Politechnika Wrocławska

Wydział Matematyki

Analiza danych rzeczywistych przy pomocy modelu ARMA

Marcelina Kosiorowska, Maria Lubneuskaya

Spis treści

1. Wstęp

1.1. Cel raportu

W niniejszym raporcie zajmiemy się dobraniem modelu ARMA do danych odzwierciedlających średnią temperaturę we Wrocławiu w poszczególnych dniach. Dzięki zdolności modelu ARMA do prognozowania przyszłych wartości szeregów czasowych, będziemy w stanie przewidywać zmiany temperatury w przyszłości.

1.2. Definicje i wzory

1. **Modelem ARMA(p, q)** nazywamy słabostacjonarny szereg czasowy $\{X_t\}$, który spełnia równanie:

$$X_t - \Phi_1 X_{t-1} - \Phi_2 X_{t-2} - ... - \Phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + ... + \theta_q Z_{t-q} \tag{1}$$

   gdzie $\{Z_t\}$ jest białym szumem.

2. **ACF** (ang. Autocorrelation Function) to funkcja mierząca korelację między wartościami X_t a X_{t-k}.

3. **PACF** (ang. Partial Autocorrelation Function) to funkcja która również mierzy ową korelację, jednak w przeciwieństwie do ACF, eliminuje wpływ pośrednich wartości szeregu $(X_t, ..., X_{t-(k-1)})$.

4. **Przedziały ufności dla ACF i PACF**:

$$\left[z_{\frac{\alpha}{2}},\ z_{1-\frac{\alpha}{2}}\right] \tag{2}$$

   gdzie $z_{\frac{\alpha}{2}}$, $z_{1-\frac{\alpha}{2}}$ to kwantyle odpowiednio rzędu $\frac{\alpha}{2}$ i $1 - \frac{\alpha}{2}$ dla próby wartości ACF/PACF.

5. **Różnicowanie szeregu czasowego** to technika, która pozwala wyodrębnić stacjonarny szereg $\{Y_t\}$ z szeregu niestacjonarnego $\{X_t\}$. Różnicowanie rzędu m usuwa sezonowość o okresie m. Technika ta polega na znalezieniu nowego szeregu czasowego postaci:

$$Y_t = X_t - X_{t-m}$$

6. **Kryterium AIC** to kryterium wyboru pomiędzy modelami statystycznymi. Najlepiej dopasowany model to model o parametrach p, q, dla których kryterium dany wzorem:

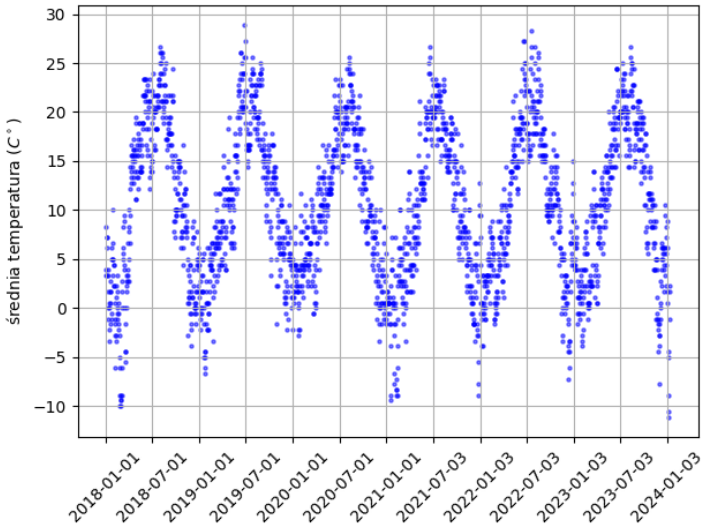
$$AIC(p, q) = -2ln(L) + 2(p + q),$$

   gdzie L - funkcja wiarygodności, przyjmuje najmniejszą wartość.

1.3. Tematyka danych

* Badany zbiór danych opisuje średnią temperaturę we Wrocławiu w poszczególnych dniach
* Źródłem jest amerykańska rządowa strona National Centers for Environmental Information ([2]), przechowująca informacje dotyczące środowiska.
* Dane pochodzą z okresu od 1 stycznia 2018 roku do 14 styczna 2024 roku i są zebrane z 2203 obserwacji.

1.4. Wizualizacja danych



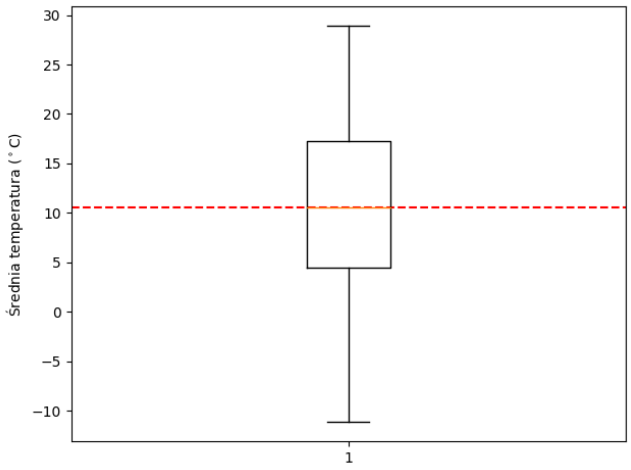
Rysunek 1: Wykres średniej temperatury od czasu

Na Wykresie 1 możemy zaobserwować wyraźną okresowość, która uświadamia nam, że badany szereg czasowy nie jest stacjonarny.

2. Przygotowanie danych do analizy

2.1. Zbadanie jakości danych

W celu oceny jakości naszych danych badamy poniższy wykres pudełkowy:



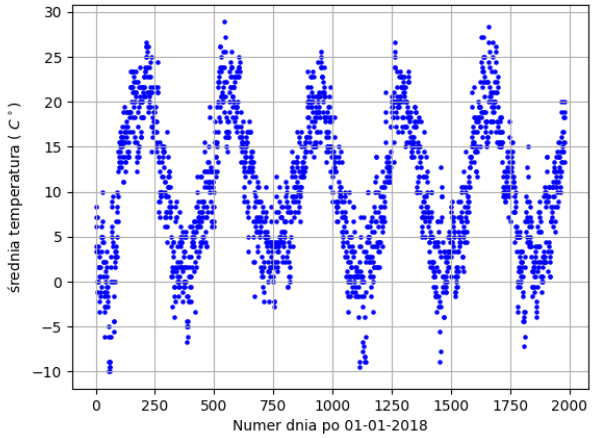
Rysunek 2: Boxplot danych temperaturowych

Możemy zauważyć, że badany zbiór danych nie posiada danych odstających, co świadczy o dobrej jakości danych. Dodatkowo przeglądając dane nie natrafiłyśmy na możliwe braki, zatem dane możemy uznać za jakościowo dobre.

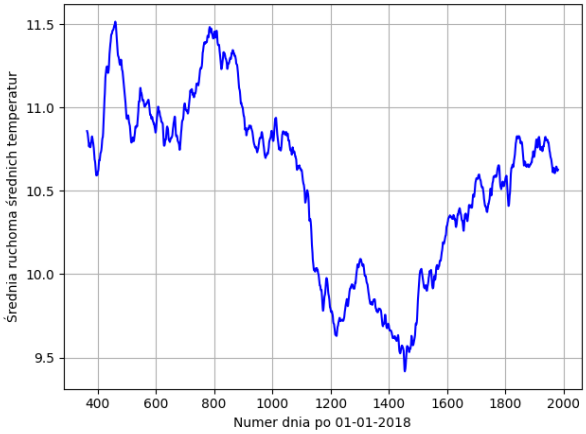
2.2. Podział danych

Z racji iż w podpunkcie 4.3. będziemy prognozować przyszłe wartości temperatur, to podzielimy zbiór danych na dwie części w następujący sposób:

* Dane treningowe: pierwsze 1982 danych (ok. 90% zbioru) - dane do których będziemy dopasowywać model ARMA
* Dane testowe: pozostałe 221 danych (ok. 10% zbioru) - dane które posłużą nam do predykcji



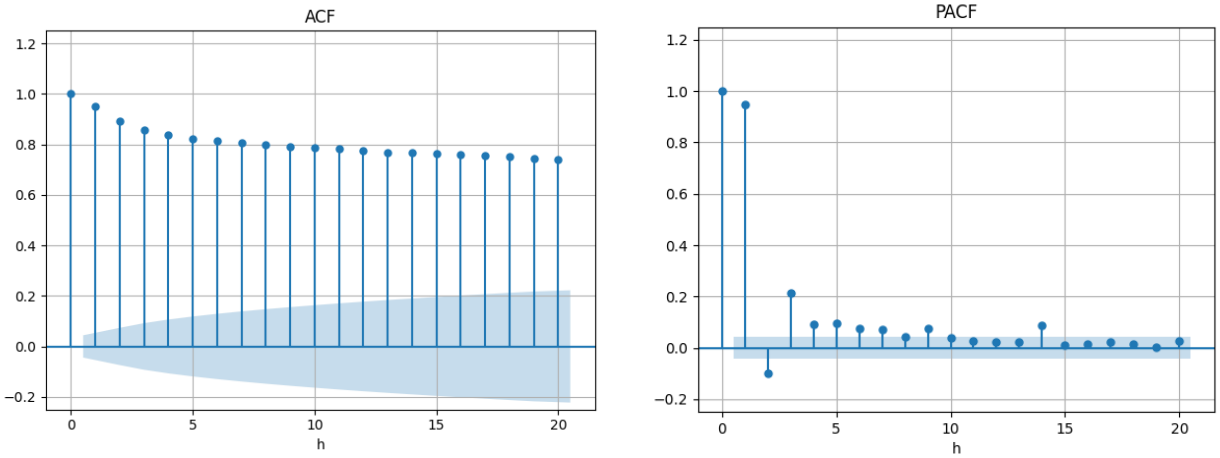
Rysunek 3: Wykres danych po podziale



Rysunek 4: Wykres średniej ruchomej dla okna 365

Na wykresie danych (Wykres 3) obserwujemy, że dane treningowe (tak jak i szereg oryginalny) są okresowe, zatem szereg posiada sezonowość. Wartości na wykresie średniej ruchomej (Wykres 4) nie mają tendencji do wyraźnego spadku ani wzrostu, zatem dane nie posiadają trendu. Jednak badany szereg czasowy nie jest stacjonarny w słabym sensie, ze względu na wspomnianą na początku sezonowość.

Tezę o niestacjonarności szeregu potwierdzają poniższe wykresy funkcji autokorelacji i częściowej autokorelacji, gdzie funkcja ACF przyjmuje dość duże wartości dla wszystkich h, a wartość PACF dla $h = 14$ wypada z przedziału ufności.



Rysunek 5: Zestawienie funkcji ACF oraz PACF dla danych

Dodatkowo przeprowadzimy test ADF (augmented Dickey-Fuller test) weryfikujący hipotezę o niestacjonarności naszych danych. Przyjmujemy poziom istotności $\alpha = 0.05$ oraz następujący zestaw hipotez:

* Hipoteza zerowa H_0: Wielomian autoregresji ma pierwiastki jednostkowe (niestacjonarność).
* Hipoteza alternatywna H_1: Wielomian autoregresji nie ma pierwiastków jednostkowych (stacjonarność).

Z testu odczytujemy, że `p-wartość` $\approx$ `0.02927`. Jest to wartosc mniejsza niż przyjęty poziom istotności, co prowadzi nas do odrzucenia hipotezy zerowej. Wynik testu potwierdza brak trendu w danych.

Pomimo tego faktu, na wykresie danych obserwujemy wyraźną okresowość danych, dlatego nadal przyjmujemy szereg za niestacjonarny.

W celu zniwelowania owej niestacjonarności, skorzystamy z metody różnicowania szeregu.

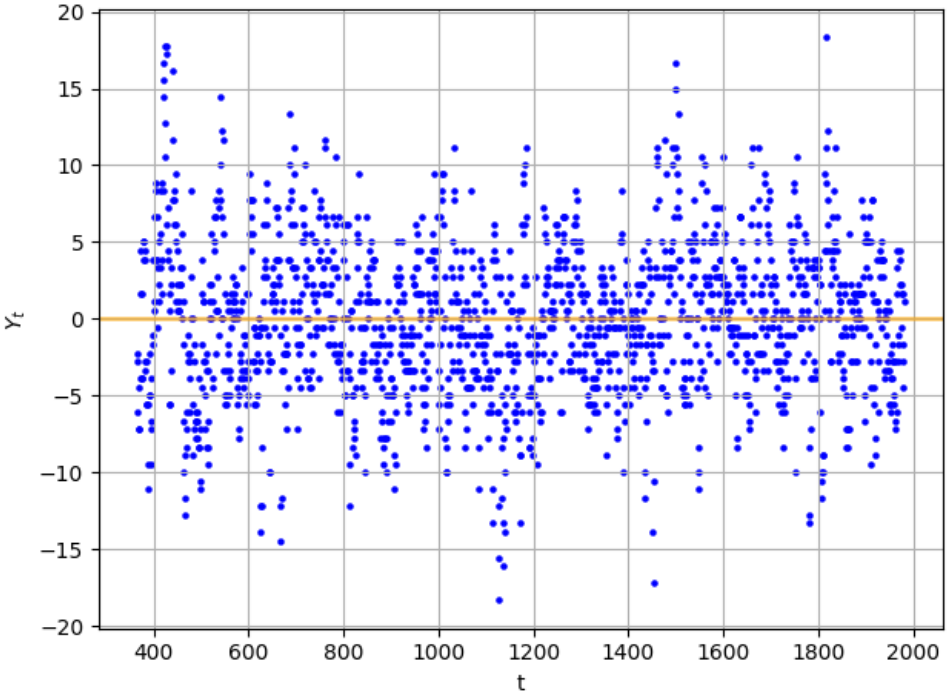
2.3. Różnicowanie szeregu czasowego

Różnicowanie szeregu czasowego polega na usuwaniu trendu i sezonowości danych (w naszym przypadku jedynie sezonowości, ponieważ nasz szereg jest pozbawiony trendu).

W danych obserwujemy sezonowość roczną. Aby ją usunąć zastosujemy metodę różnicowania z parametrem 365.

Przekształcony szereg zdefiniujemy w następujący sposób:

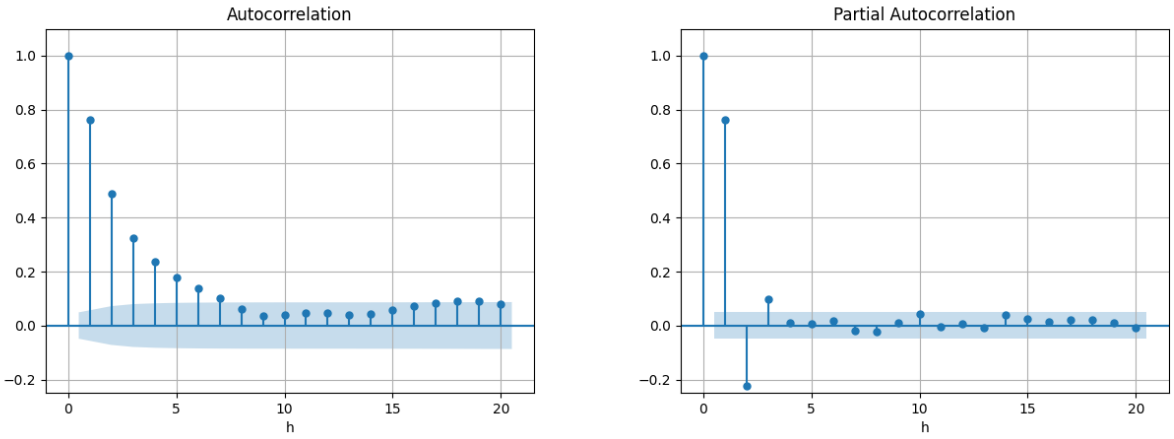
$$Y_t = X_t - X_{t-365} \quad \text{gdzie } t \in \{365, 1982\}$$



Rysunek 6: Wykres danych po usunięciu sezonowości

Z Wykresu 6 możemy odczytać, że otrzymane dane oscylują wokół zera, a ich periodyczność zaniknęła. Ponadto rozkład danych ukazuje stałość wariancji. Te cechy pozwalają nam wysunąć tezę, że otrzymany szereg czasowy jest stacjonarny w słabym sensie.

W celu potwierdzenia naszej tezy ponownie posłużymy się wykresami ACF i PACF:



Rysunek 7: Zestawienie funkcji ACF oraz PACF dla zróżnicowanych danych

Widzimy, że funkcja ACF maleje do zera w sposób wykładniczy, zaś funkcja PACF przyjmuje wartości zerowe zaraz po argumencie $h = 3$. To potwierdza naszą tezę o stacjonarności.

Dodatkowo znając cechy interpretacji wykresu PACF zauważamy, że otrzymany szereg czasowy prawdopodobnie może być opisany modelem AR(3).

Naszą tezę weryfikujemy jeszcze przy pomocy testu ADF (Augmented Dickey–Fuller test), z którego otrzymujemy `p-wartość` $\approx 7 * 10^{-27}$ (mniejsza od poziomu istotności 0.05). Zatem szereg Y_t jest stacjonarny w słabym sensie.

3. Modelowanie danych przy pomocy modelu ARMA

3.1. Dobranie rzędu p i q modelu

W naszej analizie rozważamy model ARMA ze wzoru (1), zatem dla naszego szeregu model prezentuje się nastepująco:

$$Y_t - \Phi_1 Y_{t-1} - \Phi_2 Y_{t-2} - ... - \Phi_p Y_{t-p} = Z_t + \theta_1 Z_{t-1} + ... + \theta_q Z_{t-q}$$

W celu dobrania rzędu modelu ARMA posłużymy się kryterium informacyjnym AIC (ang. Akaike Information Criterion). Owe kryterium opiera się na wskaźniku, który pokazuje jak dobrze dopasowany jest model (im mniejsza wartość statystyki tym optymalniejszy model).

Statystykę AIC obliczamy dla kombinacji $p \in \{0, 1, 2, 3, 4\}$ oraz $q \in \{0, 1, 2, 3, 4\}$.

Kryterium AIC przyjmuje najmniejszą wartość dla $p = 3$ i $q = 0$, więc najoptymalniejszym dopasowaniem do danych jest model ARMA(3,0)=AR(3).

3.2. Estymacja parametrów modelu

Nasz model będzie wyrażony poprzez poniższe równanie:

$$Y_t - \Phi_1 Y_{t-1} - \Phi_2 Y_{t-2} - \Phi_3 Y_{t-3} = Z_t$$

W celu dobrania jak najlepszych parametrów do naszego modelu (współczynniki Φ_1, Φ_2, Φ_3 i wariancja białego szumu σ^2) korzystamy z estymatorów wyznaczonych metodą największej wiarygodności. Owe estymatory mają następujące wartości:

$$\begin{cases} \hat{\Phi}_1 &= 0.956735 \\ \hat{\Phi}_2 &= -0.317180 \\ \hat{\Phi}_3 &= 0.098868 \\ \hat{\sigma}^2 &= 9.869775 \end{cases}$$

4. Ocena dopasowania modelu

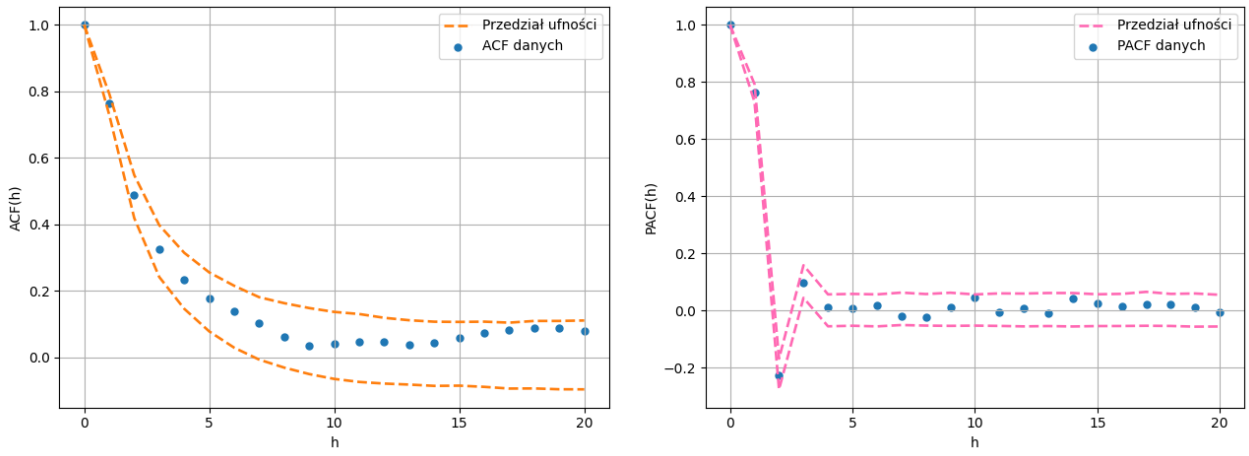
Dopasowaliśmy model, jednakże należy jeszcze zweryfikować poziom owego dopasowania. Wykorzystamy do tego trzy sposoby:
1. Wyznaczenie przedziałów ufności dla funkcji ACF oraz PACF.
2. Porównanie linii kwantylowych (przedziałów ufności) z trajektorią danych rzeczywistych.
3. Porównanie prognozy dla przyszłych obserwacji z rzeczywistymi danymi.

4.1. Przedziały ufności dla ACF/PACF

Generujemy 1000 trajektorii procesu AR(3) przy użyciu wyznaczonych w punkcie 3.2. współczynników i dla każdej trajektorii obliczamy wartości funkcji autokorelacji (ACF) oraz autokorelacji cząstkowej (PACF).

Następnie, dla każdego kroku h, gdzie $h \in \{0, ..., 20\}$, wyznaczamy kwantyle próbkowe. Korzystając ze wzoru (2), uzyskujemy przedział ufności na poziomie ufności $1 - \alpha$.

Rezultaty naszej analizy przedstawiają poniższe wykresy:



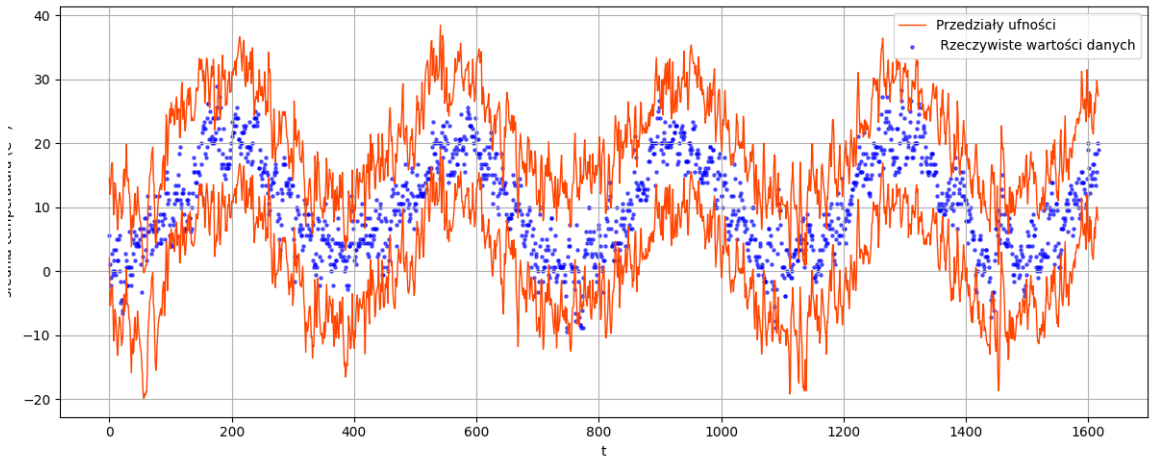
Rysunek 8: Zestawienie ACF i PACF dla szeregu Y_t z przedziałami ufności na poziomie ufności 0.95

Zauważamy, że wszyskie wartości funkcji ACF i PACF dla danych wpadają do swoich przedziałów ufności, a jednocześnie przedziały są dosyć wąskie.

4.2. Porównanie linii kwantylowych z trajektorią danych

Generujemy 10000 trajektorii z modelu ARMA (3,0) z wyestymowanymi poprzednio parametrami. Następnie do każdej wartości z owej trajektorii dodajemy temperaturę z roku wcześniejszego, tak aby przywrócić sezonowość danych. Dla tej trajektorii wyliczamy kwantyle rzędu $\frac{\alpha}{2}$ oraz $1 - \frac{\alpha}{2} (\alpha = 0.05)$, które są konieczne do wyznaczenia linii kwantylowych.

Poniższy wykres przedstawia porównanie owych lini kwantylowych z trajektorią naszych danych:

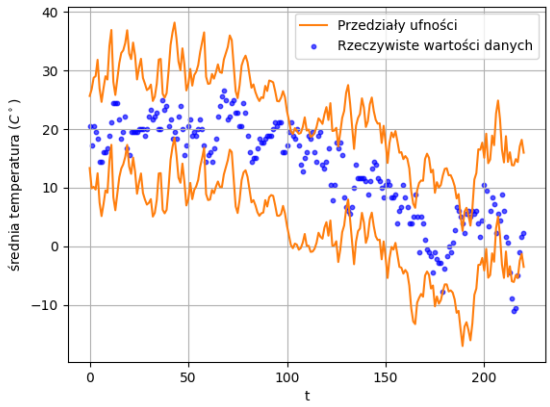


Rysunek 9: Porównanie trajektorii danych z liniami kwantylowymi

Obserwujemy, że większość danych wpada do wyznaczonych przedziałów ufności, a dodatkowo obliczyliśmy, że takich danych jest około 92%.

4.3. Porównanie prognozy przyszłych obserwacji z danymi

W tym punkcie ponownie generujemy 10000 trajektorii modelu ARMA(3,0) z parametrami wyznaczonymi wcześniej. Analogicznie do poprzedniego podpunktu, do każdej wartości trajektorii dodajemy temperaturę z roku poprzedniego (ze zbioru danych treningowych) i wyznaczamy linie kwantylowe. Następnie porównujemy linie kwantylowe z danymi testowymi:

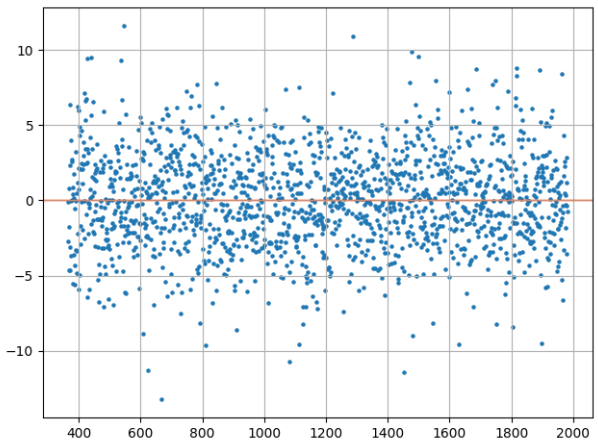


Około 90% rzeczywistych temperatur ze zbioru testowego mieści się między liniami kwantylowymi. Należy jednak zauważyć, że średnia odległość między odpowiednimi punktami linii kwantylowych wynosi około $18C°$, więć prognoza temperatury za pomocą dobranego modelu nie jest bardzo dokładna.

5. Weryfikacja założeń dotyczących szumu

W celu ostatecznego zweryfikowania poprawności dopasowania naszego modelu należy sprawdzić jeszcze założenia dotyczące białego szumu dla realizacji zmiennych losowych $\{Z_t\}$ (zwanych residuami). Założenia, które będziemy badać są następujące:

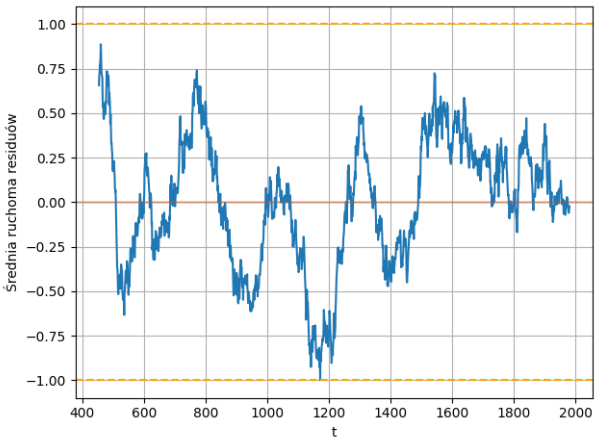
1. $E(Z_t) = 0$,
2. $Var(Z_t) < \infty$,
3. Zmienne Z_t są nieskorelowane,
4. Zmienne Z_t są z rozkładu normalnego.



Rysunek 10: Wykres rozproszenia residuów

5.1. Średnia

Aby zbadać średnią residuów posłużymy się wykresem średniej ruchomej wartości resztowych (różnice między rzeczywistymi wartościami a wartościami przewidywanymi przez model ARMA) z oknem 90.

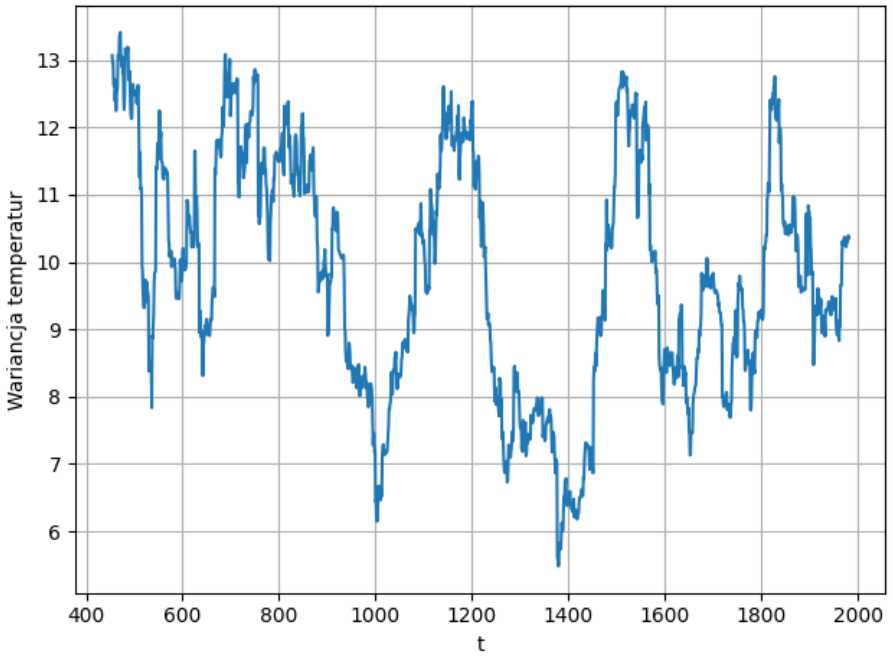


Rysunek 11: Wykres średniej ruchomej dla residuów

Na Wykresie 10 widzimy, że residua losowo układają się wokół zera. Analizując Wykres 11, zauważamy, że dla całego przedziału t wartości średniej mieszczą się w przedziale wartości [-1,1], a w dodatku głównie oscylują wokół zera. Nie obserwujemy także trendu ani sezonowości, co potwierdza stabilność średniej.

5.2. Wariancja

Na Wykresie 10 obserwujemy, że wariancja nie zmienia się w czasie. Dodatkowo obserwujemy, że na poniższym wykresie wariancja ruchoma residuów nie wykazuje wyraźnej tendencji spadkowej/wzrostowej zatem możemy stwierdzić, że dane posiadają stałą wariancję.



Rysunek 12: Wykres wariancji ruchomej residuów w oknie 90

By utwierdzić się w naszej tezie przeprowadzimy jeszcze test Modified Levene. Przyjmujemy poziom istotności $\alpha = 0.05$ oraz następującą hipotezę zerową:

$$H_0 : \text{Dane pochodzą z rozkładów o tej samej wariancji.}$$

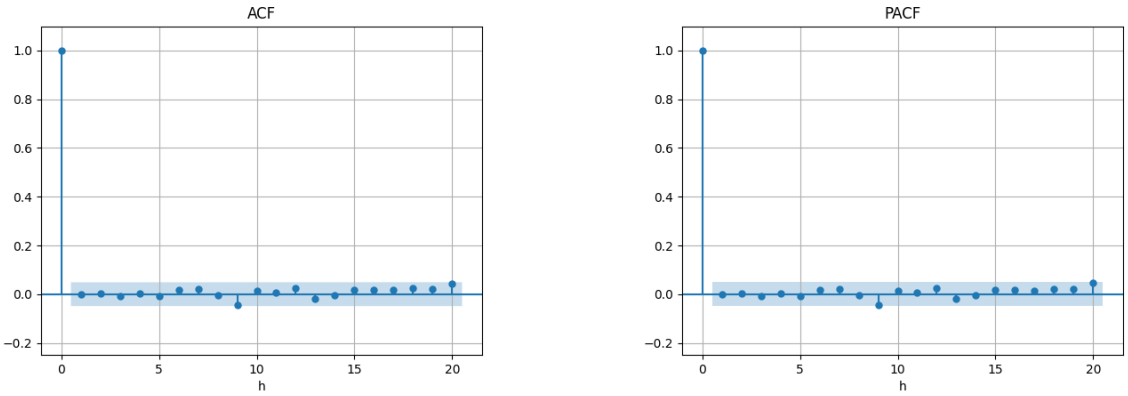
W danych pogodowych niestała wariancja residuów mogłaby wynikać z tego, że wariancja temperatur różni się w zależności od sezonów. W związku z tym, podzielimy wektor residuów na 15 części tak, aby każda z nich odzwierciedlała jeden sezon.
Po dokonanym podziale, wynik testu Levene dla próbek jest następujący:

```
LeveneResult(statistic=1.5376843772585298, pvalue=0.09075573258852325)
```

Zauważamy, że p-wartość jest powyżej ustalonego poziomu istności (0.05), zatem nie ma podstawy do odrzucenia hipotezy zerowej. Oznacza to, że residua mają stałą wariancję.

5.3. Brak korelacji

Kolejnym założeniem, które badamy, jest brak korelacji residuów.
W tym celu sporządzamy wykresy autokorelacji i częściowej autokorelacji:



Rysunek 13: Zestawienie funkcji ACF oraz PACF dla residuów

Powyższe wykresy pokazują, że dla wszystkich opóźnień $h > 0$ wartości funkcji ACF i PACF są bliskie 0, co oznacza, że dane są nieskorelowane.

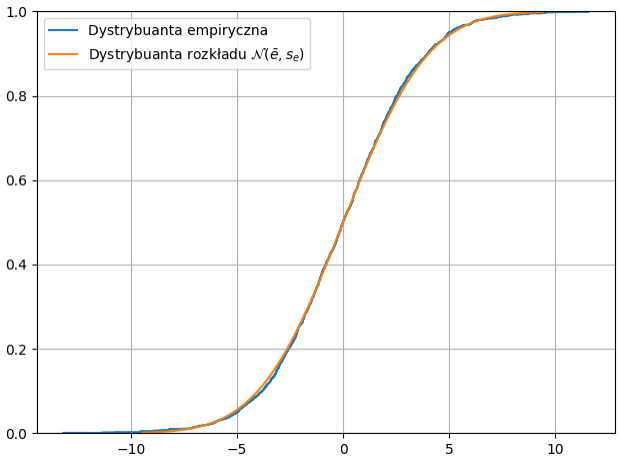
5.4. Normalność rozkładu

Ostatnim założeniem o residuach, które poddamy sprawdzeniu jest normalność ich rozkładu. Aby wypełnić ten cel posłużymy się testem Kołmogorowa-Smirnowa oraz wykresem dystrybuant.

* **Test Kołmogorowa-Smirnowa:**
  Przyjmujemy hipotezę zerową H_0: Dane pochodzą z rozkładu normalnego.

  Z testu otrzymaliśmy `p-wartość` ≈ 0.837, która jest większa niż α (nie odrzucamy H_0), czyli residua pochodzą z rozkładu normalnego.

* **Porównanie dystrybuant:** Dystrybuanty pokrywają się, co potwierdza tezę o normalności residuów.



Rysunek 14: Porównanie dystrybuanty empirycznej dla wartości residuów z dystrybuantą teoretyczną rozkładu normalnego o średniej i wariancji wyznaczonych z tych wartości

6. Wnioski

Celem tego raportu było dobranie modelu ARMA(p, q) do danych średniej temperatury we Wrocławiu. Wybrane dane nie posiadały wartości odstających ani brakujących, nie miały one również trendu. Zatem przed doborem modelu ARMA należało jedynie usunąć sezonowość, co wykonałyśmy za pomocą metody różnicowania.

Udało nam się dopasować model, który jest modelem ARMA(3,0). Stwierdzamy, że jest on dobrze dobranym modelem, ponieważ analiza dopasowania wykazała następujące własności:

* Wartości funkcji ACF i PACF wpadają do wyznaczonych dla modelu przedziałów ufności.
* Dane mieszczą się pomiędzy liniami kwantylowymi.
* Prognoza przyszłych obserwacji przebiegła pomyślnie, ponieważ większość naszych danych wpada do wyznaczonych przedziałów ufności. Jednakże warto zaznaczyć, iż owe przedziały ufności są dosyć szerokie (ok.18 C°) zatem dobrany model nie jest najlepszy do rzeczywistych prognoz temperatury.

Wysoką jakość naszego dopasowania potwierdza także przeprowadzona analiza szumu, która udowodniła wszystkie założenia białego szumu dla residuów naszych danych.

Podsumowując, przeprowadzone działania pokazały nam przydatność modelu ARMA w analizie danych rzeczywistych, jednak dokładniejsza predykcja byłaby możliwa gdybyśmy posiadali wiedzę na temat innych danych wpływających na temperaturę tj. opady deszczu, prędkość wiatru czy poziom wilgotności.

Literatura

[1] Dr hab. inż. Agnieszka Wyłomańska, prof. uczelni, *Komputerowa Analiza Szeregów Czasowych: Wykłady*, semestr zimowy 2023/2024.

[2] National Centers for Environmental Information (NCEI). (2024). Retrieved from `https://www.ncei.noaa.gov/cdo-web/datasets/GHCND/locations/CITY:PL000044/detail`