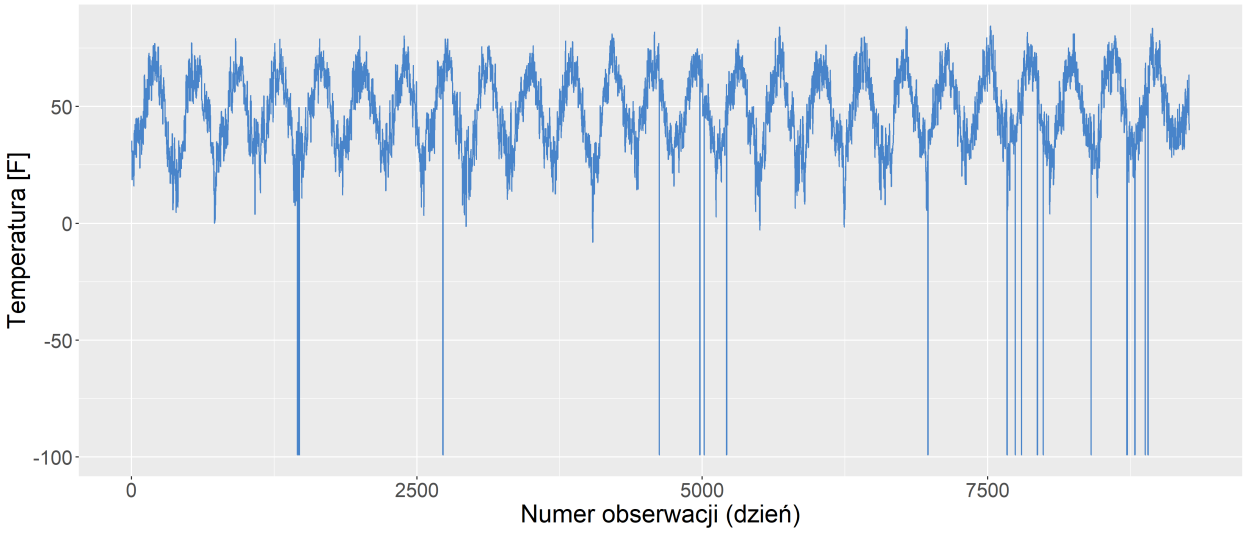
Komputerowa analiza szeregów czasowych - raport 2

Szymon Malec 262276 , Tomasz Hałas 254637

1. Wstęp i opis danych

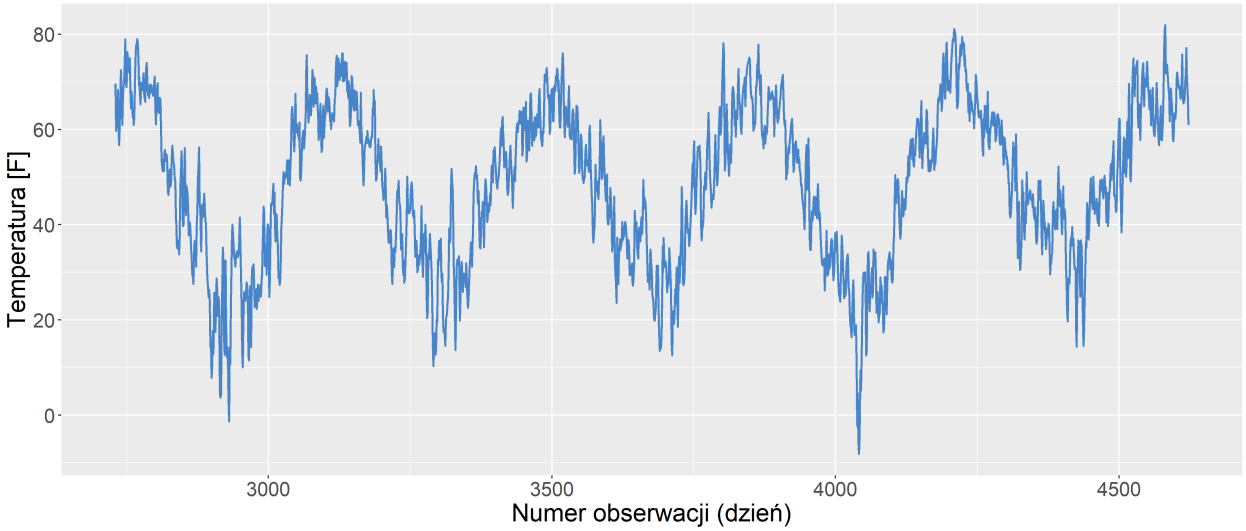
Celem niniejszej pracy jest wykorzystanie modelu ARMA do zbadania danych opisujących średnią dobową temperaturę w Warszawie. Skorzystaliśmy z danych[1] pochodzących z amerykańskiego National Climatic Data Center. Strona zawiera średnie dobowe temperatury dla 324 miast na świecie. Dane są aktualizowane na bieżąco od 1 stycznia 1995 r. do chwili obecnej. Temperatura jest mierzona w stopniach Fahrenheita. Dane pogodowe dla miasta Warszawy znadują się na wykresie 1.



Wykres 1: Średnia dobowa temperatura w latch 1995-2020 w mieście Warszawa.

Autorzy danych ustanowili konwencję, aby dla brakujących danych przypisać wartość -99 stopni Fahrenheita. Ze względu na duży rozmiar danych oraz występowanie brakujących wartości, do analizy postanowiliśmy wybrać przedział czasowy 22.06.2002 – 27.08.2007, w którym nie występują żadne braki. Wykres temperatur dla miasta Warszawa w tym okresie widoczny jest poniżej.

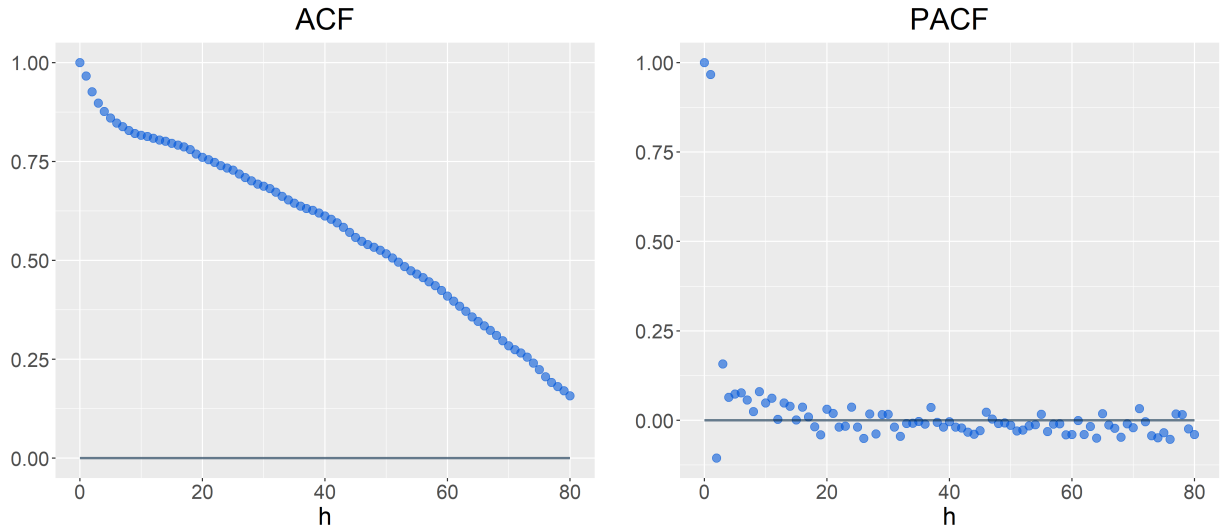
[1] https://academic.udayton.edu/kissock/http/Weather/default.htm



Wykres 2: Średnia dobowa temperatura w latch 2002-2007 w mieście Warszawa.

2. Przygotowanie danych

Na wykresie 2 zauważyć można, że w danych występuje wyraźna okresowość. Nie jest to oczywiście zaskoczeniem, ponieważ dane opisują temperaturę, która w zimie przyjmuje mniejsze wartości niż w lecie. Z powodu występującej okresowości, stwierdzamy, że szereg z pewnością nie jest stacjonarny, co potwierdzają także wykresy ACF i PACF (wykres 3). W szczególności wykres funkcji autokorelacji, która bardzo powoli zbiega do zera, wskazuje na brak stacjonarności.

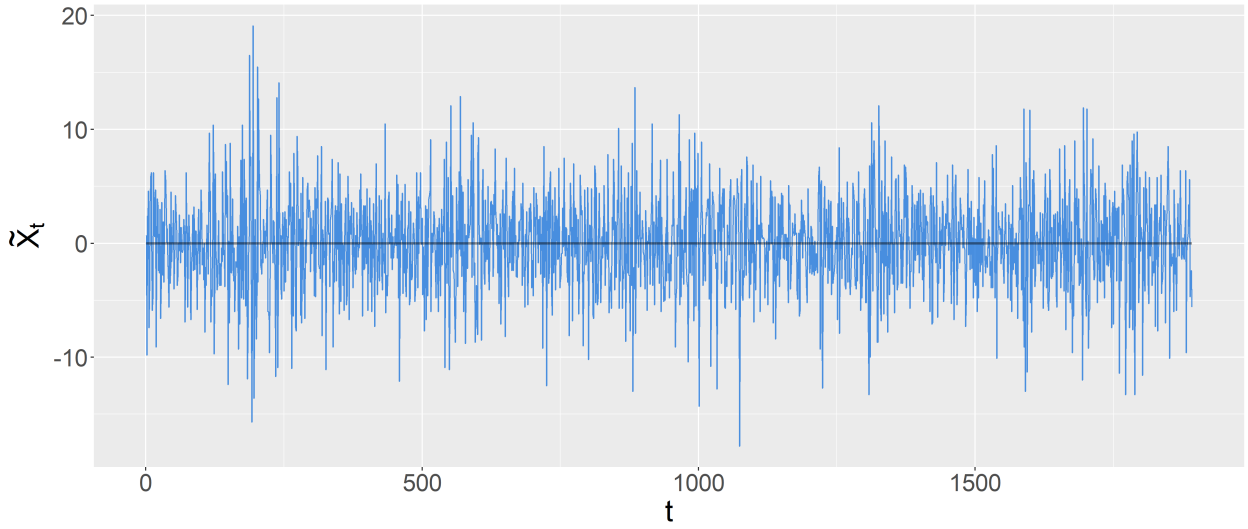


Wykres 3: Wykresy funkcji autokorelacji (ACF) i autokorelacji cząstkowej (PACF) dla surowych danych w zależności od przesunięcia h.

Oznaczmy badany szereg jako $\{X_t\},\ t \in \{1, 2, \ldots, 1893\}$. Aby uzyskać stacjonarność, skorzystamy z metody różnicowania. Definiujemy nowy szereg

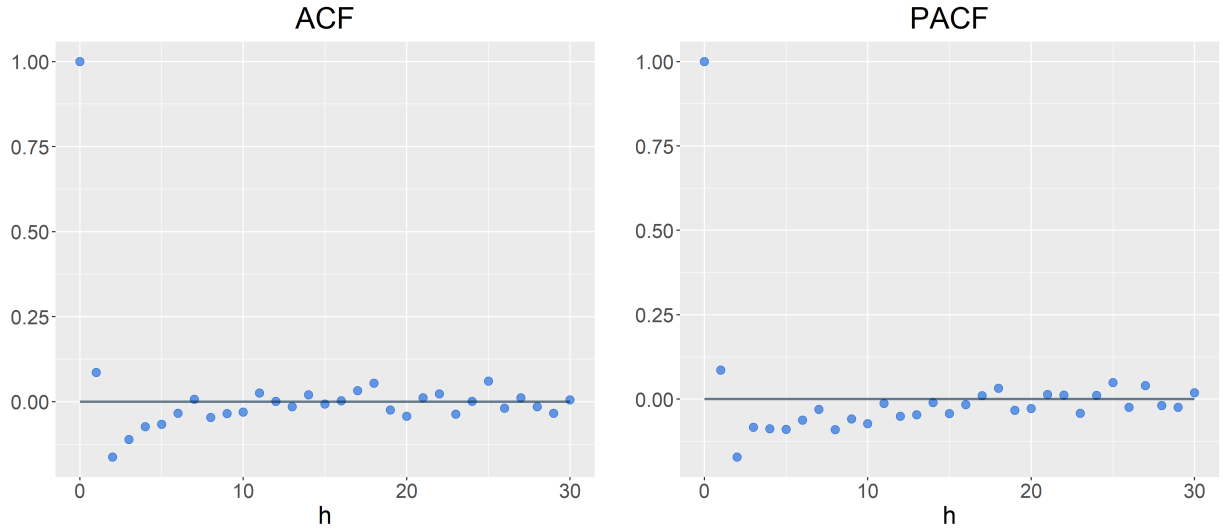
$$\tilde{X}_t = X_{t+1} - X_t, \quad t \in \{1, 2, \ldots, 1892\}.$$

Na wykresie 4 zobaczyć można, że szereg $\{\tilde{X}_t\}$ oscyluje wokół zera, a jego wariancja wydaje się być stabilna.



Wykres 4: Dane po zastosowaniu metody różnicowania.

Dodatkowo wykresy ACF i PACF (wykres 5) nie dają podstaw, by przypuszczać że, szereg nie jest stacjonarny. Sugerują one jednak, że dane są między sobą zależne (całkiem wysokie wartości korelacji dla małych h). Do zbadania tych zależności wykorzystamy model ARMA.



Wykres 5: Wykresy funkcji autokorelacji (ACF) i autokorelacji cząstkowej (PACF) dla danych po zastosowaniu metody różnicowania w zależności od przesunięcia h.

3. Dopasowanie modelu ARMA

Rozważmy ogólny model ARMA(p, q):

$$\tilde{X}_t - \varphi_1 \tilde{X}_{t-1} - ... - \varphi_p \tilde{X}_{t-p} = Z_t + \theta_1 Z_{t-1} + ... + \theta_q Z_{t-q},$$

gdzie $\{Z_t\} \sim WN(0, \sigma^2)$ jest białym szumem. W pierwszej kolejności musimy dobrać odpowiedni rząd modelu tzn. takie p i q, dla których model będzie najlepiej dopasowany do danych. W tym celu skorzystamy z kryterium informacyjnego AIC (Akaike information criterion), które wskazuje jak dobrze model jest dopasowany. Im mniejsza wartość statystyki AIC, tym lepiej jest on dopasowany. Aby wyłonić najbardziej optymalne p i q, obliczamy statystykę AIC dla wszystkich kombinacji $p \in \{1, 2, ..., 8\}$ oraz $q \in \{1, 2, ..., 8\}$. Dla naszych danych, najmniejszą wartość statystyka przyjęła dla $p = 3$ i $q = 4$. Możemy zatem uprościć model do postaci ARMA(3, 4), tj.

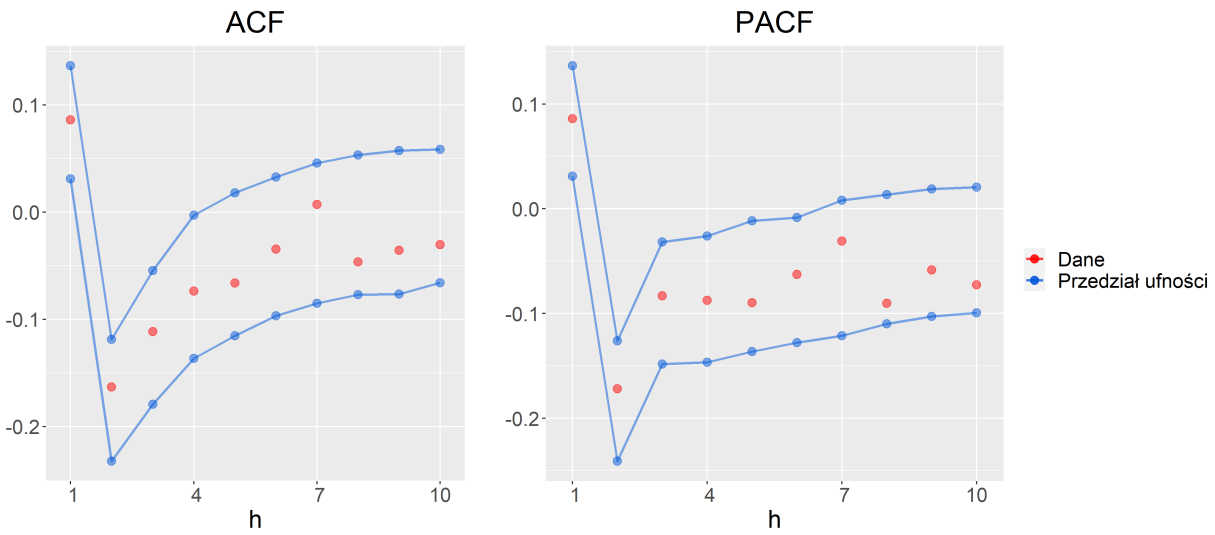
$$\tilde{X}_t - \varphi_1 \tilde{X}_{t-1} - \varphi_2 \tilde{X}_{t-2} - \varphi_3 \tilde{X}_{t-3} = Z_t + \theta_1 Z_{t-1} + \theta_2 Z_{t-2} + \theta_3 Z_{t-3} + \theta_4 Z_{t-4}.$$

Aby wyestymować współczynniki $\varphi_{1,2,3}$ i $\theta_{1,2,3,4}$ oraz σ^2, wykorzystamy pakiet Stats dostępny w języku R. Obliczone wartości parametrów widoczne są w tabeli 1. Do ich obliczenia wykorzystane zostały estymatory wyznaczone metodą największej wiarygodności.

Współczynnik	Wartość
$\hat{\varphi}_1$	0.1037
$\hat{\varphi}_2$	0.0498
$\hat{\varphi}_3$	0.1979
$\hat{\theta}_1$	-0.0583
$\hat{\theta}_2$	-0.2693
$\hat{\theta}_3$	-0.3274
$\hat{\theta}_4$	-0.0691
$\hat{\sigma}^2$	16.5

Tablica 1: Wartości współczynników badanego modelu ARMA(3, 4) wyznaczone za pomocą estymatorów największej wiarygodności.

Aby sprawdzić, czy model został dobrze dopasowany, przeprowadzimy symulację Monte Carlo w celu wyznaczenia przedziałów ufności dla funkcji ACF i PACF naszego modelu. Generujemy 10 000 trajektorii procesu ARMA(3, 4) z wyznaczonymi współczynnikami i dla każdej trajektorii obliczamy ACF i PACF. Natępnie, dla każdego przesunięcia h, wyznaczamy kwantyle próbkowe (rzędu $\frac{\alpha}{2}$ i $1 - \frac{\alpha}{2}$). W ten sposób dla każdego h otrzymujemy przedział ufności na poziomie ufności $1 - \alpha$.



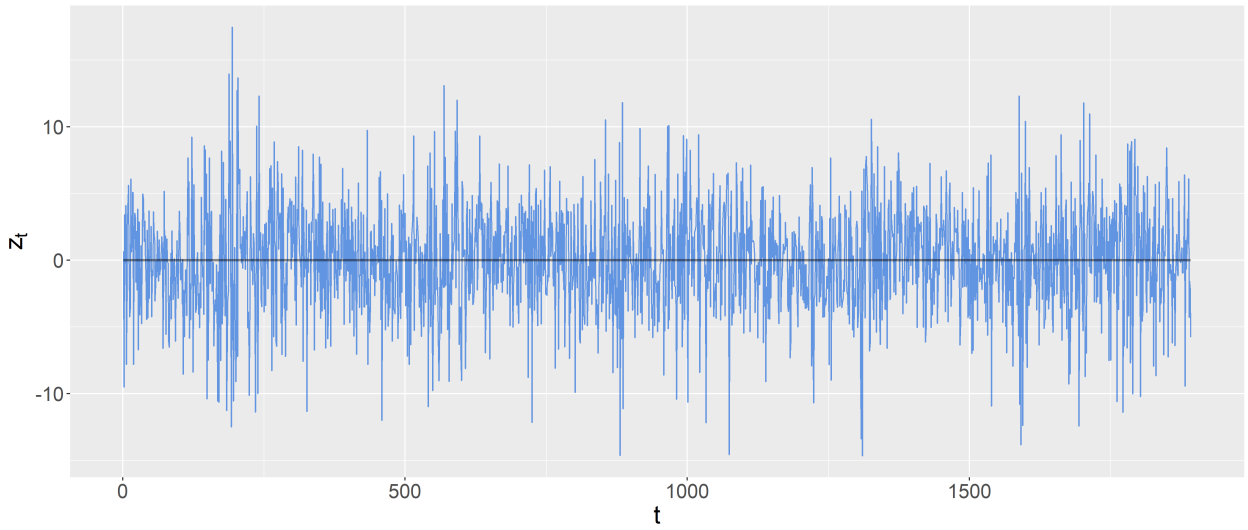
Wykres 6: Wykresy ACF i PACF dla szeregu $\tilde{X}_t$ wraz z przedziałami ufności na poziomie ufności $1 - \alpha = 0.95$.

Jak możemy zauważyć na wykresie 6, wartości ACF i PACF dla naszych danych mieszczą się w wyznaczonych przedziałach. Zatem możemy stwierdzić, że model został dobrany prawidłowo.

4. Analiza szumu

Aby zweryfikować, czy dobrany model ARMA jest w pełni poprawny, musimy jeszcze sprawdzić, czy dane spełniają wszystkie założenia modelu. Mianowicie sprawdzimy, czy residua (realizacje zmiennych losowych $\{Z_t\}$) spełniają założenia dotyczące białego szumu:

- stała średnia równa 0,
- stała i ograniczona wariancja,
- niezależność.

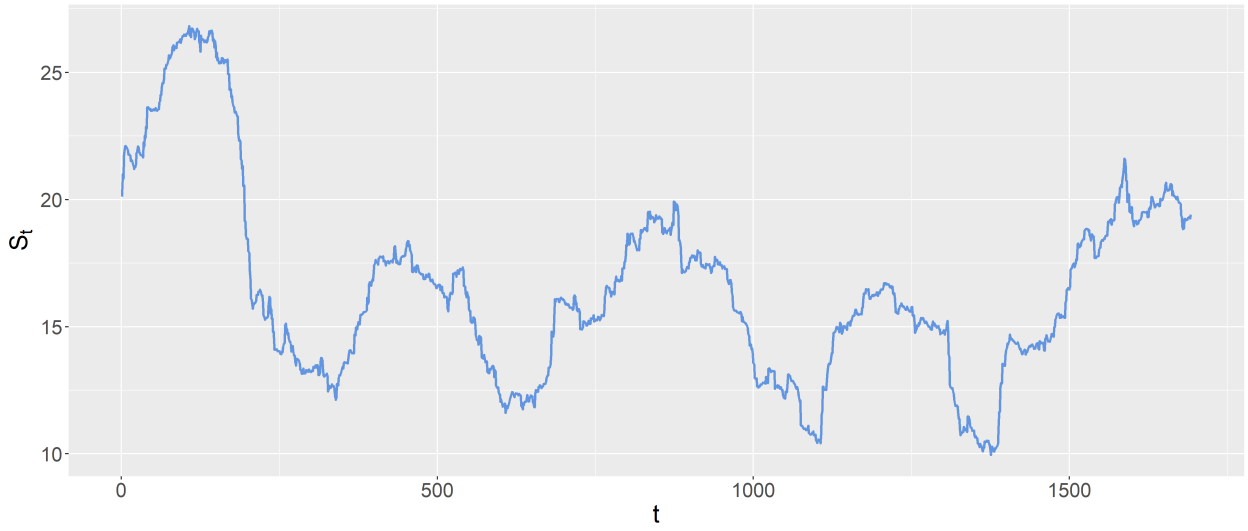


Wykres 7: Wykres wartości resztowych (residuów).

Widzimy, że na wykresie 7 nasze wartości oscylują wokół 0, z czego wnioskujemy, że średnia utrzymuje się na poziomie 0. Wśród residuów nie widać znaczących wartości odstających, więc wariancja jest raczej ograniczona. Aby odpowiedzieć na pytanie, czy wariancja jest stała, skorzystamy z wariancji cząstkowej

$$S_t^2 = \frac{1}{h-1} \sum_{i=t}^{t+h-1} (z_i - \bar{z})^2,$$

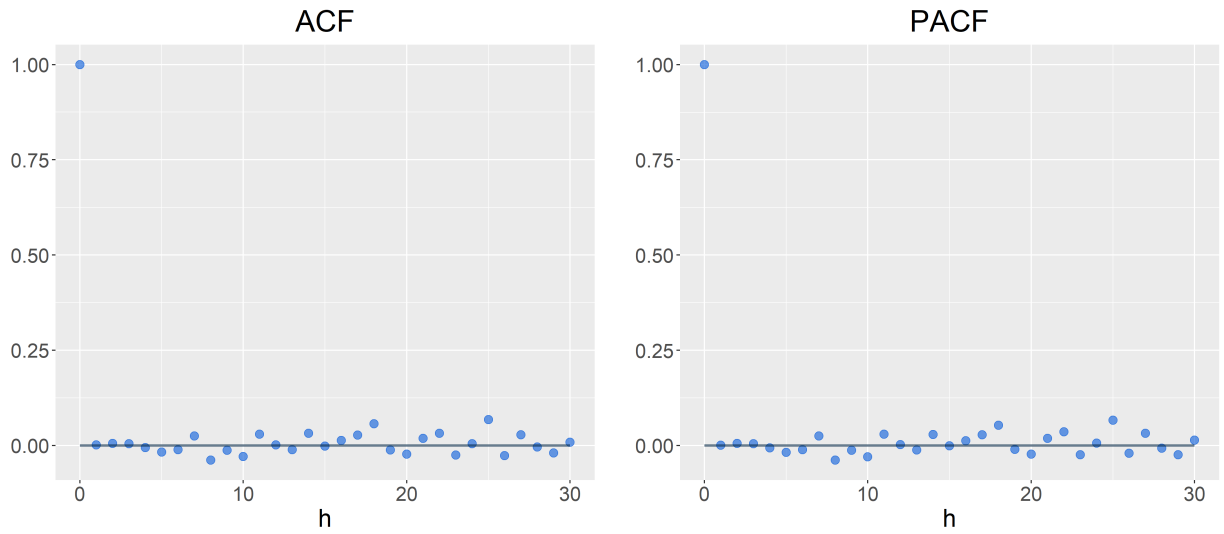
gdzie $\{z_i\}$ oznaczają residua, a h jest długością przedziału, z jakiego wyliczana jest wariancja.



Wykres 8: Wykres wariancji cząstkowej dla $h = 200$.

Na wykresie 8 wyraźnie dostrzegalne jest okresowe zachowanie wariancji. Co więcej, zachowanie to przypomina okresowość pierwotnych danych. Porównując wykresy 2 i 8 dostrzegamy,

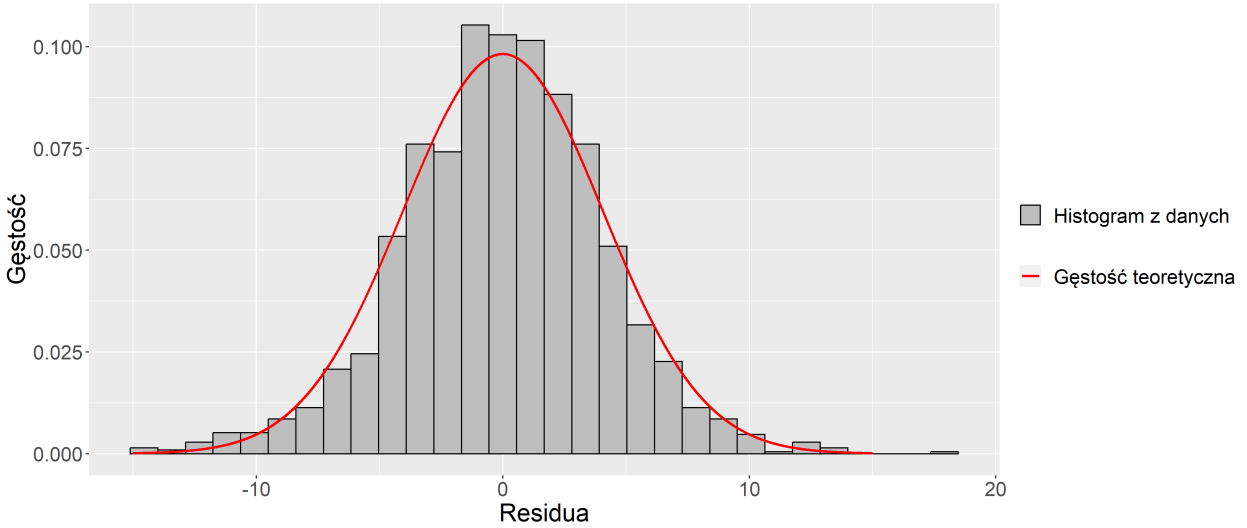
że w zimie wariancja residuów jest większa niż w lecie. Zatem założenie o stałej wariancji nie możemy uznać za spełnione. W takiej sytuacji należałoby się zastanowić nad użyciem transformacji Boxa-Coxa w celu stabilizacji wariancji.



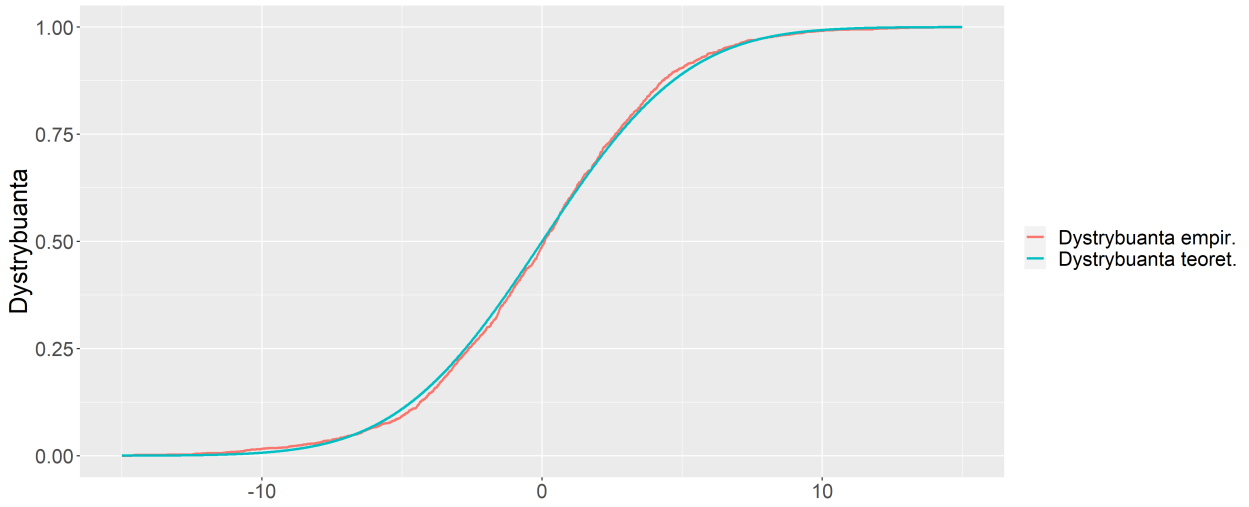
Wykres 9: Wykresy ACF i PACF dla wartości resztowych.

Następnie analizie poddaliśmy niezależność. Na podstawie wykresów autokorelacji oraz autokorelacji cząstkowej (wykres 9) możemy wnioskować, że są one niezależne od siebie, ponieważ residua oscylują blisko wokół 0 dla każdego $h \neq 0$.

Dodatkowo możemy sprawdzić, czy wartości resztowe mają rozkład normalny. W tym celu porównujemy histogram z residuów z gęstością teoretyczną rozkładu $\mathcal{N}(0, 16.5)$ oraz dystrybuantę empiryczną z dystrybuantą teoretyczną wspomnianego rozkładu.



Wykres 10: Porównanie histogramu z wartości resztowych z gęstością teoretyczną rozkładu $\mathcal{N}(0, 16.5)$.



Wykres 11: Porównanie dystybuanty empirycznej z wartości resztowych i dystrybuanty teoretycznej rozkładu $\mathcal{N}(0, 16.5)$.

Na podstawie wykresów 10 i 11 wnioskujemy, że residua najprawdopodobniej mają rozkład normalny, ponieważ dystrybuanty wyraźnie się pokrywają, z kolei kształt histogramu przypomina gęstość teoretyczną. W celu weryfikacji hipotezy dotyczącej rozkładu, wykonaliśmy test Kołmogorowa-Smirnowa. Otrzymana p-wartość wynosi 0.2905, co nie daje nam powodów do odrzucenia hipotezy.

5. Podsumowanie

Na podstawie wykresu 6 widzimy, że nasz model został poprawnie dopasowany do danych. Dodatkowo potwierdza to przeprowadzona analiza szumu. Udało nam się pokazać, że jest on z rozkładu normalnego, co pozwola na symulację badanego procesu ARMA, związku z czym także na przeprowadzenie predykcji. Jednak należy mieć na uwadze, że w naszych danych pojawia się zmienna wariancja, co może mieć negatywny wpływ na wyniki predykcji, dlatego należałoby rozważyć wykorzystanie metody Boxa-Coxa w celu stabilizacji wariancji. Poza tym cały model jest poprawnie skonstruowany.