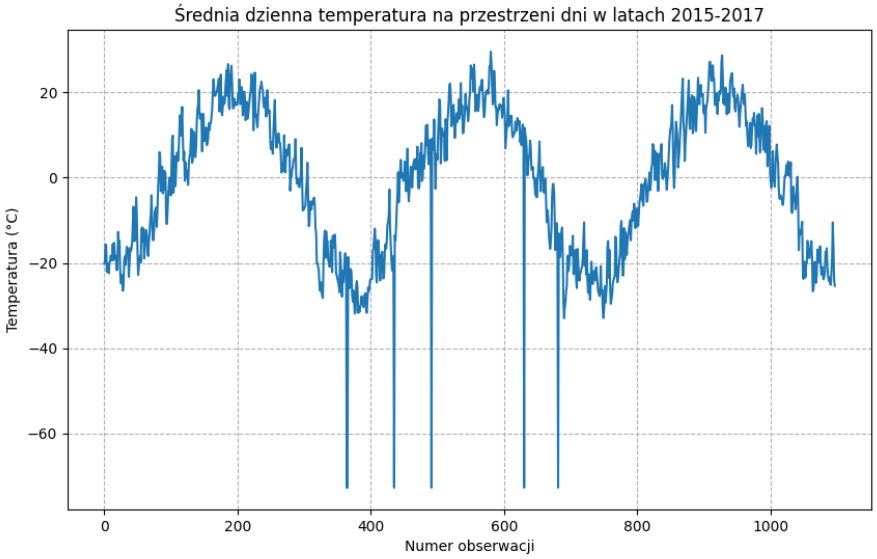
Celem naszego raportu jest przeanalizowanie danych, przy pomocy modelu ARMA, dotyczących temperatury w wybranym mieście, aby móc zrozumieć wzorce sezonowe oraz trendy długoterminowe.

1.2 Informacje o danych

Dane pozyskaliśmy ze strony kaggle. W wybranym przez nas zbiorze znajdują się średnie dzienne temperatury dla głównych miast na świecie na przełomie XX i XXI wieku. My natomiast skupimy się na analizie średniej dziennej temperatury w stolicy Mongolii - Ułan-Bator w latach 2015 - 2017. Wybraliśmy to miasto ze względu na wysoką wariancję temperatur w tym regionie. Dane liczą 1096 obserwacji.

1.3 Wizualizacja danych

Zacznijmy od wizualizacji danych, na których chcemy opierać naszą pracę.



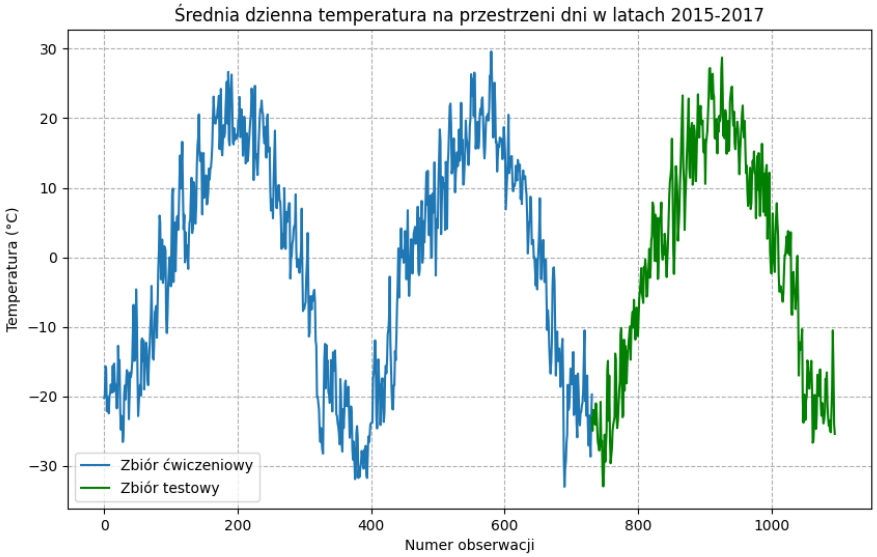
Rysunek 1: Wizualizacja danych

Jak widać na zamieszczonym wykresie, dane posiadają kilka odstających obserwacji, które uważamy za błędne, na podstawie znajomości klimatu w analizowanym przez nas mieście. Uważamy, że doszło w tych przypadkach do błędnego odczytu temperatury. Ponieważ błędne odczyty zdarzały się niekiedy dwa dni z rzędu, zdecydowaliśmy się na zastąpienie ich pomiarami z dnia poprzedniego zamiast średniej z dnia poprzedniego oraz następnego.

2 Przygotowanie danych do analizy

Po oczyszczeniu danych, zdecydowaliśmy się na podział obserwacji na testowe i treningowe. Jako dane treningowe przyjęliśmy pomiary z lat 2015-2016 (731 obserwacji), natomiast za dane testowe posłużą nam obserwacje z roku

2017 (365 obserwacji). Od tego momentu dane widoczne na wykresie poniżej będziemy nazywać surowymi.



Rysunek 2: Podział na zbiór ćwiczeniowy i testowy

2.1 Dekompozycja szeregu czasowego

Zacznijmy od zdefiniowania czym jest estymowana funkcja autokorelacji (ACF) oraz częściowej autokorelacji (PACF).

2.1.1 Estymowana funkcja autokorelacji (ACF)

Autokorelacja dla opóźnienia k ($\hat{\rho}_k$) jest definiowana jako:

$$\hat{\rho}_k = \frac{\hat{\gamma}_k}{\hat{\gamma}_0},$$

gdzie:

- $\hat{\gamma}_k$ to estymowana autokowariancja dla opóźnienia k,

- $\hat{\gamma}_0$ to wariancja (autokowariancja dla opóźnienia 0).

Estymowana autokowariancja $\hat{\gamma}_k$ obliczana jest według wzoru:

$$\hat{\gamma}_k = \frac{1}{n} \sum_{t=k+1}^{n} (X_t - \bar{X})(X_{t-k} - \bar{X}),$$

gdzie:

- X_t to wartość szeregu czasowego w chwili t,
- $\bar{X}$ to średnia arytmetyczna szeregu,
- n to liczba obserwacji.

2.1.2 Estymowana funkcja częściowej autokorelacji (PACF)

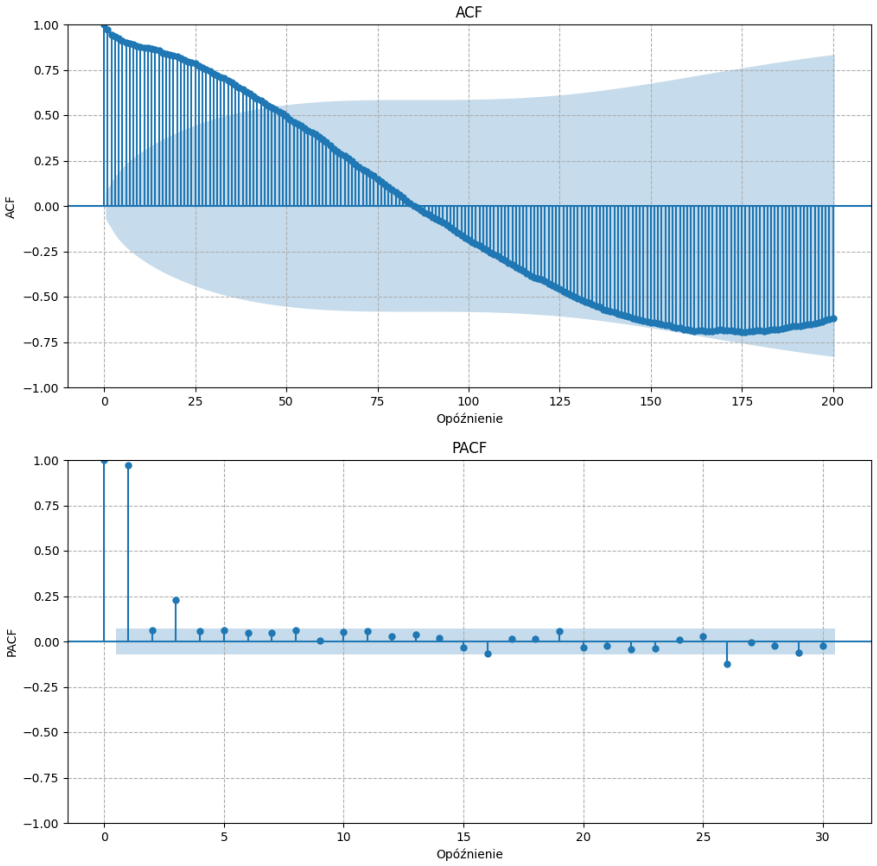
Częściowa autokorelacja dla opóźnienia k ($\hat{\phi}_{kk}$) jest wyznaczana jako współczynnik korelacji reszt między X_t i X_{t-k}, po usunięciu wpływu wszystkich pośrednich opóźnień $1, 2, \ldots, k-1$.

Wartości PACF są estymowane poprzez regresję wielokrotną:

1. Regresja X_t na $X_{t-1}, X_{t-2}, \ldots, X_{t-k}$,
2. Współczynnik przy X_{t-k} to $\hat{\phi}_{kk}$.

2.1.3 Estymowane ACF oraz PACF dla surowych danych

Zanim przystąpimy do dekompozycji szeregu czasowego, przeanalizujmy funkcję autokorelacji i częściowej autokorelacji dla surowych danych.



Rysunek 3: Wykres ACF i PACF dla surowych danych

Na podstawie wykresu autokorelacji możemy przypuszczać, że dane posiadają silną sezonowość, ponieważ wartości autokorelacji zachowują się okresowo. Co więcej, duże wartości funkcji autokorelacji mogą oznaczać, że dane są ze sobą mocno skorelowane. Z tych samych powodów możemy również zakładać, że szereg czasowy nie jest stacjonarny. Aby potwierdzić te przypuszczenia, przeprowadzimy test ADF na niestacjonarność.

Z drugiego wykresu jesteśmy w stanie wyczytać, że wartości częściowej funkcji autokorelacji bardzo szybko spadają do zera, a następnie wokół niego oscylują. Sugerować to może, że odpowiednim modelem do opisu naszych danych będzie model AR(p).

2.1.4 Test ADF dla surowych danych

Test Augmented Dickey-Fuller (ADF) służy do sprawdzenia, czy szereg czasowy jest stacjonarny. Hipotezy testu ADF są sformułowane następująco:

- H_0: Szereg czasowy ma pierwiastek jednostkowy, czyli jest niestacjonarny.
- H_1: Szereg czasowy nie ma pierwiastka jednostkowego, czyli jest stacjonarny.

Regresja wykonywana w ramach testu ADF jest opisana równaniem:

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \sum_{i=1}^{p} \phi_i \Delta y_{t-i} + \epsilon_t$$

gdzie:

- $\Delta y_t = y_t - y_{t-1}$: różnicowanie szeregu czasowego,
- α: wyraz wolny,
- βt: trend deterministyczny,
- y_{t-1}: wartość opóźniona szeregu,
- ϕ_i: współczynniki autoregresji,
- ϵ_t: składnik losowy,
- p: liczba opóźnień dobrana według kryteriów informacyjnych.

Statystyka testowa ADF wyrażona jest wzorem:

$$ADF = \frac{\hat{\gamma}}{\mathrm{SE}(\hat{\gamma})}$$

gdzie:

- $\hat{\gamma}$: estymator współczynnika γ,
- $\mathrm{SE}(\hat{\gamma})$: odchylenie standardowe estymatora $\hat{\gamma}$.

Wykonaliśmy powyższy test przy pomocy funkcji *adfuller* z pakietu *statsmodels* w pythonie. Wartość statystyki testowej wyniosła w przybliżeniu $ADF = -1{,}57$, natomiast p-wartość $p = 0{,}50$. Oznacza to, że na przyjętym przez nas poziomie istotności $\alpha = 0{,}05$ nie mamy podstaw do odrzucenia hipotezy zerowej i stwierdzamy, że szereg czasowy złożony z surowych danych nie jest stacjonarny.

2.1.5 Identyfikacja trendów deterministycznych

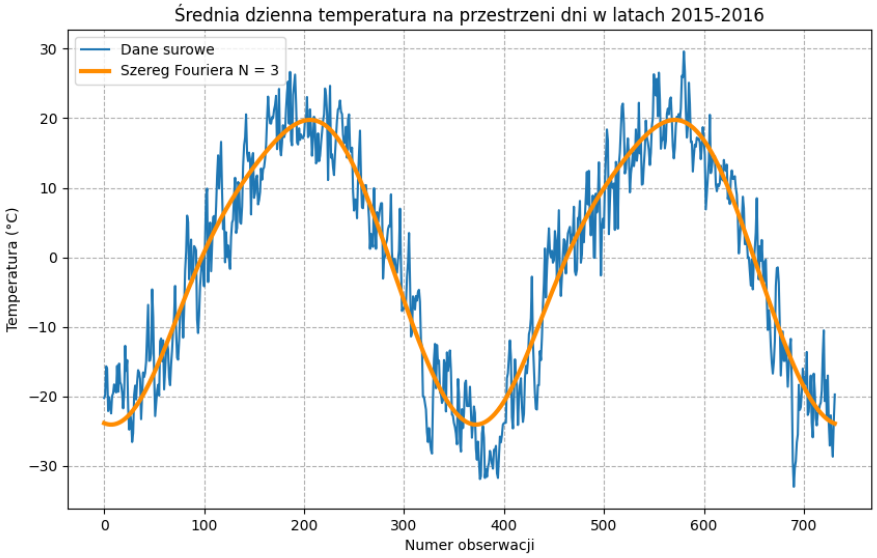
Posiadane przez nas dane mają silną sezonowość, dlatego na tym etapie skupimy się na jej usunięciu. W tym celu zdecydowaliśmy się na użycie szeregu Fouriera danego wzorem:

$$f(t) = a_0 + \sum_{k=1}^{N} \left[a_k \cos\left(\frac{2\pi k t}{T}\right) + b_k \sin\left(\frac{2\pi k t}{T}\right) \right]$$

gdzie:

- a_0 – średnia wartość sygnału (stała),
- a_k – amplituda składnika cosinusowego dla k-tej fali harmonicznej,
- b_k – amplituda składnika sinusowego dla k-tej fali harmonicznej,
- N - liczba fal,
- k – numer fali,

- T – okres sygnału (tutaj 365 dni),
- t – czas



Rysunek 4: Sezonowość danych

Na przedstawionym powyżej wykresie w szeregu Fouriera zdecydowaliśmy się na ustawienie parametru $N = 3$. Uważamy, że jest to kompromis pomiędzy zbyt słabym dopasowaniem a nadmiernym dopasowaniem i takie rozwinięcie szeregu Fouriera będzie optymalne do średniej dziennej temperatury. Parametry a_k, b_k, a_0 dobrane zostały natomiast za pomocą regresji liniowej. Tak wyliczoną sezonowość usuwamy z naszych danych, aby otrzymać szereg stacjonarny.

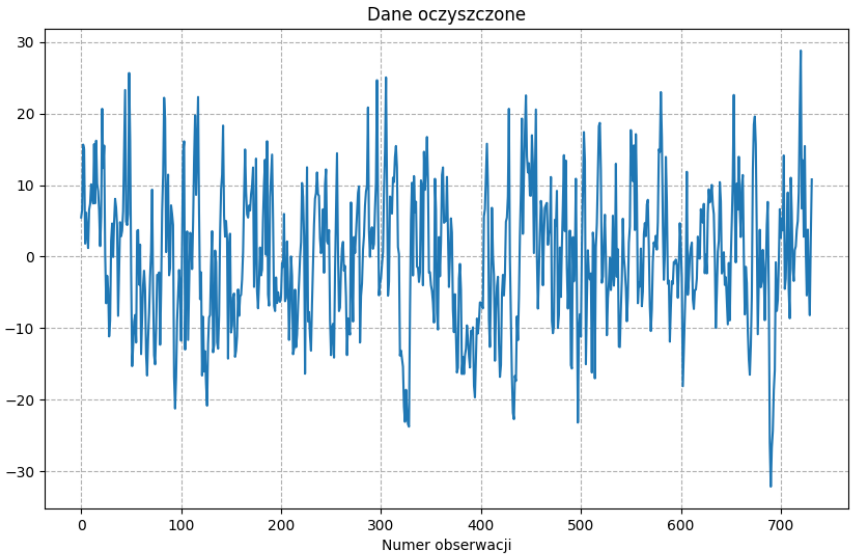
Zdecydowaliśmy się także usunąć trend, wykorzystując prostą regresji. Na całość, w celu ustabilizowania wariancji, nałożyliśmy transformację Boxa-Coxa daną następującym wzorem:

$$y_i^{(\lambda)} = \begin{cases} \frac{y_i^{\lambda}-1}{\lambda}, & \text{dla } \lambda \neq 0, \\ \ln(y_i), & \text{dla } \lambda = 0, \end{cases}$$

gdzie:

- y_i - obserwacja, która podlega transformacji
- λ - parametr transformacji, który kontroluje stopień nieliniowości.

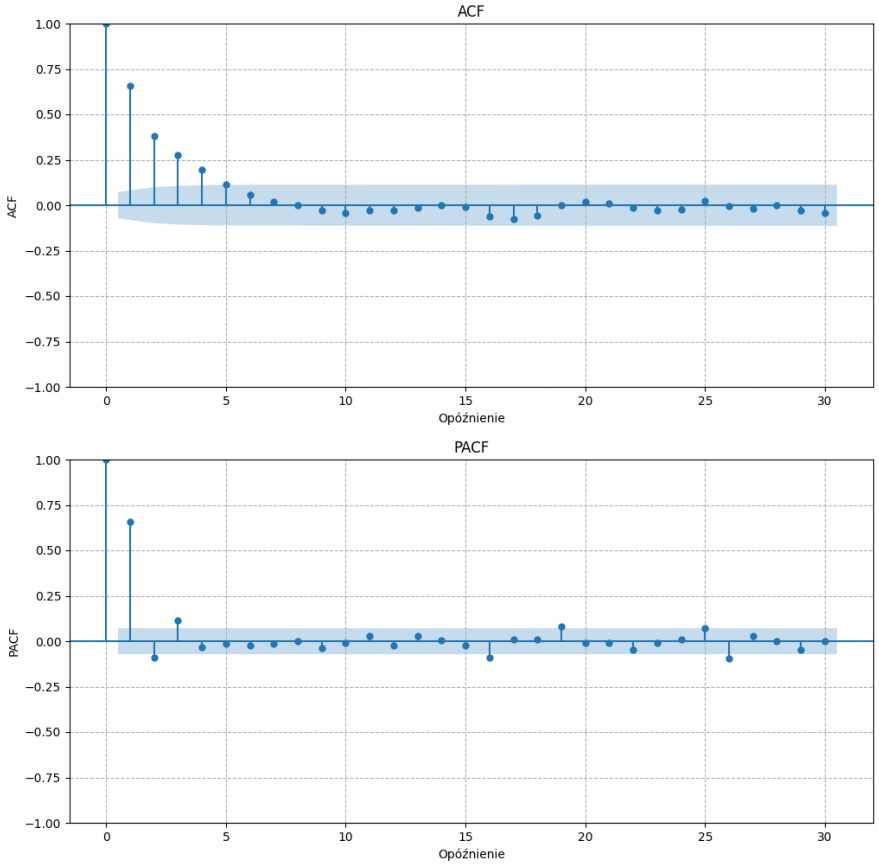
Po wykonaniu wyżej wymienionych transformacji dane prezentują się następująco.



Rysunek 5: Dane po usunięciu sezonowości

Dane z powyższego wykresu będziemy w dalszej części pracy nazywać oczyszczonymi.

2.1.6 ACF oraz PACF dla oczyszczonych danych



Rysunek 6: Wykres ACF i PACF dla oczyszczonych z sezonowości danych

Analizując wykres autokorelacji, zauważyć możemy, że dane po oczyszczeniu wykazują zdecydowanie mniejszą sezonowość oraz, ze względu na szybciej zbiegające do 0 wartości, możemy również stwierdzić, że obserwowalna jest mniejsza zależność danych między sobą.

Na podstawie wykresu autokorelacji częściowej dla danych oczyszczonych możemy spodziewać się, że najbardziej pasujący parametr autoregresyjny modelu ARMA przyjmie wartość całkowitą z przedziału $[1; 3]$. Uważamy tak, ponieważ dla tych wartości opóźnienia PACF znajduje się poza przedziałem ufności.

2.1.7 Test ADF dla oczyszczonych danych

Wykonajmy ponownie test ADF, tym razem dla danych oczyszczonych. Wartość statystyki testowej wynosi w przybliżeniu $ADF = -9{,}71$, a p-wartość $p = 1{,}03 \times 10^{-6}$. Na poziomie istotności $\alpha = 0{,}05$ odrzucamy H_0 na rzecz H_1, co prowadzi nas do wniosku, że szereg czasowy powstały z oczyszczonych danych jest stacjonarny.

Na podstawie zamieszczonych wyżej wykresów ACF i PACF oraz testu ADF, możemy stwierdzić, że oczyszczanie danych z sezonowości przyniosło zamierzone efekty.

3 Modelowanie danych przy pomocy ARMA

Modelowanie danych przy pomocy ARMA zacznijmy od dobrania rzędu modelu. W tym celu posłużymy się kryteriami informacyjnymi takimi jak AIC (Akaike Information Criterion), BIC (Bayesian Information Criterion) oraz HQIC (Hannan-Quinn Information Criterion), które dane są następującymi wzorami:

$$AIC = -2\ln(L) + 2k$$

$$BIC = -2\ln(L) + \ln(n) \cdot k$$

$$HQIC = -2\ln(L) + 2 \cdot k \cdot \ln(\ln(n))$$

gdzie:

- n – liczba punktów w danych
- L – to maksymalna funkcja wiarygodności modelu (likelihood),
- k – to liczba parametrów w modelu (p+q)

Aby wybrać najbardziej optymalne współczynniki p i q dla modelu ARMA(p,q) obliczamy podane wyżej statystyki dla ich różnych wartości z przedziału $[0; 5]$, ponieważ po analizie funkcji częściowej autokorelacji uważamy, że w tym przedziale znajdują się współczynniki najlepiej pasujące do modelu. Następnie wyniki porównujemy między sobą i zgodnie z zasadą, iż im niższa wartość kryterium, tym lepszy model (w sensie balansu pomiędzy dopasowaniem a złożonością), wybieramy ten, który najczęściej posiadał najmniejsze wartości podanych kryteriów.

p	q	AIC	BIC	HQIC
3	0	4982.82	5005.80	4991.69
1	2	4982.90	5005.88	4991.77
4	0	4984.05	5011.62	4994.68
3	1	4984.20	5011.77	4994.83
4	1	4984.60	5016.76	4997.00

Tabela 1: Kryteria informacyjne

Na podstawie powyższych statystyk decydujemy się na wybranie modelu ARMA o parametrze autoregresji równym 3 i parametrze średniej ruchomej równym 0.

Aby oszacować współczynniki naszego modelu ARMA(3,0) skorzystamy z faktu, iż jest to model AR(3). Z tego powodu wykorzystamy metodę Yule-Walkera estymacji parametrów szeregu autoregresyjnego.

3.1 Metoda Yule-Walkera

Metoda Yule'a-Walkera opiera się na równaniach Yule'a-Walkera, które są wyprowadzone na podstawie własności stacjonarnego procesu autoregresyjnego.

Rozważmy stacjonarny proces autoregresyjny rzędu p (AR(p)) opisany równaniem:

$$X_t = \sum_{k=1}^{p} \varphi_k X_{t-k} + \varepsilon_t,$$

gdzie ε_t to biały szum o zerowej wartości oczekiwanej i wariancji σ^2.

Dla procesu AR(p) autokowariancja γ_h spełnia równania Yule'a-Walkera:

$$\gamma_h = \sum_{k=1}^{p} \varphi_k \gamma_{h-k}, \quad \forall h \geqslant 0.$$

Dzieląc przez γ_0 otrzymujemy równania w postaci współczynników autokorelacji:

$$\rho_h = \sum_{k=1}^{p} \varphi_k \rho_{h-k}, \quad \forall h \geqslant 0.$$

Dla $h = 1, 2, \ldots, p$ powyższe równania można zapisać w formie układu równań liniowych:

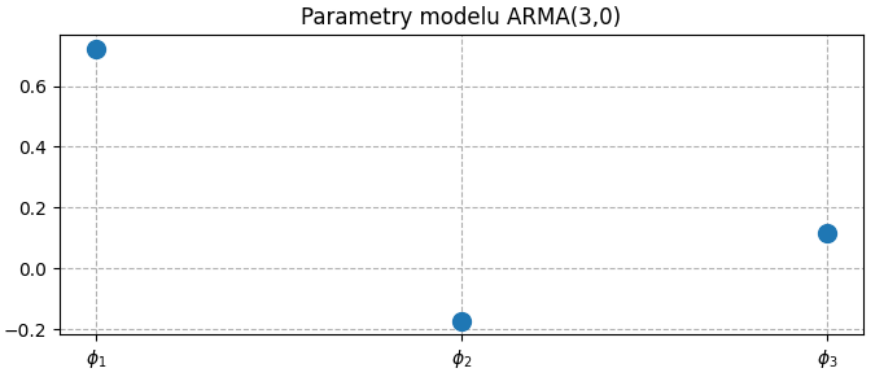
$$\begin{bmatrix} \rho_1 \\ \rho_2 \\ \vdots \\ \rho_p \end{bmatrix} = \begin{bmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{p-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{p-2} \\ \rho_2 & \rho_1 & 1 & \cdots & \rho_{p-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{p-1} & \rho_{p-2} & \rho_{p-3} & \cdots & 1 \end{bmatrix} \begin{bmatrix} \varphi_1 \\ \varphi_2 \\ \vdots \\ \varphi_p \end{bmatrix}.$$

Macierz układu jest dodatnio określona i nieosobliwa, dzięki czemu możemy jednoznacznie wyznaczyć wektor parametrów:

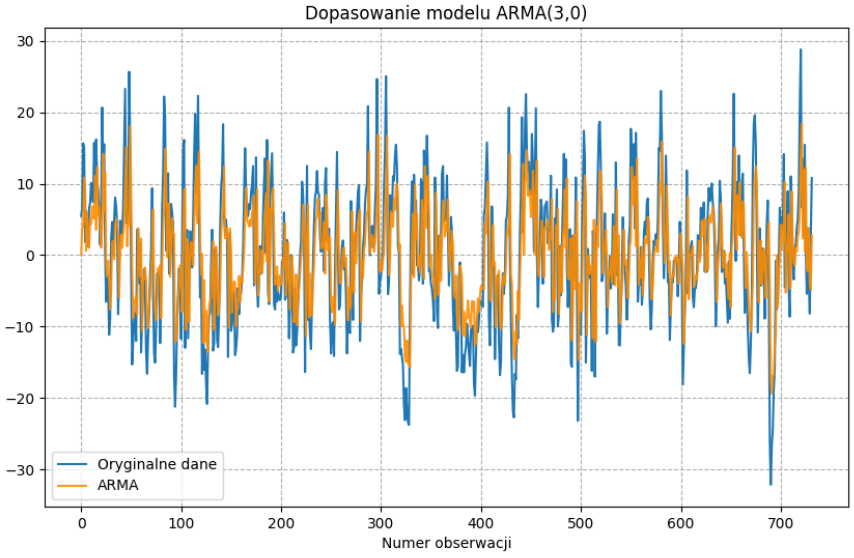
$$\varphi = R_\rho^{-1} \tilde{\rho},$$

gdzie R_ρ jest macierzą autokorelacji, a $\tilde{\rho}$ to wektor autokorelacji.

Wykorzystując funkcję *model.fit(method='yule_walker')* z pakietu *statsmodels* wyestymowaliśmy parametry naszego modelu, które widoczne są na poniższym wykresie.



Rysunek 7: Współczynniki modelu ARMA(3,0)

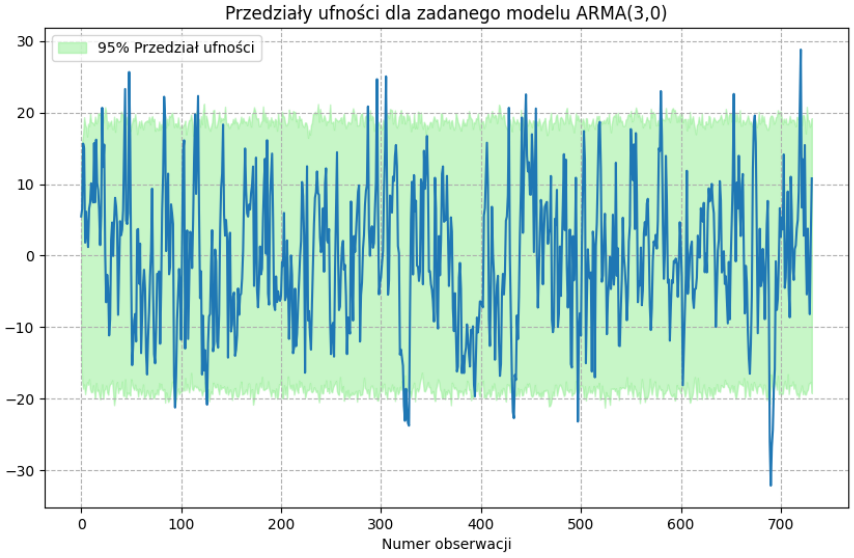


Rysunek 8: Wykres modelu ARMA

Z wykresu widzimy, że model AR(3), o znalezionych przez nas współczynnikach, wydaje się dobrze dopasowywać do oczyszczonych danych. Przejdźmy teraz do bardziej szczegółowej oceny jakości tego dopasowania.

4 Ocena dopasowania modelu

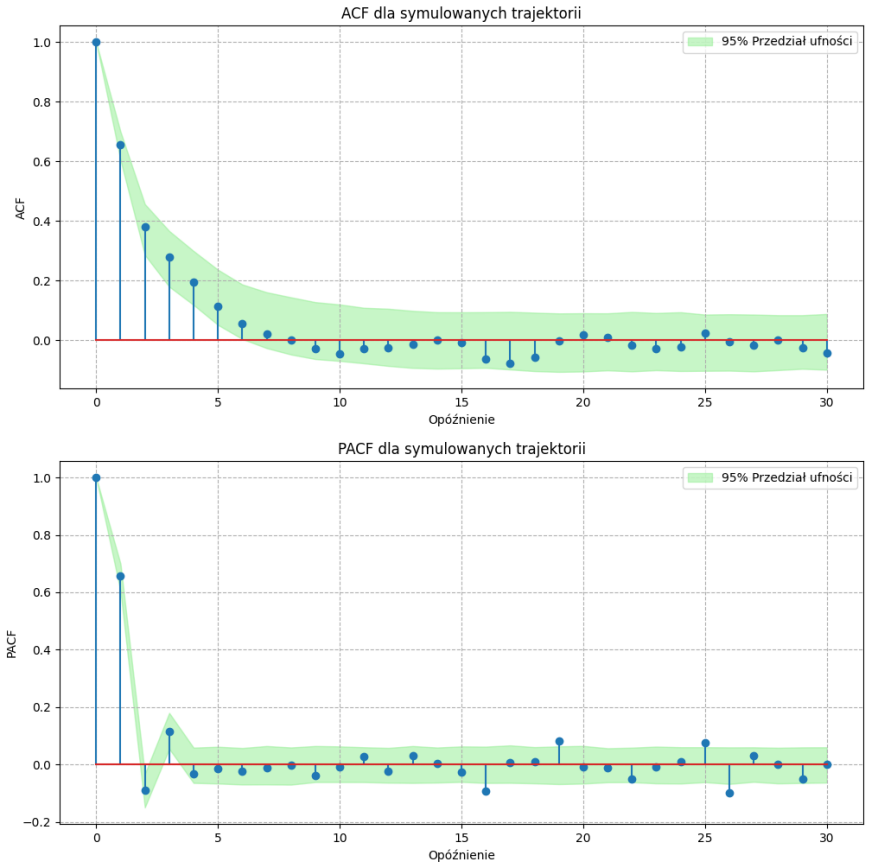
Ocenę dopasowania modelu zaczniemy od sprawdzenia, czy oczyszczone dane mieszczą się w przedziale ufności dla modelu ARMA(3,0) o zadanych wyżej parametrach.



Rysunek 9: Przedział ufności dla modelu ARMA(3,0), dla $\alpha = 0{,}05$

Jak odczytać możemy z wykresu, zdecydowana większość obserwacji mieści się w wyznaczonym przedziale ufności. Jednakże wśród danych znajdują się także te, które nie mieszczą się w przedziale. Przyczyny tego doszukiwalibyśmy się w dużej wartości wariancji analizowanego modelu.

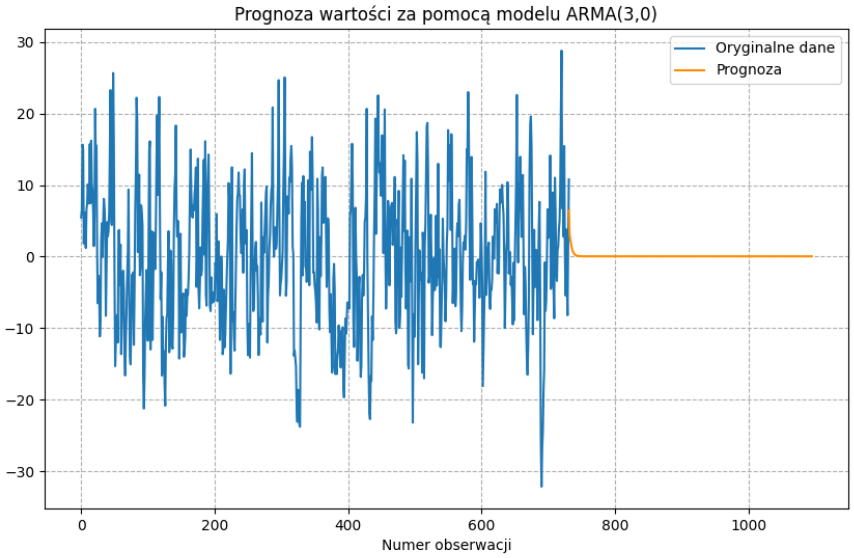
Następnym krokiem badania dopasowania modelu będzie sprawdzenie, czy dla danych oczyszczonych wartości autokorelacji i częściowej autokorelacji mieszczą się w przedziałach ufności wyznaczonych dla naszego modelu.



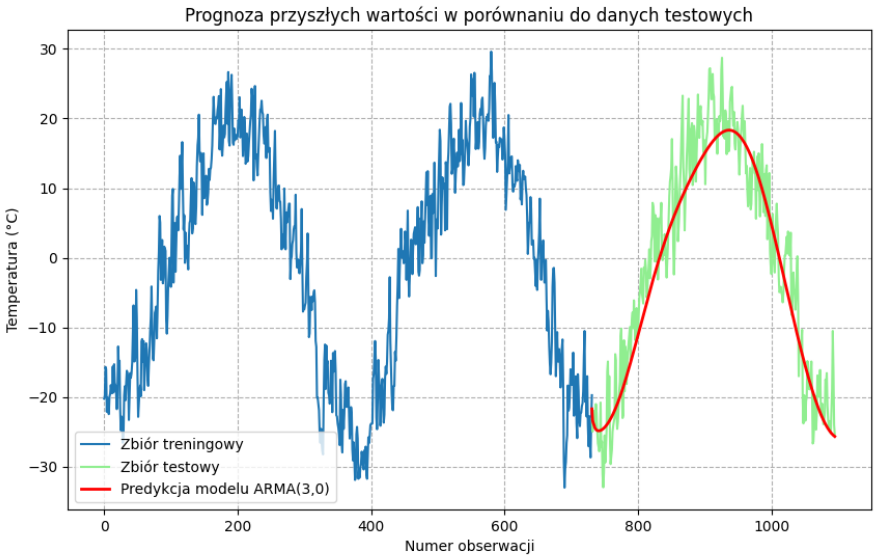
Rysunek 10: Wykres ACF i PACF dla modelu ARMA

W obu przypadkach wartości funkcji ACF i PACF w większości mieszczą się w wyznaczonych przedziałach ufności dla $\alpha = 0.05$. Na podstawie tych trzech wykresów możemy szacować, że model został prawidłowo dobrany do danych.

Dla większej pewności sprawdźmy, jak model radzi sobie z predykowaniem przyszłych wartości.



Rysunek 11: Wykres wartości prognozowanych dla modelu ARMA



Rysunek 12: Wykres wartości prognozowanych dla modelu ARMA

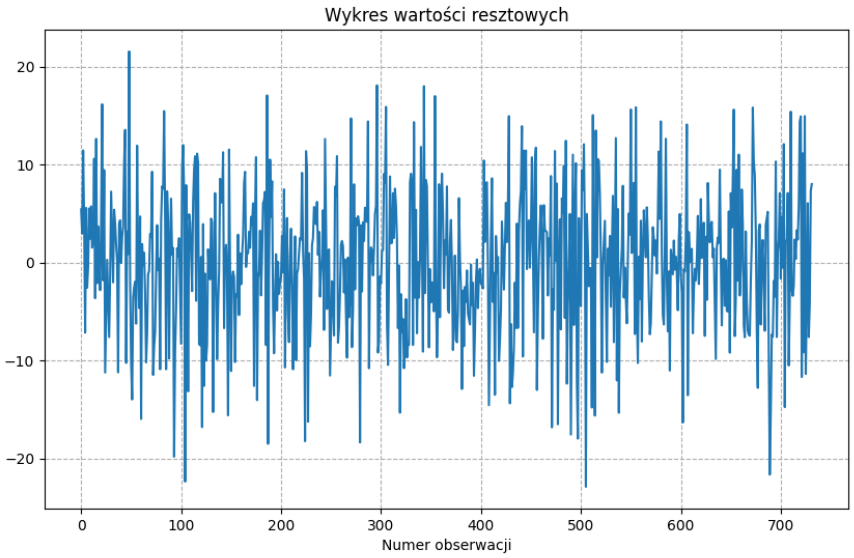
Na rysunku 11 przedstawiającym prognozę modelu dla danych oczyszczonych, obserwujemy prostą $y = 0$. Dzieje się tak, ponieważ próbkowa wartość oczekiwana przyjmuje wartość w okolicach 0, a model rozpoznaje wahania w danych jako szum, który nie zmienia ogólnego trendu.

Na wykresie 12 przedstawione są dane z poprzedniego rysunku, na które ponownie nałożyliśmy trend, sezonowość oraz odwrotną transformację Boxa-Coxa. Widzimy, że przekształcony w taki sposób model dobrze radzi sobie z wychwytywaniem ogólnego sezonowego trendu.

5 Weryfikacja założeń dotyczących szumu

5.1 Wykres wartości resztowych

Analizę wartości resztowych rozpocznijmy od narysowania ich trajektorii.



Rysunek 13: Wykres wartości resztowych

5.2 Sprawdzenie średniej

Jednym z założeń modelu ARMA jest to, że residua pochodzą z rozkładu, którego średnia wynosi 0. W celu zbadania tej hipotezy przeprowadźmy test t.

5.2.1 T - test

Test t-Studenta służy do weryfikacji, czy średnia populacji różni się istotnie od wartości teoretycznej μ_0. Hipotezy testu są sformułowane następująco:

- H_0: $\mu = \mu_0$, czyli średnia populacji jest równa wartości teoretycznej,

- H_1: $\mu \neq \mu_0$, czyli średnia populacji różni się od wartości teoretycznej.

Statystyka testowa T jest wyrażona wzorem:

$$T = \frac{\overline{X} - \mu_0}{\frac{S}{\sqrt{n}}}$$

gdzie:

- $\bar{x}$: średnia z próby,
- μ_0: teoretyczna wartość średniej,
- S: odchylenie standardowe z próby, wyrażone jako

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{X})^2},$$

- n: liczebność próby.

Korzystając z funkcji *ttest_1samp* z modułu *scipy.stats* otrzymaliśmy wartość statystyki testowej T równą $-0{,}04$. Na podstawie wartości p wynoszącej 0,97 wykonanej dla powyższego testu na poziomie istotności $\alpha = 0{,}05$ nie możemy odrzucić hipotezy zerowej mówiącej, że wartość oczekiwana dla reszt wynosi 0. Przyjmujemy ją zatem jako prawdziwą.

5.3 Sprawdzenie wariancji

Kolejnym z podstawowych założeń modelu ARMA jest stała wariancja szumu. Sprawdźmy zatem, czy analizowane residua są homoskedastyczne.

5.3.1 Arch Test

Test ARCH (Autoregressive Conditional Heteroskedasticity) służy do wykrywania warunkowej heteroskedastyczności, a jego hipotezy są następujące:

- H_0: Brak efektu ARCH (reszty mają stałą wariancję w czasie, homoskedastyczność),

- H_1: Obecny efekt ARCH (wariancja reszt zależy od przeszłych wartości).

Test ARCH opiera się na regresji reszt podniesionych do kwadratu (ε_t^2) względem ich opóźnień. Regresja ma postać:

$$\varepsilon_t^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2 + u_t,$$

gdzie:

- ε_t: reszty modelu,
- ε_{t-i}^2: kwadraty opóźnionych reszt,
- α_0: wyraz wolny,
- α_i: współczynniki regresji,
- q: liczba opóźnień,
- u_t: składnik losowy (biały szum).

Statystyka testowa opiera się na współczynniku determinacji R^2 z tej regresji i wyrażona jest wzorem:

$$LM = n \cdot R^2,$$

gdzie:

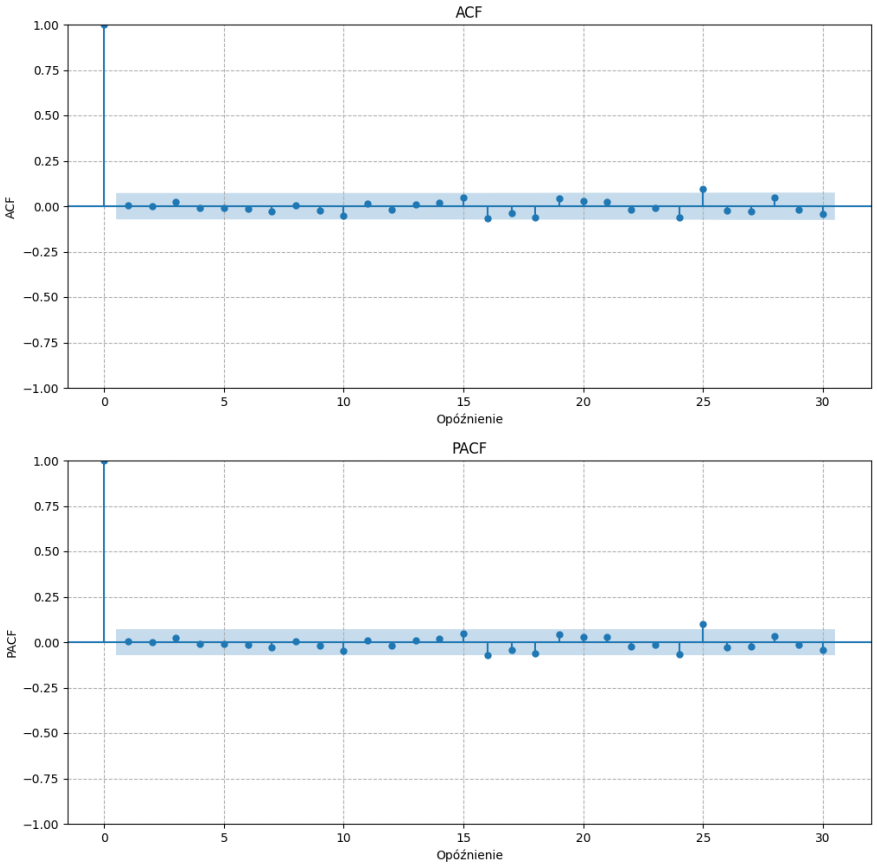
- n: liczba obserwacji w regresji,
- R^2: współczynnik determinacji regresji.

Test wykonaliśmy używając funkcji *het_arch* z biblioteki *stats.models.diagnostic*. Wartość p wyniosła 0,36 na poziomie istotności $\alpha = 0{,}05$. Z tego powodu hipotezę mówiącą o jednolitości wariancji potraktujemy jako prawdziwą.

5.4 Sprawdzenie niezależności

Następnym krokiem analizy wartości resztowych będzie sprawdzenie, czy są one od siebie niezależne. W pierwszym etapie skonstruujemy wykresy ACF i PACF.

5.4.1 Wykres ACF i PACF dla residuów



Rysunek 14: ACF i PACF dla residuów

Z powyższych wykresów spodziewamy się, że residua nie zależą od siebie, ponieważ oba wykresy zanikają do 0 już dla opóźnienia równego 1. Aby potwierdzić hipotezę o niezależności wartości resztowych, przeprowadźmy jednak odpowiednie testy statystyczne.

5.4.2 Test Ljunga-Boxa

Test Ljunga-Boxa służy do sprawdzania, czy w resztach szeregu czasowego występuje autokorelacja na wybranej liczbie opóźnień. Jest często stosowany do oceny, czy reszty są niezależne i przypominają biały szum. Hipotezy testu są następujące:

- H_0: Reszty są niezależne (brak autokorelacji, reszty przypominają biały szum),
- H_1: Reszty nie są niezależne (występuje autokorelacja na wybranym poziomie opóźnień).

Statystyka testowa Ljunga-Boxa jest wyrażona wzorem:

$$Q = n(n+2)\sum_{k=1}^{m}\frac{\hat{\rho}_k^2}{n-k},$$

gdzie:

- n: liczba obserwacji w próbie,
- m: liczba opóźnień, które są uwzględniane w teście,
- $\hat{\rho}_k$: estymator współczynnika autokorelacji dla opóźnienia k.

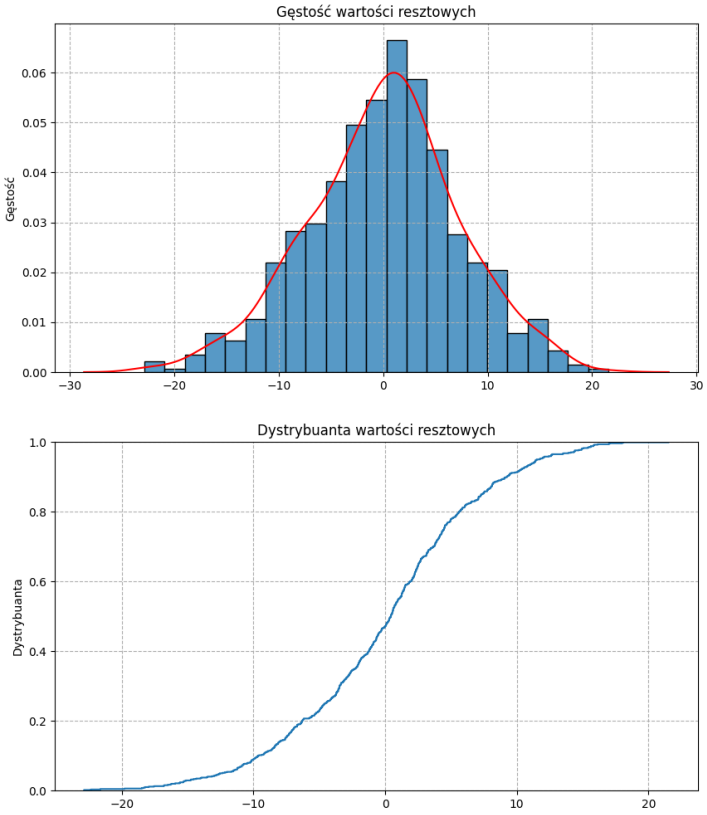
Wykorzystując funkcję *acorr_ljungbox* z modułu *statsmodels.api* otrzymaliśmy statystykę testową $Q = 25{,}63$ na poziomie istotności $\alpha = 0.05$ oraz p-wartość $p = 0{,}43$. Tym samym otrzymujemy potwierdzenie naszej hipotezy zerowej o niezależności residuów.

5.5 Sprawdzenie normalności

Ostatnim etapem sprawdzania wartości resztowych jest analiza, czy pochodzą one z rozkładu normalnego.

W pierwszym kroku narysujmy dystrybuantę oraz gęstość dla naszych danych.

5.5.1 Dystrybuanta i gęstość

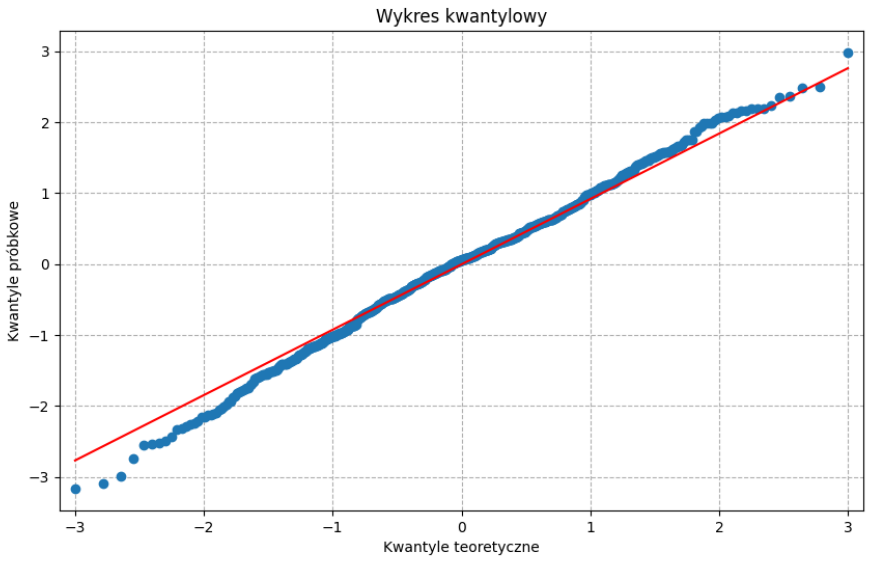


Rysunek 15: Dystrybuanta i gęstość

Na podstawie wykresów gęstości oraz dystrybuanty, widzimy, że residua pochodzić mogą z rozkładu normalnego.

Aby się upewnić, porównajmy także, czy wartości kwantyli próbkowych zgadzają się z kwantylami teoretycznymi. W tym celu skonstruujmy wykres kwantylowy.

5.5.2 Wykres kwantylowy



Rysunek 16: Wykres kwantylowy

Kwantyle próbkowe i teoretyczne w większości pokrywają się ze sobą tworząc prostą. Jednakże dla posiadania większej pewności czy reszty modelu pochodzą z rozkładu normalnego wykonajmy odpowiednie testy na normalność.

5.5.3 Test Shapiro-Wilka

Test Shapiro-Wilka służy do sprawdzenia, czy próbka danych pochodzi z populacji o rozkładzie normalnym. Jest to jeden z najbardziej popularnych testów normalności, szczególnie efektywny dla małych i średnich prób. Hipotezy testu są następujące:

- H_0: Dane pochodzą z rozkładu normalnego,
- H_1: Dane nie pochodzą z rozkładu normalnego.

Statystyka testowa W testu Shapiro-Wilka wyrażona jest wzorem:

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

gdzie:

- n: liczba obserwacji w próbie,
- $x_{(i)}$: uporządkowane dane (w porządku rosnącym),
- $\bar{x}$: średnia arytmetyczna próby,
- a_i: wagi zależne od macierzy kowariancji i średnich wartości teoretycznych danych o rozkładzie normalnym.

Do wykonania testu sprawdzającego czy szum naszego modelu pochodzi z rozkładu normalnego, wykorzystaliśmy funkcję *shapiro* z biblioteki *scipy.stats*. Statystyka testowa wyniosła $W = 0{,}996$, a p-wartość $p = 0{,}090$, co prowadzi nas do wniosku, że na poziomie istotności testu $\alpha = 0{,}05$ nie możemy odrzucić hipotezy zerowej mówiącej o ich normalności. Przyjmujemy zatem, że jest ona prawdziwa.

6 Zakończenie

Analizując dane średniej dziennej temperatury na przestrzeni dni w stolicy Mongolii w latach 2015-2017, po usunięciu wszelkich sezonowości otrzymaliśmy model ARMA(3,0) dany wzorem: $X_t - 0{,}72X_{t-1} + 0{,}17X_{t-2} - 0{,}11X_{t-3} = Z_t$. Model ten następnie porównaliśmy z danymi rzeczywistymi, a także stworzyliśmy prognozę wartości na następny rok. Prognoza ta w naszej ocenie dobrze odzwierciedlała sezonowość w zbiorze testowym. Następnie przeprowadziliśmy analizę dopasowania modelu na podstawie zawierania się w przedziałach ufności. Uważamy, że model przeszedł test pomyślnie i przystąpiliśmy do badania wartości resztowych. Na podstawie testów na niezależność, stałą wariancję oraz normalność stwierdzamy, że uzyskany przez nas model ARMA(3,0) spełnił wszystkie wymienione warunki. W związku z powyższym uważamy, że model AR(3), który wykorzystaliśmy, nadaje się do estymacji przyszłych wartości średniej dziennej pogody w tym rejonie.