Sprawozdanie nr 1

Analiza danych rzeczywistych przy pomocy modelu ARMA

Dominik Hołoś, 275995
Weronika Jaszkiewicz, 275990

1. Wstęp

Cel pracy

Celem raportu jest analiza danych dotyczących liczby plam słonecznych z zamiarem sprawdzenia ich dopasowania do modelu ARMA. Przeprowadzona analiza pozwoli ocenić możliwość zastosowania tego modelu do predykcji liczby plam słonecznych w przyszłości. Szczególną uwagę poświęcono identyfikacji odpowiednich parametrów modelu oraz walidacji przyjętych założeń.

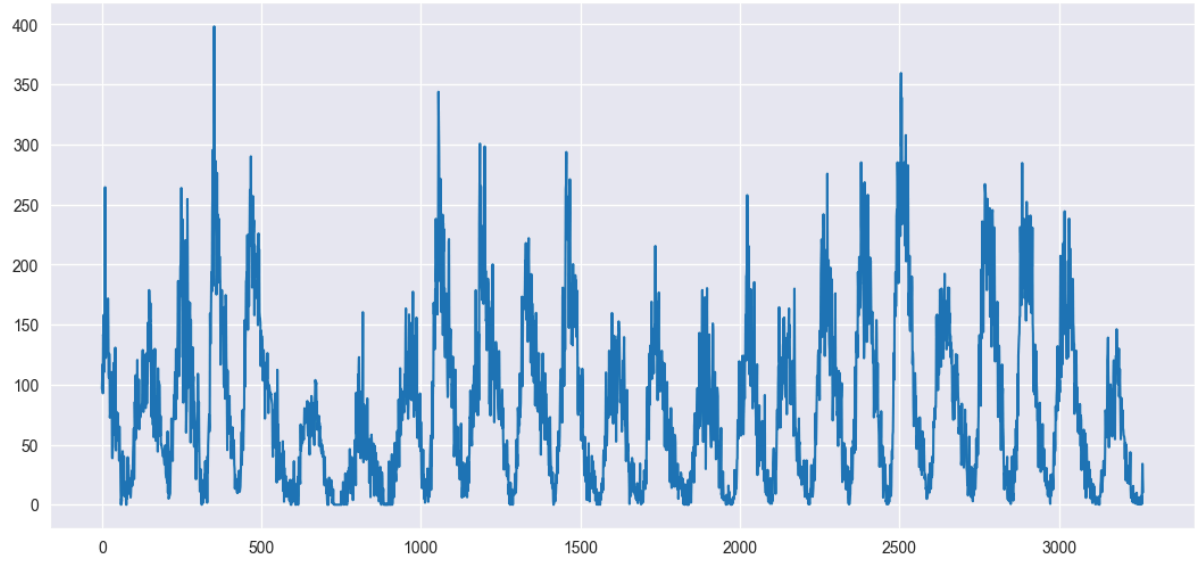
Dane

Analizowane dane dotyczą miesięcznych średnich liczb plam słonecznych, które są zjawiskami obserwowanymi na powierzchni Słońca. Plamy słoneczne charakteryzują się obniżoną temperaturą i występują w wyniku koncentracji strumienia pola magnetycznego, co ogranicza konwekcję. Liczba plam zmienia się w czasie w cyklu słonecznym o długości około 11 lat.

Zmienna uwzględniona w analizie to Monthly Mean Total Sunspot Number. Dane obejmują próbę o długości 3264, a ich źródłem jest baza danych SIDC - Solar Influences Data Analysis Center. Zostały one pobrane z platformy Kaggle pod adresem: Sunspots dataset on Kaggle. Zostały one udostępnione na prawach domeny publicznej. Zawierają one informacje o średniej miesięcznej liczbie plam słonecznych od 1749/01/01 do 2017/08/31.

Wizualizacja danych

Poniżej przedstawiono wykres surowych danych, czyli miesięcznych średnich liczby plam słonecznych:



Wykres nr 1: Wizualizacja danych

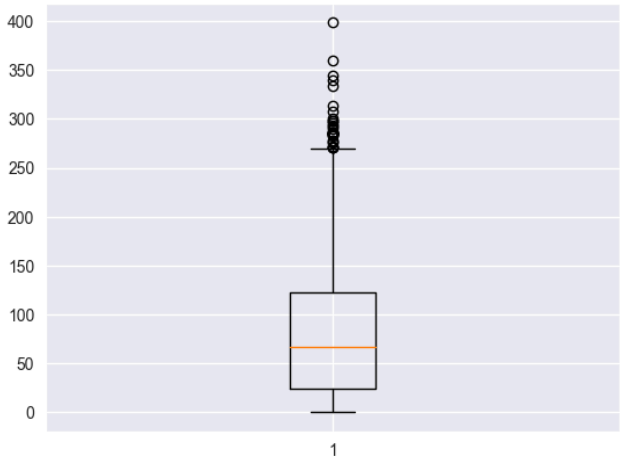
Na wykresie nr 1 wyraźnie widać pewną periodyczność danych.

2. Przygotowanie danych do analizy

Na początku przeprowadzono weryfikację jakości danych pod kątem:

- Wartości spoza zakładanego przedziału - liczba plam słonecznych powinna być liczbą nieujemną.
- Braków w danych - analizowano występowanie brakujących obserwacji w kolumnie Monthly Mean Total Sunspot Number.
- Błędów w próbkowaniu - sprawdzono, czy dane są równomiernie rozłożone w czasie.

Def. Wykres pudełkowy (boxplot) – Jest to to graficzna prezentacja rozkładu cechy statystycznej. Prostokąt wykresu rozciąga się od pierwszego (Q1) do trzeciego kwartylu (Q3), a jego wysokość odpowiada rozstępowi międzykwartylowemu (IQR). Wewnątrz znajduje się linia oznaczająca medianę. "Wąsy" wykresu rozciągają się od Q1 - 1,5 IQR do Q3 + 1,5 IQR.



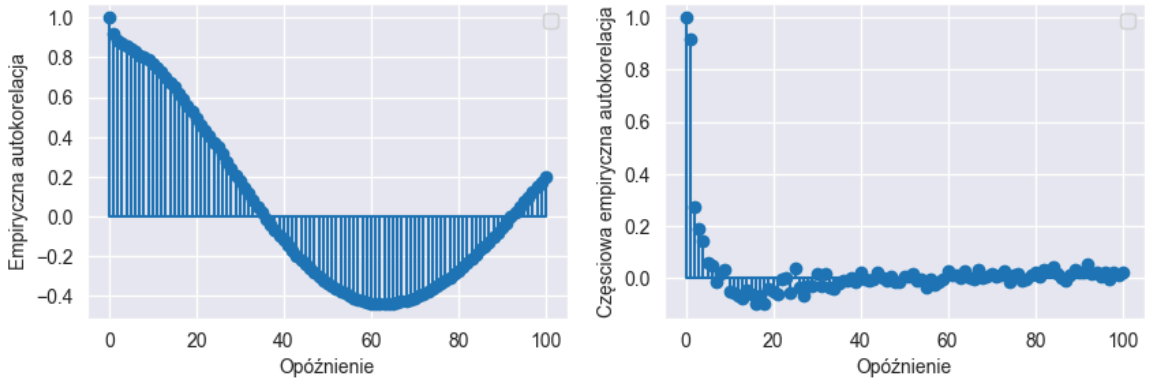
Wykres nr 2: Wykres pudełkowy danych

Próbka nie zawiera braków danych. Mediana danych znajduje się w dolnej połowie wykresu pudełkowego, co sugeruje lekką asymetrię rozkładu w kierunku wyższych wartości. Większość danych mieści się w przedziale od dolnego do górnego kwartylu, czyli pomiędzy ok. 25 a 100. Na wykresie widać wiele wartości odstających powyżej górnej granicy wąsów, co wskazuje na obecność ekstremalnych wartości w górnym zakresie.

Def. Autokorelacja (ACF) – Funkcja opisująca zależności pomiędzy obserwacjami szeregu czasowego a jego opóźnionymi wartościami. Jest wykorzystywana do identyfikacji wzorców sezonowych, trendów oraz do oceny stacjonarności szeregu czasowego.

Def. Częściowa autokorelacja (PACF) – Funkcja określająca bezpośrednią zależność pomiędzy wartością szeregu a jego opóźnionymi wartościami, eliminując wpływ pośrednich opóźnień. Jest szczególnie użyteczna w identyfikacji rzędu procesów autoregresyjnych (AR), ponieważ wskazuje liczbę istotnych opóźnień w modelu.

W celu oceny predykcji wyodrębniono część danych do zbioru testowego. Zbiór testowy obejmuje dane z ostatnich 100 miesięcy:



Wykres nr 3: Empiryczna autokorelacja i częściowa autokorelacja

Na wykresie nr 3 znajdują się się wykresy empirycznej autokorelacji oraz częściowej autokorelacji. Wykres autokorelacji wskazuje na znaczną autokorelację dla niskich opóźnień, która powoli maleje w sposób sinusoidalny. Takie zachowanie sugeruje, że dane mogą mieć charakter cykliczny lub sezonowy, co jest zgodne z oczekiwaniami, jeśli analizowane dane dotyczą plam słonecznych. Częściowa autokorelacja wykazuje znaczące wartości dla pierwszych kilku opóźnień, a następnie gwałtownie maleje i oscyluje wokół zera. Wyraźny pik na początkowych opóźnieniach sugeruje, że w danych dominuje proces autoregresyjny. Rząd procesu AR można oszacować na podstawie liczby istotnych lagów na wykresie PACF.

Identyfikacja trendów deterministycznych

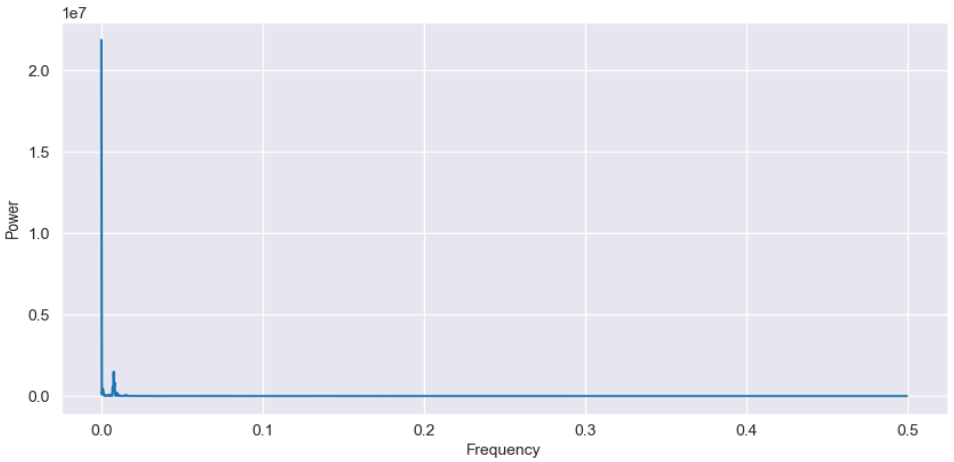
Def. Periodogram to narzędzie w analizie szeregów czasowych, które pozwala na oszacowanie gęstości widmowej mocy sygnału, czyli rozkładu energii (mocy) w funkcji częstotliwości. Jest to metoda stosowana do identyfikacji częstotliwości, które dominują w danym sygnale lub szeregu czasowym.

Periodogram dla dyskretnego szeregu czasowego $x(t)$ jest oszacowaniem widma mocy tego szeregu, które jest definiowane jako kwadrat amplitudy transformaty Fouriera sygnału. Dla próbki $\{x_1, x_2, \dots, x_N\}$ periodogram $P(f)$ dla częstotliwości f wyraża się wzorem:

$$P(f) = \frac{1}{N} \left| \sum_{n=1}^{N} x_n e^{-i2\pi f n} \right|^2$$

gdzie:

- x_n to wartość sygnału w czasie n,
- f to częstotliwość,
- N to liczba próbek.



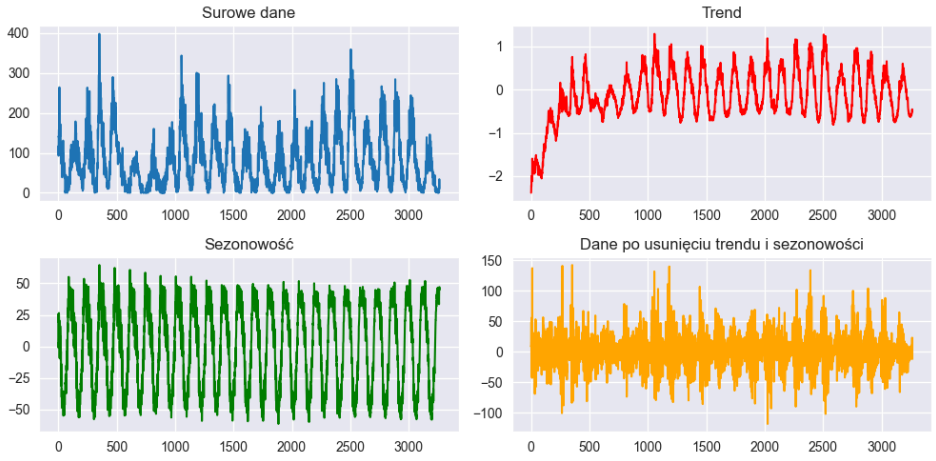
Wykres nr 4: Periodogram

Na podstawie sporządzonego periodogramu (wykres nr 4) zidentyfikowaliśmy największe skoki w widmie częstotliwościowym danych. Największa wartość mocy widma występuje dla okresu ~131, co odpowiada częstotliwości 0.0076. Wartość ta została wykorzystana jako parametr **seasonal_periods** w funkcji **ExponentialSmoothing**, aby skutecznie modelować i uwzględnić komponent sezonowy.

Def. Trend - Długoterminowa zmiana wartości zmiennej, która wskazuje na jej systematyczny wzrost lub spadek w czasie.

Def. Sezonowość - Regularne, powtarzające się zmiany w danych, które występują w określonych cyklach. Wynika z powtarzających się wpływów czynników, które mają wpływ na analizowaną zmienną.

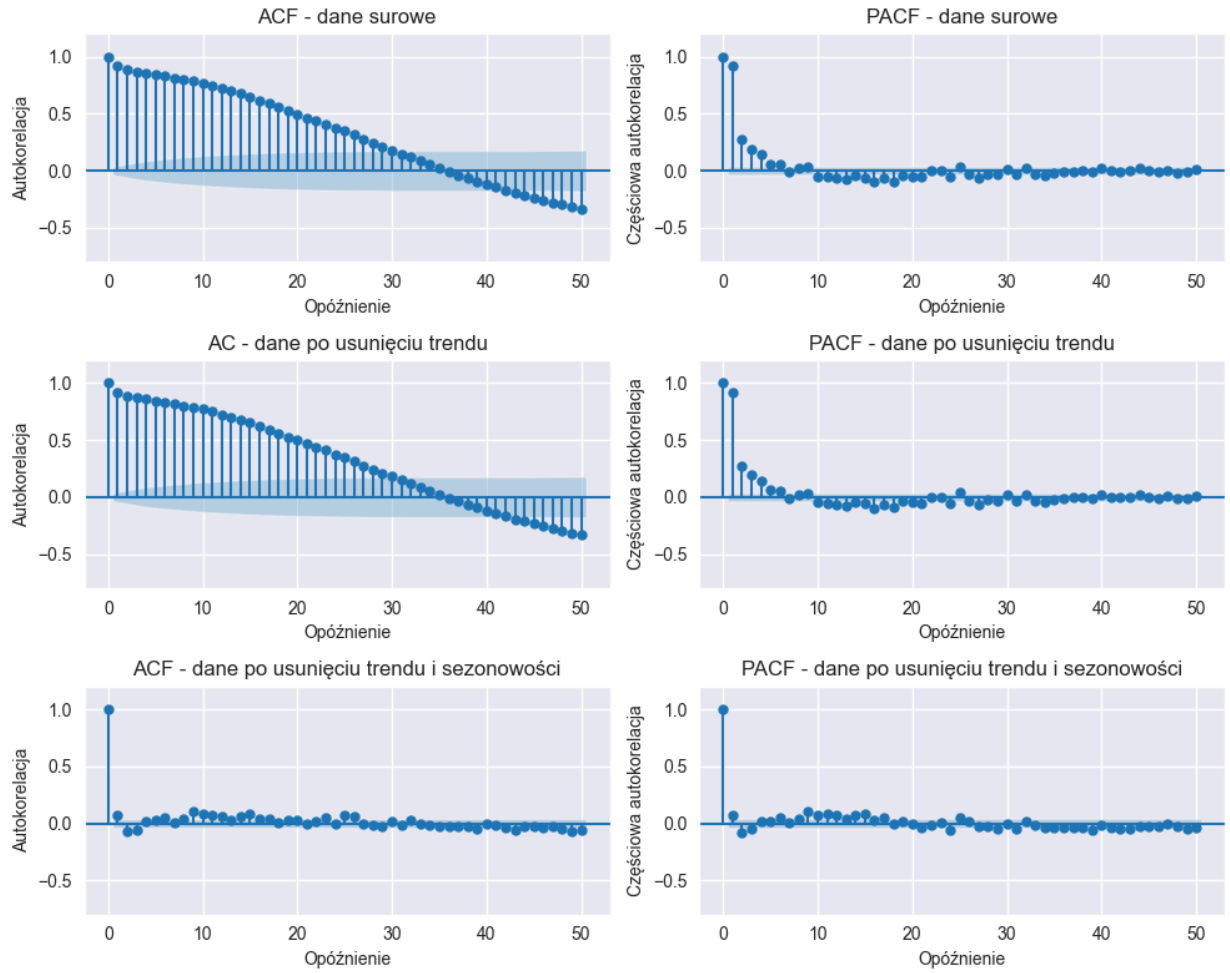
Na wykresie nr 5 przedstawiono proces dekompozycji czasowej danych:



Wykres nr 5: Dekompozycja danych

Surowe dane mają wyraźny wzorzec cykliczny o regularnych fluktuacjach. Widoczne są również zmiany w amplitudzie, co może sugerować obecność trendu lub zmienność w czasie. Po dekompozycji widoczny jest trend malejący w pierwszej części szeregu czasowego, który następnie stabilizuje się. Sugeruje to, że dane zawierają długoterminowe zmiany, które mogą być związane z powolnymi zmianami w analizowanym procesie. Sezonowość jest wyraźnie widoczna i stabilna w czasie, co wskazuje na regularnie powtarzające się wzorce. Charakterystyka sezonowości (amplituda i częstotliwość) pozostaje względnie stała. Dane po odjęciu trendu i sezonowości wydają się być szeregiem z losowym rozkładem reszt. Mogą one być dalej modelowane za pomocą modelu ARMA.

Ostatecznie wykonano wykresu empirycznej autokorelacji i częściowej autokorelacji, dla surowych danych oraz dla danych po usunięciu komponentów deterministycznych.



Wykres nr 6: ACF i PACF dla nowych szeregów

3. Modelowanie danych przy pomocy ARMA

W celu określenia optymalnego rzędu modelu ARMA zastosowano kryteria informacyjne: AIC, BIC oraz HQIC. Kryteria te uwzględniają zarówno jakość dopasowania modelu do danych, jak i liczbę parametrów, co pozwala na wybór najlepszego modelu.

Def. Kryteria informacyjne

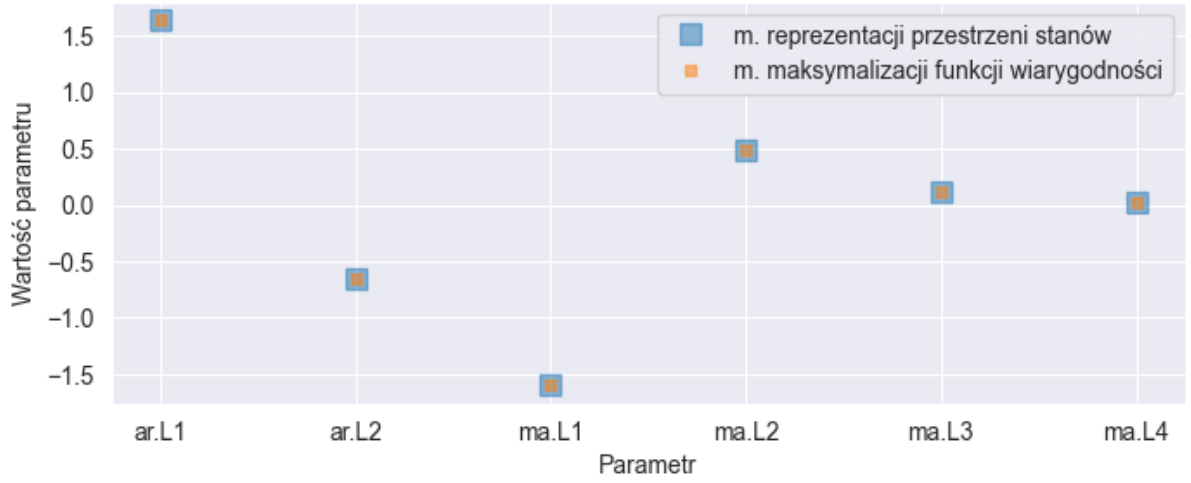
1. **AIC (Kryterium Informacyjne Akaike'a)**
   Kryterium AIC ocenia model, balansując pomiędzy dopasowaniem do danych a liczbą parametrów. Im mniejsza wartość AIC, tym lepszy model. AIC preferuje modele, które dobrze odzwierciedlają strukturę danych, jednocześnie unikając nadmiernego uproszczenia.

2. **BIC (Kryterium Informacyjne Bayesa)**
   Kryterium BIC, podobnie jak AIC, ocenia jakość dopasowania modelu, ale kładzie większy nacisk na ograniczenie liczby parametrów. Dzięki temu BIC preferuje prostsze modele, co jest szczególnie istotne w przypadku dużych zbiorów danych.

3. **HQIC (Kryterium Informacyjne Hannana-Quinna)**
   HQIC to kompromis pomiędzy AIC a BIC. Jest bardziej wyważone, jeśli chodzi o uwzględnianie liczby parametrów, co sprawia, że szczególnie dobrze sprawdza się w analizach dużych zbiorów danych. HQIC wybiera modele, które oferują dobrą równowagę między dopasowaniem a złożonością.

Dla wszystkich możliwych kombinacji parametrów pp i qq w przedziale od 0 do 5 obliczono wartości AIC, BIC oraz HQIC. Wszystkie trzy kryteria wskazały model **ARMA(2,4)** jako najlepszy, co oznacza, że model ten zapewnia optymalne dopasowanie przy jednoczesnym zachowaniu odpowiedniej prostoty.

Parametry modelu ARMA(2,4) oszacowano z wykorzystaniem dwóch różnych metod: reprezentacji przestrzeni stanów oraz maksymalizacji funkcji wiarygodności. Oba zestawy estymacji przedstawiono graficznie na wykresie nr 7.

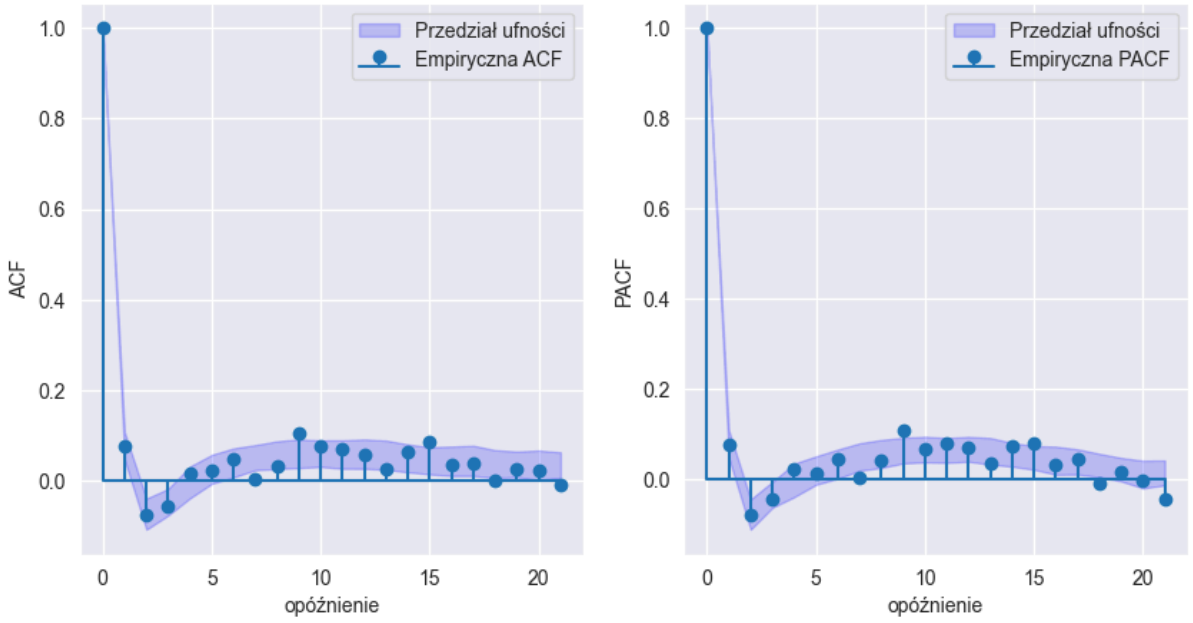


Wykres nr 7: Porównanie parametrów

Wartości wyestymowane dwoma różnymi sposobami są do siebie zbliżone. Metoda maksymalizacji funkcji wiarygodności dodatkowo zakłada, że residua są białym szumem tj. niezależność oraz jednolitość rozkładu residuów. Zatem ostatecznie wybrano analizę modelu ze współczynnikami wyznaczonymi metodą pierwszą.

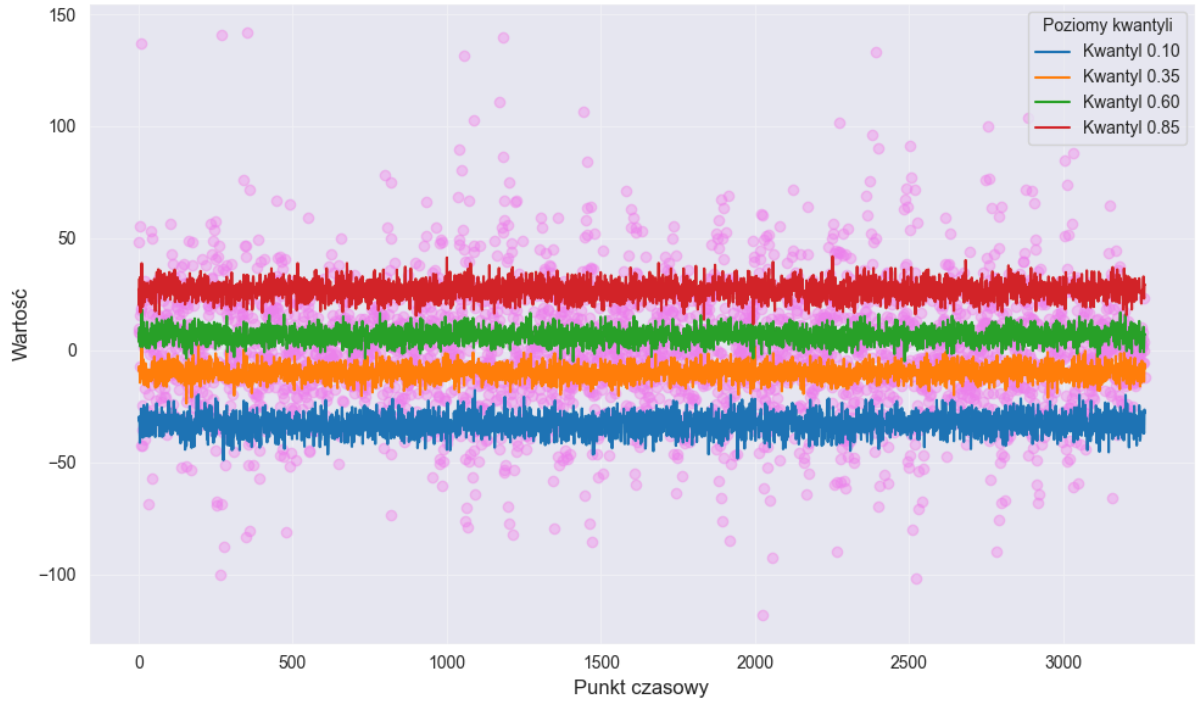
4. Ocena dopasowania modelu

Aby ocenić dopasowanie modelu, użyto przedziałów ufności dla empirycznego wykresu autokorelacji i autokorelacji częściowej dla residuów.



Wykres nr 8: Przedziały ufności oraz empiryczna autokorelacja i częściowa autokorelacja

Empiryczne wykresy pokrywają się z przedziałami ufności na poziomie 95%. Nie obserwuje się większych odchyleń od przedziałów. Dla większości opóźnień wartości autokorelacji zawierają się w przedziałach ufności.



Wykres nr 9: Linie kwantylowe oraz wizualizacja residuów

kwantyle	Teoretyczny udział residuów	Empiryczny udział residuów
0,1 - 0,35	25%	23,001%
0,1 - 0,6	50%	54,028%
0,1 - 0,85	75%	77,642%
0,35 - 0,6	25%	31,026%
0,35 - 0,85	50%	54,640%
0,6 - 0,85	25%	23,614%

Tabela nr 1: Procentowy udział residuów między liniami kwantylowymi.

Rozkład danych między liniami kwantylowymi pokazuje, w jakim zakresie znajdują się reszty modelu w stosunku do różnych percentyli. W wąskich zakresach kwantylowych procentowy udział reszt wynosi około 23%, co oznacza, że reszty są równomiernie rozproszone w węższych zakresach kwantylowych. To wskazuje na brak znacznych

koncentracji wartości w tych regionach, co może być efektem stabilnego działania modelu dla większości danych. Około 54% reszt mieści się między kwantylami 0.10 a 0.60, co sugeruje, że większość danych znajduje się bliżej mediany, ale z lekką tendencją do rozkładu w kierunku niższych wartości. Prawie 78% reszt mieści się między kwantylami 0.10 a 0.85, co świadczy o tym, że zdecydowana większość danych znajduje się w stosunkowo szerokim zakresie wokół wartości centralnych. Zakres między kwantylami 0.35 a 0.60 obejmuje około 31% reszt, co oznacza, że środkowy segment danych jest mniej zagęszczony w porównaniu z szerszym zakresem.

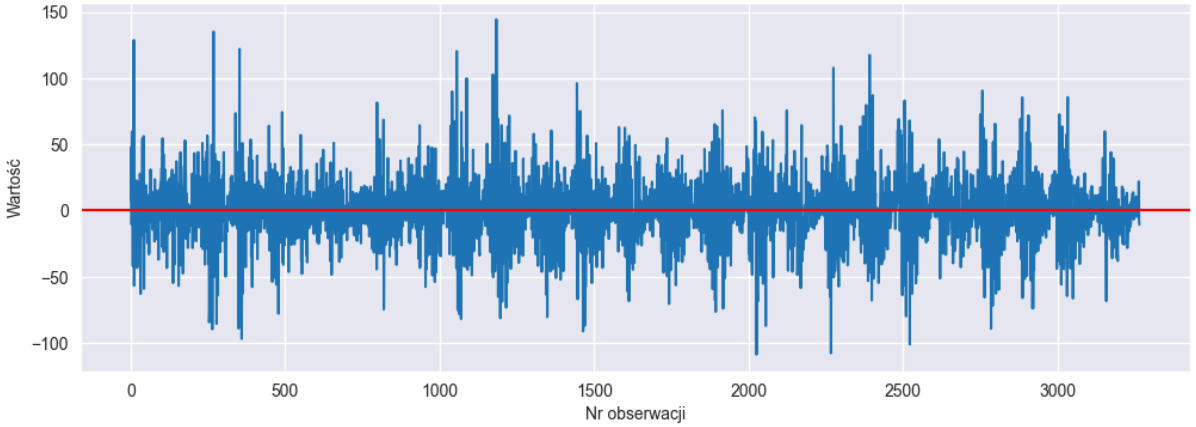
Rozkład reszt wygląda na symetryczny wokół mediany, gdyż wskazuje na brak dużych błędów systematycznych. Większość reszt znajduje się w granicach między kwantylami 0.10 a 0.85, co oznacza, że model przewiduje dane stabilnie, a reszty są umiarkowanie rozproszone.

5. Weryfikacja założeń szumu

Aby zweryfikować założenia szumu, wyznaczymy residua na podstawie dopasowanego modelu i sprawdzimy ich zachowania.

Założenia szumu są następujące:

- residua są nieskorelowane,
- residua mają wartość oczekiwaną równą 0,
- residua mają stałą wartość wariancji,
- (dodatkowe) residua mają ten sam rozkład, w szczególności rozkład normalny.



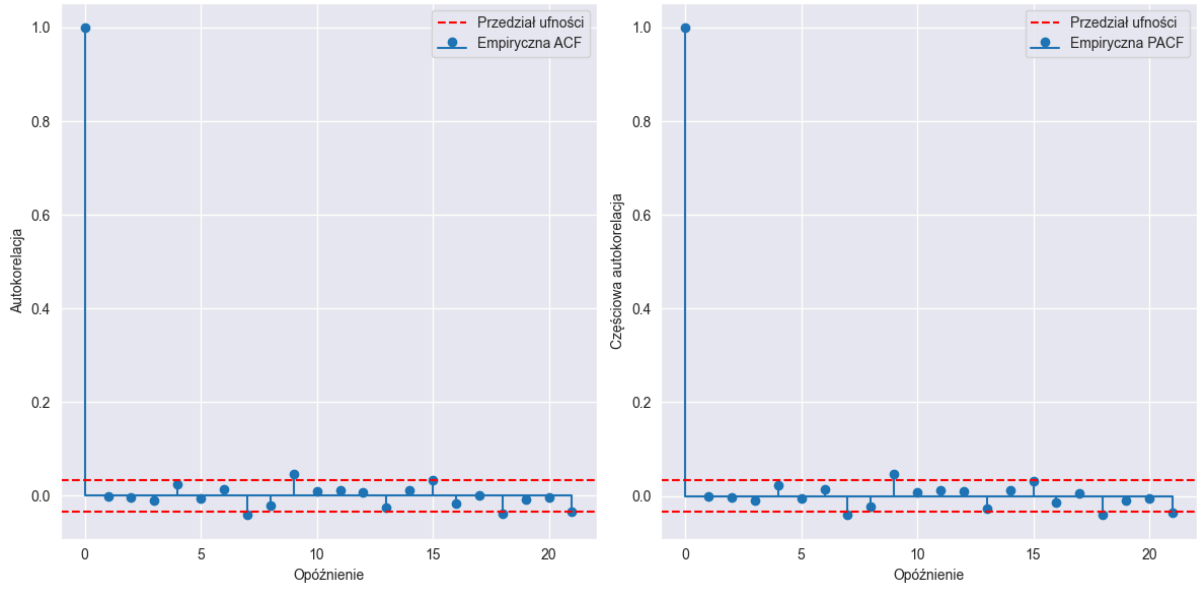
Wykres nr 10: Wizualizacja reszt

Wartość oczekiwana reszt wyniosi -0,0413.

Do sprawdzenia stałości wariancji wykorzystano test jednorodności wariancji Levene`a.

Def. Test Levene'a (jednorodności wariancji): Dla każdej zmiennej zależnej wykonywana jest analiza wariancji wartości bezwzględnych odchyleń od średniej w odpowiedniej grupie. Jeżeli test Levene'a daje wynik statystycznie istotny, to należy odrzucić hipotezę o jednorodności wariancji.

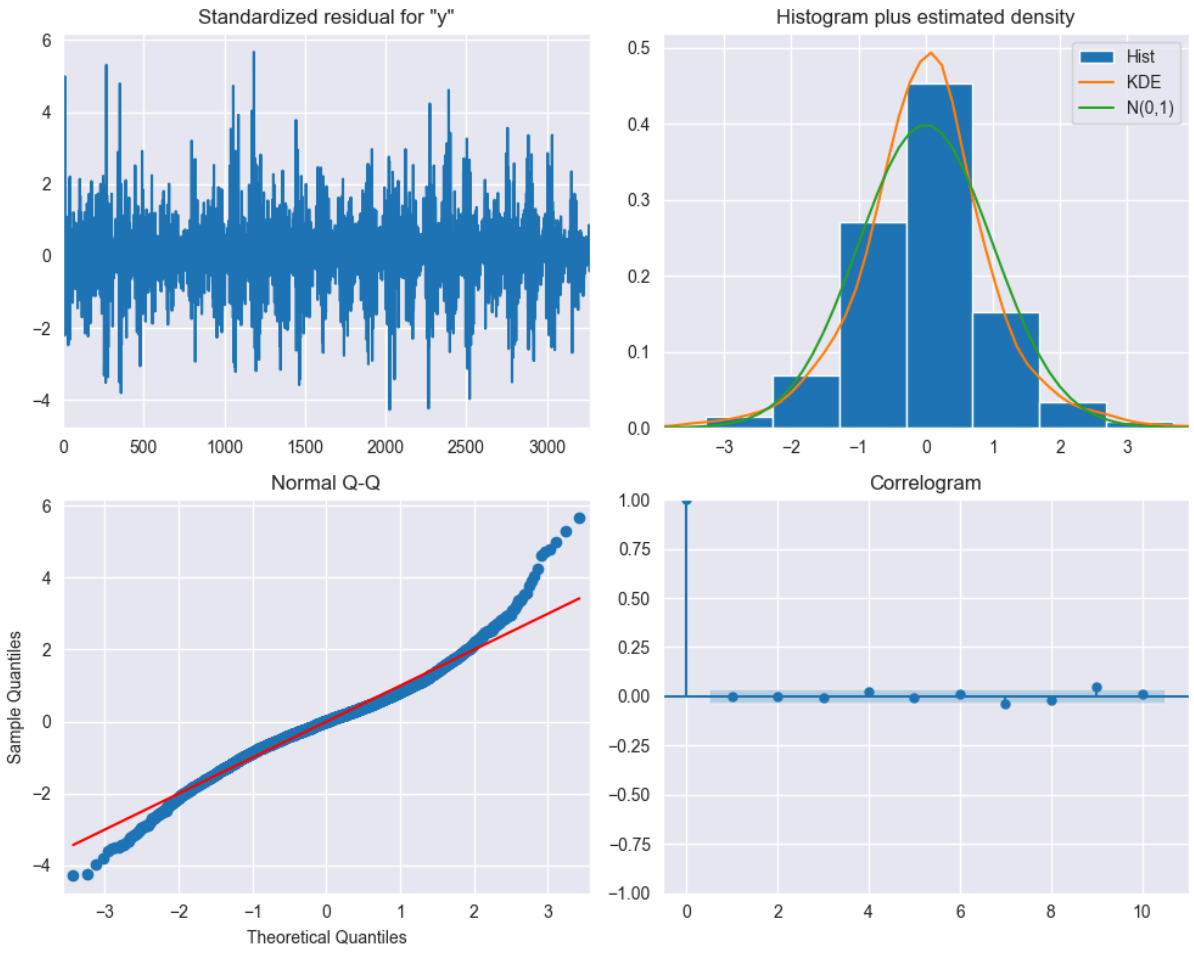
Z powyższego testu statystycznego, wynika, że wariancja jest wartością stałą dla residuów na poziomie ufności 95%.



Wykres nr 11: Empiryczne ACF i PACF oraz przedziały ufności

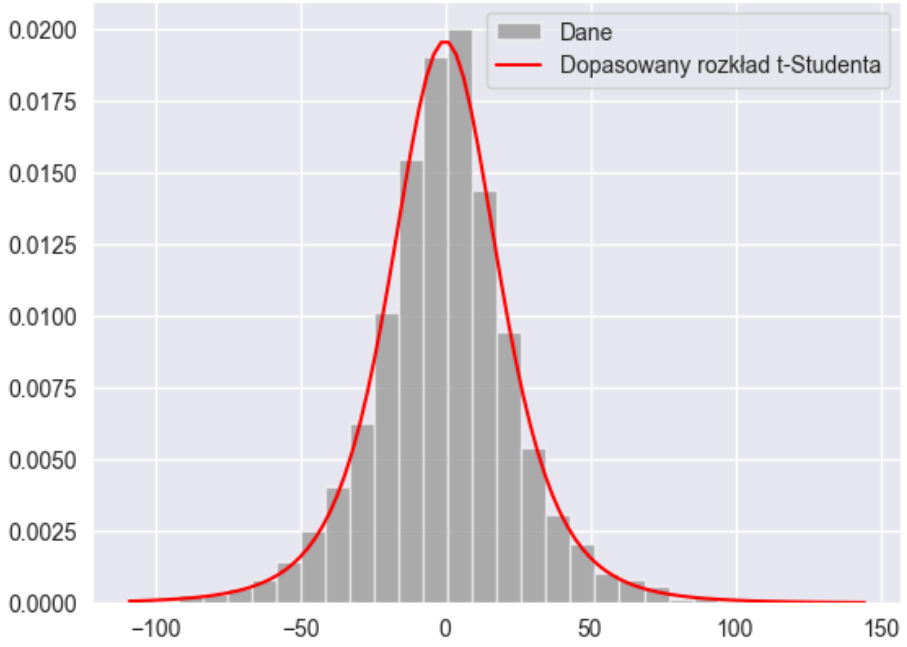
Na wykresie ACF widzimy, że większość punktów znajduje się w granicach przedziału ufności. To sugeruje, że autokorelacje na dalszych opóźnieniach są bliskie zeru i brak jest istotnych powiązań między kolejnymi wartościami residuów. Podobnie jak w przypadku ACF, na wykresie PACF także niemal wszystkie wartości mieszczą się w przedziale ufności. Oznacza to brak istotnych częściowych autokorelacji. Residua można uznać za niezależne, ponieważ nie ma dowodów na występowanie znaczącej autokorelacji.

Def. QQ plot (Quantile-Quantile plot) - Wykres porównujący rozkład empiryczny z teoretycznym lub dwa rozkłady empiryczne. Oś X przedstawia kwantyle jednego, a oś Y drugiego rozkładu. Jeśli dane są zgodne, punkty układają się wzdłuż y= x. Służy do oceny normalności danych i wykrywania odchyleń.



Wykres nr 12: Testy na normalność rozkładu.

Z powyższych wykresów, nietrudno zauważyć, że residua modelu mają rozkład zbliżony do normalnego. Jednak na Q-Q plotcie, widać wartości odstające. Zatem dopasowano do danych rozkład t-Studenta i oszacowano parametry.



Wykres nr 13: Gęstość rozkładu t-Studenta oraz histogram reszt

Rozkład t-Studenta z parametrami: 4 stopni swobody, $\sigma = 19,25$, $\mu = -0,05$ zdecydowanie lepiej opisuje residua, niż rozkład normalny.

6. Wnioski

W przeprowadzonej analizie zastosowano model ARMA w celu dopasowania danych do modelu szeregów czasowych. Podczas procesu modelowania zbadano założenia modelu oraz cechy danych, co miało na celu znalezienie odpowiedniego modelu opisującego dane. Model ARMA został dopasowany do danych na podstawie kryteriów informacyjnych oraz analizy wykresów autokorelacji i częściowej autokorelacji.

Reszty modelu wykazują charakterystykę zbliżoną do białego szumu. Nie zaobserwowano istotnych autokorelacji w resztach, co prowadzi do wniosku o niezależności residuów. Rozkład reszt jest symetryczny i skoncentrowany wokół zera, a ich wariancja jest stała w czasie na poziomie istotności 95%. Empiryczne wykresy autokorelacji oraz częściowej autokorelacji zawierają się w przedziałach ufności dla rozważanego modelu. Analiza rozkładu reszt w odniesieniu do linii kwantylowych doprowadziła również do uznania ich jako symetrycznych do mediany i rozproszonych równomiernie.

Powyższe wnioski wskazują, że opisywane dane można modelować za pomocą modelu ARMA(2,4). Możliwe, że potrzebne jest rozważenie rozszerzonych klas modeli, takich jak

SARIMA oraz usunięcia dodatkowego deterministycznego okresowego komponentu, widocznego na wykresie nr 12, aby jeszcze dokładniej móc uwzględnić specyficzne cechy danych.

Źródła:

Autoregressive moving-average model - Wikipedia

Autokorelacja - Pogotowie Statystyczne

Periodogram – Wikipedia, wolna encyklopedia