Politechnika Wrocławska

Wydział Matematyki

Matematyka Stosowana

Analiza emisji CO przy pomocy modelu ARMA

Autorzy: Antoni Krzak, Filip Miśkiewicz

Styczeń 2025

Spis treści

1 Wstęp

1.1 Cel pracy

Raport powstał w ramach kursu Komputerowa Analiza Szeregów Czasowych. Naszym celem była analiza danych emisji tlenku węgla w Stanach Zjednoczonych przy użyciu języka *Python*. Korzystamy z modelu ARMA, który służy do opisu szeregu za pomocą dwóch wielomianów – autoregresji oraz średniej ruchomej. Przed dopasowaniem modelu, dane należało odpowiednio przetworzyć, tak by spełniały jego warunki. Dla ARMA(p, q) szereg X_t opisujemy za pomocą równania (1).

$$X_t - \phi_1 X_{t-1} - ... - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} +_{...} + \theta_q Z_q \tag{1}$$

gdzie: $\phi_1, ..., \phi_p$ - współczynniki wielomianu autoregresji; $\theta_1, ..., \theta_q$ - współczynniki wielomianu średniej ruchomej; $\phi, \theta \in \mathbb{R}$ oraz $p, q \in \mathbb{Z}$.

W pracy skupiamy się na dobraniu odpowiednich parametrów p, q (dalej zwanych także rzędem modelu) oraz wykonujemy szereg testów, mających na celu ocenę wykonanego przez nas dopasowania oraz weryfikację założeń dotyczących szumu.

1.2 Dane

Dane [1], które wykorzystujemy w naszej pracy dotyczą zanieczyszczeń powietrza na terenie Stanów Zjednoczonych w latach 2000 – 2023. Otwierając plik za pomocą biblioteki *pandas* obserwujemy, że dane składają się z 22 kolumn, reprezentujących różne rodzaje zanieczyszczeń powietrza oraz 665414 wierszy z podziałem na poszczególne stany oraz dni, w których wykonano pomiary. W analizie skupiamy się na emisji CO w latach 2000–2022. Wszystkie wartości poziomu tlenku węgla podane zostały w ppm (parts per milion – części na milion).

2 Przygotowanie danych do analizy

2.1 Zbadanie jakości danych

Początkowo edytujemy plik poprzez pozostawienie dwóch interesujących nas kolumn – daty wykonania pomiaru oraz wykazanej emisji CO. W obu kolum-

Statystyki	Emisja CO
liczba obserwacji	1201
średnia wartość	0.354764
odchylenie standardowe	0.154514
minimalna wartość	0.158127
kwantyl 25%	0.249492
mediana	0.318545
kwantyl 75%	0.407180
maksymalna wartość	1.164770

Tabela 1: Podstawowe statystyki analizowanych danych.

nach nie występują braki w danych, ani wartości spoza możliwego przedziału, dzięki czemu możemy przejść do odpowiedniego sformatowania danych. Z powodu dużej liczby pomiarów przypadających na każdy dzień, wyznaczamy średnią emisję CO w każdym tygodniu.

Wstępne statystyki danych uzyskujemy za pomocą funkcji *describe* we wspomnianej bibliotece *pandas*.

2.2 Dekompozycja szeregu czasowego

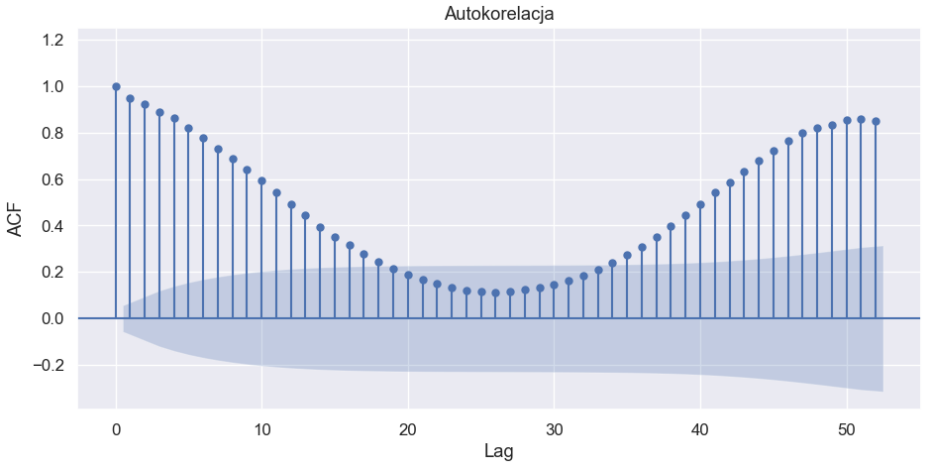
Wykres danych przed dekompozycją i odpowiednim formatowaniem możemy zobaczyć na Rysunku 1. W latach 2000 – 2013 obserwujemy wyraźny trend malejący, a na przestrzeni całego badanego okresu występuje sezonowość.

Wahania sezonowe dobrze obrazuje autokorelacja danych w okresie jednego roku przedstawiona na wykresie Rysunku 2. ARMA służy do modelowania szeregów stacjonarnych, zatem w celu przeprowadzenia analizy, sformatujemy dane poprzez wykonanie odpowiedniej transformacji i usunięcie trendu oraz sezonowości.

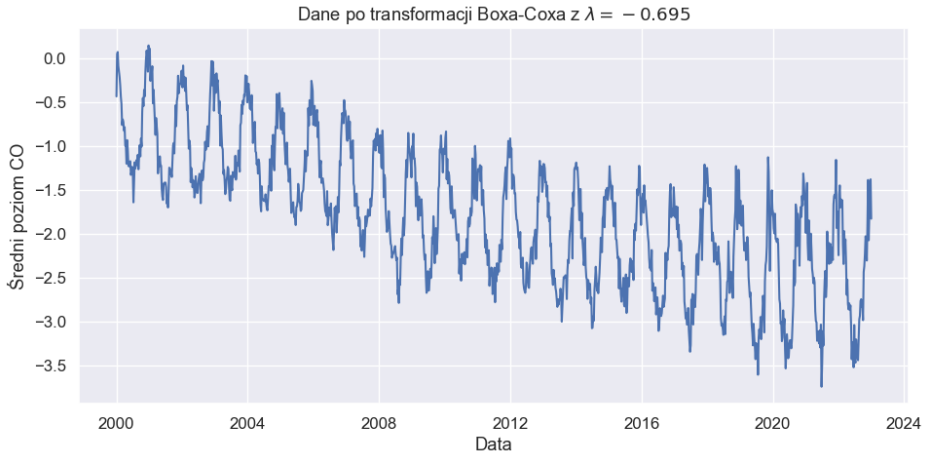
2.2.1 Transformacja Boxa–Coxa.

Transformacja Boxa–Coxa [2] sprowadza dane do rozkładu normalnego poprzez przekształcenie oparte na współczynniku λ. Aby zostosować transformację Boxa–Coxa transformowane wartości muszą być dodatnie, co jest prawdą dla badanego przez nas zbioru (minimalna wartość = 0.158127).

Rysunek 1: Wykres danych przed dekompozycją.



Rysunek 2: Autokorelacja danych przed usunięciem trendu i sezonowości.



Rysunek 3: Dane po transformacji Boxa–Coxa.

Transformację Boxa–Coxa wyrażamy wzorem (2).

$$x' = \begin{cases} \frac{x^\lambda - 1}{\lambda} \text{ dla } \lambda = 0 \\ \ln x \text{ dla } \lambda \neq 0 \end{cases} \tag{2}$$

gdzie λ to maksymalny argument logarytmu funkcji największej wiarygodności wyrażanego wzorem (3).

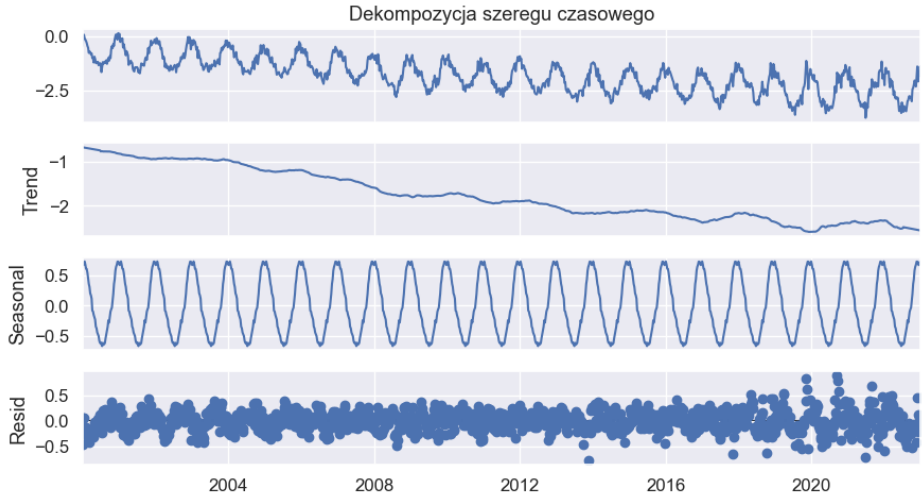
$$LL = -\frac{n}{2} \ln \sigma^2 + (\lambda - 1) \sum_{i=1}^{n} \ln (x_i) \tag{3}$$

gdzie n – liczność próby, σ – odchylenie standardowe danych, $x_1, x_2, ..., x_n$ – obserwacje. Transformację przeprowadzamy za pomocą funkcji *boxcox* z biblioteki *scipy.stats* z wyznaczonym parametrem $\lambda = -0.695$. Na Rysunku 3 przedstawiamy wykres szeregu po zastosowaniu transformacji Boxa–Coxa.

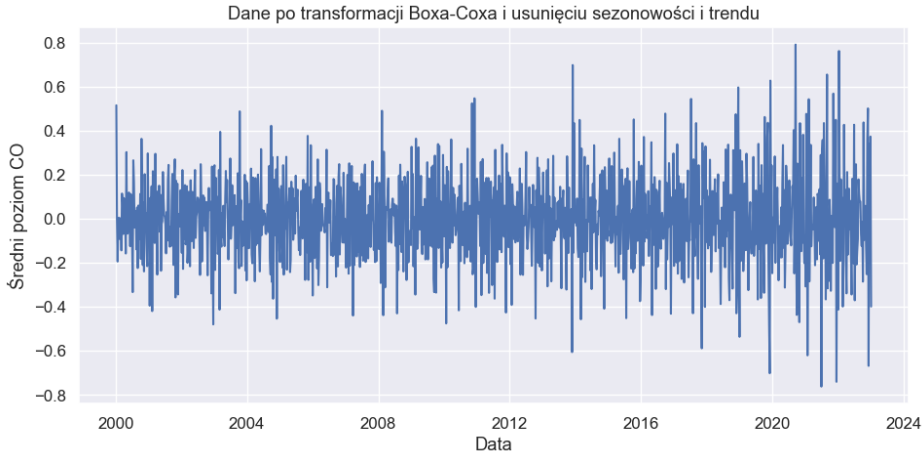
2.2.2 Usunięcie sezonowości i trendu

Po wykonanej transformacji możemy dokonać dekompozycję szeregu czasowego (Rysunek 4). Przedstawia ona kolejno: przebieg szeregu w czasie, trend, sezonowość oraz przebieg residuów.

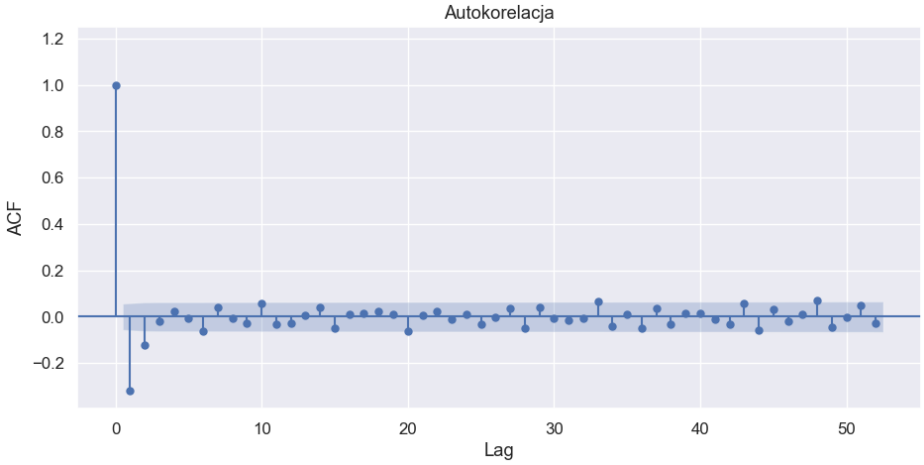
Odejmując sezonowość, a następnie usuwając trend metodą *diff* otrzymujemy wykres danych w czasie przedstawiony na Rysunku 5. Na uzyskanym szeregu możemy wykonać testy, które zweryfikują czy tak zmodyfikowane dane możemy analizować za pomocą modelu ARMA.



Rysunek 4: Dekompozycja szeregu czasowego.



Rysunek 5: Dane po transformacji i usunięciu trendu oraz sezonowości.

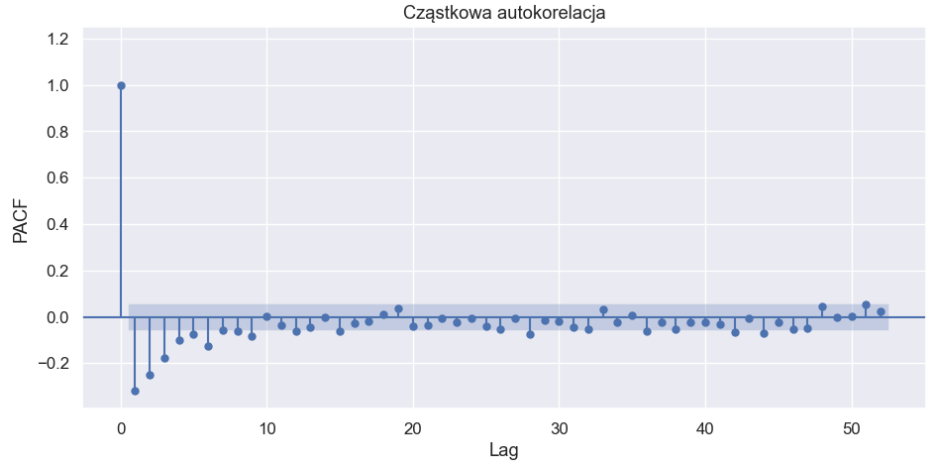


Rysunek 6: Autokorelacja szeregu.

2.2.3 Wzmocniony Test Dickeya-Fullera na stacjonarność szeregu.

Aby zweryfikować czy otrzymany szereg jest stacjonarny wykonujemy Wzmocniony Test Dickeya–Fullera, który sprawdza obecność pierwiastka jednostkowego w wielomianie autoregresyjnym. Hipoteza zerowa zakłada obecnośc takiego pierwiastka, natomiast alternatywna jego brak. Zakładając, że taki pierwiastek nie występuje, przyjmujemy, że szereg jest stacjonarny. Za poziom istotności w tym teście (a także we wszystkich następnych, które wykonamy) przyjmujemy $\alpha = 0.05$, Po wykonaniu testu na naszym szeregu (korzystając z biblioteki *statsmodels*) otrzymujemy, że statystyka testowa = -9.911, p-wartość = 0.0 $\Rightarrow$ szereg jest stacjonarny.

Możemy również zobaczyć jak wyglądają autokorelacja (Rysunek 6) oraz cząstkowa autokorelacja (Rysunek 7) szeregu po przeprowadzonych modyfikacjach. Widzimy, że sezonowość, którą obserwowaliśmy na Rysunku 2 już nie występuje. Możemy rozpocząć analizę uzyskanego szeregu za pomocą modelu ARMA.



Rysunek 7: Cząstkowa autokorelacja szeregu.

3 Modelowanie danych przy pomocy ARMA

Po udanej dekompozycji przejdziemy teraz do znalezienia szeregu ARMA, który odda jak najwierniej trajektorię oczyszczonych danych. Dzięki temu będziemy mogli, chociażby, spróbować przewidzieć przybliżone wartości danych w przyszłości, a także powiedzieć więcej na temat tych, które już posiadamy.

3.1 Dobranie rzędu modelu

Cały proces, rozpoczniemy od dopasowania rzędu modelu, czyli znalezienia wartości p i q. Za część autoregresyjną (AR) naszego szeregu czasowego odpowiada liczba p, która wskazuje też od ilu poprzednich wartości zależy aktualna wartość X_t. Z kolei q odpowiada za średnią ruchomą (MA), czyli to ilu niezależnych zmiennych Z_t generowanych z tego samego rozkładu o średniej 0 i stałej wariancji użyjemy do obliczenia X_t. Ostatecznie szereg czasowy ARMA(p,q) definiuje równanie (1).

Do znalezienia p i q użyjemy klasy *ARIMA* z biblioteki *statsmodels*. Przy pomocy powiązanej z nią funkcji *fit*, dla podanych wcześniej danych oraz rzędów AR i MA, stworzymy model ARMA, z wyestymowanymi także wszystkimi parametrami ϕ_i i θ_j (przy pomocy domyślnej metody *statespace*), dla $i = 1, 2, \ldots, p$ i $j = 1, 2, \ldots, q$. Najlepszy możliwy model wybierzemy porównując wartości kryteriów informacyjnych dla różnych p i q. Kryteria, których

p	q	AIC	BIC	HQIC
4	1	-778.20	-742.57	-764.78
5	1	-776.80	-736.09	-761.46
5	3	-776.23	-725.36	-757.05
6	3	-775.20	-719.21	-754.11
5	2	-774.68	-728.87	-757.42

Tabela 2: 5 najmniejszych wartości dla kryteriów informacyjnych (zaokrąglone do dwóch miejsc po przecinku).

użyjemy to:

- kryterium informacyjne Akaikego (Akaike informtion criterion),

$$AIC = 2(p+q) - 2lnL(p, q, x_1, \ldots, x_n),$$

- Bayesowskie kryterium informacyjne (Bayesian informtion criterion),

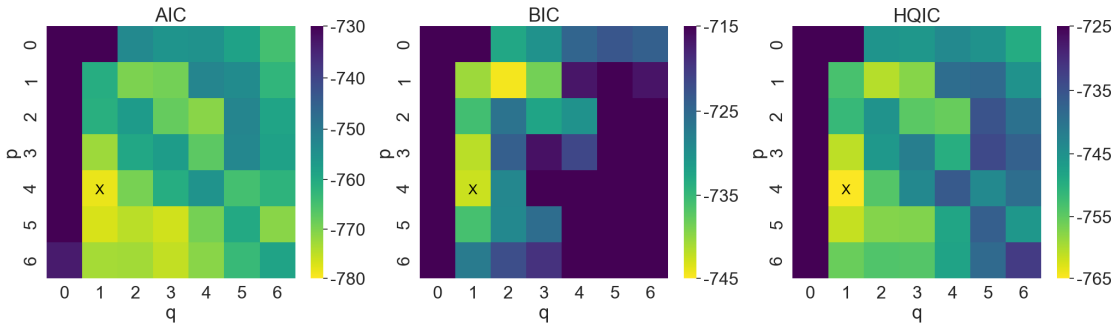
$$BIC = (p+q)ln(n) - 2lnL(p, q, x_1, \ldots, x_n),$$

- kryterium informacyjne Hannana-Quinna(Hannan-Quinn Information Criterion)

$$HQIC = 2(p+q)ln(ln(n)) - 2lnL(p, q, x_1, \ldots, x_n),$$

gdzie $L(p, q, x_1, \ldots, x_n)$ to funkcja największej wiarygodności, n to liczba obserwacji.

Ostatecznie po przeprowadzeniu testów, jako optymalne, czyli najmniejsze co do wartości kryteriów informacyjnych (Tabela 2) okazują się $p = 4$ oraz $q = 1$.



Rysunek 8: Heatmapy wartości kryteriów informacyjnych w zależności od p i q.

Wyniki dodatkowo potwierdzają heatmapy wartości poszczególnych kryteriów (Rysunek 8). Widzimy tutaj, że dla BIC ARMA(4,1) wypada trochę gorzej od kilku innych wartości dla części autoregresyjnej i średniej ruchomej, ale za to w przypadku AIC i HQIC wartości $p = 4$ i $q = 1$ spisują się zauważalnie najlepiej.

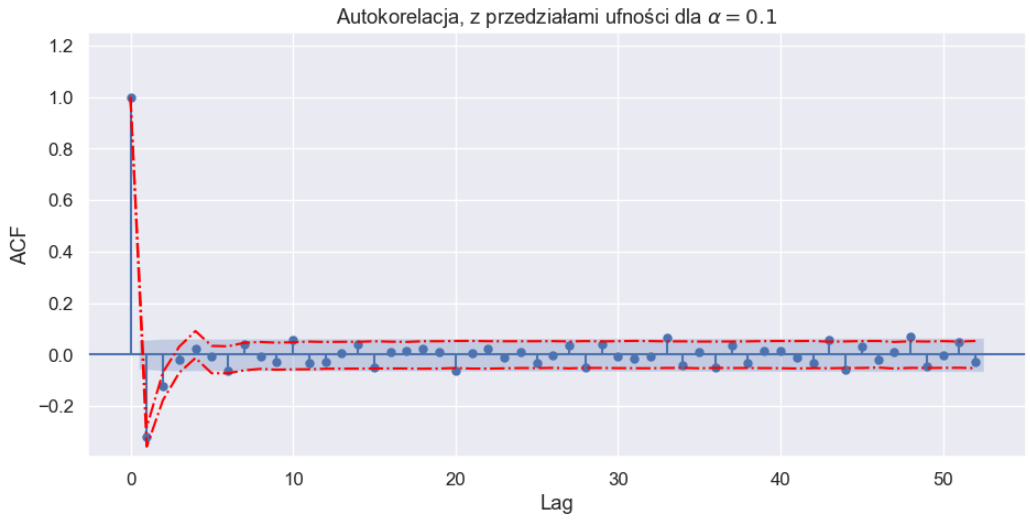
3.2 Estymacja parametrów modelu

W momencie, gdy dobraliśmy już rząd modelu możemy wyestymować teraz wartości parametrów ϕ i θ.

ϕ_1	ϕ_2	ϕ_3	ϕ_4	θ_1
0.459	0.060	0.078	0.083	-0.983

Tabela 3: Wartości wyestymowanych parametrów szeregu ARMA(4,1) (z dokładnością do trzech cyfr po przecinku).

Ostateczne wartości parametrów ϕ i θ znajdują się w Tabeli 3, i to na nich będziemy operować w dalszej części. Zostały one wyestymowane przy pomocy wcześniej już wspominanej funkcji *fit* (metodą *statespace*).



Rysunek 9: Autokorelacja szeregu czasowego z nałożonymi przedziałami ufności.

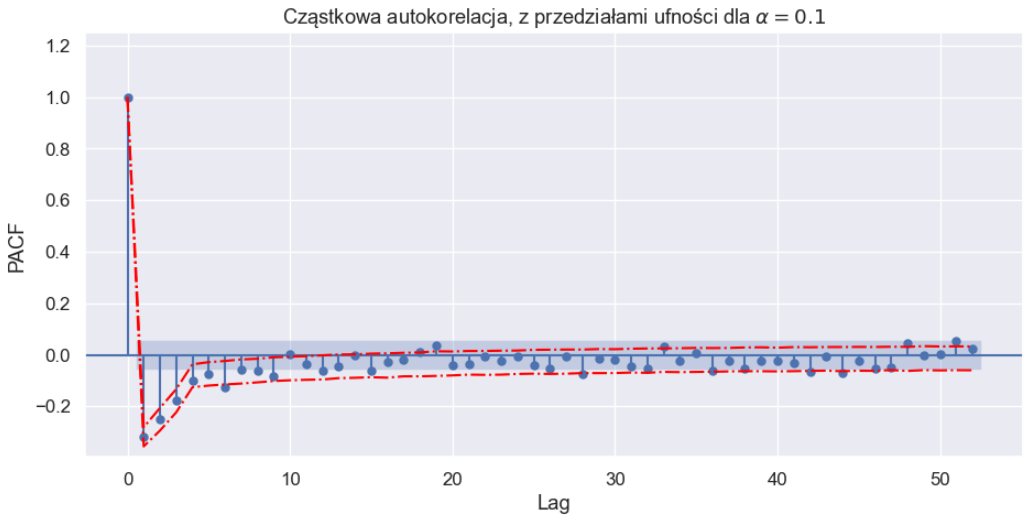
4 Ocena dopasowania modelu

W tym momencie przyszedł czas na weryfikację całego poprzedniego rozdziału, czyli odpowiedź na pytanie — czy wszystko to, co zrobiliśmy przed chwilą, ma pokrycie w danych. Dokładniej mówiąc, czy dopasowaliśmy model dobrego rzędu z odpowiednimi parametrami.

4.1 Przedziały ufności dla PACF i ACF

Zaczniemy od sprawdzenia, czy wcześniej już narysowane przez nas funkcje autokorelacji i cząstkowej autokorelacji mieszczą się w przedziałach ufności dla szeregu ARMA(4,1).

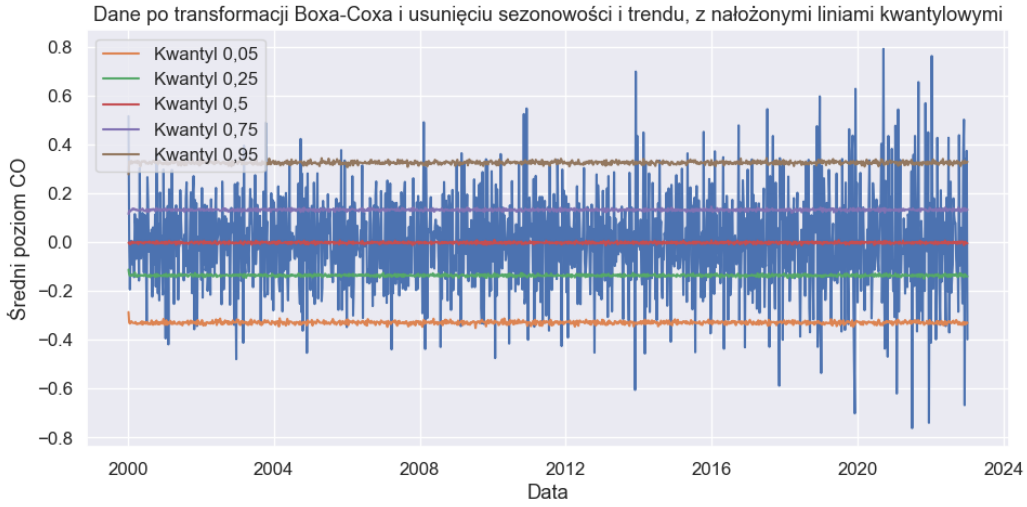
Dla autokorelacji (Rysunek 9) widać, że około 90% zakładanych wartości, przy dobranym poziomie ufności $\alpha = 0.1$, zawiera się w wyznaczonych przedziałach. Możemy już w takim razie, przynajmniej częściowo, wnioskować, że nasz model został dobrany poprawnie. Sprawdźmy jednak jeszcze, jak zachowuje się funkcja cząstkowej autokorelacji.



Rysunek 10: Cząstkowa autokorelacja szeregu czasowego z nałożonymi przedziałami ufności.

Dla cząstkowej autokorelacji (Rysunek 10) podobnie nie powinniśmy mieć zastrzeżeń co do liczby obserwacji poza wyznaczonymi przedziałami.

Ostatecznie, biorąc pod uwagę wykresy ACF i PACF z nałożonymi przedziałami ufności, stwierdzamy, że dokonany w poprzedniej części dobór modelu był trafny.



Rysunek 11: Trajektoria szeregu czasowego z liniami kwantylowymi.

4.2 Porównanie linii kwantylowych z trajektorią

W celu weryfikacji porównamy jeszcze wyznaczone przy pomocy symulacji Monte Carlo linie kwantylowe z trajektorią naszych danych.

Jak widzimy na Rysunku 11, zakładane części wszystkich wartości mieszczą się w odpowiadających przedziałach — około 10% obserwacji nie znajduje się między kwantylami rzędu 0.05 i 0.95, a mniej więcej połowa nie zawiera się pomiędzy pierwszym i trzecim kwartylem.

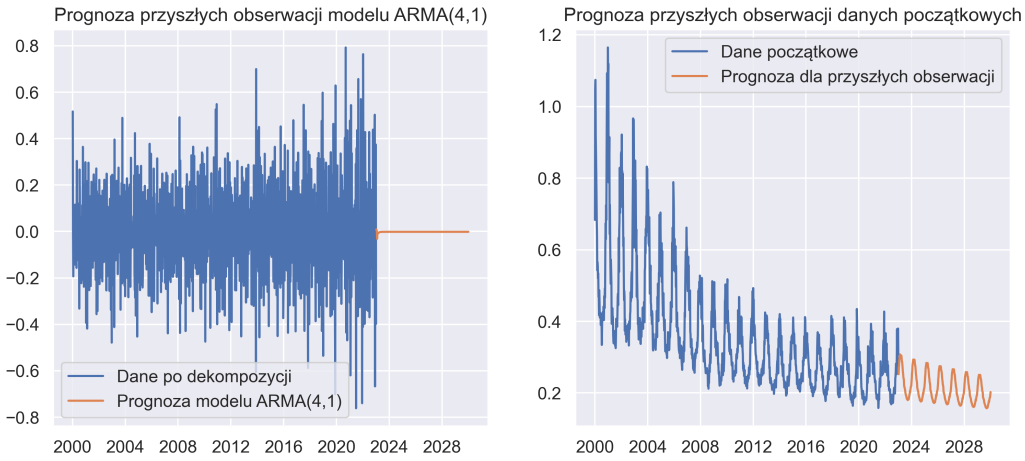
Kwantyl	0.05	0.25	0.5	0.75	0.95
Część danych mniejszych	0.049	0.227	0.506	0.768	0.952

Tabela 4: Części danych będące poniżej konkretnych linii kwantylowych.

W Tabeli 4 obserwujemy ile danych jest co do wartości mniejszych od pięciu przykładowych kwantyli szeregu ARMA(4,1) z poprzedniego wykresu. Widać, że otrzymane ułamki pokrywają się z dużą dokładnością z kwantylami. Na podstawie otrzymanych kwantylów, po raz kolejny wnioskujemy, że wybrany model ARMA jest poprawny.

4.3 Prognoza dla przyszłych obserwacji

W ramach końcowej weryfikacji sprawdzimy, czy predykcja wykonana przy pomocy metody *predict*, związanej z klasą *ARIMA*, da racjonalne rezultaty dla naszych danych. Jeżeli okaże się, że tak, będziemy już niemalże pewni o prawidłowości doboru parametrów i rzędu naszego modelu.



Rysunek 12: Prognozy przyszłych obserwacji do końca 2030 roku.

Jak widzimy zarówno dla szeregu ARMA, jak i danych początkowych predykcje wydaja się względnie logiczne wyniki. Szereg autoregresyjnej średniej ruchomej z założenia ma wartość oczekiwaną równą zero, co widać na lewym wykresie prognozy. Przewidywania widoczne na prawym wykresie powstały po nałożeniu na perdykcję z metody *predict* po kolei trendu i sezonowości, których pozbyliśmy się przy oczyszczaniu danych. Na sam koniec na tak powstałych danych przeprowadziliśmy transformację odwrotną do wcześniej wspominanej Boxa-Coxa, z tym samym parametrem λ. Końcowo nie wyglądają może one idealnie, ale na pewno w sporym stopniu przybliżają one wartości, które mogą wystąpić w przyszłości.

Sumując cały poświęcony ocenie modelu rozdział, możemy stwierdzić z dużą dozą pewności, że model dobrany w ramach rozdziału 3 jest poprawny. Jednakże predykcja dla danych początkowych wygląda nie do końca poprawnie. W następnej części sprawdzimy, czy nie jest to może spowodowane czymś innym niż źle dobranym rzędem modelu ARMA.

5 Weryfikacja założeń dotyczących szumu

Ostatnim z procesów, który musimy przeprowadzić, jest weryfikacja założeń dotyczących szumu naszych danych, przy uwzględnieniu, że są one z szeregu ARMA(4,1) o obliczonych już wcześniej parametrach ϕ i θ. Zrobimy to, aby mieć pewność, że autoregresywna średnia ruchoma na pewno oddaje poprawnie nasze dane i może służyć do pracy z nimi.

5.1 Założenie dotyczące średniej

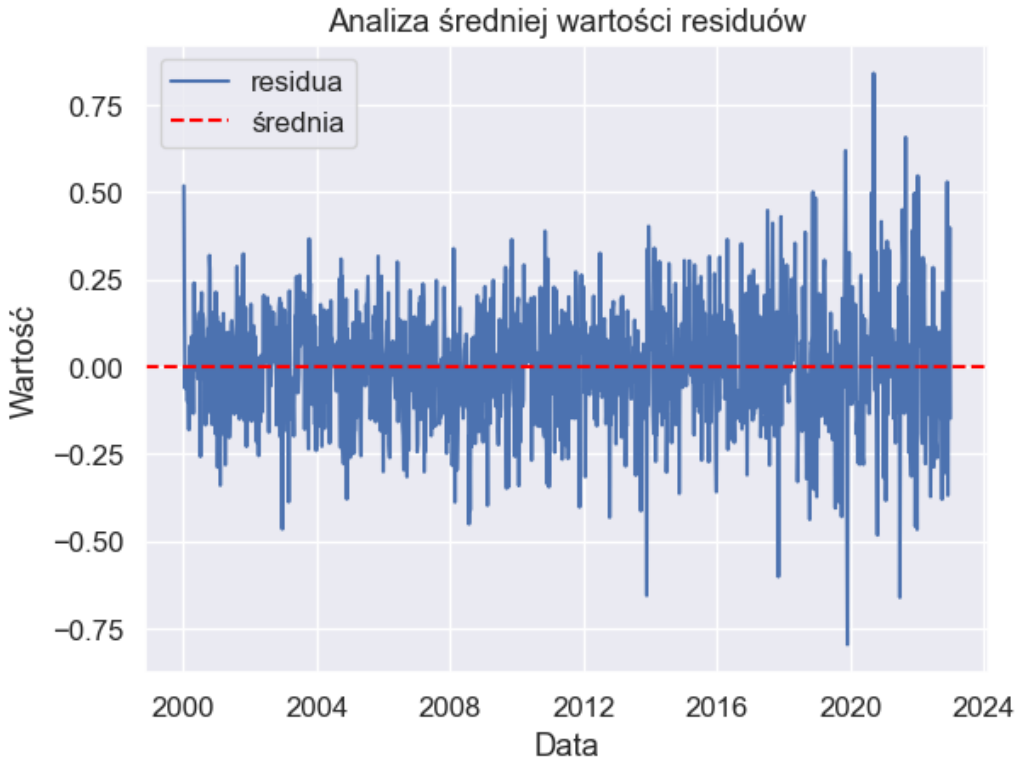
Pierwsze założenie dotyczące średniej mówi, że residua modelu powinny mieć średnią 0.

Jak widzimy, linia wyznaczająca średnią jest bardzo bliska zera, a residua oscylują równomiernie w jej okolicach. Po policzeniu średnia wynosi dokładnie 0.00123, z dokładnością do trzech cyfr znaczących. Jest to wartość tak bliska zeru, że możemy być niemal pewni, że residua pochodzą z rozkładu o zakładanej średniej. Dla pewności sprawdzimy jednak to jeszcze przy pomocy testu t Studenta dla jednej średniej, z biblioteki *statsmodels*.

Wspomniany test t Studenta jako hipotezę zerową przyjmuje, że średnia jest równa μ_0, a jako alternatywną, że nie jest równa tyle. Statystyka testowa T jest dana wzorem

$$T = \frac{\bar{X} - \mu_0}{S}\sqrt{n} \tag{4}$$

gdzie $barX$ to średnia badanych danych, μ_0 to domniemana średnia (w naszym wypadku 0), n to liczba danych. Ostatecznie jej wartość jest dla naszych danych równa około 0.245, czyli blisko zera, co oznacza, że średnie są blisko siebie (im bliżej zera tym lepiej świadczy to hipotezie zerowej), przy czym nasza (jak już zauważyliśmy jest lekko większa). Natomiast p-wartość jest równa 0.806, czyli na ogólnie przyjętym poziomie istotności $\alpha = 0.05$ nie ma podstaw by odrzucić hipotezę zerową.



Rysunek 13: Trajektoria residuów z ich średnią.

Patrząc na wszystkie wykorzystane przed chwilą metody, jesteśmy w stanie stwierdzić, że analizowane residua rzeczywiście pochodzą z rozkładu o średniej 0.

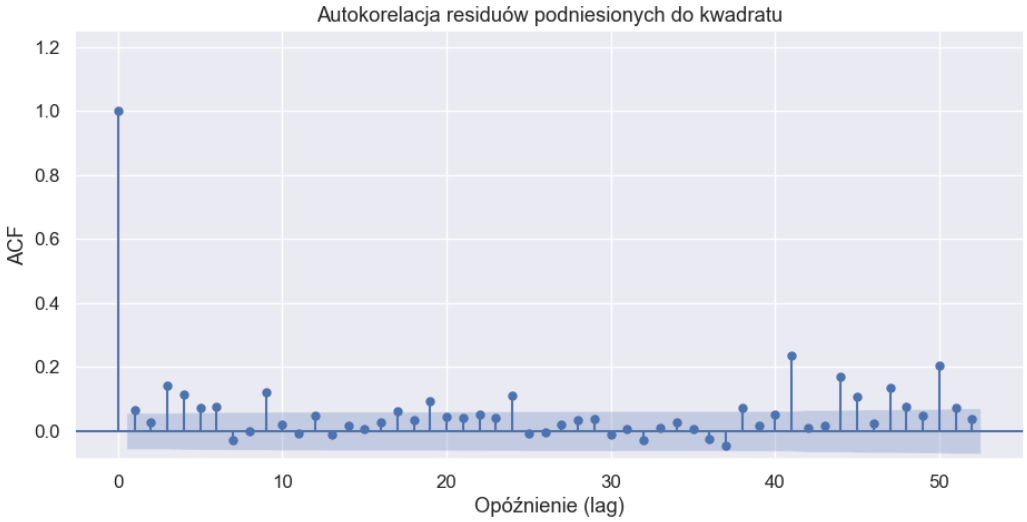
5.2 Założenie dotyczące wariancji

Kolejnym założeniem jest, że szum analizowanego modelu ARMA(4,1) jest z rozkładu o stałej wariancji. Podejdziemy do tego tematu podobnie jak w poprzednim podpunkcie i zaczniemy ponownie od analizy wykresu residuów.

Patrząc na Rysunek 13, możemy zauważyć, że dane z późniejszych lat zdają się mieć większą tendencję do osiągania wartości skrajnych, odbiegających od średniej o nawet 0.5 w kwestii wartości bezwzględnej. Nie wystarczy jednak to, żeby stwierdzić o zmienności wariancji rozkładu, z którego pochodzą dane. W celu dalszej analizy przeprowadzimy dwa testy statystyczne.

Pierwszym z nich będzie ARCH test, czyli Autoregressive Conditional Heteroskedasticity test. Jest on stworzony do testowania właśnie zmienności wariancji w czasie w przypadku reszt szeregów czasowych. Opiera się on na kwadratach residuów i ich zależności między sobą. Jako hipotezę zerową przyjmuje stałość wariancji, a jako alternatywną jej brak. Dla naszego szumu wartość statystyki w przypadku tego testu wyniosła aż 66.6, a p-wartość jest niemalże równa 0 (rzędu 10^{-10}). Statystyka, więc bardzo wyraźnie odbiega od 0, a p-wartość prowadzi do odrzucenia hipotezy zerowej.

Drugi test to zmodyfikowany test Levene'a (Modified Levene Test) porównujący wariancję między dwoma grupami danych. Podejmujemy się zmodyfikowanego testu, ponieważ jest on bardziej odporny na dane niebędące z rozkładu normalnego (mimo że i tak jest to jednym z naszych założeń).W jego ramach podzielimy nasze dane na pół — mamy ich 1200, pierwsze 600 będzie stanowiło pierwszą grupę, reszta — drugą. Jako hipotezę zerową mamy tutaj przypuszczenie, że obie grupy mają taką samą wariancję, jako alternatywną, że wariancje się różnią. Obliczając statystykę testową, porównuje się tutaj odchylenia pojedynczych wartości od średnich wewnątrz podanych grup, a także biorąc pod uwagę wszystkie dane. Ostatecznie wynosi ona tutaj 19.9. Natomiast, p-wartość ponownie jest bliska zeru i przyjmuje wartość rzędu 10^{-6}. Na podstawie tych dwóch faktów odrzucamy po raz kolejny hipotezę zerową i coraz bardziej skłaniamy się do wniosku na temat braku spełnienia założenia dotyczącego wariancji.



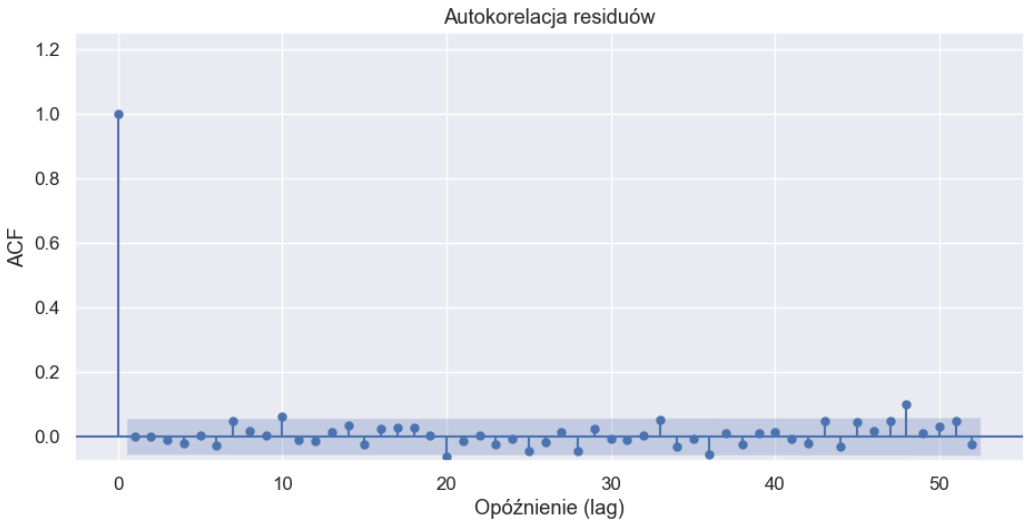
Rysunek 14: Wykres autokorelacji residuów podniesionych do kwadratu.

W kwestii stałej wariancji bada się też, wykres autokorelacji residuów (Rysunek 14). Wartości wykraczają na nim dość często poza przedziały ufności na poziomie $\alpha = 0.1$, z całą pewnością częściej niż teoretycznie powinny. Jest to kolejne spostrzeżenie sugerujące, że rozkład, z którego pochodzą nasze dane, nie ma stałej wariancji.

Patrząc na wszystko, co uwzględniliśmy w tej części, jesteśmy w stanie stwierdzić, że założenie dotyczące stałej wariancji szumu nie jest w tym wypadku spełnione.

5.3 Założenie dotyczące niezależności

Następne założenie dotyczy niezależności residuów. Po raz kolejny nasze podejście będzie podobne jak w poprzednich częściach.



Rysunek 15: Wykres autokorelacji residuów.

Patrząc na wykres, autokorelacji wartości resztowych (Rysunek 15) nie widzimy wyraźnych zależności dla żadnych opóźnień, a dodatkowo same wartości autokorelacji w wyraźnej większości zawierają się w przedziałach ufności. Dane nie powinny być na tej podstawie zależne.

Zweryfikujemy ten fakt jeszcze testem Ljunga-Boxa dla różnych wartości opóźnień. Test ten jako hipotezę zerową przyjmuje brak autokorelacji w danym szeregu czasowym, a jako alternatywę jej występowanie. Opiera się na współczynniku korelacji i rozkładzie chi-kwadrat, w przypadku liczenia statystyki testowej.

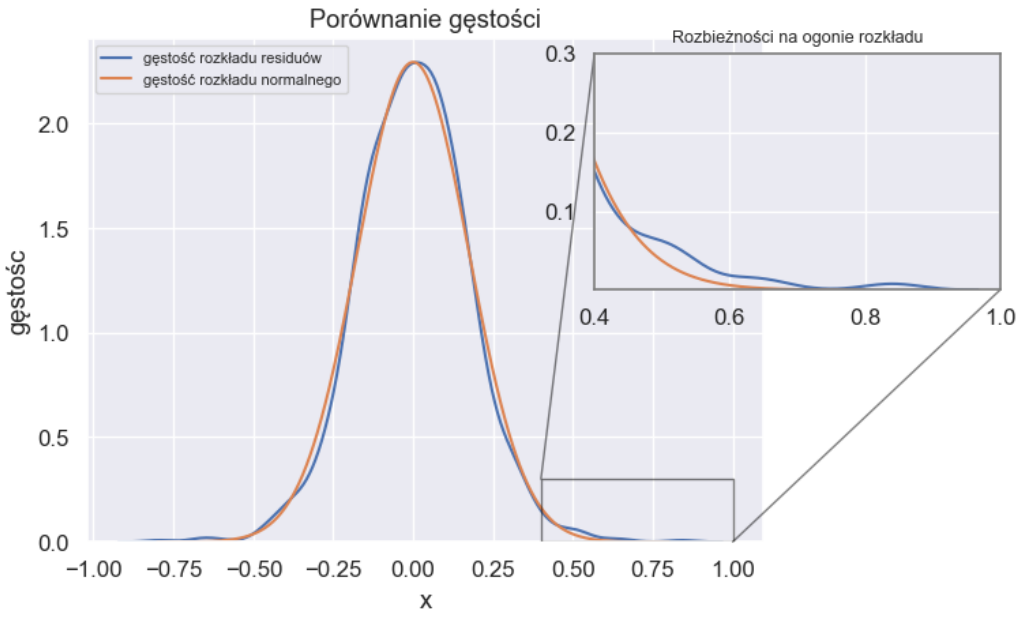
Rysunek 16: P-wartość testu Ljunga-Boxa w zależności od opóźnienia.

Jak widzimy na Rysunku 16, dla żadnej z wartości opóźnień p-wartość naszego testu nie spada poniżej wyjątkowo założonego tutaj poziomu istotności $\alpha = 0.15$. Oznacza to, że dla wszystkich wartości opóźnień nie znajdujemy podstaw by odrzucić hipotezę zerową o braku autokorelacji. Jest to kolejny fakt wskazujący na niezależność residuów.

Biorąc pod uwagę cały powyższy podrozdział, jesteśmy w stanie stwierdzić, że dla naszego szumu założenie o niezależności zostało spełnione.

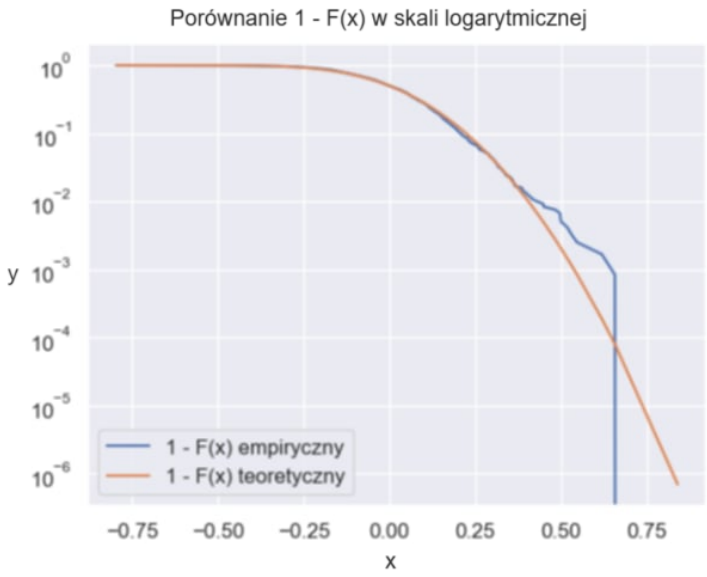
5.4 Założenie dotyczące normalności rozkładu

Sprawdzamy, czy residua zachowują się zgodnie z rozkładem gaussowskim, poprzez porównanie gęstości rozkładów (Rysunek 17).



Rysunek 17: Rozkład normalny oraz rozkład wartości residuów.

Na Rysunku 17 szczególną uwagę zwracają różnice na ogonach rozkładu. Obserwujemy je również na Rysunku 18 przy wartościach $1 - F(x)$ gdzie $F(X)$ to dystrybuanta rozkładu. Wykres wykonany został w skali logarytmicznej.

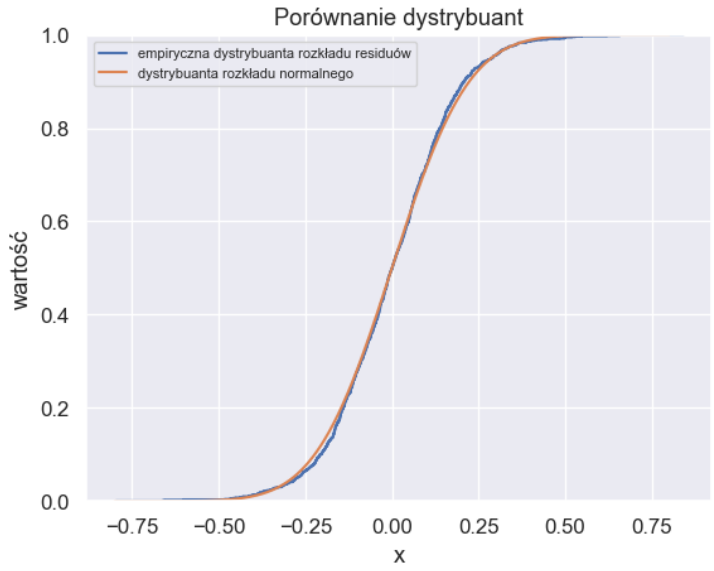


Rysunek 18: Rozbieżności na ogonach w skali logarytmicznej.

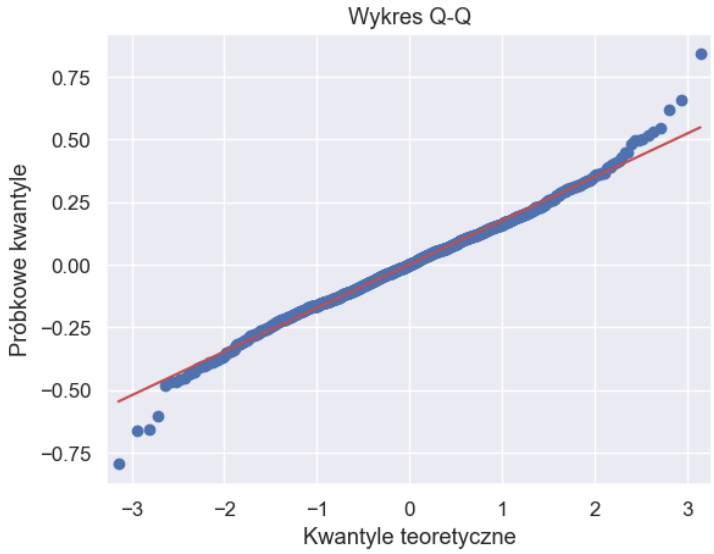
Dystrybuanty obu rozkładów poza ogonami są do siebie bardzo zbliżone, co widzimy na Rysunku 19.

Na wykresie kwantyl–kwantyl (Rysunek 20) również obserwujemy znaczące różnice na ogonach rozkładu.

Patrząc na parametry naszego szumu, widzimy, że niezupełnie jest on z rozkładu normalnego. Mimo że jego skośność jest bliska 0 (dokładnie 0.0275), to kurtoza nadmiarowa odbiega wyraźnie od wartości zerowej i wynosi 1.212. Potwierdza to też test Jarque-Bera oparty na tych dwóch wartościach. Po jego przeprowadzeniu otrzymaliśmy statystykę testową na poziomie 72.3 i p-wartość rzędu 10^{-16}. Oznacza to, że na poziomie istotności takim jak w przypadku innych testów tutaj też odrzucamy hipotezę zerową o normalności tego rozkładu.



Rysunek 19: Dystrybuanty rozkładów.



Rysunek 20: Wykres kwantyl–kwantyl (QQ plot).

6 Podsumowanie i wnioski

W raporcie przedstawiliśmy analizę dostępnych publicznie danych dotyczących emisji tlenku węgla za pomocą modelu ARMA. Po odpowiedniej transformacji obserwowanych wartości, na podstawie znanych kryteriów informacyjnych mogliśmy dobrać rząd i odpowiednie parametry modelu, które z dużą dokładnością odzwierciedlały zachowanie się danych. Przeanalizowaliśmy również residua szeregu, których rozkład okazał się bliski normalnemu, o średniej 0 i niezależnych kolejnych wartościach. Szum nie spełnił jednak założenia o stałej wariancji. Może być to powodem, przez który utworzony model predykcyjny, dotyczący wartości emisji CO w Stanach Zjednoczonych w latach przyszłych nie wydaje się być do końca poprawny. Bardziej poprawnym podejściem byłoby z pewnością użycie modelu uwzględniającego zmienność wariancji, np. GARCH (powiązany ze wspominanym już ARCH testem). Mimo wszystko model ARMA może służyć w uproszczony sposób do analizy naszych danych. Dalsze badania mogłyby obejmować znalezienie danych z lat 2023-2024 i porównanie rzeczywistych wartości wraz z przyjętym przez nas modelem.

Bibliografia i licencje

[0] Brina Blum. *Zdjęcie ze strony tytułowej: architecture building Free Stock Image*. URL: https://stocksnap.io/photo/architecture-building-LGRZBOM0HK. Licencja: CC0.

[1] GusLovesMath. *U.S. Pollution Data 2000 - 2023*. URL: https://www.kaggle.com/datasets/guslovesmath/us-pollution-data-200-to-2022. Licencja: U.S. Government Works.

[2] *Transformacja Boxa–Coxa*. URL: http://manuals.pqstat.pl/statpqpl:usepl:arkpl:normstandpl#fnt__1.