Jędrzej Bojanowski, Magdalena Ryś

Analiza danych rzeczywistych przy pomocy modelu ARMA

Średnia temperatura powietrza w Warszawie

styczeń 2025

1. Wstęp

1.1. Cel raportu

Poniższa praca ma na celu analizę rzeczywistych danych dot. średnich temperatur powietrza w Warszawie korzystając z modelu ARMA (Autoregressive Moving Average). Raport obejmuje przygotowanie danych, modelowanie szeregu czasowego, a także ocenę i interpretację uzyskanych wyników.

1.2. Analizowane dane

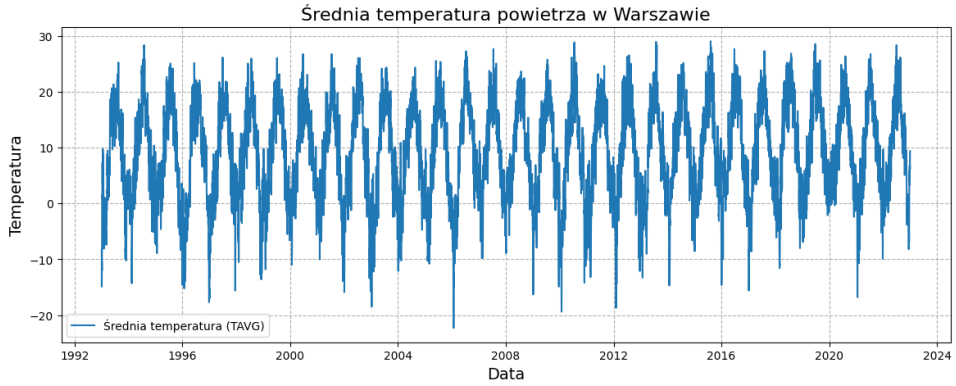
Analizowane dane dotyczą codziennych średnich temperatur powietrza w okresie od 1 stycznia 1993 roku do 31 grudnia 2022 odnotowanych na stacji meteorologicznej Okęcie w Warszawie. Długość próby wynosi 10954. Dane pochodzą ze strony kaggle.com z pliku Warsaw Daily Weather Dataset, zawierającej wyniki codziennych pomiarów meteorologicznych w latach 1993 - 2022 w wyżej wymienionej stacji pomiarowej. Link do powyższych danych zamieszczono w sekcji *7. Źródła.*

1.3. Dane Meteorologiczne w Warszawie

Warszawa, jako miasto w centralnej części Polski, charakteryzuje się klimatem umiarkowanym przejściowym. Oznacza to, że zimy są tu stosunkowo chłodne, a lata umiarkowanie ciepłe, natomiast średnie temperatury roczne oscylują zazwyczaj w okolicach 8-9°C. W okresie zimowym wahają się one od -2°C do 0°C, choć w trakcie szczególnie mroźnych dni temperatura może spaść poniżej -20°C. Natomiast latem, są to temperatury w granicach 18-20°C, a w bardzo upalne dni nawet 30°C. Ponadto obserwowane w ostatnich latach zmiany klimatyczne wpłynęły również na klimat w Warszawie - zimy stają się coraz łagodniejsze, lata natomiast znacznie bardziej upalne.

1.4. Wizualizacja danych

Poniższy wykres przedstawia codzienne średnie temperatury powietrza w Warszawie w latach 1993–2022. Dane ukazują charakterystyczny sezonowy przebieg, z regularnymi wahaniami związanymi ze zmianami pór roku, w tym wyższymi temperaturami latem i niższymi zimą.



Rysunek 1. Wykres średnich temperatur w Warszawie w latach 1993-2022

2. Przygotowanie danych do analizy

2.1. Badanie jakości danych

Badanie jakości danych jest istotną częścią procesu przygotowywania danych do analizy. W poniższej części przeprowadzono kontrolę danych temperaturowych, obejmującą identyfikację danych spoza przewidywanego przedziału, a także detekcję braku danych.

2.1.1. Detekcja wartości spoza zakładanego przedziału

W ramach analizy jakości danych przeprowadzono kontrolę odnotowanych średnich temperatur w celu wykrycia wartości spoza oczekiwanego przedziału. Ze względu na umiarkowany klimat w miejscu pomiaru opisany w sekcji *1.3. Dane Meteorologiczne w Warszawie*, za dopuszczalny zakres wartości przyjęto temperatury w granicach od -20°C do 30°C.

Wykryto 3 pomiary wykraczające poza przewidywany przedział:

— 2006-01-22: -20.8°C
— 2006-01-23: -22.3°C
— 2006-01-24: -20.4°C

Mimo, że wykryte pomiary są poniżej dolnej granicy określonej jako -20°C, nie podjęto decyzji o ich usunięciu, ponieważ zostały uznane za ekstremalne, ale realne w kontekście warunków pogodowych. Wskazuje na to między innymi fakt, że odczyty te następują po sobie dzień po dniu. Ponadto, pomiar został odnotowany zimą 2005/6, która to jest uznawana za jedną z najdotkliwszych w ciągu ostatnich dziesięcioleci.

2.1.2. Identyfikacja braków danych

Przeprowadzono również detekcję braków danych w pliku. Otrzymane wyniki wskazują na to, że w kolumnie TAVG (średnia temperatura powietrza) nie wykryto żadnych braków danych (wartości NaN). Zatem, dla wszystkich dat w zakresie od 1993 do 2022 roku, przypisano wartość temperatury, co świadczy o kompletności danych w tym zakresie.

Mimo, że dane temperaturowe są kompletne, po głębszej analizie wykryto trzy brakujące daty: 1999-01-10, 2021-02-11, 2021-02-12. Brak tych dat może wskazywać na niedostępność pomiarów w tych dniach, co może wynikać z

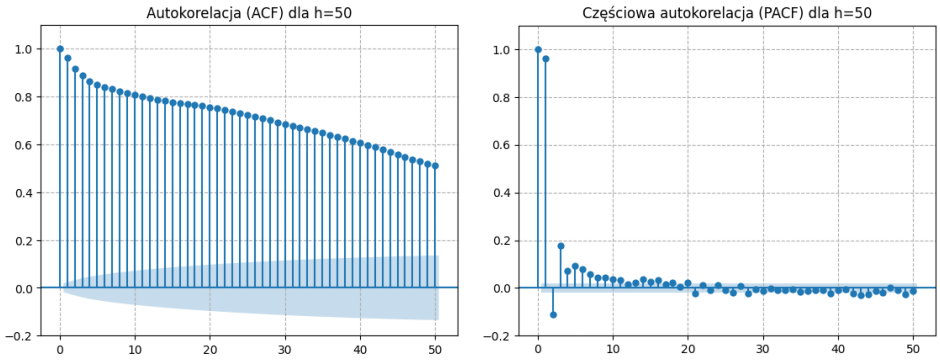
różnych czynników, takich jak przerwy w zbieraniu danych lub problemy techniczne z systemem pomiarowym.

Brakujące daty stanowią minimalny ułamek całego zbioru danych, obejmującego okres 30 lat. Ze względu na marginalny wpływ tych braków na wyniki analizy, zdecydowano się na pozostawienie tych dat jako brakujących bez podejmowania dalszych działań uzupełniających.

2.2. Dekompozycja szeregu czasowego

2.2.1. Wykresy ACF i PACF dla surowych danych

W celu analizy struktury danych wygenerowano wykresy funkcji autokorelacji (ACF) oraz cząstkowej funkcji autokorelacji (PACF). Wykres ACF pokazuje zależności między wartościami szeregu czasowego a ich opóźnionymi wartościami, podczas gdy PACF pozwala na zidentyfikowanie bezpośrednich zależności pomiędzy zmiennymi, eliminując wpływ wartości pośrednich. Poniższe wykresy wygenerowane zostały dla surowych danych, co pozwala zauważyć obecność trendów i sezonowości w szeregu czasowym.



Rysunek 2. Wykresy ACF i PACF dla surowych danych

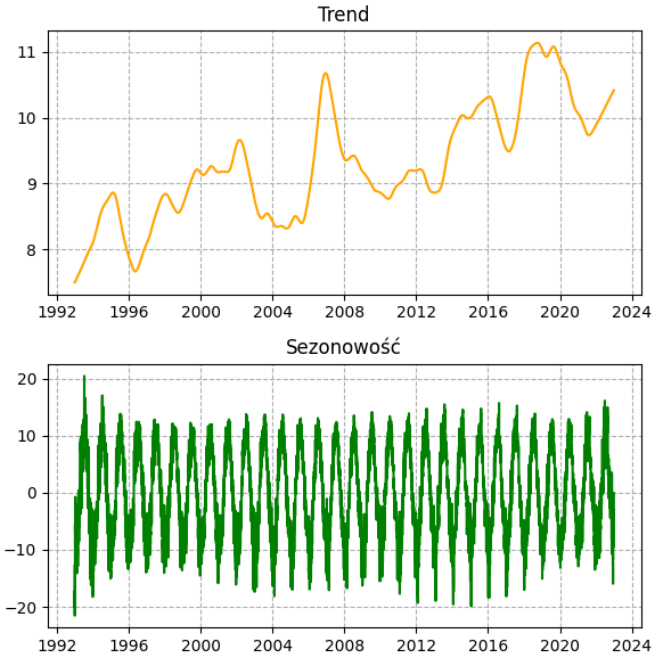
Wygenerowany wykres ACF opada bardzo wolno, co sugeruje obecność trendów lub sezonowości. Ponadto wartości znacznie przekraczają zakres zaznaczony na wykresie, co oznacza, że są one istotne statystycznie.

Wykres PACF dla surowych danych pokazuje wyraźną istotność pierwszego opóźnienia, kolejne opóźnienia mają mniejsze znaczenie, ale nadal są istotne.

2.2.2. Identyfikacja trendów deterministycznych

Wykresy przedstawione w sekcji *2.2.1. Wykresy ACF i PACF dla surowych danych* sugerowały obecność trendu lub sezonowości, które w poniższej części zidentyfikowano i usunięto. W tym celu przeprowadzono dekompozycję szeregu czasowego, która pozwoliła na wyodrębnienie komponentu trendu, sezonowości oraz składnika resztowego.

Wykorzystano w tym celu metodę dekompozycji STL (Seasonal-Trend decomposition using Loess), będącej jedną z lepszych metod w przypadku danych o silnym sezonowym charakterze, jakim są dane pogodowe. STL oparta jest na technice Loess (ang. Local Regression), która pozwala na wygładzenie danych w sposób elastyczny, dostosowując krzywą wygładzającą do lokalnych danych.

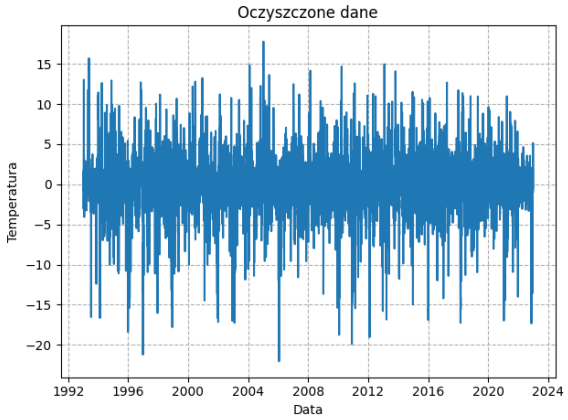


Rysunek 3. Wykryte w danych trend i sezonowość

Jak widać na wygenerowanych wykresach, w danych temperaturowych zaobserwowano wyraźny trend wzrostowy, co może wiązać się ze zjawiskiem globalnego ocieplenia oraz współczesnymi zmianami klimatycznymi. Ponadto wykryto ewidentną regularną sezonowość w skali roku. Odpowiada ona naturalnym cyklom pogodowym w Warszawie i zmieniającym się porom roku na tym obszarze.

2.2.3. Szereg uzyskany po usunięciu trendu i sezonowości

Przedstawiony na poniższym wykresie szereg czasowy został uzyskany poprzez usunięcie trendu deterministycznego oraz komponentu sezonowego.



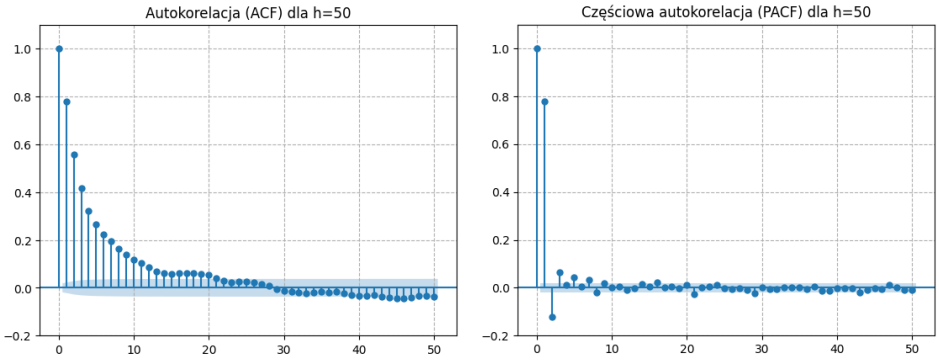
Rysunek 4. Szereg uzyskany po usunięciu trendu i sezonowości

Oczyszczony szereg można analizować za pomocą modeli ARMA. Otrzymay szereg, jeśli jest stacjonarny, powinien mieć średnią bliską zeru oraz stałą wariancję. Powyższa wizualizacja sugeruje, że usuwanie komponentów

deterministycznych przebiegło skutecznie, a szereg wykazuje cechy szeregu stacjonarego.

2.2.4. Wykresy ACF i PACF dla uzyskanego szeregu

Po dekompozycji ponowiono analizę funkcji ACF i PACF w celu sprawdzenia, czy uzyskany szereg czasowy wykazuje cechy stacjonarności.



Rysunek 5. Wykresy ACF i PACF dla uzyskanego szeregu

Wykres ACF wykazuje znacznie szybszy spadek niż w przypadku danych surowych, co wskazuje na poprawny przebieg dekompozycji. Wartości zbiegają do 0 po kilku opóźnieniach, co jest typowe dla szeregów stacjonarnych. Natomiast wykres PACF potwierdza, że istotne zależności następują przy pierwszych kilku opóźnieniach, natomiast reszta okazuje się nieistotna statystycznie. Otrzymane wyniki są zgodne z założeniem stacjonarności.

3. Modelowanie danych przy pomocy ARMA

3.1. Dobranie rzędu modelu

W celu dobrania jak najodpowiedniejszego modelu ARMA, wyszukano najlepszy rząd modelu na podstawie trzech kryteriów informacyjnych:

— **AIC** (Akaike Information Criterion), który szuka równowagi między dopasowaniem a złożonością. Obliczany ze wzoru $AIC = 2k - 2\ln(L)$
— **HQIC** (Hannan-Quinn Information Criterion), będący bardziej restrykcyjny w kwestii złożoności. Obliczany ze wzoru $HQIC = 2k\ln(\ln(n)) - 2\ln(L)$
— **BIC** (Bayesian Information Criterion), który karze bardziej złożone modele, zwłaszcza te o znacznej długości. Obliczany ze wzoru $BIC = \ln(n)k - 2\ln(L)$

gdzie:

— n - liczba próbek
— k - liczba parametrów
— L - funkcja prawdopodobieństwa

Aby wybrać jak najlepszy rząd modelu należy dobrać takie parametry p i q, żeby obliczona wartość powyższych kryteriów była jak najmniejsza. Obliczono więc wartości AIC, BIC oraz HQIC dla różnych parametrów, i posortowano względem kolejno każdego z kryteriów.

	p	q	AIC	BIC	HQIC
8	2	2	4232.996346	4262.442877	4244.188083
7	2	1	4254.649671	4279.188447	4263.976118
6	2	0	4294.899720	4314.530741	4302.360878
5	1	2	4342.372260	4366.911037	4351.698708
2	0	2	4344.838105	4364.469127	4352.299263

Rysunek 6. Sortowanie po wartości AIC oraz HQIC

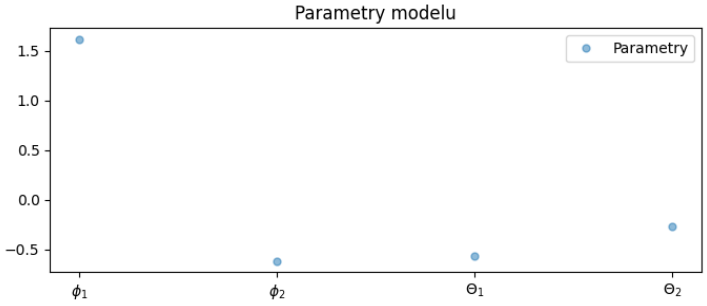
	p	q	AIC	BIC	HQIC
8	2	2	4232.996346	4262.442877	4244.188083
7	2	1	4254.649671	4279.188447	4263.976118
6	2	0	4294.899720	4314.530741	4302.360878
2	0	2	4344.838105	4364.469127	4352.299263
5	1	2	4342.372260	4366.911037	4351.698708

Rysunek 7. Sortowanie po wartości BIC

W zamieszczonych powyżej tabelach zamieszczono posortowane względem AIC oraz BIC (*Rysunek 6.*) oraz względem BIC (*Rysunek 7.*) wartości. Jak widać, dla każdego z badanych kryteriów, najniższą wartość, a zarazem najlepsze dopasowanie, osiągamy dla modelu o parametrach $p = 2$ oraz $q = 2$. Wykorzystane kryteria zapewniają, że wynik bierze pod uwagę równocześnie dopasowanie do danych oraz złożoność modelu. Otrzymane wyniki sugerują więc, że model ARMA(2,2) zapewnia najbardziej optymalne dopasowanie do badanych danych temperaturowych.

3.2. Estymacja parametrów modelu

Na podstawie modelu ARMA(2,2) wyznaczone zostały parametry $\phi_1, \phi_2, \theta_1, \theta 2$ przy użyciu metody największej wiarygodności.



Rysunek 8. Parametry modelu

Nasze wartości to $\phi_1 \approx 1.62, \phi_2 \approx -0.62, \theta_1 \approx -0.57, \theta_2 \approx -0.27$

4. Ocena dopasowania modelu

4.1. Przedziały ufności dla PACF/ACF

W tej sekcji przeprowadzono analizę funkcji autokorelacji (ACF) i częściowej autokorelacji (PACF) dla oczyszczonych danych, czyli danych po usunięciu trendu i sezonowości.
Wykresy ACF i PACF pokazują, że wartości autokorelacji dla danych oscylują wokół zera i mieszczą się w granicach istotności, co sugeruje, że dane z modelu są stacjonarne i nie wykazują istotnej autokorelacji, co sugeruje, że spełniają założenie stacjonarności i są bliskie białemu szumowi.
Przedziały ufności dla ACF i PACF zostały wyznaczone na podstawie symulacji bootstrapowej, co pozwala na lepsze zrozumienie rozkładu autokorelacji.

Autokorelacja mierzy zależność pomiędzy wartością szeregu czasowego a jego przesuniętymi wartościami. Współczynnik autokorelacji rzędu h definiujemy jako:

$$\gamma_h = \frac{1}{n} \sum_{t=1}^{n-h} (X_t - \bar{X})(X_{t+h} - \bar{X}) \tag{1}$$

gdzie:

— γ_h – kowariancja dla opóźnienia h,
— X_t – wartość szeregu w czasie t,
— $\bar{X}$ – średnia wartość szeregu czasowego,
— n – liczba obserwacji.

Znormalizowana autokorelacja (ACF):

$$\rho_h = \frac{\gamma_h}{\gamma_0} \tag{2}$$

gdzie:

$$\gamma_0 = \frac{1}{n} \sum_{t=1}^{n} (X_t - \bar{X})^2. \tag{3}$$

Częściowa autokorelacja eliminuje wpływ pośrednich wartości i odpowiada współczynnikom w modelu AR(h).

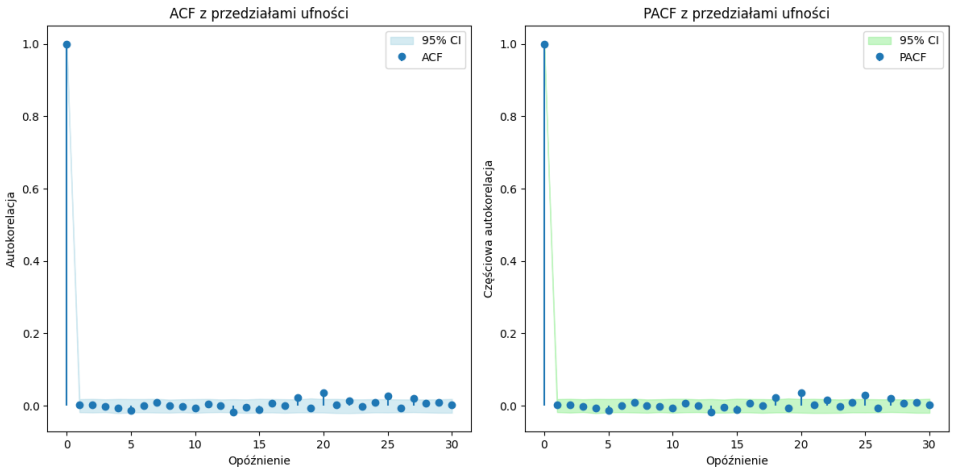
$$\phi_{h,h} = \frac{\rho_h - \sum_{j=1}^{h-1} \phi_{h-1,j} \rho_{h-j}}{1 - \sum_{j=1}^{h-1} \phi_{h-1,j} \rho_j} \tag{4}$$

dla $h \geqslant 1$, gdzie:

— $\phi_{h,h}$ – współczynnik autoregresji dla modelu AR(h),
— ρ_h – autokorelacja rzędu h,
— $\phi_{h-1,j}$ – współczynniki PACF dla wcześniejszych rzędów.

Dla $h = 1$:

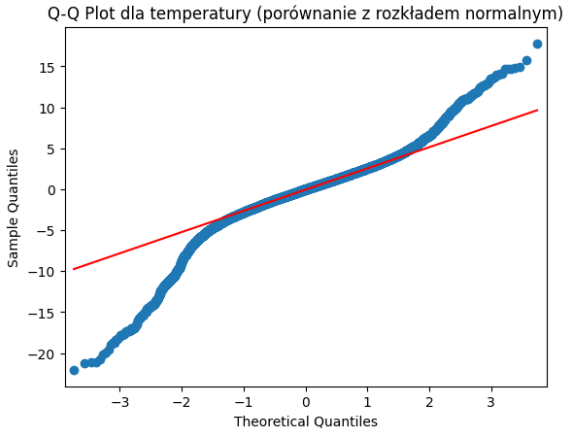
$$\phi_{1,1} = \rho_1. \tag{5}$$



Rysunek 9. ACF i PACF z przedziałami ufńści

4.2. Porównanie linii kwantylowych z trajektorią

W tej części przeprowadzono porównanie rozkładu danych oczyszczonych z rozkładem normalnym za pomocą wykresów kwantylowych (Q-Q plot) oraz histogramów.



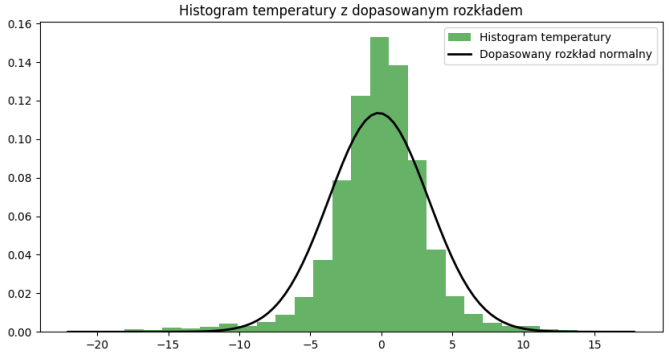
Rysunek 10. Porównanie oczyszczonych danych do Rozkładu normalnego

Wykres Q-Q przedstawia porównanie kwantyli empirycznych oczyszczonych danych z kwantylami teoretycznego rozkładu normalnego.

— Większość punktów leży blisko linii referencyjnej, co sugeruje, że rozkład oczyszczonych danych jest zbliżony do normalnego.
— Pewne odchylenia na krańcach wykresu mogą wskazywać na lekką skośność rozkładu

Histogram oczyszczonych danych porównano z dopasowanym rozkładem normalnym. Histogram pokazuje, że oczyszczone dane mają rozkład bliski normalnemu, ale mogą występować małe odstępstwa w ogonach rozkładu. Możliwe przyczyny odchyleń

— Obecność nieliniowych wzorców, których model ARMA nie wychwytuje.
— Asymetria w rozkładzie temperatur (np. ekstremalne wartości temperatur mogą być częstsze zimą niż latem)



Rysunek 11. Histogram dla oczyszczonych danych do rozkładu normalnego

5. Weryfikacja założeń dotyczących szumu

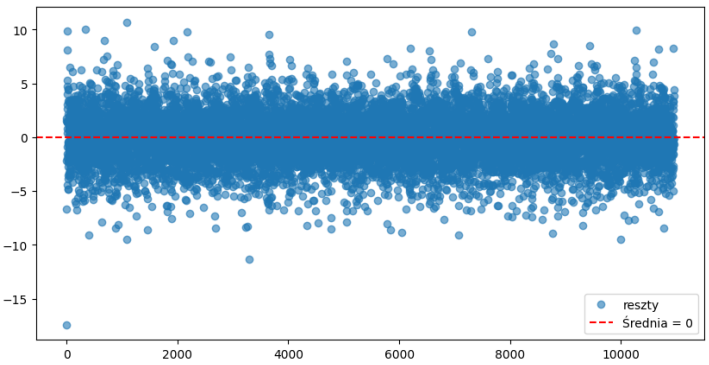
W celu oceny jakości reszt modelu przeprowadzono weryfikację następujących założeń:

— średnia reszt równa zeru,
— stałość wariancji reszt,
— niezależność reszt,
— normalność rozkładu reszt.

5.1. Średnia reszt

5.1.1. Wykres wartości resztowych

Poniższy wykres przedstawia wartości resztowe, natomiast zaznaczona linia reprezentuje oczekiwaną średnią resztową, której wartość wynosi $\mu = 0$. Wartości rozkładają się regularnie wokół linii, co sugeruje brak wyraźnych błędów w modelu.



Rysunek 12. Wykres wartości resztowych z oznaczoną średnią $\mu = 0$

5.1.2. Test t-Studenta

Hipoteza zerowa H_0: Średnia reszt jest równa zeru ($\mu = 0$),
Hipoteza alternatywna H_1: Średnia reszt nie jest równa zeru ($\mu \neq 0$).
Statystyka t-Studenta jest obliczana za pomocą wzoru:

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

gdzie:

— $\bar{x}$ to średnia z próby,
— μ_0 to hipotetyczna średnia (w tym przypadku 0),
— s to odchylenie standardowe próby,
— n to liczebność próby.

P-wartość oblicza się na podstawie rozkładu t-Studenta z $(n-1)$ stopniami swobody. Dla testu dwustronnego p-wartość wyraża się wzorem:

$$p = 2 \cdot (1 - T_{n-1}(|t|))$$

P-wartość odzwierciedla prawdopodobieństwo uzyskania wyniku co najmniej tak ekstremalnego jak obserwowany, zakładając, że hipoteza zerowa jest prawdziwa. Jeżeli p-wartość jest mniejsza od poziomu istotności α, odrzucamy hipotezę zerową, co oznacza, że średnia reszt różni się od zera w sposób statystycznie istotny.

Wyniki testu:

— Statystyka t: $t \approx 0.0565$,
— p-wartość: $p \approx 0.95494$.

Ponieważ p-wartość jest znacznie większa od przyjętego poziomu istotności $\alpha = 0.05$, nie mamy podstaw do odrzucenia hipotezy zerowej H_0. Oznacza to, że średnia reszt statystycznie nie różni się od zera.

5.1.3. Wnioski

Biorąc pod uwagę wynik testu, a także wygenerowany wykres, można wnioskować, że reszty modelu mają średnią równą zeru.

5.2. Stałość wariancji reszt

5.2.1. Wykres wartości resztowych

Aby ocenić stałość wartości restowych wykorzystano wykres przedstawiony w sekcji *5.1.1. Wykres wartości resztowych*. Można zauważyć, że przedstawione na wykresie reszty są rozproszone równomiernie wokół zera. Nie wykazują również żadnych wzorców czy wyraźnych skupisk. Wykres sugeruje więc, że wariancja wartości resztowych jest stała.

5.2.2. Modified Levene Test

Hipoteza zerowa H_0: Wariancje w różnych grupach są równe,
Hipoteza alternatywna H_1: Wariancje w różnych grupach są różne.

Statystyka testowa obliczana jest ze wzoru:

$$W = \frac{(N-k)}{(k-1)} \cdot \frac{\sum_{i=1}^{k} n_i(\bar{Z}_i - \bar{Z})^2}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Z_{ij} - \bar{Z}_i)^2}$$

gdzie:

— N - łączna liczba obserwacji
— k - liczba grup
— Z_{ij} - wartość zmiennej w j-tej obserwacji w i-tej grupie

Przeprowadzono test na dwóch grupach - od 0 do 5000 obserwacji, a także od 5001 do 10954.

Wyniki testu:

— Statystyka w: $W \approx 0.343$,
— p-wartość: $p \approx 0.558$.

Ponieważ p-wartość jest większa od przyjętego poziomu istotności $\alpha = 0.05$, nie mamy podstaw do odrzucenia hipotezy zerowej H_0. Oznacza to, że wariancja reszt statystycznie jest stała.

5.2.3. Arch Test

Hipoteza zerowa H_0: Wariancje w różnych grupach są równe,
Hipoteza alternatywna H_1: Wariancje w różnych grupach są różne.

Statystyka testu Arch jest obliczana na podstawie regresji kwadratów reszt:

$$\hat{y}_t^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i \hat{y}_{t-i}^2 + \epsilon_t$$

Wyniki testu:

— Statystyka w: $W \approx 163.4$,
— p-wartość: $p \approx 6.416 \times 10^{-30}$.

Ponieważ p-wartość jest mniejsza od przyjętego poziomu istotności $\alpha = 0.05$, odrzucamy hipotezę zerową H_0. Oznacza to, że wariancja reszt statystycznie nie jest stała.

5.2.4. Goldfeld-Quandt Test

Hipoteza zerowa H_0: Wariancje w dwóch grupach są równe.
Hipoteza alternatywna H_1: Wariancje w dwóch grupach są różne.

Statystyka testowa obliczana jest na podstawie stosunku sum kwadratów reszt (RSS) z dwóch grup:

$$F = \frac{RSS_2/df_2}{RSS_1/df_1}$$

gdzie:

— RSS_1, RSS_2 - suma kwadratów reszt dla grup 1 i 2,
— df_1, df_2 - liczba stopni swobody dla grup 1 i 2.

Wyniki testu:

— Statystyka testowa: $F \approx 0.806258$,
— p-wartość: $p \approx 0.999999$.

Ponieważ p-wartość jest większa od przyjętego poziomu istotności $\alpha = 0.05$, nie mamy podstaw do odrzucenia hipotezy zerowej H_0.

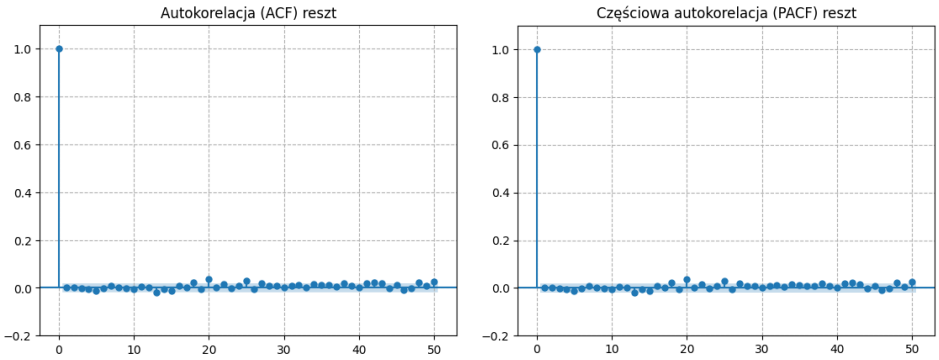
5.2.5. Wnioski

Po przeprowadzeniu trzech testów statystycznych - Modified Levene Test, Arch Test i Goldfeld-Quandt Test - dwa z nich wykazały, że wariancja reszt statystycznie może być uznana za stałą. Potwierdza to również przeanalizowany wykres wartości resztowych.

5.3. Niezależność reszt

5.3.1. Wykresy ACF i PACF dla wartości resztowych

Przedstawione poniżej wykresy, przedstawiają funkcję autokorelacji i częściowej autokorelacji dla wartości resztowych.



Rysunek 13. Wykresy ACF i PACF dla wartości resztowych

Na wykresie ACF można zauważyć, że poza opóźnieniem 0, wartości autokorelacji dla reszt oscylują wokół zera i mieszczą się w granicach istotności. Wskazuje to na fakt, że w resztach nie występuje znacząca autokorelacja.

Analogicznie, na wykresie PACF poza opóźnieniem 0, wartości oscylują wokół zera, co sugeruje brak zależności w wyższych rzędach opóźnień.

5.3.2. Test Ljunga-Boxa

Hipoteza zerowa H_0: Brak autokorelacji w resztach dla danego opóźnienia.

Hipoteza alternatywna H_1: Istnieje autokorelacja w resztach dla przynajmniej jednego opóźnienia.

Statystyka testu Ljunga-Boxa jest obliczana jako suma kwadratów autokorelacji reszt z uwzględnieniem odpowiedniej liczby lagów:

$$Q = n(n+2)\sum_{k=1}^{m}\frac{\rho_k^2}{n-k}$$

gdzie:

— n – liczba obserwacji,
— m – liczba opóźnień (lagów),
— ρ_k – autokorelacja reszt dla laga k.

Dla większości lagów (1–24) p-wartości są większe od $\alpha = 0.05$, co oznacza, że reszty w tych lagach nie wykazują istotnej autokorelacji. Jednakże dla wyższych lagów (25–40), p-wartości są mniejsze od $\alpha = 0.05$. Ponieważ dla niektórych lagów p-wartości są mniejsze od przyjętego poziomu istotności $\alpha = 0.05$, odrzucamy hipotezę zerową H_0.

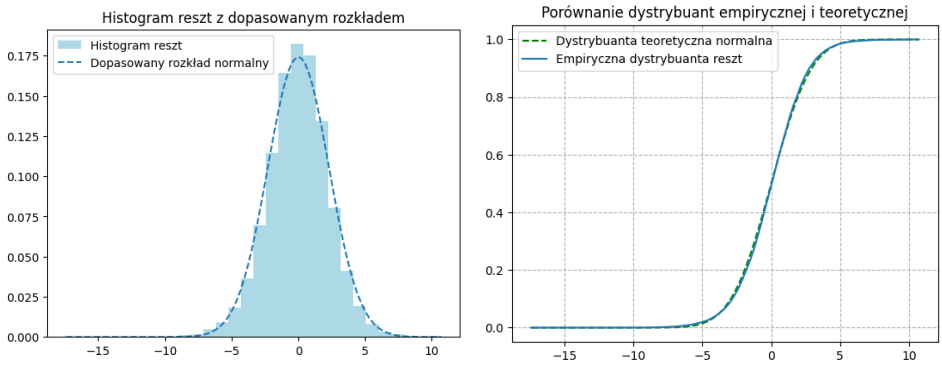
5.3.3. Wnioski

Mimo, że wykresy sugerują, że reszty są niezależne, po przeprowadzeniu testu statystycznego należy odrzucić tą hipotezę na rzecz hipotezy alternatywnej - istnieje autokorelacja w resztach dla przynajmniej jednego opóźnienia.

5.4. Normalność rozkładu reszt

5.4.1. Wykresy gęstości i dystrybuanty

W celu zbadania normalności rozkładu reszt wygenerowano wykresy gęstości i dystrybuanty.

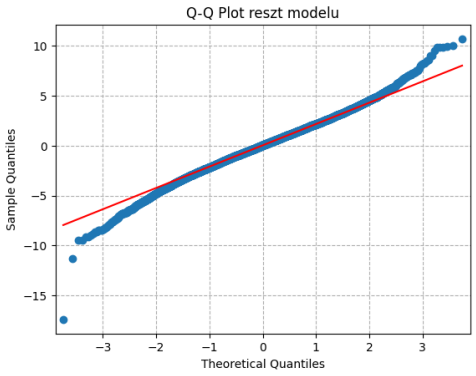


Rysunek 14. Gęstość i dystrybuanta rozkładu reszt

Histogram rozkładu reszt w dużym stopniu pokrywa się z wykresem gęstości dopasowanego do niego rozkładu normalnego, natomiast symetryczny kształt sugeruje brak skośności. Dystrybuanta empiryczna również w dużym stopniu pokrywa się z teoretyczną. Sugeruje to, że rozkład reszt może być zgodny z rozkładem normalnym.

5.4.2. Wykres kwantylowy

W ramach kolejnej metody zbadania normalności rozkładu reszt wygenerowano wykres kwantylowy. Jeśli reszty są normalnie rozłożone, punkty powinny rozkładać się na czerwonej linii.



Rysunek 15. Wykres kwantylowy rozkładu reszt

Na wyżej przedstawionym wykresie możemy zauważyć, że środkowa część leży blisko linni, jednakże na krańcach pojawiają się pewne odchylenia, co może sugerować skośność lub obecność wartości odstających.

5.4.3. Test Shapiro-Wilka

Hipoteza zerowa H_0**:** Rozkład danych w grupie jest normalny.

Hipoteza alternatywna H_1**:** Rozkład danych w grupie nie jest normalny.

Statystyka testowa obliczana jest ze wzoru:

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

gdzie:

— n - liczba obserwacji w próbce
— x_i - wartość i-tej obserwacji
— $\bar{x}$ - średnia arytmetyczna próbki
— a_i - współczynniki, które zależą od rang obserwacji w próbce

Wyniki testu:

— Statystyka testowa: $W \approx 0.99369$,
— p-wartość: $p \approx 1.54 \times 10^{-21}$.

Ponieważ p-wartość jest mniejsza od przyjętego poziomu istotności $\alpha = 0.05$, odrzucamy hipotezę zerową H_0. Oznacza to, że rozkład reszt nie jest normalny.

5.4.4. Test Jaque-Bera

Hipoteza zerowa H_0: Rozkład danych jest normalny.
Hipoteza alternatywna H_1: Rozkład danych nie jest normalny.

Statystyka testowa obliczana jest ze wzoru:

$$JB = \frac{n}{6}\left(S^2 + \frac{1}{4}(K-3)^2\right)$$

gdzie:

— n - liczba obserwacji,
— S - skośność rozkładu,
— K - kurtoza rozkładu.

Wyniki testu:

— Statystyka testowa: $JB \approx 557.492$,
— p-wartość: $p \approx 8.752 \times 10^{-122}$.

Ponieważ p-wartość jest mniejsza od przyjętego poziomu istotności $\alpha = 0.05$, odrzucamy hipotezę zerową H_0. Oznacza to, że rozkład reszt nie jest normalny.

5.4.5. Wnioski

Pomimo, że testy wizualne sugerowały rozkład normalny, przeprowadzone testy statystyczne odrzuciły hipotezę o normalności rozkładu. Otrzymane wyniki mogą sugerować, że lepszym dopasowaniem byłby rozkład t-Studenta.

5.4.6. Rozkład t-Studenta - Test Kolmogorowa-Smirnowa

Hipoteza zerowa H_0: Reszty pochodzą z rozkładu t-Studenta,
Hipoteza alternatywna H_1: Reszty nie pochodzą z rozkładu t-Studenta.

Statystyka testowa Kolmogorowa-Smirnowa obliczana jest jako:

$$D = \sup_x |F_n(x) - F(x)|$$

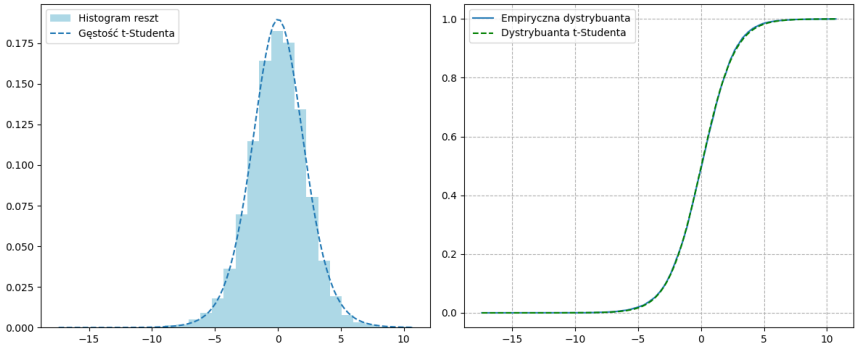
gdzie:

— $F_n(x)$ – funkcja empiryczna rozkładu reszt,
— $F(x)$ – funkcja dystrybuanty rozkładu t-Studenta, który jest dopasowywany do reszt.

Wyniki testu:

— Statystyka testowa: $D \approx 0.0078$,
— p-wartość: $p \approx 0.5092$.

Ponieważ p-wartość jest większa od przyjętego poziomu istotności $\alpha = 0.05$, nie mamy podstaw do odrzucenia hipotezy zerowej H_0. Oznacza to, że reszty mogą pochodzić z rozkładu t-Studenta.



Rysunek 16. Porównanie gęstości i dysrybuanty rozkładu reszt i t-Studenta

Wygenerowano również powyższe wykresy porównujące gęstość oraz dystrybuantę rozkładu reszt oraz rozkładu t-Studenta. Jak widać, są one nieznacznie lepiej dopasowane niż wykresy dla rozkładu normalnego. Można więc wnioskować, że reszty mogą pochodzić z rozkładu t-Studenta.

6. Podsumowanie

W niniejszym raporcie przeprowadzono analizę szeregu czasowego średnich temperatur powietrza w Warszawie w latach 1993–2022 przy użyciu modelu ARMA.

W pierwszej kolejności dokonano wstępnej analizy danych, obejmującej identyfikację wartości odstających oraz brakujących obserwacji. Podjęto decyzję o zachowaniu wykrytych ekstremalnych temperatur oraz brakujących dat, ponieważ dane te były ekstremalne, ale realne w kontekście pogodowym, natomiast ich liczba była marginalna względem całości zbioru.

Następnie przeprowadzono dekompozycję szeregu czasowego, w celu wyodrębnienia trendu i sezonowości, co pozwoliło na wyodrębnienie stacjonarnych reszt. Dekompozycja szeregu czasowego ujawniła trend wzrostowy, co jest zgodne z obserwowanymi globalnymi zmianami klimatycznymi. Wykryto również silną sezonowość, odzwierciedlającą cykliczne zmiany pór roku.

Kolejna część raportu skupiła się na modelowaniu danych przy pomocy szeregu ARMA. Na podstawie kryteriów informacyjnych (AIC, BIC, HQIC) dobrano optymalny rząd modelu. Następnie metodą największej wiarogodności wyestymowano parametry modelu.

W sekcji *4. Ocena dopasowania modelu* skupiono się na analizie wykresów ACF i PACF dla oczyszczonych danych, a także wyznaczeniu dla nich przedziałów ufności metodą bootstrapową.

Następnie model został poddany szczegółowej ocenie, w tym analizie reszt pod kątem ich średniej, wariancji, niezależności oraz normalności rozkładu. Wyniki testów statystycznych wskazują, że reszty mają średnią bliską zeru i stałą wariancję, jednak wykazują pewną autokorelację oraz odchylenia od rozkładu normalnego. Test Kolmogorowa-Smirnowa wskazał, że rozkład t-Studenta może być lepszym dopasowaniem niż rozkład normalny. Autokorelacja reszt dla wyższych lagów może sugerować konieczność dalszej optymalizacji modelu. Większość testów nie wykazała problemów ze stałością wariancji, choć istnieją pewne przesłanki wskazujące na taką możliwość.

Podsumowując, model ARMA(2,2) dobrze opisuje dane temperaturowe, jednakże istnieją pewne ograniczenia związane z niezależnością reszt i ich rozkładem.

7. Źródła

— Warsaw Daily Weather Dataset link.