Raport 2

Komputerowa analiza szeregów czasowych

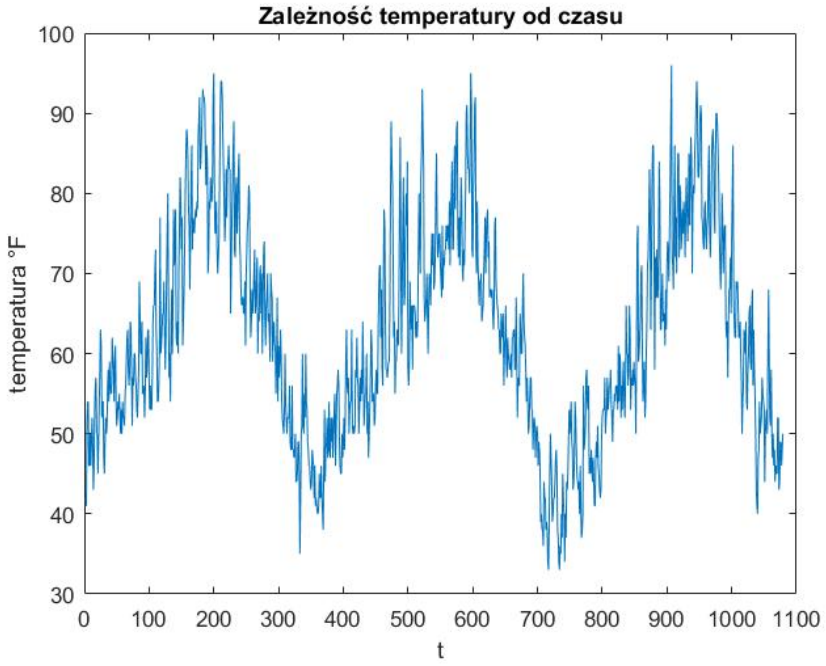
Analiza danych rzeczywistych przy pomocy modelu ARMA

Patrycja Ozgowicz 249795
Martyna Boniatowska 249763

08.02.2021r.

1 Przedstawienie i opis danych

Dane wybrane do analizy przedstawiają najwyższą dzienną temperaturę zanotowaną w mieście Seattle. Jest to największe miasto w stanie Waszyngton w Stanach Zjednoczonych. Pomiarów dokonywano na międzynarodowym porcie lotniczym Seattle-Tacoma. Dane obejmują okres czasu od 01.01.2015 r. do 14.12.2017 r. Pomiary były wykonywane codziennie, zarówno w dni robocze jak i weekendy, stąd wniosek, że można je traktować jako szereg czasowy. Zbiór danych został opracowany przez National Oceanic and Atmospheric Administration i udostępniony do domeny publicznej. Głównym celem wykonywanych pomiarów było zbadanie tego, jak często pada i jak duże są opady w Seattle. Problem badawczy powstał w związku z tym, że jest to miasto znane ze swojej deszczowej pogody. Oprócz pomiarów ilości opadów dokonywano pomiarów temperatury. Tabela, z której pochodzą analizowane dane, zawierała również najwyższą i najniższą temperaturę osiągniętą w ciągu dnia. Do wykonania raportu wykorzystane zostały najwyższe zanotowane dzienne temperatury podawane w skali Fahrenheita. Dane zostały znalezione na stronie kaggle.com [1]. Z kolei Kaggle mieni się największą na świecie społecznością grupująca specjalistów zajmujących się analizą danych i statystyką, inżynierów pracujących nad uczeniem maszynowym, stąd można znaleźć tam dużą ilość zbiorów danych.



Rysunek 1: Wykres dziennych maksymalnych temperatur w Seattle od czasu

[1] Link: https://www.kaggle.com/rtatman/did-it-rain-in-seattle-19482017

Na wykresie 1 przestawiono najwyższe dzienne temperatury odnotowane w ciągu dnia w mieście Seattle od numeru obserwacji. Na osiach dni, wartości 1 odpowiada data 01.01.2015 r., natomiast wartości ostatniej - dzień 14.12.2017 r. Można zauważyć, że trajektoria zachowuje się w sposób okresowy. Regularnie następują wzrosty i spadki temperatury, co w naturalny sposób wynika ze zmiany pór roku. Rozpatrywanym okresem czasu są 3 lata, zatem mamy trzy wzrosty w miesiącach wakacyjnych oraz spadki w miesiącach zimowych. Dane zostaną odpowiednio przygotowane i zostanie przeprowadzona ich analiza pod względem dopasowania modelu szeregu czasowego ARMA.

2 Przygotowanie danych do analizy

2.1 Wstep teoretyczny

Szereg czasowy - jest to ciąg obserwacji pokazujący kształtowanie się badanego zjawiska w kolejnych okresach czasu. Oznaczany jest jako X_t, a kolejne obserwacje indeksowane są momentami w czasie równooddalonymi od siebie ($t \in \mathcal{Z}$).

Średnia szeregu czasowego - jeśli $\{X_t\}$ jest szeregiem czasowym, to jego średnia jest zdefiniowana następująco:

$$\mu_x(t) = EX_t < \infty$$

gdzie $t \in \mathcal{Z}$

Funkcja autokowariancji - jest to wielkość równa kowariancji pomiędzy procesem stochastycznym a tym samym procesem przesuniętym o pewien odcinek czasu. Jeśli $\{X_t\}$ jest szeregiem czasowym (procesem stochastycznym) oraz EX_t^2 jest skończona, funkcję autokowariancji definiowana jest następująco:

$$\gamma_x(t,h) = cov(X_t, X_{t+h}) = E[(X_t - \mu_x(t))((X_{t+h} - \mu_x(t+h))]$$

gdzie h - lag (przesunięcie)

Funkcja autokorelacji (ACF) - jest to statystyka opisująca, w jakim stopniu dany wyraz szeregu zależy od wyrazów poprzednich w szeregu czasowym. Jeśli $\{X_t\}$ jest szeregiem czasowym oraz EX_t^2 jest skończona, funkcję autokorelacji definiowana jest następująco:

$$\rho_x(t,t+h) = \frac{cov(X_t, X_(t+h))}{\sqrt{Var(X_t)Var(X_{t+h})}}$$

Funkcja autokorelacji cząstkowej (PACF) - jest to korelacja między wartościami szeregu oddalonymi o h przedziałów od siebie, z jednoczesną rejestracją wartości z przedziałów znajdujących się pomiędzy.

Szereg czasowy stacjonarny w słabym sensie - szereg czasowy X_t ($t \in \mathcal{Z}$) jest stacjonarny w słabym sensie gdy spełnia następujące warunki:

- $\mu_x(t) = const, \forall t$
- $\gamma_x(t,h) = cov(X_t, X_{t+h}) = \gamma_x(h)$
- $\rho_x(t, t+h) = \frac{\gamma_x(h)}{\gamma_x(0)}$

Dekompozycja Wolda - jest to proces mający na celu usunięcie trendu deterministycznego nieokresowego i sezonowości z szeregu czasowego. Innymi słowy wyznaczenie X_t^{**}, zdefiniownego ponizej.

$$X_t = m(t) + s(t) + X_t^{**}$$

gdzie:

X_t - dane rzeczywiste wybrane do analizy
$m(t)$ - trend deterministyczny nieokresowy
$s(t)$ - składowa deterministyczna okresowa
X_t^{**} - szereg czasowy stacjonarny w słabym sensie

2.2 Przeprowadzenie dekompozycji

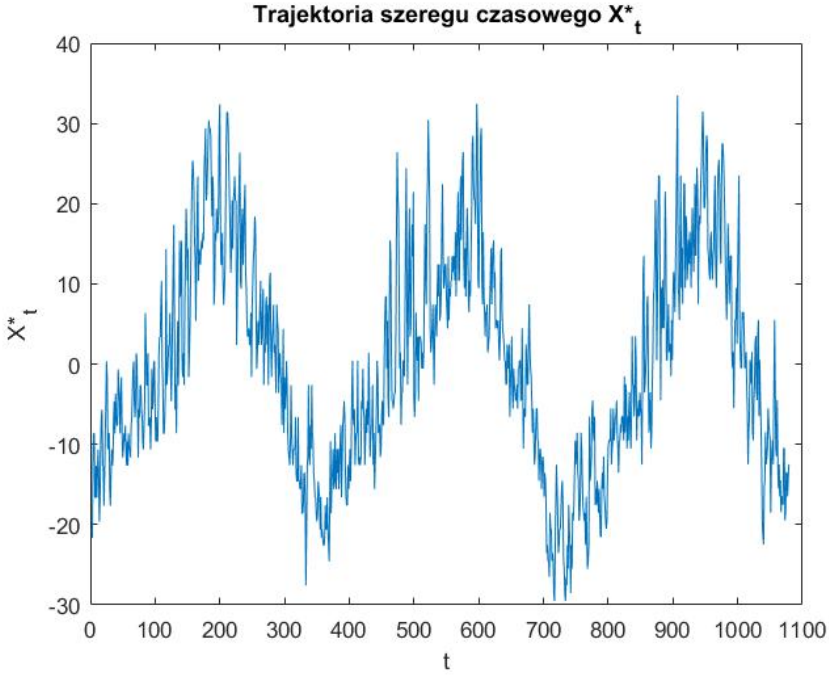
W celu usunięcia trendu liniowego z danych (oznaczonych jako X_t) dopasowano do nich prostą regresji przy użyciu metody najmniejszych kwadratów. Następnie usunięto liniowość, odejmując prostą i tworząc szereg czasowy X_t^*:

$$X_t^* = X_t - (b_1 \cdot t + b_0),$$

gdzie b_0 i b_1 są współczynnikami prostej regresji wyznaczonej z metody najmniejszych kwadratów. Mamy:

$$X_t^* = X_t - (-0.00015706t + 62.6455),$$

Poniżej przedstawiono wykres nowego szerego czasowego X_t^*



Rysunek 2: Wykres przedstawiający trajektorię szeregu czasowego X_t^*, który powstał po usunięciu trendu liniowości z danych

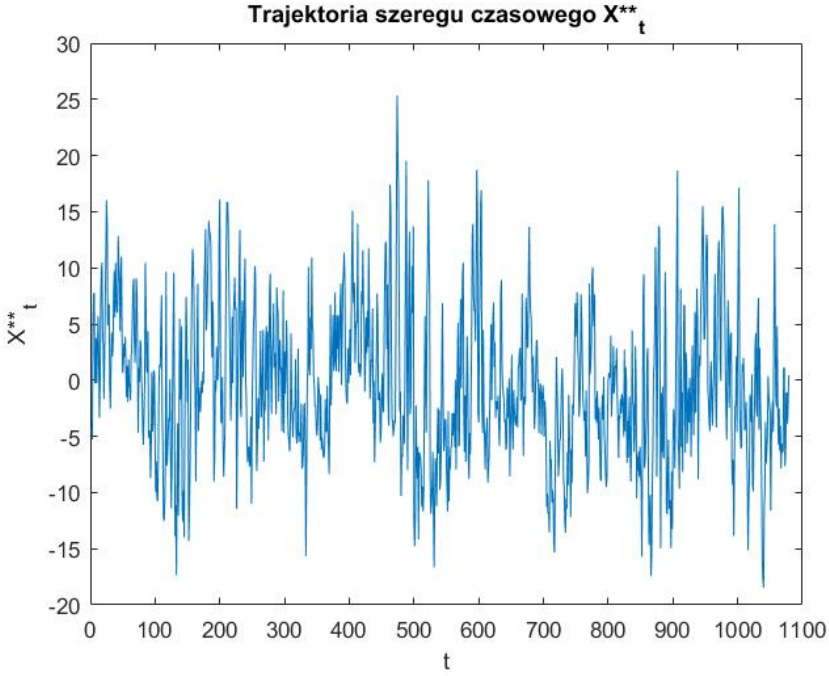
Wykres 2 przedstawia trajektorię szeregu powstałego po usunięciu liniowości z danych. Warto zauważyć, że kształt wykresu nie uległ zmianie, jedynie nastąpiło jego przeskalowanie. Nowe dane przyjmują wartości dodatnie i ujemne, natomiast dane oryginalne przyjmowały tylko wartości dodanie. Dlatego teraz możliwe jest usunięcie sezonowości. W następnym kroku usunięta została okresowość z szeregu bez trendu liniowego, tzn. powstał nowy szereg X_t^{**}:

$$X_t^{**} = X_t^* - \hat{s}(t),$$

gdzie $\hat{s}(t)$ jest funkcją postaci $a \cdot sin(b \cdot t + c)$ dopasowaną do wygererowanych danych. Mamy

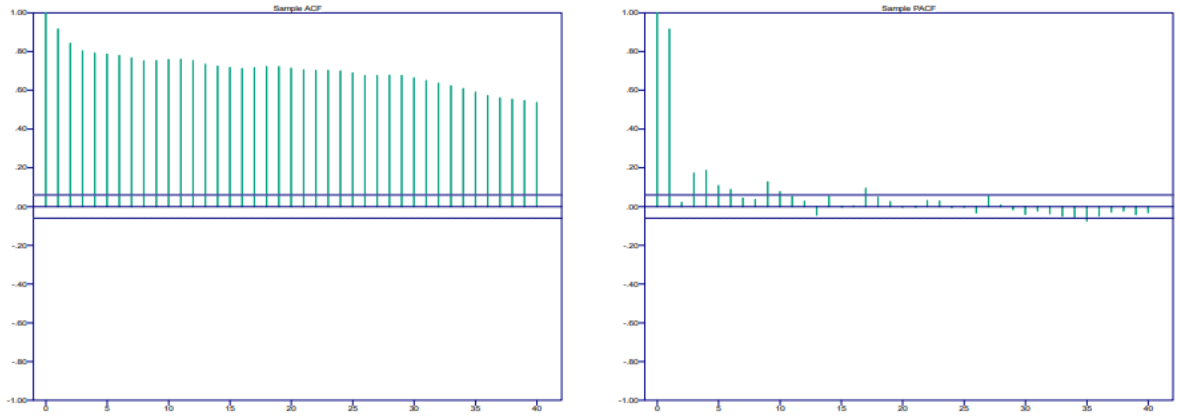
$$X_t^{**} = X_t^* - (16.4 \cdot sin(0.01697 \cdot t - 1.694))$$

Wykres przedstawiający szereg po usunięiu okresowości i liniowości:

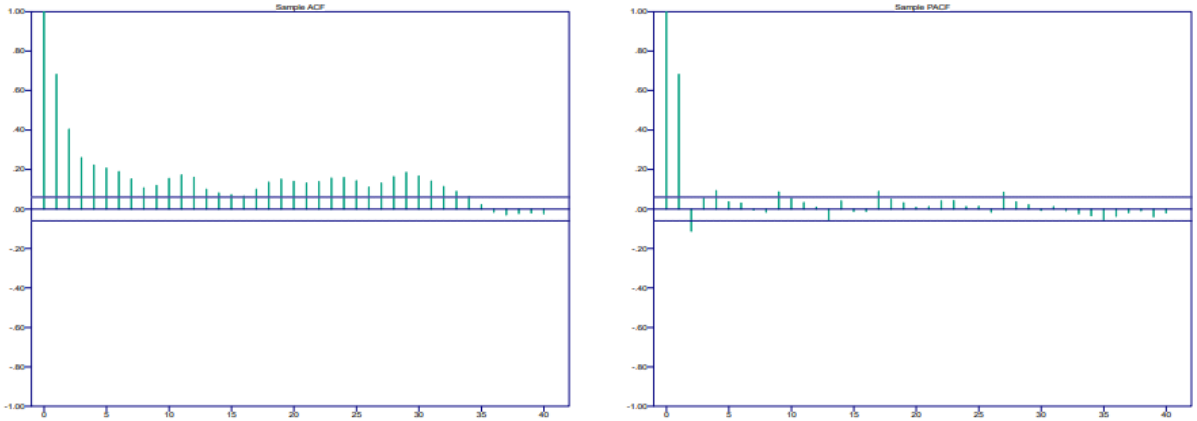


Rysunek 3: Wykres przedstawiający trajektorię szeregu czasowego X_t^{**}, który powstał po usunięciu trendu liniowości i okresowości z szeregi X_t

Powyższy wykres 3 przedstawia szereg czasowy gotowy do dalszej analizy. Nie występują już w nim czynniki deterministyczne, które zostały usunięte po przeprowadzeniu dekompozycji Wolda. Aby sprawdzić, czy została ona wykonana poprawnie, poniżej przedstawiono wykresy funkcji autokorelacji i częściowej autokorelacji. Porównano funkcję ACF i PACF dla szeregu czasowego będącego oryginalnymi danymi, a także nowego szeregu, powstałego po dekompozycji.



Rysunek 4: Wykres funkcji ACF i PACF dla oryginalnych danych



Rysunek 5: Wykres funkcji ACF i PACF dla szeregu czasowego po dekopozycji Wolda

Patrząc na wykresy 4 i 5 można zauważyć, że po przeprowadzeniu dekompozycji Wolda funkcja ACF mocno uległa zmianie. Dla oryginalnych danych malała bardzo powoli i nie było wartości, które mieściłyby się w zaznaczonych na niebiesko przedziałach ufności. Po przeprowadzeniu dekompozycji dane maleją znacznie szybciej i pojawiły się wartości mieszczące się w przedziale ufności. Mimo wszystko widać, że większa część warotści nie wpada w przedział ufności, a także zauważalna jest pewna powtarzalność. Może to oznaczać, że dekopozycja nie została wykonana w pełni poprawnie i w danych mógł pozostać pewnien deterministyczny trend okresowy. Jednak poznane metody nie umożliwiły lepszego

oczyszczenia danych. Stąd dalsza analiza zostanie przeprowadzona dla otrzymanych danych. Dla funkcji PACF róznica na wykresach nie jest aż tak znacząca, ale można zaobserwować, że po dekompozycji dane przyjmują mniejsze wartości i oscylują wokół zera w przedziałach ufności.

3 Modelowanie przy pomocy ARMA

3.1 Wstęp teoretyczny

Szereg czasowy $\{X_t\}$ dla $t \in \mathcal{Z}$ jest **szeregiem ARMA(p,q)** jeśli jest stacjonarny w słabym sensie oraz dla każdego t spełnione jest równanie:

$$X_t - \phi_1 X_{t-1} - \phi_2 X_{t-2} - ... - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \theta_2 Z_{t-2} + ... + \theta_q Z_{t-q},$$

gdzie $\{Z_t\} \sim$WN(0,σ^2) oraz

$$\phi(z) = 1 - \phi_1 z - \phi_2 z^2 - ... - \phi_p z^p$$

$$\theta(z) = 1 + \theta_1 z + \theta_2 z^2 + ... + \theta_q z^q$$

nie mają wspólnych pierwiastków.

Kryteria informacyjne - opierają się na entropii informacji niesionej przez model (niepewności modelu) tzn. szacują utraconą informację, gdy dany model jest używany do opisu badanego zjawiska. Powinno się zatem wybierać model o minimalnej wartości danego kryterium informacyjnego.

- **AIC** - Kryterium informacyjne Akaikego

$$ACC = -2ln(L_{FM}) + 2(k)$$

- **AICC** - Poprawione kryterium informacyjne Akaikego

$$AICC = AIC + \frac{2k(k+1)}{n-k-1}$$

- **BIC** - Bayesowskie kryterium informacyjne Schwartza

$$BIC = -2\ln L_{FM} + k\ln(n),$$

gdzie:

L_{FM} - maksimum funkcji wiarygodności pełnego modelu
$k = p + q$ - liczba zmiennych w modelu
n - liczność próby

Metoda największej wiarygodności - metoda wykorzystywana do estymacji współczynników modelu. Polega na wyznaczeniu funkcji wiarygodności, zależnej od estymowanego parametru. Następnie szuka się jej maksimum poprzez przyrównanie pochodnej do zera. Wyznaczona w ten sposób wartość to szukany współczynnik.

3.2 Dobór modelu do danych

Aby dopasować odpowiedni model do danych, posłużono się programem ITSM 2000. Użyto funkcji autofit model, która dopasowuje model ARMA do danych. Funkcja przeanalizowała wszystkie możliwe modele dla p z zakresu 0-6 oraz q z zakresu 0-6. Oceniła, który model jest najlepszy na podstawie dwóch kryteriów informacyjnych AICC oraz BIC. Najlepszym modelem jest ten, dla którego wartości obu kryteriów są jak najmniejsze. Parametry estymowała metodą największej wiarygodności. I tak najlepiej dobranym modelem dla analizowanych danych okazał się być model ARMA(6,6). Wartości kryteriów informacyjnych wynosiły odpowiednio:

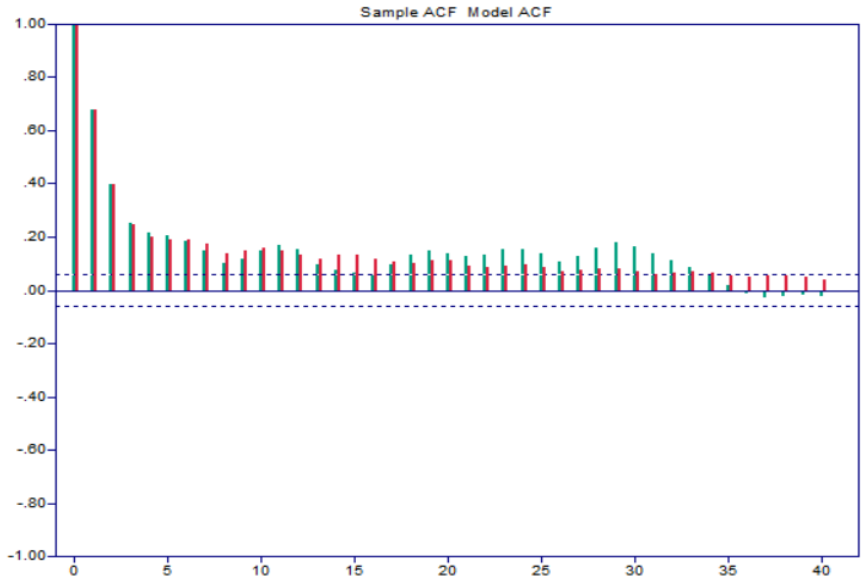
$$AICC = 6493.89$$

$$BIC = 6530.58$$

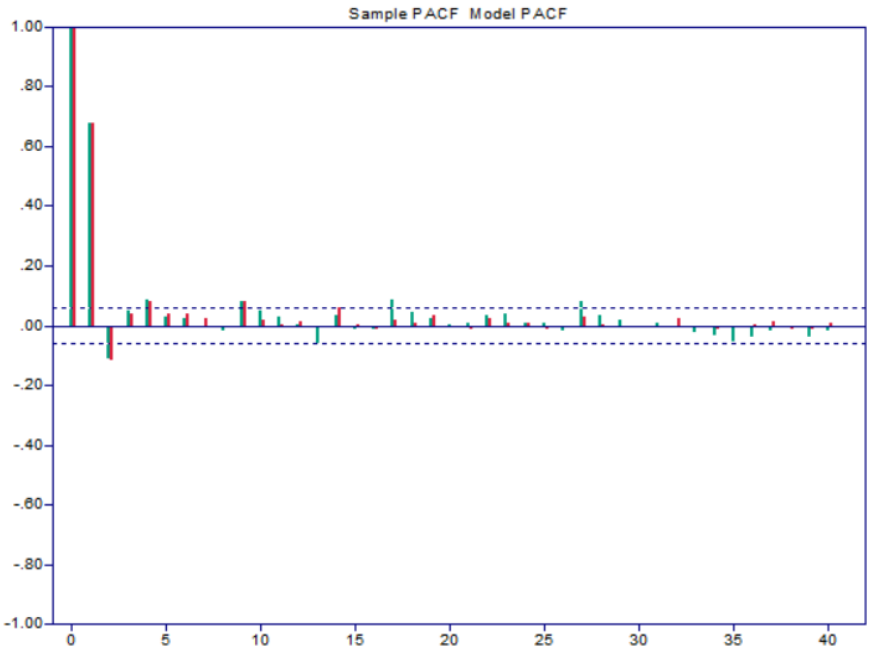
Przyglądając się innym wartościom kryteriów informacyjnych, można wyciągnąć wniosek, że dla modelu ARMA(6,6) rzeczywiście osiągnęły one najmniejszą wartość. Oznacza to, że ten model najlepiej oddaje charakter analizowanych danych dla wybranego zakresu. Otrzymany model wygląda następująco:

$$X_t - 0.2316X_{t-1} + 0.1259X_{t-2} - 0.2086X_{t-3} - 0.1742X_{t-4} - 0.7104X_{t-5} + 0.3129X_{t-6} = Z_t + 0.5195Z_{t-1} + 0.3838Z_{t-2} - 0.01620Z_{t-3} - 0.1994Z_{t-4} - 0.8428Z_{t-5} - 0.2372Z_{t-6} \quad (1)$$

Poniżej przedstawiono wykres funkcji ACF i PACF dla danych po wykonaniu dekompozycji i danych z modelu teoretycznego.



Rysunek 6: Porównanie funkcji ACF dla szeregu czasowego po wykonaniu dekopozycji i dobranego modelu teoretycznego ARMA(6,6)

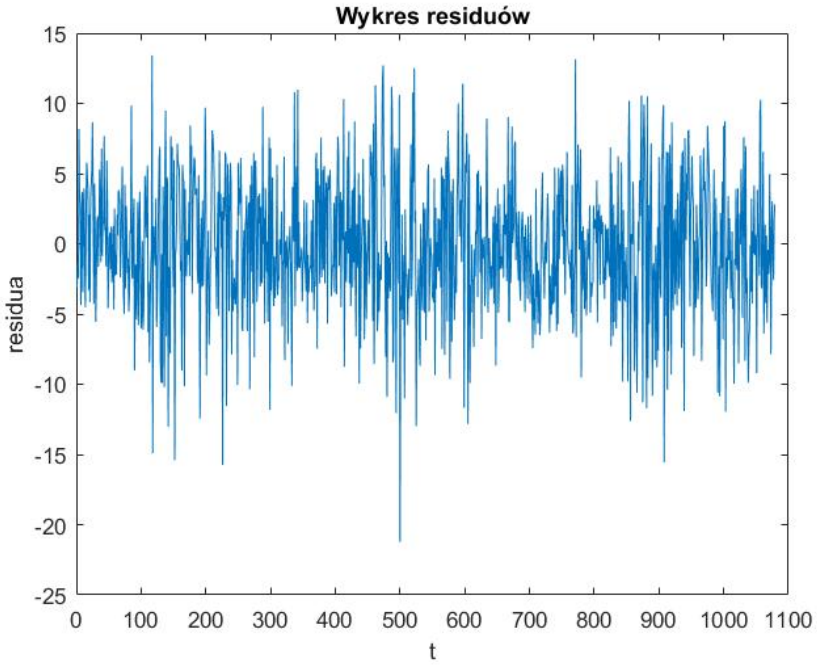


Rysunek 7: Porównanie funkcji PACF dla szeregu czasowego po wykonaniu dekopozycji i dobranego modelu teoretycznego ARMA(6,6)

Powyższe wykresy 6 i 7 porównują funkcję ACF i PACF dla szeregu czasowego po wykonaniu dekompozycji i dobranego modelu teoretycznego ARMA(6,6). Można zauważyć, że wartości empiryczne pokrywają się dość dobrze z wartościami teoretycznymi. Widoczne są drobne odstępstwa dla większych lagów w funkcji autokorelacji, gdzie wartości empiryczne są większe od tych teoretycznych. Ze względu na te odstępstwa podjęto próbę zwiększenia zakresu współczynników p i q w celu znalezienia jeszcze lepszego dopasowania modelu teoretycznego. Okazało się, że następnym wytypowanym modelem jest ARMA (6,15). Dla tak dobranych parametrów można zobaczyć lepsze dopasowanie funkcji autokorelacji i częściowej autokorelacji, jednak taki dobór parametrów powoduje znaczne wydłużenie czasu obliczeń, w związku z czym do dalszej analizy wykorzystano pierwszy dopasowany model, tj. ARMA(6,6).

4 Weryfikacja poprawności modelu

Model jest poprawnie dobrany do danych jeśli residua są białym szumem, czyli ciągiem nieskorelowanych zmiennych losowych $\{Z\}_{t\in\mathcal{Z}} \sim WN(0,\sigma)$. Wykres przedstawiający wartości resztowe zamieszczono poniżej:



Rysunek 8: Wykres przedstawiający residua

Wykres 8 wizualizuje wartości residuów od numeru obserwacji. Przyjmują one na zmianę wartości dodatnie i ujemne. Ich własności zbadano poniżej.

4.1 Analiza średniej

Zgodnie z teorią wartość średnia residuów, powinna wynosić 0. W rozważanym przypadku jest to

$$E[Z_t] = -0.0881$$

Wartość średniej jest mała i bliska 0. Dodatkowo, przyglądając się wykresowi 7, przedstawiającemu wartości residuów można dostrzec, że znajdują się one powyżej i poniżej zera. Można więc przypuszczać, że wartości te się znoszą, a średnia wynosi 0. By to potwierdzić, przeprowadzono t test. Jest to test, który zwraca decyzję testową dla hipotezy zerowej, że dane pochodzą z rozkładu normalnego o średniej równej zero i nieznanej wariancji. Alternatywna hipoteza głosi, że rozkład populacji nie ma średniej równej zero. Wynik h wynosi 1, jeśli test odrzuca hipotezę zerową na poziomie istotności 5%, a 0 w innym przypadku. Wynik:

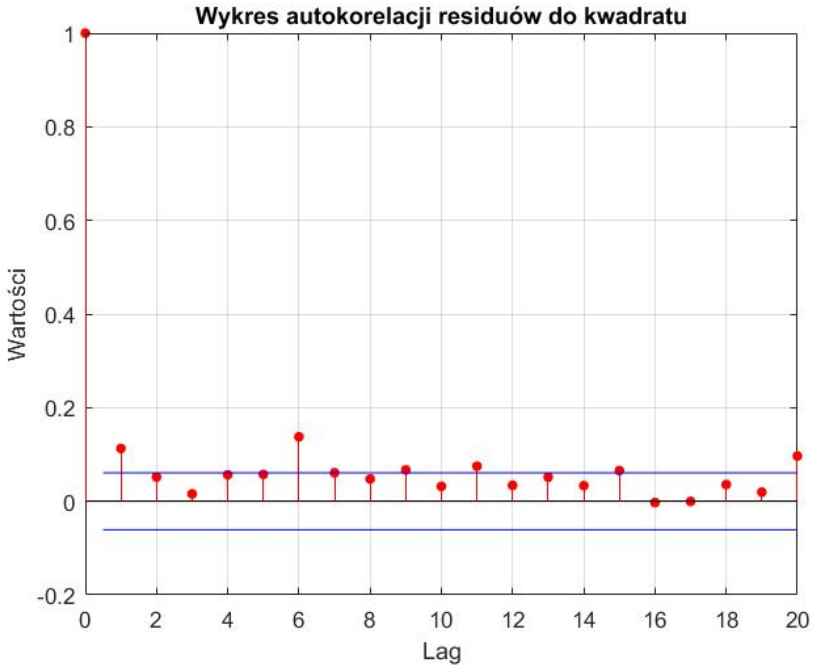
$$h = 0, p = 0.5503$$

Oznacza to, że hipoteza zerowa została przyjęta na poziomie istotności 0.05,a p-wartość jest wyższa od tego poziomu, czyli nie ma podstaw do odrzucenia

hipotezy zerowej. Stąd wniosek, że średnia wartości resztowych wynosi 0 i jest zgodna z założeniami.

4.2 Analiza wariancji

Według teoretycznych założeń wartość wariancji residuów powinna być stała dla wszystkich wartości. W celu sprawdzenia tego założenia narysowano wykres autokorelacji residuów podniesionych do kwadratu. Jeśli na wykresie residua zachowują się tak, jak zachowują się dane niezależne można stwierdzić, że wariancja jest stała, ponieważ nie waha się znacząco.

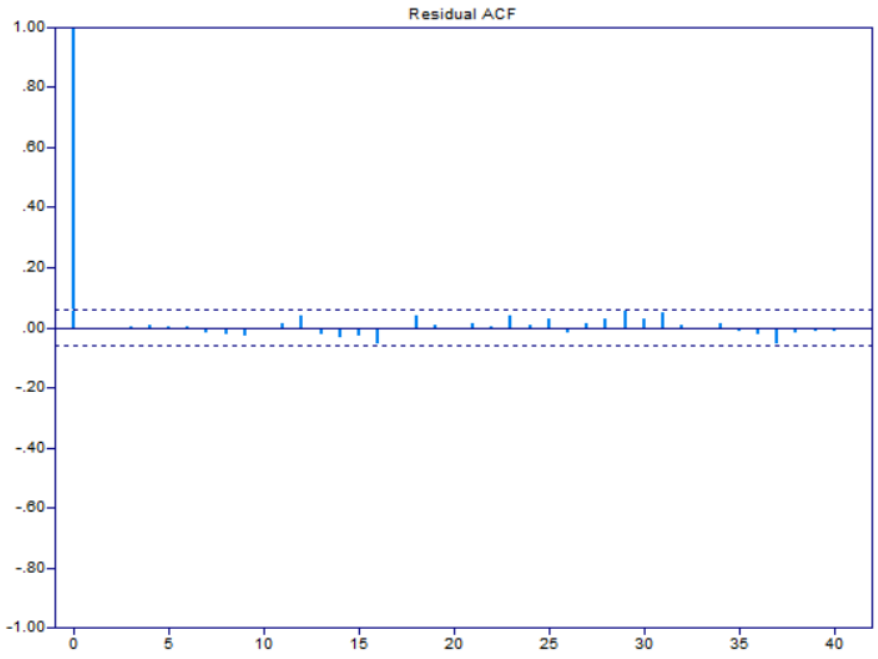


Rysunek 9: Wykres autokorelacji residuów do kwadratu

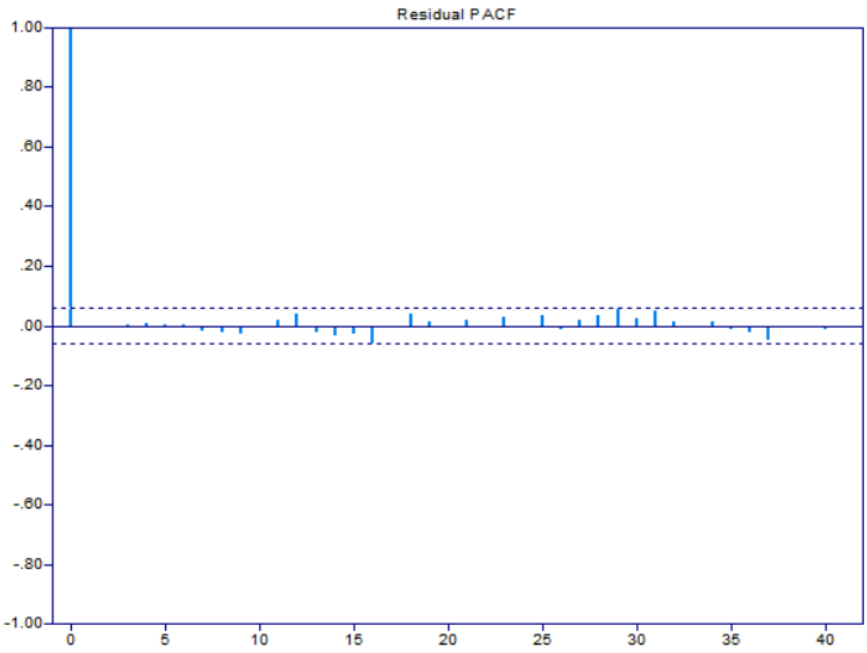
Przyglądając się wykresowi 9, można zauważyć, że nie różni się on znacząco od wyników, które byłyby otrzymane dla wartości niezależnych. W związku, z czym można przypuszczać, że wariancja jest stała, ponieważ widoczne na wykresie 9 wartości autokorelacji w znacznej większości mieszczą się w zaznaczonych na niebiesko przedziałach ufności. Jest kilka wartości, nieznacznie wystających poza przedział. Zatem wykres ten pozwala sądzić, że wariancja jest stała. Wyliczono, że wynosi ona 23.4387.

4.3 Analiza niezależności

Otrzymane residua zgodnie z teorią powinny być nieskorelowane. Warto dodać, że jeśli residua są niezależne, są także nieskorelowane. Dlatego najpierw sprawdzona zostanie ich niezależność. Pierwszą próbą sprawdzenia tego założenia może być spojrzenie na wykres 8 przedstawiający residua. Dla wartości niezależnych chmura powinna być równomiernie rozłożona. Tak też wyglądają otrzymane residua, w których nie są zauważalne żadne tendencje. Przeprowadzając dokładniejszą analizę w sposób nieformalny, można narysować wykres funkcji autokorelacji dla residuów. Jeśli wartości będą bliskie 0 dla całej próby, i tylko w zerze otrzymana zostanie wartość 1 można mówić o niezależności. Bardziej formalnym sposobem sprawdzenia jest wykonanie testu Ljung'a-Box'a, który używany jest do oceny zależności między danymi. Za hipotezę zerową test ten przyjmuje, że korelacja między obserwacjami równa jest 0. Hipoteza alternatywna natomiast mówi, że są one zależne.



Rysunek 10: Wykres autokorelacji residuów



Rysunek 11: Wykres częściowej autokorelacji residuów

Aby dane rzeczywiste były zgodne z teorią, wartości residuów dla konkretnych lagów powinny oscylować w okolicach zera, pomiędzy niebieskimi liniami oznaczającymi przedziały ufności. Na wykresie 10 i 11 widać, że wartości właśnie tak się zachowują, zarówno w przypadku autokorelacji jak i częściowej autokorelacji. Wartości nie przewyższają zaznaczonego poziomu ufności dla obydwu funkcji. Stąd też można wnioskować, że wartości resztowe są niezależne. W celu formalnego sprawdzenia należy zobaczyć jaki wynik zwraca lbqtest. Wyniki testu:

$$h = 0, p = 0.8393$$

Otrzymane $h = 0$, oznacza, że nie ma podstaw do odrzucenia hipotezy zerowej na rzecz hipotezy alternatywnej. Wartość $p = 0.8393$ większa od poziomu ufności 0.05 nie świadczy za hipotezą alternatywną. Stąd wniosek, że wartości resztowe z modelu są niezależne.

4.4 Analiza rozkładu

Zgodnie z założeniami, residua powinny być nieskorelowanymi zmiennymi losowymi. Dodatkowo można sprawdzić, czy pochodzą one z rozkładu normalnego, ponieważ takie właśnie założenie zostało poczynione przy wyznaczaniu przedziałów ufności w następnym podrozdziale. Aby to zbadać, przeprowadzono

dwa rodzaje testów. Sprawdzenie mniej formalne, na które składało się wyliczenie charakterystyk, gęstości, dystrybuanty i kwantyli oraz sprawdzenie formalne, które zawiera testy statystyczne. Poniżej zaprezentowano wyniki przeprowadzonej analizy

SPRAWDZENIE NIEFORMALNE

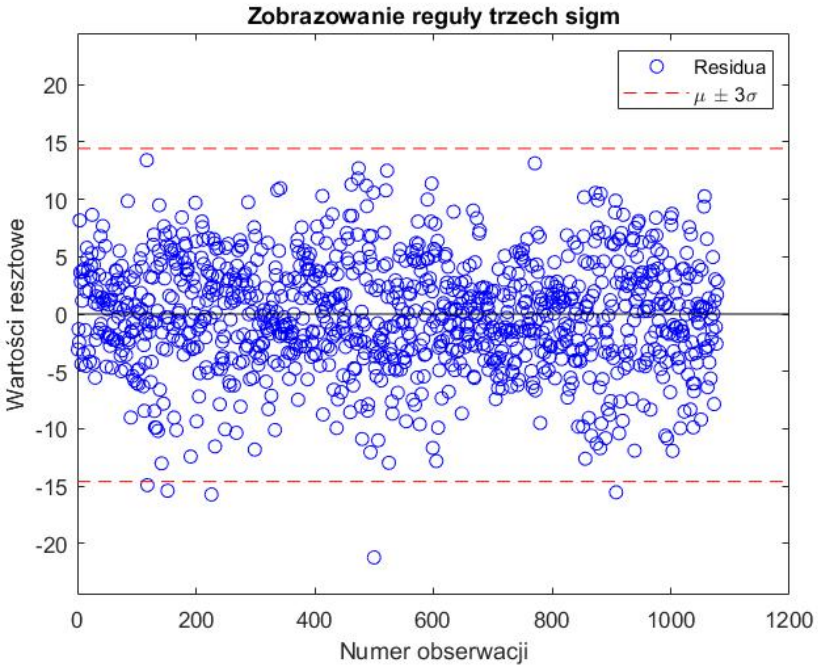
- Statystyki residuów:

Statystyka	Wartość dla residuów	Wartość dla N(0,σ)
Kutrtoza	3.3177	3
Współczynnik skośności	0.0802	0

Tabela 1: Tabela z porównaniem kurtozy i współczynnika skośności dla residuów i teoretycznego rozkładu normalnego N(0,σ)

Powyższa tabela 4 porównuje skośność i spłaszczenie rozkładu residuów z rozkładem normalnym. Można z niej odczytać, że wartość kurtozy 3.3177 dla wartości resztowych jest bliska wartości 3 dla rozkładu normalnego. Podobnie ze współczynnikiem skośności, którego wartość 0.0802 jest bliska 0.

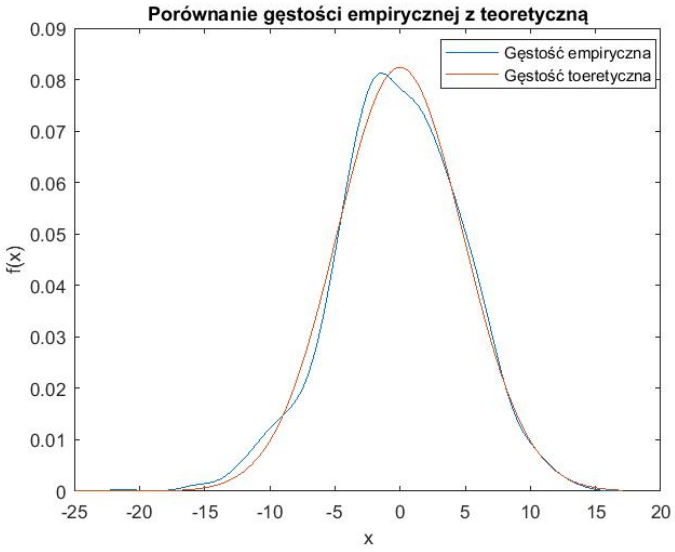
- Reguła trzech sigm mówi, że jeżeli zmienna ma rozkład normalny bądź zbliżony do rozkładu normalnego to $99,7\%$ obserwacji znajduje się w zakresie pomiędzy ± 3 odchylenia standardowe od średniej. Dla residuów wykres obrazujący tę regułę wygląda następująco:



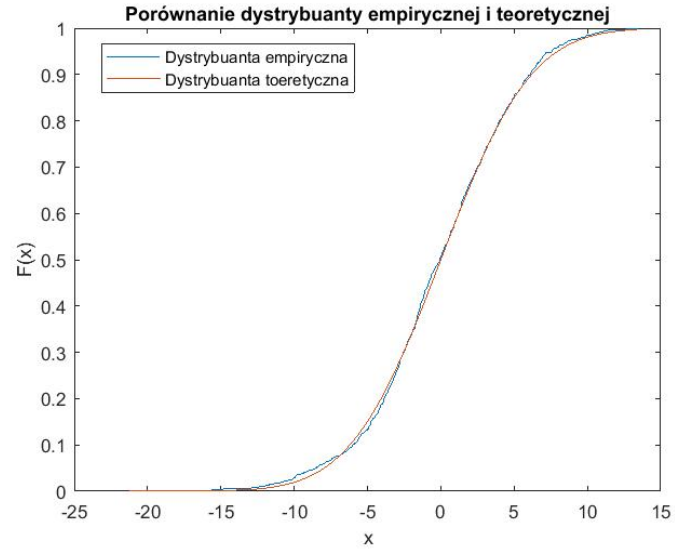
Rysunek 12: Wykres obrazujący regułę 3 sigm dla rozkładu residuów

Na wykresie 12 przedstawiono wartości resztowe od numeru obserwacji. Dodatkowo czerwoną przerywaną linią zaznaczono wartości $\mu \pm 3\sigma$, gdzie μ jest średnią z próby, a σ odchyleniem standardowym z próby. Na wykresie możemy zobaczyć, że 5 wartości resztowych znalazło się poza przedziałem. Oznacza to, że 99.5366% danych znalazło się w skonstruowanym zakresie. Jest to wartość bliska $99,7\%$, tak jak jest w teorii, stąd przypuszczenie, że dane moga pochodzić z rozkładu normalnego.

- Porównanie gęstości i dystrybuanty empirycznej z teoretyczną



Rysunek 13: Wykres porównujący gęstość empiryczną residuów z dystrybuantą rozkładu normalnego $\mathcal{N}(0, 4.8414)$

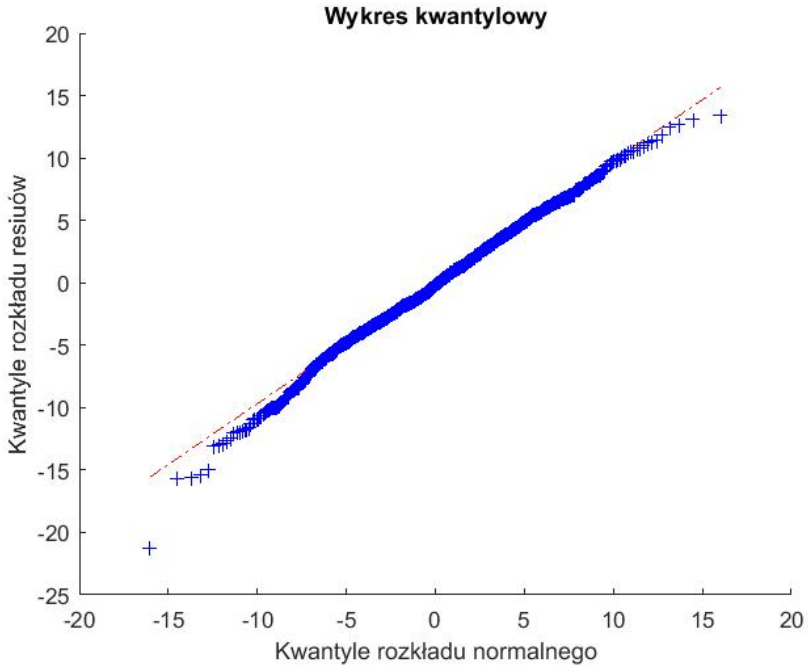


Rysunek 14: Wykres porównujący dystrybuantę empiryczną residuów z dystrybuantą rozkładu normalnego $\mathcal{N}(0, 4.8414)$

Na powyższych wykresach 13, 14 porównano gęstości i dystrybuanty roz-

kładu wartości resztowych z rozkładem normalnym z średnią 0 i wariancją wyliczoną z próby $\mathcal{N}(0, 4.8414)$. Na pierwszym z nich widać, że gęstości mają podobny przebieg. Ich kształt jest zbliżony, choć gęstość teoretyczna jest bardziej symetryczna u góry. Funkcje z wykresu 12 mają podobny przebieg, dystrybuanta empiryczna przyjmuje wartości wokół tej teoretycznej.

- Wykres kwantylowy



Rysunek 15: Wykres porównujący kwantyle residuów z kwantylami rozkładu normalnego $\mathcal{N}(0, 4.8414)$

Wykres 15 jest wykresem kwantylowym, który porównuje kwantyle rozkładu residuów z kwantylami rozkładu normalnego $\mathcal{N}(0, 4.8414)$. Z wykresu można odczytać, że kwantyle się pokrywają. Niebieskie plusy oznaczające kwantyle wartości resztowych nieco odbiegają od prostej jedynie na końcach przedziału. Jest to zauważalne, jednak nie odrzuca kategorycznie możliwości, że dane pochodzą z rozkładu normalnego. Statystyki rozkładu empirycznego i teoretycznego z tabeli 4 przyjmowały podobne wartości, w przybliżeniu można powiedzieć, że zachowana została zasada trzech sigm, dystrybuanty oraz gęstości miały podobny kształt. Także wykres kwantylowy przybierał wartości podobne. Nie ma więc podstaw, by sądzić, że dane nie pochodzą z rozkładu normalnego. Poniższe formalne sprawdze-

nie zależności między tymi rozkładami pozwoli na podjęcie ostatecznej decyzji, czy rozkład residuów jest rozkładem normalnym.

SPRAWDZENIE FORMALNE

- Test Kołmogorowa - Smirnowa to najczęściej stosowany test statystyczny sprawdzający, czy próba pochodzi z rozkładu normalnego. Bazuje on na różnicy pomiędzy dystrybuantą teoretyczną a empiryczną. Hipoteza zerowa zakłada właśnie należenie próby do rozkładu normalnego, natomiast alternatywna odrzuca to założenie. Zastosowano modyfikację powyższego testu i analizowane dane przyrównano do rozkładu normalnego ze średnią zero i wariancją z danych.

- Test Anderson-Darling jest podobny do K-S testu, ponieważ również bazuje na różnicy pomiędzy dystrybuantą empiryczną a teoretyczną. Jest jednak bardziej wrażliwy na różnice w ogonach. Test ten również za hipotezę zerową przyjmuje, że wektor pochodzi z rozkładu normalnego, a hipoteza alternatywna odrzuca to założenie.

- Test Jarque-Bera bazuje na empirycznych parametrach rozkładu, takich jak skośność i kurtoza. Dzięki niemu można ocenić, czy próba pochodzi z rozkładu normalnego o nieznanej średniej i wariancji. Pochodzenie z takiego rozkładu zakłada hipoteza zerowa, natomiast alternatywna odrzuca to założenie.

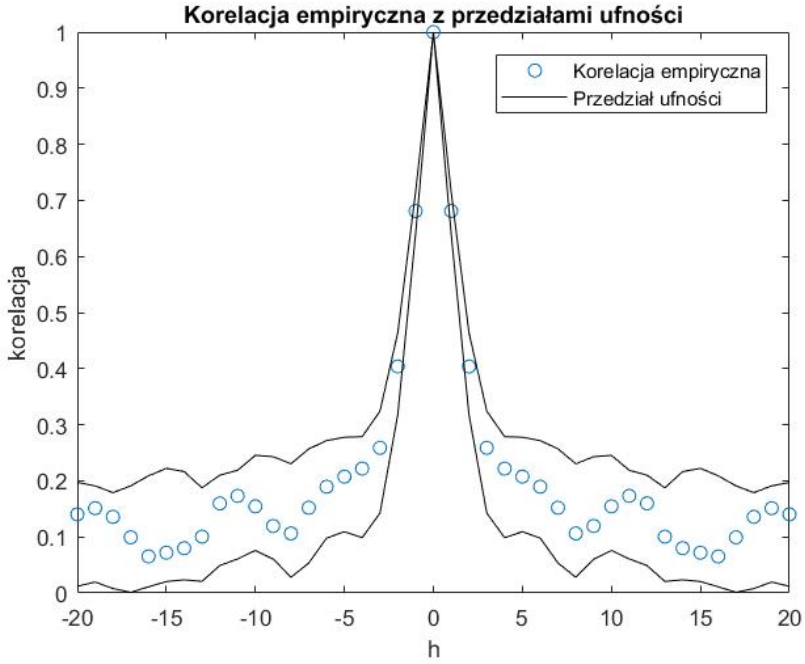
Test	h	p
Kołmogorow - Smirnow	0	0.7515
Anderson-Darling	0	0.5046
Jarque-Bera	1	0.0036

Tabela 2: Tabela z wynikami testów normalności

Jeżeli test przyjmuje wartość $h = 1$, świadczy to o odrzuceniu jego hipotezy zerowej na rzecz hipotezy alternatywnej. Wartość p mówi o tym, jak bardzo bliska odrzucenia jest hipoteza zerowa. Im mniejsze p, tym większe prawdopodobieństwo odrzucenia hipotezy zerowej. Jak widać w tabeli 2, jeden test odrzucił hipotezę zerową, natomiast dwa testy jej nie odrzuciły. Analizując wartość p dla testu Kołmogorowa - Smirnowa oraz Andersona-Darlinga widać, że jest ona większa od poziomu ufności 0.05, co sugeruje, że nie ma powodów do odrzucenia hipotezy zerowej. Mimo, że jeden z testów sugeruje odrzucenie hipotezy zerowej, to nieformalne sprawdzenie rozkładu, a także dwa pozostałe testy pozwalają wnioskować, że dane mogą pochodzą z rozkładu normalnego $\mathcal{N}(0, 4.8414)$

4.5 Porównanie empirycznej funkcji korelacji z teoretycznymi przedziałami ufności

W celu weryfikacji, czy model został poprawnie dobrany do analizowanych danych stworzono wykres empirycznej autokorelacji z zaznaczonymi teoretycznymi przedziałami ufności na poziomie ufności 0.05 i 0.95. Metodą Monte Carlo stworzono 100 modeli ARMA(6,6) ze współczynnikami ze wzoru (1), gdzie $\{Z\}_t \sim N(0, \sqrt{46.0147})$, gdzie wartość 46.0147 jest wariancją z próby. Dla każdego modelu wyliczono funkcję autokorelacji dla lagów od -20 do 20. Dla każdego lagu wyznaczono przedział ufności. Funkcję autokorelacji dla danych z zaznaczonymi przedziałami przedstawiono na wykresie 16



Rysunek 16: Wykres obrazujący funkcję korelacji danych z zaznaczonymi teoretycznymi przedziałami ufności dla dobranego modelu wyrażonego wzorem 1

Z wykresu 16 można odczytać, że wartości empiryczne w całości mieszczą się w symulacyjnie wyznaczonych przedziałach ufności. Pozwala to wnioskować, że model został poprawnie dopasowany do danych.

5 Podsumowanie i wnioski

W raporcie analizowany był zbiór danych dotyczący dziennych maksymalnych temperatur zanotowanych w porcie lotniczym w Seattle, pochodzących z okre-

su od 01.01.2015 r. do 14.12.2017 r. W pierwszej części przedstawiono krótki opis danych oraz ich wizualizację. Na podstawie otrzymanego wykresu (wykres 1) stwierdzono, że w wybranych danych istnieją pewne deterministyczne trendy. Na pierwszy rzut oka można było stwierdzić obecność trendu okresowego, ponieważ dane obrazowały dzienne temperatury obejmujące okres trzech lat, w podobnych okresach notowano wzrosty i spadki temperatur spowodowane zmianami pór roku. W kolejnej części podjęto próbę usunięcia zaobserwowanych trendów zarówno liniowego jak i okresowego. Następnie sprawdzono, czy dane, które otrzymano po oczyszczeniu, mogą być traktowane jako szereg czasowy. Sprawdzenia tego dokonano poprzez analizę funkcji ACF i PACF. Porównano wykres 4, na którym przedstawiono funkcje przed oczyszczeniem i wykres 5, na którym widać funkcje dla danych po oczyszczeniu i stwierdzono, że po dekompozycji można spróbować dobrać model ARMA do oczyszczonych danych. Kolejnym etapem było dopasowanie odpowiedniego modelu ARMA. W tym celu wykorzystano program ITSM 2000, który umożliwił automatyczny dobór parametrów modelu, na podstawie poznanych kryteriów informacyjnych. Otrzymano model ARMA(6,6), dla którego wartości kryteriów informacyjnych były najmniejsze ze wszystkich testowanych parametrów. Dodatkowo przy użyciu tego samego programu wyznaczono współczynniki modelu, wykorzystując metodę największej wiarygodności. W celu sprawdzenia, czy model został dopasowany poprawnie i dobrze oddaje charakter analizowanych danych na wykresach 6 i 7 porównano ACF i PACF zarówno dla próby, jak i dla dopasowanego modelu. Zauważono, że wykres teoretyczny jest bardzo zbliżony kształtem do tego empirycznego. W związku z czym uznano, że wybrany model może być poprawny. Jednak w celu upewnienia się, czy na pewno tak jest, wykonano kilka kroków mających na celu sprawdzenie założeń dotyczących residuum. Zgodnie z założeniami modelu residua miały wartość średnią zbliżoną do 0 oraz na podstawie wykresu 9 stwierdzono, że wykres residuum do kwadratu zachowuje się bardzo podobnie jak dla danych niezależnych, co pozwala wnioskować stałość wariancji. Następnie przeanalizowano wykresy autokorelacji i częściowej autokorelacji, dzięki którym można potwierdzić teorię o niezależności wartości resztowych. Sprawdzono również, czy pochodzą one z rozkładu normalnego, ponieważ do wyznaczenia przedziałów ufności dla funkcji korelacji, trzeba było założyć, z jakiego rozkładu pochodzi szum. Po przeprowadzeniu nieformalnego i formalnego sprawdzenia można wnioskować, że residua mogą pochodzić z rozkładu normalnego. Przyjmując takie założenia, wyznaczono przedziały ufności dla funkcji autokorelacji i sprawdzono, czy empiryczna korelacja wybranych danych po dekompozycji mieści się w tych przedziałach. Na wykresie 16 dobrze widać, że wartości w 100% wpadają w wyznaczony przedział, a co za tym idzie - model został dopasowany poprawnie.