

Analiza danych rzeczywistych przy pomocy modelu ARMA

Autorzy

Kacper Budnik, 262286
Maciej Karczewski, 262282



Politechnika Wrocławska

Wydział Matematyki 9 lutego 2023

Spis treści

1	Wprowadzenie	2
2	Przygotowanie danych do analizy	2
2.1	Sprawdzenie stacjonarności	2
3	Dekompozycja szeregu czasowego	3
3.1	Wykresy dla surowych danych	3
3.2	Różnicowanie sezonowe	3
4	Dobranie modelu ARMA	4
4.1	Rząd modelu	4
4.2	Estymacja parametrów modelu	5
5	Ocena dopasowania modelu	5
6	Analiza szumu	6
6.1	Stała średnia równa 0	6
6.2	Stała wariancja	7
6.3	Niezależność szumu	8
6.4	Założenie o normalności rozkładu	8
7	Wnioski autorów	10

1 Wprowadzenie

Analizowane dane pochodzą ze strony Kaggel. Zawierają one dane pogodowe z Indyjskiego miasta Delhi od dnia 1 listopada 1996 do 24 kwiecień 2017. Dane były pobierane w przynajmniej ośmiu ustalonych momentach dnia w odstępach trzygodzinnych. W pierwszych latach przewarżał odstęp godzinny między kolejnymi pomiarami. Zawierają one informacje między innymi o wilgotności, temperaturze punktu rosy, zjawiskach atmosferycznych, czyli informacje czy wystąpiły opady, burze, mgły i tym podobne oraz dacie pomiaru. Przykładowe dane prezentują się następująco.

<code>datetime_utc</code>	<code>_dewptm</code>	<code>_fog</code>	<code>_hail</code>	<code>_hum</code>	<code>_tempm</code>
19961101-11:00	9	0	0	27	30
19961101-12:00	10	0	0	32	28
19961101-13:00	11	0	0	44	24
19961101-14:00	10	0	0	41	24
19961101-16:00	11	0	0	47	23
19961101-17:00	12	0	0	56	21

Tabela 1: Przykładowe dane które poddamy analizie.

W analizie interesują nas jedynie kolumny `datetime_utc` przechowująca informację o dacie pomiaru oraz `_tempm` zawierająca odnotowaną temperaturę. Będziemy analizować zachowanie średniej temperatury dziennej w zależności od średnich z poprzednich dni.

2 Przygotowanie danych do analizy

Analizowane dane obejmują okres od listopada 1996 roku, jednak w pierwszych latach dane były pobierane w nieregularnie. Dlatego rozpatrujemy dane z okresu od 1 stycznia 2008 do 31 grudnia 2015 roku, równo 8 lat. Pozostałe dane tj. pochodzące z okresu 1 styczeń 2016 – 24 kwiecień 2017 pozostawiliśmy jako dane testowe. W tych okresach pomiary były robione 8 razy dziennie, co 3 godziny każdy, rzadkie przypadki pomiarów w dodatkowych porach zostały pominięte. Za obserwacje odstające uznaliśmy te, które spełniają poniższą własność

$$y \notin (Q_1 - 1.5IQR, Q_3 + 1.5IQR),$$

gdzie y jest kandydatem na obserwacje odstającą, Q_1 i Q_3 to odpowiednio pierwszy i trzeci kwantyl, a IQR jest rozstępem międzykwantylowym. W wyniku tej obserwacji odrzuciliśmy jedynie 4 obserwacje rzędu 90°C. Pozostałe wyniki należą do przedziału od 1°C do 47°C. Naszym celem jest analiza średniej dziennej temperatury, zatem w kolejnym kroku obliczyliśmy tą średnią. Jej wartości mieszczą się w przedziale od 6°C do 40°C. Uzyskane średnie temperatury rzeczywiście były osiągane w Indiach w odpowiednich czasach.

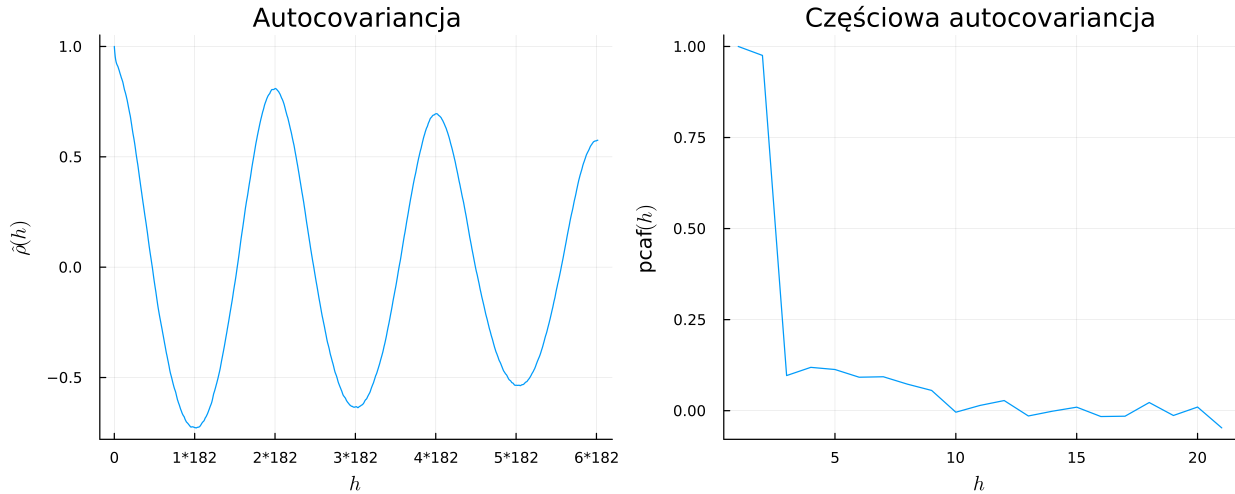
2.1 Sprawdzenie stacjonarności

W celu sprawdzenia, czy nasze dane w postaci surowej są stacjonarne, posłużymy się testem ADF (Augmented Dickey-Fuller Test). W tym celu skorzystaliśmy z funkcji `ADFTest` z pakietu `HypothesisTests` dla języka Julia. Ponieważ nasze dane są pobierane codziennie, rozpatrywaliśmy lag do wartości 365. W ten sposób $p\text{-value}=0.2619$, zatem nie mamy podstaw do odrzucenia hipotezy zerowej głoszącej, że szereg nie jest stacjonarny.

3 Dekompozycja szeregu czasowego

3.1 Wykresy dla surowych danych

Dekompozycję zaczniemy od analizy wykresów funkcji autokowariancji oraz częściowej autokowariancji. Prezentują się one następująco



Rysunek 1: Funkcje empirycznej autokowariancji i częściowej autokowariancji.

Na powyższym wykresie widać okresowe zachowanie funkcji autokowariancji z okresem około 365. Liczba ta pokrywa się z liczbą dni w roku, co wydaje się intuicyjne. Z powodu na globalne ocieplenie możemy podejrzewać, że będzie istniał dodatni trend. Rzeczywiście, korzystając z funkcji `polyfit`, otrzymaliśmy trend

$$m(h) = 24.8564 + 9 \cdot 10^{-5}h,$$

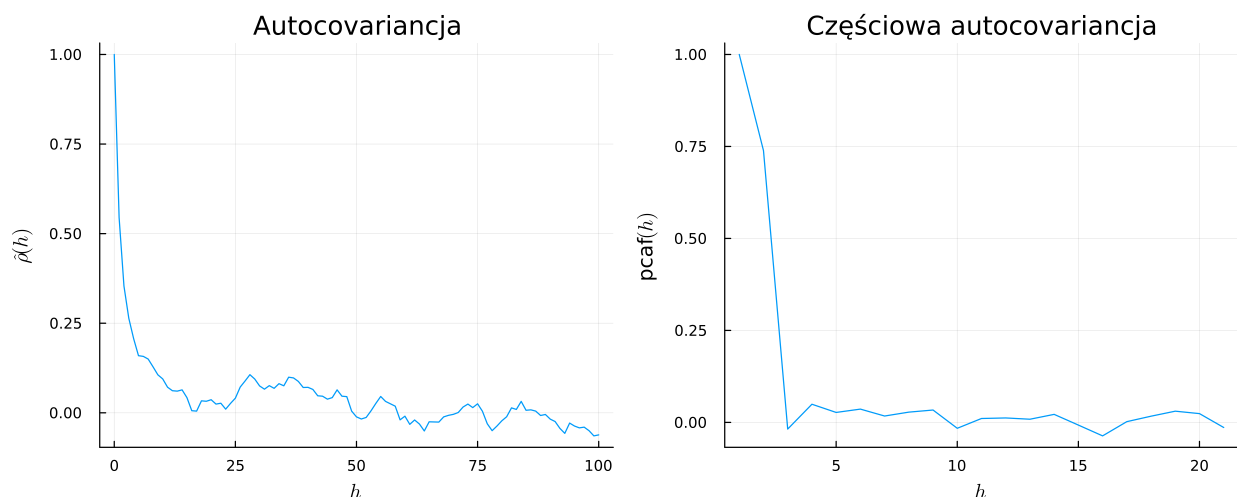
gdzie m oznacza zmianę średniej temperatury w ciągu dnia. Zatem w ciągu 10 lat średnia temperatura w Delhi, według naszych obliczeń, wzrosła w przybliżeniu o $m(10 \cdot 365) - m(0) \approx 0.3287^\circ\text{C}$. Według danych w tym samym okresie średnia temperatura na ziemi wzrosła, w zależności od regionu i źródeł, o więcej niż 0.2°C , zatem nasze wyniki częściowo pokrywają się z obserwowanymi zdarzeniami. Trend ten obliczyliśmy jedynie w celu sprawdzenia poprawności danych. W celu uodrębnienia części stacjonarnej z danych posłużyliśmy się metodą różnicowania sezonowego.

3.2 Różnicowanie sezonowe

Tak jak już wspomnieliśmy, przy analizie wykresu 1, nasze dane wykazują okresowość z okresem ok. 365 dni. By pozbyć się tych sezonowości zastosowaliśmy różnicowanie sezonowe, czyli analizowaliśmy dalej dane w postaci

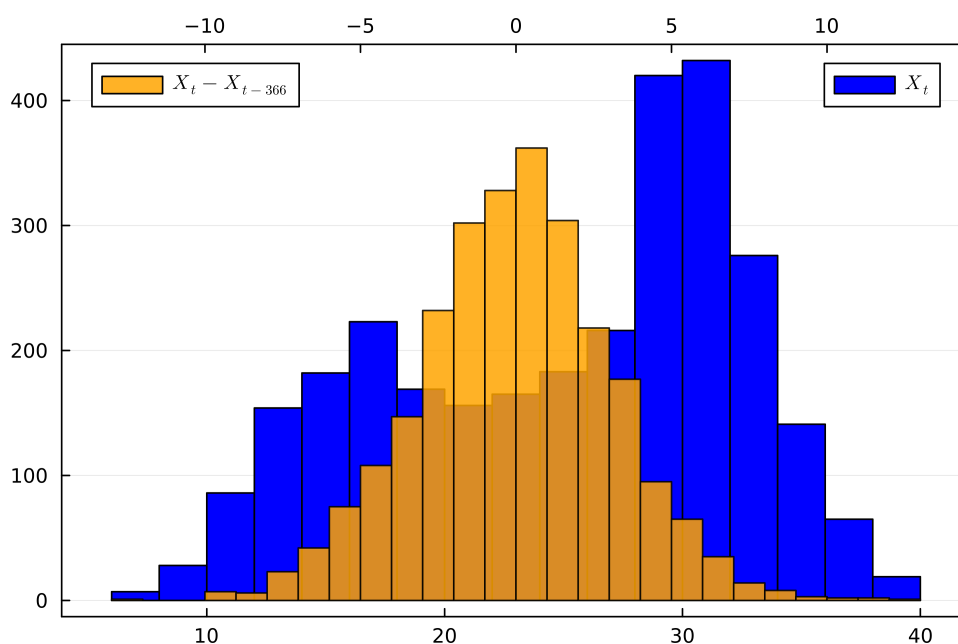
$$W_t = Y_t - Y_{t-a}.$$

W naszym przypadku a powinno być blisko 365. Dokładną wartość $a = 366$ dobraliśmy w taki sposób, by funkcja autokowariancji była jak najmniejsza dla początkowych 10 wartości. Dla szeregu $W_t = Y_t - Y_{t-366}$ wygenerowaliśmy ponownie wykresy autokowariancji i częściowej autokowariancji.



Rysunek 2: Funkcje empirycznej autokowariancji i częściowej autokowariancji.

Na wykresach widać, że funkcja częściowej i zwykłej autokowariancji szybko zbiega do wartości bliskich zero i pozostaje w jego otoczeniu. Zatem możemy założyć, że nasz szereg jest już stacjonarny. By sprawdzić to założenie ponownie wykonaliśmy ADF-Test. W tym przypadku $p\text{-value} = 0.002$, zatem możemy odrzucić hipotezę głoszącą, że szereg nie jest stacjonarny. Histogram naszych danych po transformacji wygląda następująco.



Rysunek 3: Dane przed i po transformacji. Dolna oś wskazuje na dane oryginalne, na górze widnieje oś dla danych po transformacji.

4 Dobranie modelu ARMA

4.1 Rząd modelu

W celu znalezienia rzędu modelu skorzystaliśmy z kryterium informacyjnego Akaike. Jest to metoda największej wiarygodności z karą dla danego p i q w postaci

$$AIC = -2l(x; \phi; \theta, \sigma^2) + 2(p + q + 1),$$

gdzie l jest funkcją wiarygodności. Zatem dla różnych wartości p oraz q otrzymamy różne wyniki. Wybieramy wtedy ten z najmniejszą wartością. W symulacji sprawdziliśmy kryterium AIC dla wartości p, q od 0 do 10. Poniżej przedstawiliśmy jedynie część tabeli zawierającą wartość najmniejszą.

	$q = 0$	$q = 1$	$q = 2$	$q = 3$	$q = 4$
$p = 0$	13053.2	11716.4	11333.5	11180.9	11124.3
$p = 1$	11039.6	11040.7	11035.9	11034.7	11032.6
$p = 2$	11040.8	11040.5	11027.3	11028.3	11030.2
$p = 3$	11036.6	11027.6	11030.0	11029.9	11031.7
$p = 4$	11036.7	11028.3	11030.4	11032.8	11030.8

Tabela 2: Wartości kryterium informacyjnego AIC dla wartości p i q nie większych niż 5. Na czerwono znaczone najmniejszą wartość ze wszystkich rozpatrywanych wartości.

Zatem naszym modelem będzie model $ARMA(2, 2)$.

4.2 Estymacja parametrów modelu

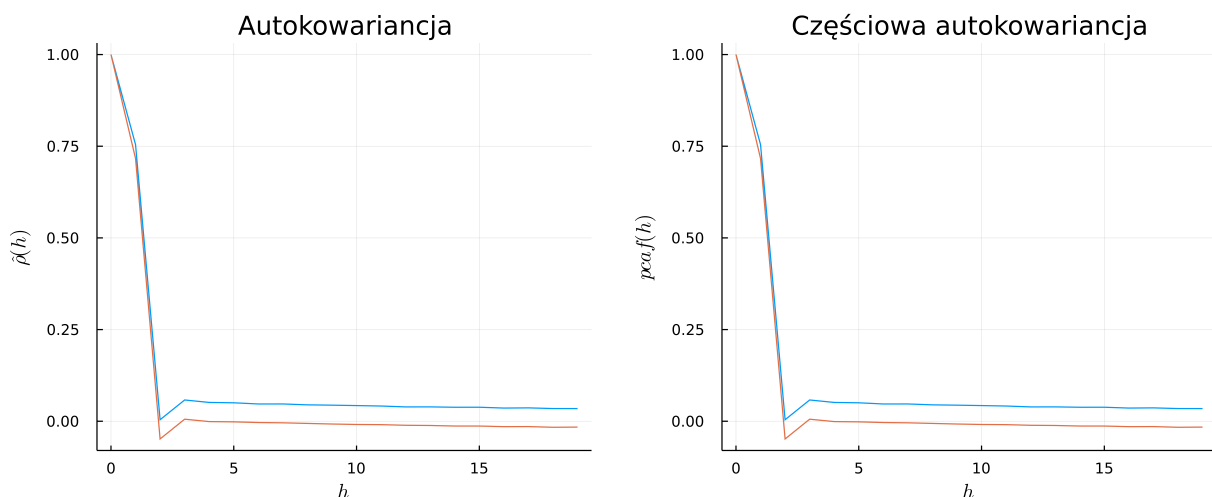
Przy znanych parametrach p i q możemy zacząć estymować współczynniki modelu. Do estymacji korzystaliśmy z metody największej wiarygodności. Skorzystaliśmy z numerycznego przybliżenia przy pomocy funkcji `ARIMA.fit()` z pakietu `statsmodels.tsa.arima.model` w języku Python. W wyniku dostaliśmy model

$$Y_t - 1.6118 Y_{t-1} + 0.6232 Y_{t-2} = Z_t - 0.8640 Z_{t-1} - 0.0688 Z_{t-2},$$

gdzie Z_t jest białym szumem o wariancji $\sigma^2 = 4.3478$.

5 Ocena dopasowania modelu

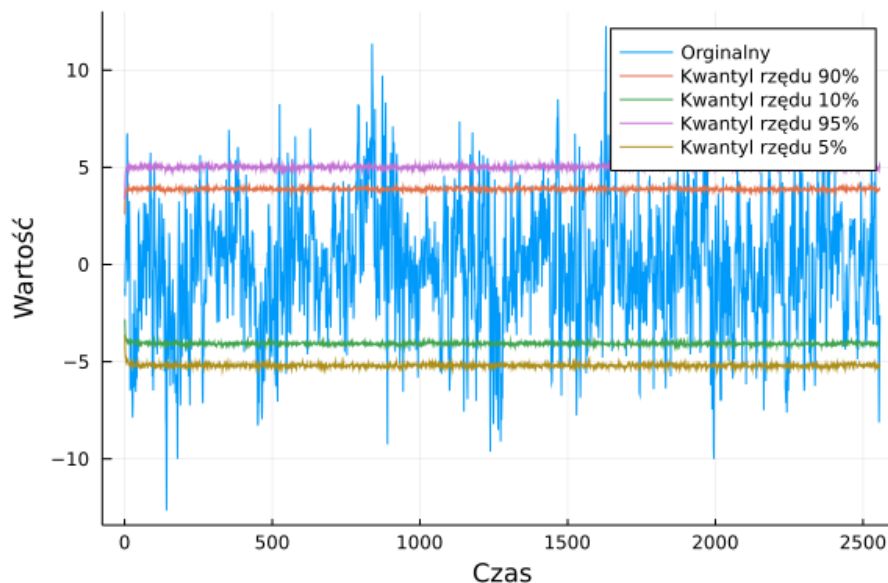
W celu oceny dopasowania modelu do danych zobaczymy jak wyglądają przedziały ufności, wyznaczone za pomocą Monte Carlo, na poziomie $\alpha = 5\%$ dla ACF i PACF



Rysunek 4: Przedziały ufności dla ACF i PACF dla naszego modelu

Jak możemy zobaczyć na wykresie 4 istnieje tylko krótko okresowa zależność pomiędzy danymi w modelu. Ta zależność pokrywa się z naszymi oczekiwaniami, ponieważ używamy modelu $ARMA(2,2)$ o parametrach 4.2 gwarantujących racjonalność.

Następnie porównamy trajektorię z liniami kwantylowymi dla naszego modelu.



Rysunek 5: Porównanie oryginalnych danych z liniami kwantylowymi modelu

Analizując linie kwantylowe naszego modelu, wyznaczone również za pomocą symulacji Monte Carlo, pokazane na wykresie 5 możemy zauważyć, że model w miarę dobrze się sprawdza dla naszych danych. 80% danych leży pomiędzy linią kwantylową rzędu 0.9 i 0.1. Natomiast 89% danych leży pomiędzy linią kwantylową rzędu 0.95 i 0.05

Patrząc na wykres zależności w modelu 4 oraz na wykres z liniami kwantylowymi 5 dochodzimy do wniosku, że model dobrze się sprawdza dla naszych danych.

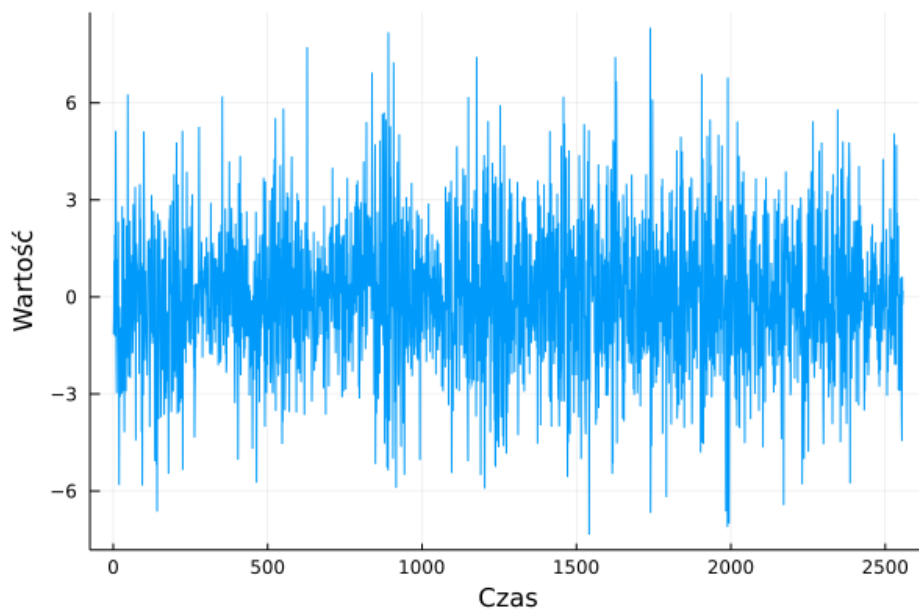
6 Analiza szumu

Podczas tworzenia modelu ARMA i późniejszej analizy zakładaliśmy następujące warunki odnośnie szumu

1. $\mathbb{E}\xi_i = 0 \quad \forall i$,
2. $Var\xi_i = \sigma^2 < \infty \quad \forall i$,
3. $\xi_i \perp \xi_j$ dla $i \neq j$,
4. ξ_i mają rozkład normalny,

6.1 Stała średnia równa 0

Sprawdzimy, czy residua naszego modelu mają stałą średnią równą 0. W tym celu zobaczymy jak wyglądają nasze wartości resztowe.



Rysunek 6: Residua naszego modelu

Jak możemy zobaczyć na wykresie 6 średnia jest stała w czasie i wynosi w przybliżeniu 0 a dokładnie $7.9 \cdot 10^{-4}$. W celu upewnienia się wykonamy też t test dla jednej zmiennej.

```
One sample t-test
-----
Population details:
  parameter of interest:  Mean
  value under h_0:       0
  point estimate:        0.000791916
  95% confidence interval: (-0.08009, 0.08167)

Test summary:
  outcome with 95% confidence: fail to reject h_0
  two-sided p-value:         0.9847

Details:
  number of observations:  2557
  t-statistic:             0.019200134473658807
  degrees of freedom:      2556
  empirical standard error: 0.04124531616232871
```

Rysunek 7: Test t dla naszych wartości resztowych

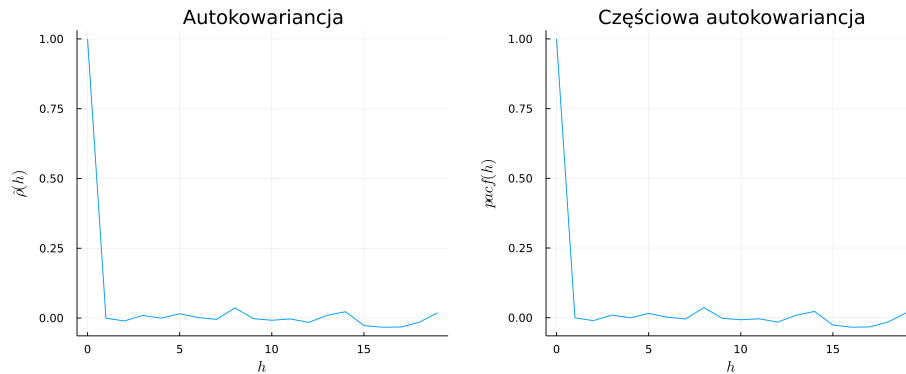
Jak możemy zobaczyć test t 7 nie miał podstaw do odrzucenia hipotezy zerowej, która mówiła ,że średnia wynosi 0. Test zwrócił p -value równe 0.9847 co jest bardzo dużą wartością. Tak więc z pewnością możemy przyjąć, że średnia jest stała i równa 0

6.2 Stała wariancja

Z wykresu 6 możemy odczytać,także że wariancja jest stała. Dodatkowo możemy obliczyć ,że należy ona do przedziału $[4.141; 4.554]$ na poziomie ufności $\alpha = 5\%$

6.3 Niezależność szumu

Sprawdźmy teraz założenie dotyczące niezależności szumu w czasie. Sprawdzimy na początek wykres autokowariancji i częściowej autokowariancji.



Rysunek 8: ACF i PACF dla wartości reszkowych modelu

Na wykresie 8 możemy zobaczyć, że mamy szum jest zależny tylko od siebie w tym samym momencie, ponieważ tylko dla $h = 0$ ACF oraz PACF przyjmują wartość niebędącą w otoczeniu zera.

Dodatkowo wykonamy test Ljunga-Boxa by upewnić się, że szum jest niezależny od siebie.

```
Ljung-Box autocorrelation test
-----
Population details:
  parameter of interest: autocorrelations up to lag k
  value under h_0:      "all zero"
  point estimate:       NaN

Test summary:
  outcome with 95% confidence: fail to reject h_0
  one-sided p-value:      0.9788

Details:
  number of observations: 2557
  number of lags:        1
  degrees of freedom correction: 0
  Q statistic:            0.000705594
```

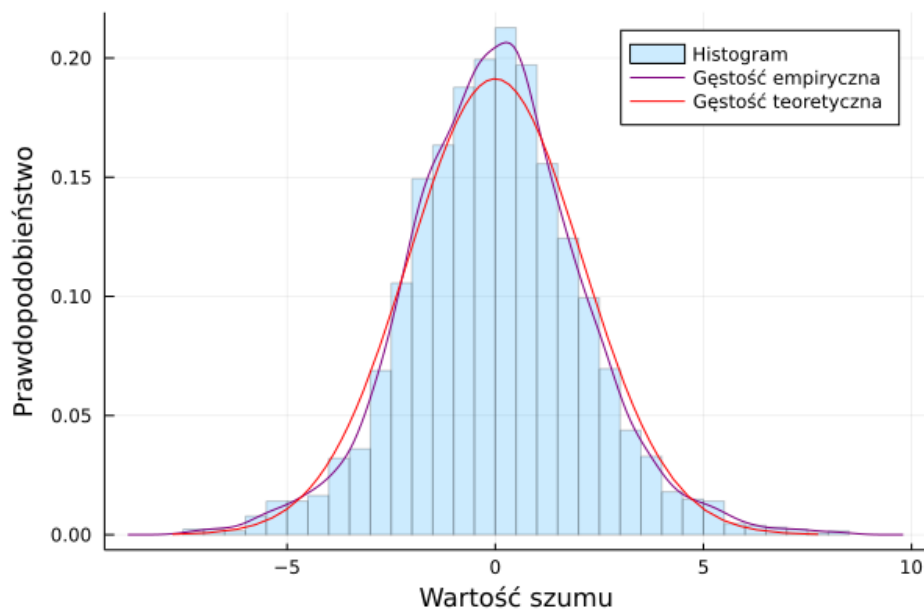
Rysunek 9: Test Ljunga Boxa dla naszych wartości reszkowych

Test nam potwierdza, że nie mamy podstaw do odrzucenia hipotezy mówiącej, że szum jest niezależny od siebie. Otrzymaliśmy p -value równe 0.9847 co jest bardzo dużą wartością.

Łącząc wykres 4 oraz test Ljunga Boxa stwierdzamy, że szum jest niezależny.

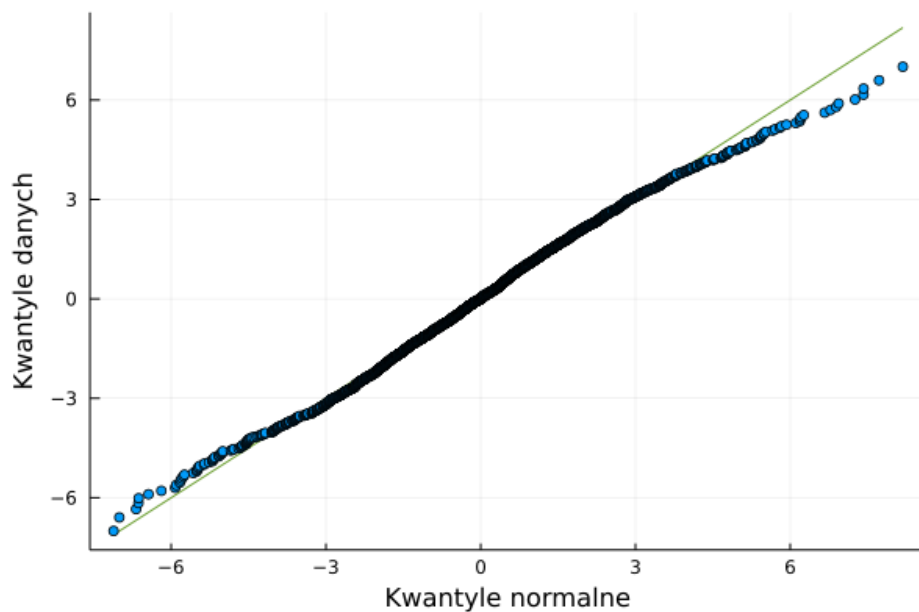
6.4 Założenie o normalności rozkładu

Sprawdźmy teraz założenie mówiące o tym, że szum ma rozkład normalny. Na początku sprawdzimy histogram wraz z gęstością empiryczną i teoretyczną rozkładu normalnego.



Rysunek 10: Histogram, gęstość rozkładu szumu wraz z rozkładem normalnym

Patrząc na histogram wraz z gęstościami 10 możemy załuważyć, że rozkład wartości resztowych może być normalny ale nie musi. W celu dalszego sprawdzenia normalności zobaczmy jak będzie wyglądał wykres kwantylowy.



Rysunek 11: Qqplot dla wartości resztowych

Analizując wykres kwantylowy 11 widzimy, że potencjalnie ogony rozkładu nie zgadzają się z rozkładem normalnym. W celu ostatecznej weryfikacji normalności wykonamy test Andersona Darlinga.

```

One sample Anderson-Darling test
-----
Population details:
  parameter of interest:  not implemented yet
  value under h_0:       NaN
  point estimate:         NaN

Test summary:
  outcome with 95% confidence: reject h_0
  one-sided p-value:      0.0334

Details:
  number of observations:  2557
  sample mean:             0.0007919156167252843
  sample SD:               2.0856431385378396
  A2 statistic:          2.830020981780576

```

Rysunek 12: Test Andersona Darlinga dla naszych wartości resztowych

Test Andersona Darlinga 12 odrzuca hipotezę o normalności rozkładu. Dla naszych danych p -value wynosi 3.3%.

Podsumowując rozkład szumu nie ma rozkładu normalnego.

7 Wnioski autorów

Pogodę w Indiach po odpowiednich przekształceniach można modelować modelem ARMA. W naszym przypadku udało się za modelować modelem o wartościach $p, q = 2$ 4.2. Parametry modelu mówią nam, że jest on stacjonarny i przyczynowy co pokrywa się z dalszymi testami. Model spełnia założenia, lecz biały szum nie ma rozkładu normalnego. Dużym atutem modelu jest fakt, że wykorzystuje on tylko dwie średnie temperatury z dwóch ostatnich dni. Warto pamiętać, że nadal jest to proces losowy więc nigdy temperatura nie będzie dokładnie się zgadzać z modelem.