
Raport II
Komputerowa Analiza Szeregów Czasowych

Natalia Lach 262303, Alicja Myśliwiec 262275

Matematyka Stosowana
Wydział Matematyki Politechniki Wrocławskiej

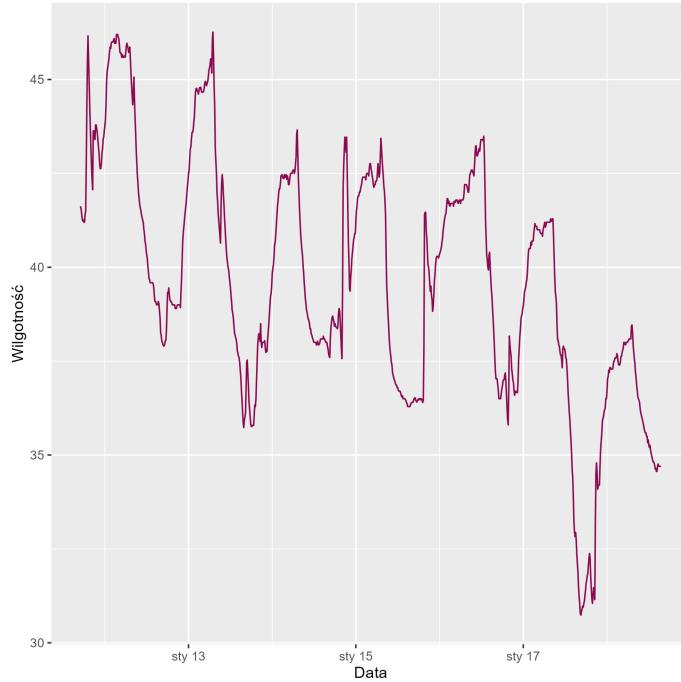
Spis treści

1. Wstęp	2
2. Pojęcia teoretyczne	2
2.1. Szereg stacjonarny w słabym sensie	2
2.2. Transformacje danych	2
2.2.1. Dekompozycja Walda	2
2.2.2. Różnicowanie danych rzędu d	3
2.3. Model ARMA(p, q)	3
2.4. Kryterium informacyjne Akaikego	3
3. Przygotowanie danych do analizy	3
3.1. Dekompozycja danych	4
3.2. Różnicowanie danych rzędu 1	5
4. Modelowanie przy pomocy ARMA(p, q)	7
4.1. Dopasowanie parametrów p i q	7
4.2. Estymacja parametrów ϕ i θ	7
5. Ocena dopasowania modelu	7
6. Analiza szumu	8
6.1. Wizualizacja	8
6.2. Nieskorelowanie	9
6.3. Analiza średniej	10
6.4. Analiza wariancji	10
6.5. Analiza normalności rozkładu	11
7. Podsumowanie	12

1. Wstęp

W niniejszej pracy zostanie przedstawiona analiza danych opisujących wilgotność w pomieszczeniu przeznaczonym do prasowania.¹ W oparciu o dane rzeczywiste zaproponowany zostanie model ARMA, następnie uzyskane wyniki zostaną sprawdzone pod względem odpowiedniego dopasowania. Wszelkie działania na zadanym zbiorze danych oraz wizualizacja wyników odbędzie się w języku R.

W wybranym zbiorze danych, pomiary były wykonywane w dniach 11.01 - 18.01 2016 roku - daje to 1000 wartości, ponieważ odczyt wilgotności aktualizowano co 10 minut.



Rys. 1: Wykres wilgotności w czasie

2. Pojęcia teoretyczne

Przyjmijmy, że $\{X_t\}_{t \in \mathbb{Z}}$ to pewien szereg czasowy.

2.1. Szereg stacjonarny w słabym sensie

Mówimy, że szereg czasowy $\{X_t\}_{t \in \mathbb{Z}}$ jest stacjonarny w słabym sensie, jeśli spełnione są następujące warunki

1. $\mu_x(t) = \mathbb{E}X_t = const.$
2. $\gamma_x(t,s) = \text{Cov}(X_t, X_s) = \mathbb{E}[X_t X_s] - \mathbb{E}X_t \mathbb{E}X_s = \gamma_x(t-s)$

2.2. Transformacje danych

2.2.1. Dekompozycja Walda

Jest to jedna z metod, mająca na celu przekształcenie procesu w proces stacjonarny. Chcemy dany proces $\{Y_t\}_{t \in \mathbb{Z}}$ przedstawić w postaci

$$Y_t = m_t + s_t + X_t, \quad (1)$$

¹ Dane pochodzą ze strony kaggle

<https://www.kaggle.com/datasets/loveall/appliances-energy-prediction!>

gdzie

- m_t - trend, funkcja deterministyczna (np. wielomian),
- s_t - sezonowość, funkcja deterministyczna (np. funkcja *sinus*).

Celem tej metody jest wyeliminowanie wymienionych deterministycznych komponentów szeregu $\{Y_t\}$, z intencją otrzymania stacjonarnego szeregu $\{X_t\}$.

2.2.2. Różnicowanie danych rzędu d

Metoda różnicowania służy głównie do usunięcia trendu liniowego. Polega ona na skonstruowaniu nowego szeregu $\{\tilde{X}_t\}_{t \in \mathbb{Z}}$ złożonego z przyrostów konkretnej wielkości d .

$$\tilde{X}_t = X_t - X_{t-d} \quad (2)$$

2.3. Model ARMA(p, q)

Model ten jest modelem autoregresyjnym średniej ruchomej. Mówimy, że szereg czasowy jest szeregiem ARMA(p, q), jeśli spełnia następujące równanie.

$$X_t - \phi_1 X_{t-1} - \dots - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \dots + \theta_q Z_{t-q}, \quad (3)$$

gdzie

- $\{Z_t\}_{t \in \mathbb{Z}} \sim \mathcal{WN}(0, \sigma^2)$ - biały szum, czyli ciąg nieskorelowanych zmiennych losowych,
- wielomiany

$$\begin{cases} \phi(z) = 1 - \phi_1 z - \dots - \phi_p z^p \\ \theta(z) = 1 + \theta_1 z + \dots + \theta_q z^q \end{cases}$$

nie mają wspólnych pierwiastków.

2.4. Kryterium informacyjne Akaikiego

Kryterium AIC służy do optymalnego dobrania parametrów p i q dla modelu ARMA. Opisane jest ono następującym wzorem

$$AIC = -2 \ln \hat{\Theta} + 2k, \quad (4)$$

gdzie

- $\hat{\Theta}$ - maksimum funkcji największej wiarogodności.
- k - liczba estymowanych parametrów modelu.

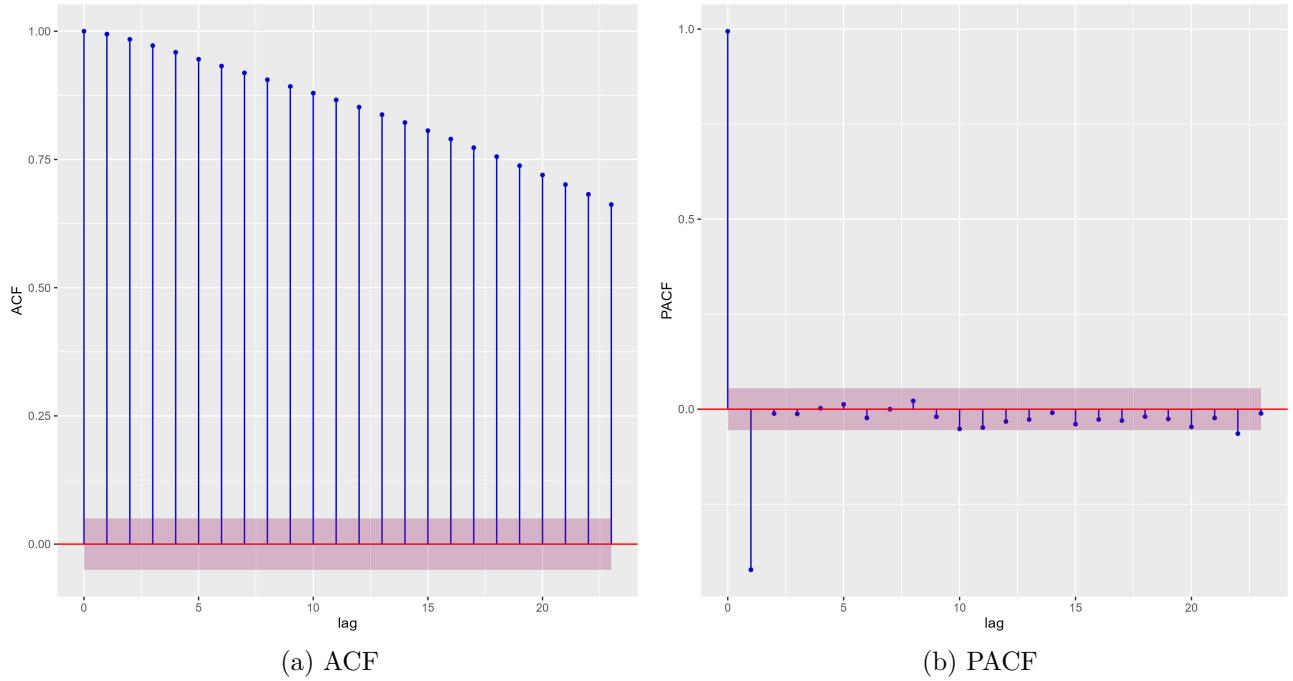
3. Przygotowanie danych do analizy

Celem raportu jest analiza danych rzeczywistych przy pomocy modelu ARMA(p, q). Aby tego dokonać, potrzebny jest jednak szereg stacjonarny. Przeprowadzony zostanie test weryfikujący hipotezę o niestacjonarności dla wybranych przez nas surowych danych, tj. test *ADF* (Augmented Dickey-Fuller Test).

```
Augmented Dickey-Fuller Test
data: data
Dickey-Fuller = -3.1215, Lag order = 9, p-value = 0.1036
alternative hypothesis: stationary
```

Rys. 2: Wynik testu *ADF* dla surowych danych

Otrzymana p -wartość jest nieco ponad dwa razy większa niż założony poziom istotności $\alpha = 0.05$. Oznacza to brak podstaw do odrzucenia hipotezy zerowej, jaką jest niestacjonarność rozważanych danych. Wyklucza to stacjonarność rozpatrywanego szeregu. W ramach dodatkowego potwierdzenia wysuniętej tezy narysowana funkcje autokorelacji.



Rys. 3: Wykresy funkcji ACF i PACF dla surowych danych

Z powyższych wykresów wynika, że badany szereg nie jest stacjonarny. Potwierdza to bardzo wolno zbiegający do zera wykres funkcji autokorelacji. Zatem koniecznym będzie przeprowadzenie transformacji danych w celu ich "ustacjonarnienia". Wykorzystane zostaną wspomniane wcześniej metody.

3.1. Dekompozycja danych

Pierwszym krokiem dekompozycji będzie usunięcie trendu m_t . Zostanie on znaleziony przy pomocy metody najmniejszych kwadratów, dzięki której można dopasować prostą regresji

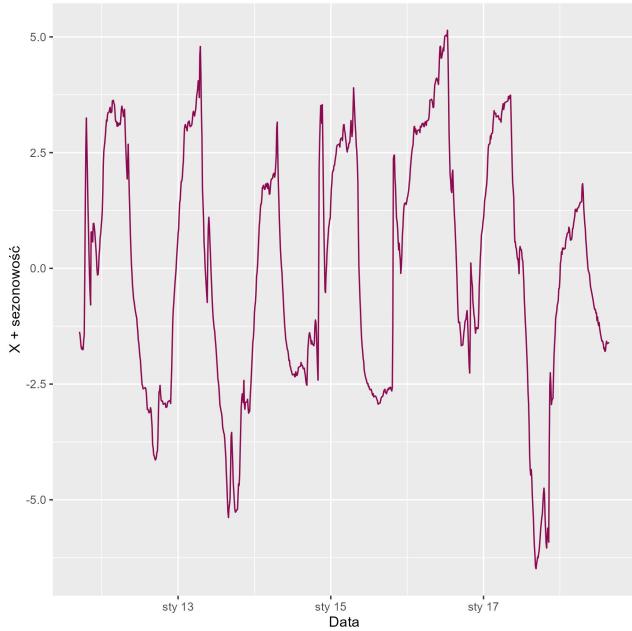
$$y = ax + b.$$

Znaleziono współczynniki o następujących wartościach: $a \approx -0.0067$, $b \approx 43.0046$. Tym samym możemy wyeliminować liniowy komponent szeregu, co zostało przedstawione na wykresie (a) rysunku (3). Na sąsiadującym wykresie (b) przedstawiono następny etap, czyli usunięcie sezonowości. Tym razem należało użyć nieliniowego odpowiednika wcześniej użytej MNK . Wykorzystana do tego została wbudowana funkcja języka R - *nls*.² Przy jej użyciu znalezione zostały najbardziej optymalne współczynniki funkcji

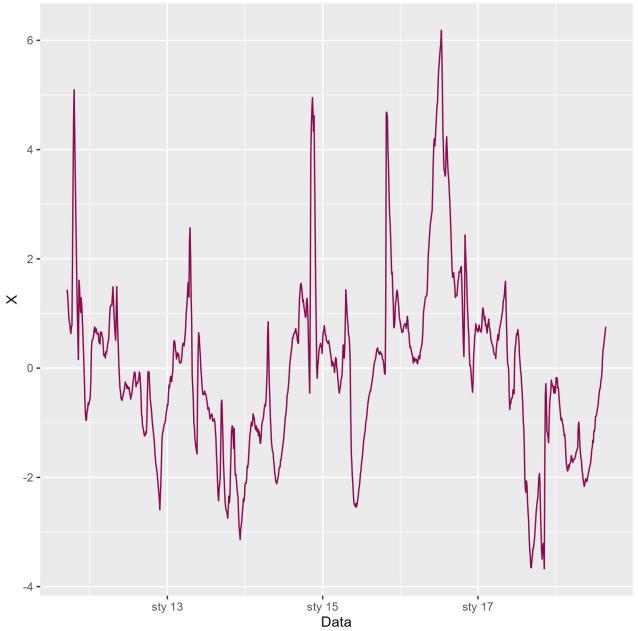
$$y = c \cdot \sin(dx + e).$$

Otrzymano kolejno: $c \approx 2.98$, $d \approx 0.04$ oraz $e \approx -20.12$.

² <https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/nls>



(a) Usunięcie trendu



(b) Usunięcie sezonowości

Rys. 4: Dekompozycja Walda

Powyższe wykresy różnią się wyglądem od wartości surowych danych. Dane z usuniętym trendem liniowym oraz sezonowością przypominają nieco bardziej szereg stacjonarny, lecz po tak przeprowadzonej transformacji, w celu upewnienia się co do stanu otrzymanego szeregu, ponownie wykonany zostanie test *ADF*.

Augmented Dickey-Fuller Test

```
data: data3
Dickey-Fuller = -3.4654, Lag order = 9, p-value = 0.04563
alternative hypothesis: stationary
```

Rys. 5: Wynik testu *ADF* po dekompozycji

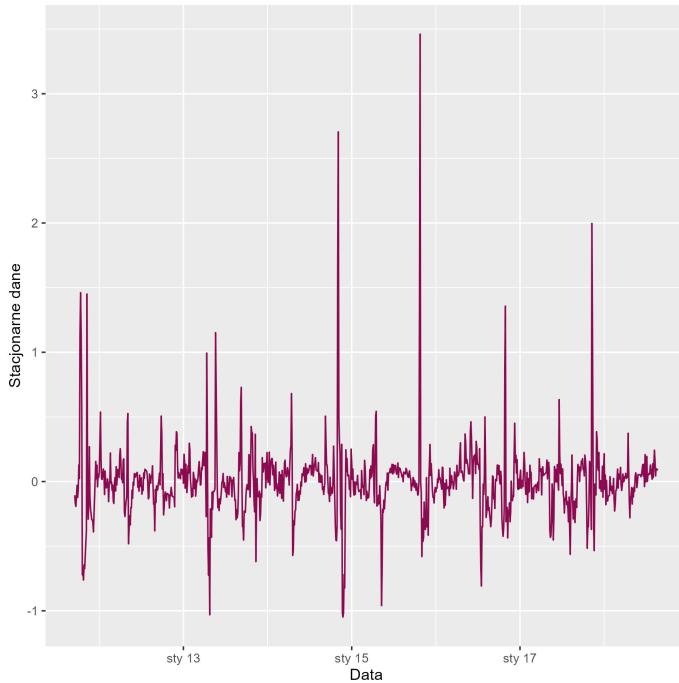
Otrzymana p -wartość wyniosła tym razem około 0.0453, co daje nam podstawy do odrzucenia hipotezy o niestacjonarności szeregu. Jest to mimo wszystko wynik bliski ustalonemu wcześniej poziomowi $\alpha = 0.05$, dlatego też zastosowana zostanie jeszcze jedna metoda transformacji danych, a mianowicie różnicowanie z krokiem $d = 1$.

3.2. Różnicowanie danych rzędu 1

Korzystając z wcześniej wspomnianej teorii, dokonane zostanie różnicowanie rzędu 1, czyli zmierzone zostaną przerosty następujących po sobie wartości według następującej formuły.

$$\tilde{X}_t = X_t - X_{t-1} \quad (5)$$

Uzyskane wyniki przedstawiono na poniższym rysunku.



Rys. 6: Różnicowanie zdekomponowanych danych

Otrzymany wykres na rysunku (6) zdecydowanie bardziej przypomina szereg stacjonarny, porównując go do wykresu surowych danych (1) czy po zastosowaniu dekompozycji (4). Ponownie można sprawdzić przy pomocy testu *ADF*, czy nałożona transformacja przybliżyła wynik do tego pożądanego.

```
Warning message in adf.test(df2$data):
"p-value smaller than printed p-value"

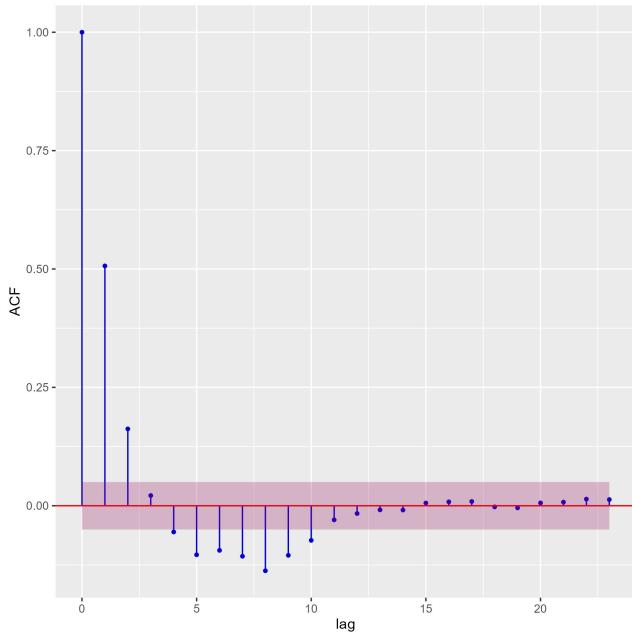
Augmented Dickey-Fuller Test

data: df2$data
Dickey-Fuller = -11.645, Lag order = 9, p-value = 0.01
alternative hypothesis: stationary
```

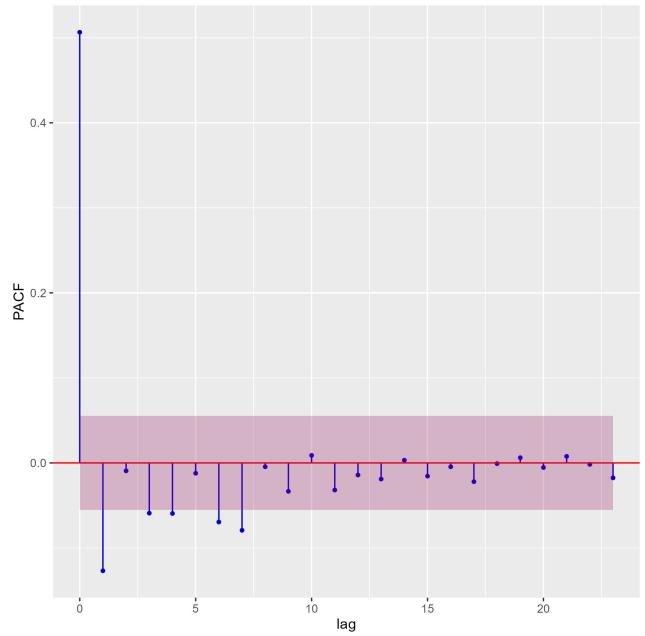
Rys. 7: Wynik testu *ADF* dla zróżnicowanych danych

Zauważalny na powyższym rysunku *Warning*, informuje o osiągnięciu bardzo niskiej wartości *p*-wartości. W takim przypadku bez zawahania odrzucamy hipotezę o niestacjonarności szeregu na rzecz hipotezy alternatywnej. Jak już wspomniano, przetransformowany szereg jest bliższy wyglądem do szeregu stacjonarnego. Potwierdza to nie tylko wynik testu *ADF*, ale również zauważalne jest to na wykresach autokorelacji przedstawionych na rysunku (8).

Tym razem wartości ACF znacznie szybciej zbiegają do zera. Już dla $lag = 10$ zatrzymują się one w przedziale $(-0.05, 0.05)$. Także PACF nieco zmienił swoją trajektorię. Ostatecznie zatem, możemy uznać nasze dane za stacjonarne i kontynuować dopasowywanie modelu ARMA(p, q).



(a) ACF



(b) PACF

Rys. 8: Wykresy funkcji ACF i PACF dla przetransformowanych danych

4. Modelowanie przy pomocy ARMA(p,q)

4.1. Dopasowanie parametrów p i q

Dopasowanie szukanych parametrów modelu zostanie przeprowadzone przy użyciu kryterium informacyjnego AIC . Spośród siatki parametrów p i q , spodziewać się można najefektywniejszej ich kombinacji. Dla rozpatrywanej siatki 5×5 , najmniejszą wartość $AIC \approx 31.81$ otrzymano dla $p = 5$ i $q = 1$.

4.2. Estymacja parametrów ϕ i θ

Korzystając z dobranych wcześniej parametrów dla modelu ARMA, można znaleźć odpowiednie parametry ϕ i θ ;

- | | |
|--------------------------|----------------------------|
| — $\phi_1 \approx 1.44$ | — $\phi_4 \approx -0.05$ |
| — $\phi_2 \approx -0.62$ | — $\phi_5 \approx -0.001$ |
| — $\phi_3 \approx 0.13$ | — $\theta_1 \approx -0.88$ |

Model ten jest zatem postaci

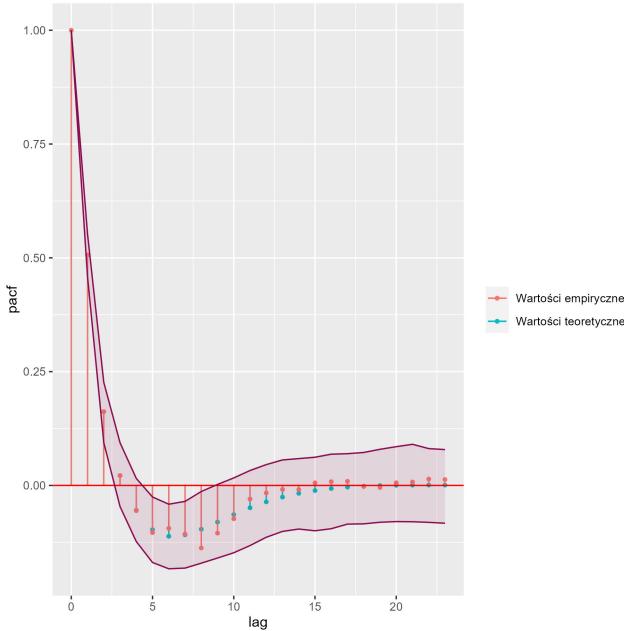
$$X_t - 1.44X_{t-1} + 0.62X_{t-2} - 0.13X_{t-3} + 0.05X_{t-4} + 0.001X_{t-5} = Z_t - 0.88Z_{t-1} \quad (6)$$

Tym samym sposobem została wyznaczona wartość estymatora parametru σ^2 , która wyniosła

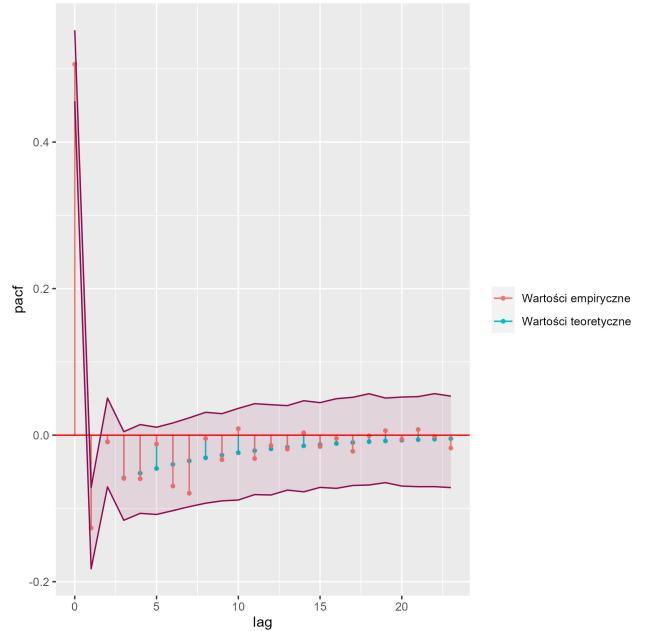
$$\sigma^2 \approx 0.05946$$

5. Ocena dopasowania modelu

Ocenę jakości dopasowania modelu można rozpocząć od porównania wykresów ACF oraz PACF.



(a) ACF



(b) PACF

Rys. 9: Porównanie wartości empirycznych z teoretycznymi

Zarówno dla klasycznej, jak i częściowej funkcji autokorelacji, wartości empiryczne są bardzo zbliżone do dopasowanych wartości teoretycznych, co można zauważyć na powyższym rysunku (9). Po dokładnym przyjrzeniu się, teoretyczne wykresy stopniowo zbiegają do 0, co jest pożądanym wynikiem. Odpowiedniki empiryczne również oscylują wokół 0, powoli zbliżając się do osi. Jednak co najważniejsze, wartości empiryczne jak i teoretyczne mieszczą się w obu przypadkach w symulacyjnie skonstruowanych przedziałach ufności. Ich wartości otrzymano, obliczając wartości kwantyli rzędu $\frac{\alpha}{2}$ oraz $1 - \frac{\alpha}{2}$ dla 1000 wygenerowanych funkcji ACF oraz PACF ze znalezionej modelu ARMA. Na podstawie wszystkich wspomnianych obserwacji można zatem stwierdzić, iż wybrany przez nas model ARMA(5,1) został dość dobrze dopasowany.

6. Analiza szumu

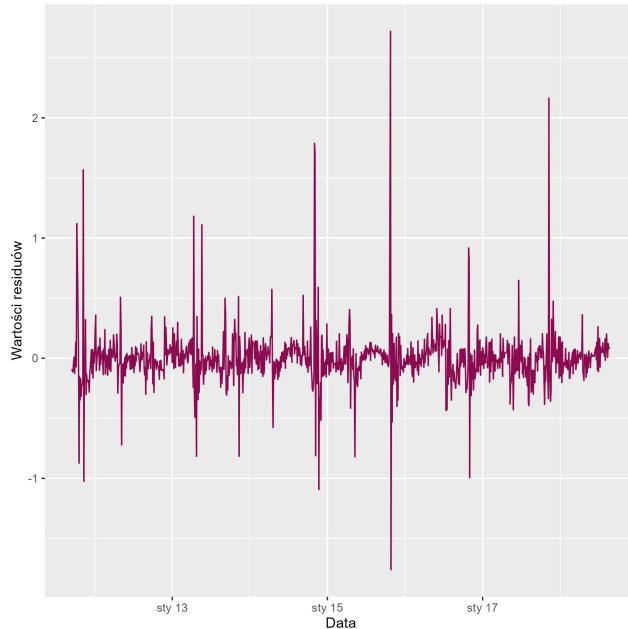
W niniejszej sekcji przedstawiona zostanie weryfikacja założeń dotyczących szumu znalezioneego modelu ARMA. Rozważane residua powinny pochodzić z rozkładu $\mathcal{WN}(0, \sigma^2)$, to jest powinny spełniać poniższe warunki.

1. Średnia jest stała i wynosi 0.
2. Wariancja jest stała, skończona i wynosi σ^2 .
3. Residua są nieskorelowane.

Ponadto można dodać założenie o normalności rozkładu, to jest niech residua pochodzą z rozkładu $\mathcal{N}(0, \sigma^2)$. W poniższych sekcjach sprawdzone zostaną wszystkie wymienione warunki, wraz z tym dodatkowym, definiującym rozkład residuów.

6.1. Wizualizacja

Wpierw narysujmy wartości residuów w czasie utworzonego modelu. Przedstawione one zostały na rysunku (10).

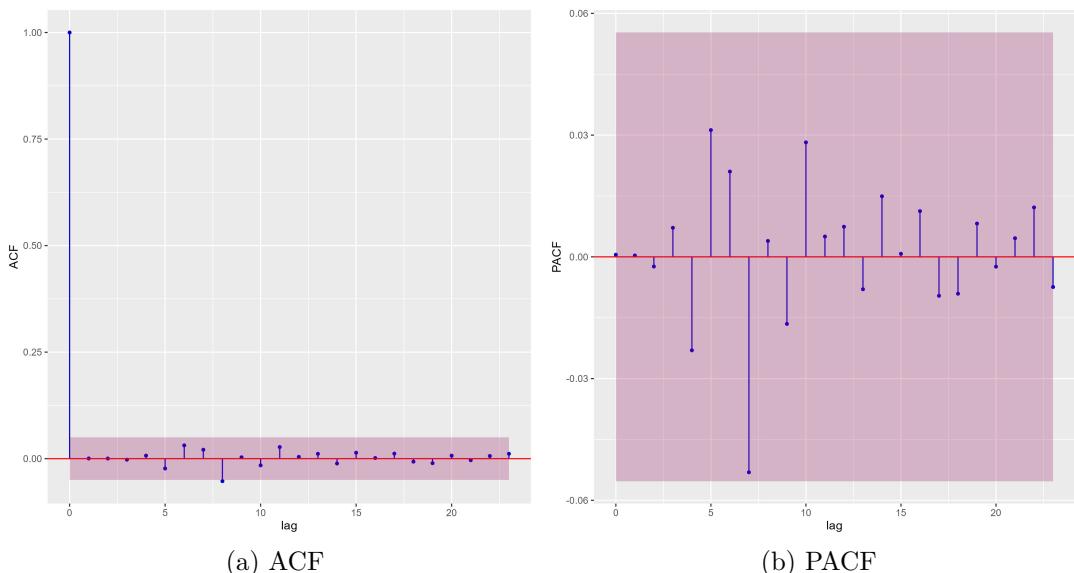


Rys. 10: Residua

Już na podstawie wykresu można wyciągnąć parę wniosków. Przykładowo można przypuszczać spełnienie założenia dotyczącego średniej oraz zmienność wariancji residuów. Obserwacje układają się wokół zera, lecz okazjonalnie pojawiają się dość wysokie skoki wartości. Są to wartości odstające, które wskazują właśnie na niespełnienie założenia o stałej i skończonej wariancji.

6.2. Nieskorelowanie

Wpierw sprawdzone zostanie założenie dotyczące nieskorelowania wartości resztowych. W tym celu ponownie posłużymy się funkcją autokorelacji oraz częściowej autokorelacji.

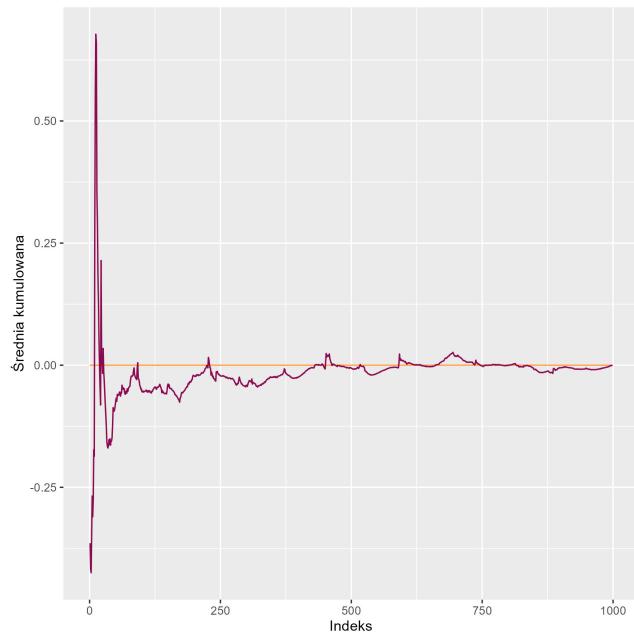


Rys. 11: Funkcje autokorelacji dla residuów

Z rysunku (11) wynika, że zarówno wartości ACF jak i PACF są bliskie zeru dla większości wartości lag . Oba wykresy mieszczą się w przedziale $(-0.05, 0.05)$ poza jednym punktem początkowym. Jest to dobry znak, świadczący o nieskorelowaniu rozważanych danych. Omawiane założenie jest zatem spełnione.

6.3. Analiza średniej

Teraz sprawdzone zostanie założenie o średniej residuów. Spodziewamy się, że będzie ona stała i równa zeru. Przypuszczenie zostanie potwierdzone dzięki wykorzystaniu akumulowanej funkcji średniej, czyli średniej tworzonej na podstawie kolejnych $1, 2, 3, \dots, 999, 1000$ danych. Wykres przedstawiono poniżej.



Rys. 12: Średnia kumulowana residuów

Wartości, tak jak przypuszczano, oscylują wokół zera. Im większy indeks, czyli większa ilość wziętych pod uwagę residuów, tym bliżej osi znajduje się otrzymana średnia. Świadczy to zatem o spełnieniu omawianego założenia o stałej i zerowej średniej.

6.4. Analiza wariancji

Aby zbadać założenie dotyczące wariancji, posłużymy się testem Boxa-Ljunga o nazwie arch test³. Jest to test, którego hipotezą zerową jest homoskedastyczność rozkładu badanych danych. Oznacza to stałą i skończoną wariancję wszystkich zmiennych losowych składających się na badany szereg. Zatem jeśli w ramach testu otrzymamy p -wartość mniejszą od ustalonego $\alpha = 0.05$, da nam to podstawy do odrzucenia hipotezy zerowej i tym samym stwierdzenia, że omawiane założenie nie jest spełnione.

```
Box-Ljung test  
data: y^2  
X-squared = 173.68, df = 2, p-value < 2.2e-16  
alternative hypothesis: y is heteroscedastic
```

Rys. 13: Test Boxa-Ljunga dla residuów

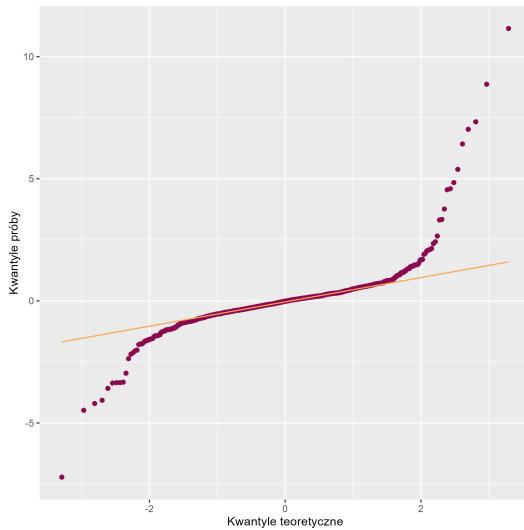
Na podstawie p -wartości widocznej w otrzymanym podsumowaniu przeprowadzonego testu, odrzucamy hipotezę zerową i stwierdzamy heteroskedastyczność rozkładu residuów. Oznacza to, że

³ Test Boxa-Ljunga
<https://search.r-project.org/CRAN/refmans/nortsTest/html/arch.test.html>!

zmienne nie pochodzą z rozkładu o tej samej skończonej wariancji. Potwierdziło się tym samym wcześniejsze przypuszczenie wysunięte na podstawie wykresu (10).

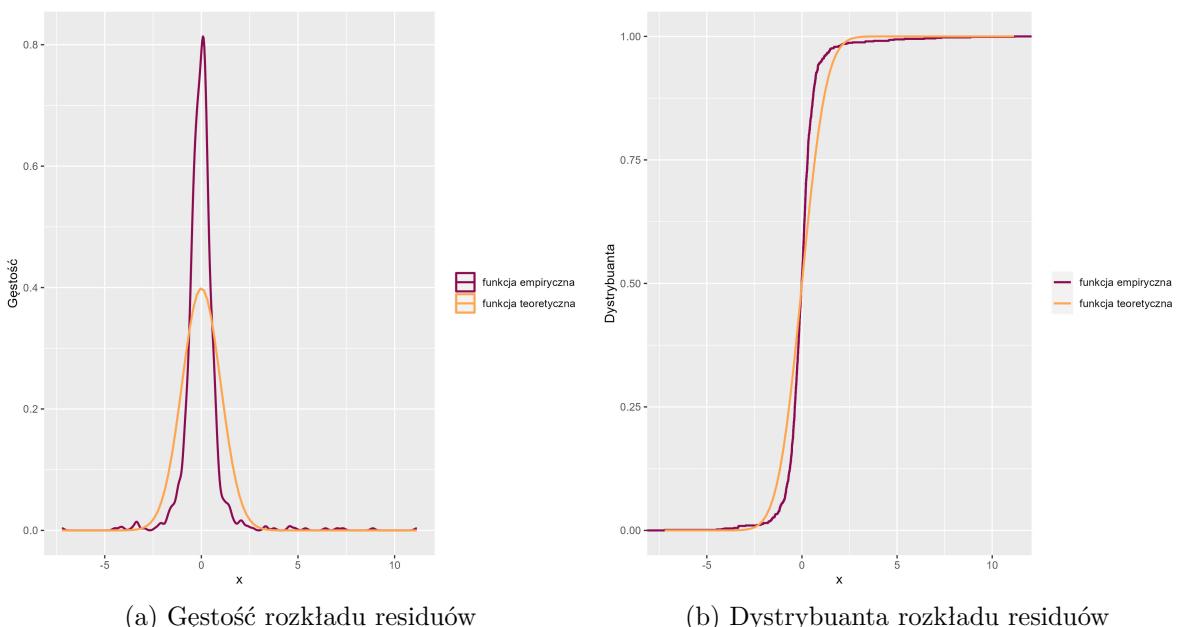
6.5. Analiza normalności rozkładu

Ostatnim założeniem jakie zweryfikujemy, jest to dotyczące normalności rozkładu residiów. Analizę przeprowadzimy na ustandaryzowanych wartościach resztowych. Wpierw zbadany zostanie wykres kwantylowy.



Rys. 14: Wykres kwantylowy

Jak wynika z powyższego rysunku (14), kwantyle empiryczne mocno odstają od teoretycznych. Punkty nie przypominają prostej linii, co świadczy o złym dopasowaniu rozkładów. Wnioskować zatem można, że dane nie pochodzą z rozkładu normalnego.



Rys. 15: Gęstość oraz dystrybuanta rozkładu residiów

Także gęstości oraz dystrybuanty empiryczne i teoretyczne znacznie od siebie odstają, co tylko potwierdza wcześniejsze przypuszczenia. Ustandaryzowane wartości resztowe nie pokrywają

się z rozkładem normalnym $\mathcal{N}(0,1)$. Ostatecznie, potwierdzić nasze rozważania możemy wykonując testy sprawdzające normalność rozkładu. Wykonane testy i uzyskane p -wartości przedstawiono w poniższej tabeli.

Test	p -wartość
Shapiro-Wilk	$2.29 \cdot 10^{-39}$
Kołmogorow-Smirnow	0
Lilliefors	$3.56 \cdot 10^{-70}$
Jarque-Bera	0

Tab. 1: Tabela przeprowadzonych testów na normalność rozkładu

W każdym przypadku z tabeli 1 możemy odczytać, że otrzymane p -wartości są bliskie zeru. Powoduje to odrzucenie hipotez zerowych dla każdego z przeprowadzonych testów, tym samym ostatecznie potwierdzenie postawionej wcześniej tezy. Residua z całą pewnością nie pochodzą z rozkładu normalnego $\mathcal{N}(0, \sigma^2)$.

7. Podsumowanie

Celem tego raportu było dopasowanie modelu ARMA do przetransformowanych surowych danych, oraz późniejsza analiza tego dopasowania. Ponieważ wybrane dane nie posiadały braków w zadanym okresie, można było przejść do dekompozycji (usunięcia trendu i sezonowości) oraz różnicowania powstałego szeregu. Jego stacjonarność została potwierdzona przy pomocy testu *ADF*, dlatego też można było przejść do głównej części sprawozdania.

Korzystając z kryterium informacyjnego Akaike (*AIC*), wyestymowano parametry dla względnie najlepszego modelu ARMA, tj. ARMA(5,1). Dopasowanie to musiało zostać zweryfikowane, co rozpoczęto od porównania wykresów teoretycznych i empirycznych funkcji klasycznej i częściowej autokorelacji. Stwierdzono zadowalające dobranie modelu na podstawie zgodności teoretycznych i empirycznych wartości ACF i PACF. Wszelkie odstające od teorii obserwacje wciąż znajdowały się w wyznaczonych przedziałach ufności.

Ponadto zweryfikowano wszelkie założenia dotyczące wartości resztowych. Niestety nie wszystkie z nich okazały się spełnione. Pomimo stałej zerowej średniej oraz nieskorelowania residuów, ich wariancja okazała się nie być stała. Także dodatkowe założenie o normalności rozkładu nie zostało spełnione. Residua z całą pewnością nie pochodziły z rozkładu $\mathcal{N}(0, \sigma^2)$. Zapewne miała na to wpływ wspomniana wcześniej zmienna wariancja.

Na podstawie otrzymanych wyników i bazujących na nich wnioskach, można stwierdzić, że znaleziony przez nas model ARMA(5,1) został bardzo dobrze dopasowany do omawianych danych. Przeprowadzona została poprawna transformacja surowych danych oraz znalezione zostały odpowiednie współczynniki ϕ oraz θ , przez co szereg był zbliżony do swojego teoretycznego odpowiednika. Jedynym niespełnionym warunkiem okazał się ten o stałej skończonej wariancji wartości resztowych. Tym samym też residua nie mogły pochodzić z założonego rozkładu normalnego. Podsumowując, powyższa analiza i przeprowadzone operacje wydają się być poprawne i zgodne z zakładaną teorią.