Analiza danych meteorologicznych za pomocą modelu ARMA

Wojciech Haładewicz, Rafał Głodek

Styczeń 2025

1 Wstęp

1.1 Wprowadzenie do tematu pracy

Szeregiem czasowym nazywany ciąg obserwacji, dotyczących pewnego zjawiska lub cechy, zebranych w różnych momentach czasu. W ujęciu matematycznym, interpretujemy to pojęcie jako realizacja pewnego procesu stochastycznego, którego dziedziną jest czas. Przy pomocy odpowiednich narzędzi matematycznych i informatycznych możemy przeprowadzić analizę danego szeregu, która pozwoli nam nie tylko zagłębić się w naturę zjawiska opisanego przez zebrane obserwacje, ale także prognozować jego przyszłe stany. Dobre prognozowanie jest następstwem zastosowania odpowiedniego modelu, który dobrze dopasuje się do zebranych danych. Poprawna analiza szeregu czasowego rozpoczyna się od wykonania potrzebnych pomiarów oraz zbadania struktury posiadanych danych. Dopiero właściwa realizacja tych kroków pozwoli nam przejść do etapu modelowania.

W naszej pracy wykorzystamy popularny model autoregresji i średniej ruchomej - ARMA - do przeprowadzenia analizy danych meteorologicznych, opisujących średnie miesięczne wartości temperatury na powierzchni Ziemi. Przejdziemy przez wszystkie opisane wyżej kroki, by w pełni poznać naturę tego zjawiska. Wykryjemy obecne w nim tendencje rozwojowe (trendy) oraz wahania sezonowe, a następnie dopasujemy do danych właściwy model, który pozwoli nam nakreślić możliwy scenariusz kształtowania się temperatur w przyszłości. Jest to temat szczególnie ważny w obecnych czasach, gdy świadomość społeczna

o stale postępujacych zmianach klimatycznych rośnie, a w przestrzeni publicznej toczy się głośna dyskusja o wpływie, jaki działania człowieka wywierają na globalne ocieplenie klimatu.

1.2 Informacje o analizowanych danych

Tak jak już wcześniej wspomnieliśmy, dane, których analizą będziemy się zajmować, określają zależność średniej miesięcznej temperatury na powierzchni Ziemi od czasu. Pochodzą one z bazy danych *Climate Change: Earth Surface Temperature Data*, którą znaleźliśmy na serwisie *Kaggle.com*. Cała baza zawiera także informacje m.in. o średniej temperaturze na powierzchni Ziemi z podziałem na państwa i duże miasta, wartościach minimalnych i maksymalnych temperatury w danym miesiącu, a także średnie miesięczne tempertury dla lądów i oceanów. Wyodrębnione przez nas dane zawierają 3159 niepustych rekordów podzielonych na dwie kolumny: *dt* (data w formacie "YYYY-MM") oraz *LandAverageTemperature* (średnia temperatura na lądzie).

Kolumna zawierająca daty rozpoczyna się od wartości 1752-12, a kończy na 2015-12. Za pomocą samodzielnie zaimplementowanej funkcji sprawdziliśmy ciągłość danych w kolumnie *dt* - czy wektor dat zawiera wszystkie miesiące po kolei. Wykazała ona, że rozważana baza zawiera wszystkie daty począwszy od października roku 1752, aż do grudnia 2015 roku.

Informacji o danych zawartych w kolumnie *LandAverageTemperature* dostarczyła nam metoda `describe`. Wyliczone przez nią statystyki opisowe zawierają się w poniższej tabeli

Statystyka	Wartość
Średnia próbkowa	8.378
Ochylenie standardowe	4.379
Minimum	-2.080
Maksimum	19.021
Pierwszy kwartyl	4.319
Mediana	8.618
Trzeci kwartyl	12.549

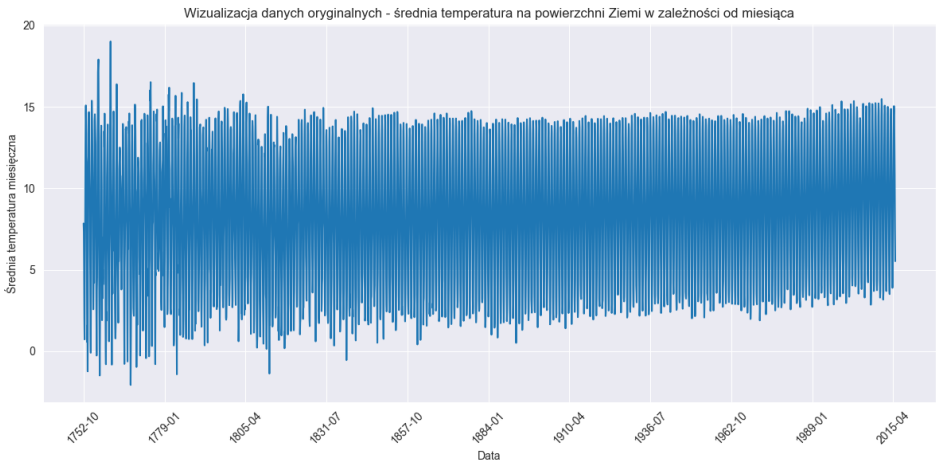
Tabela 1: Statystyki opisowe rozważanych obserwacji

Posiadamy zatem wektor 3159 obserwacji rozłożonych na osi czasu w równych odstępach. Stanowią one oryginalny szereg czasowy, który w dalszych częściach

pracy będziemy transformować i analizować.

1.3 Wizualizacja danych

Zwizualizujemy teraz nasz szereg na wykresie i zastanowimy się, ile możemy powiedzieć na temat naszych danych przedstawionych w surowej formie.



Rysunek 1: Wizualizacja oryginalnego szeregu

Na pierwszy rzut oka wydaje się, że powyższy wykres nie dostarcza wielu informacji o analizowanych danych. Można jednak zauważyć, że zawiera on pewne powtarzające się okresowo wzorce. Ten efekt nazywany jest sezonowością. Można także dopatrzeć się subtelnej tendencji wzrostowej w danych, choć nie jest to obserwacja oczywista. To zjawisko określamy mianem trendu. Trend i sezonowość dokładniej przeanalizujemy w kolejnych krokach analizy. W przeważającej większości dane mają stosunkowo stabilny rozrzut w czasie, co sugeruje, że amplituda zmian temperatury w skali miesięcznej nie podlega drastycznym wahaniom. Rozrzut jest zdecydowanie większy w początkowej części wykresu, gdzie zauważyć można najwięcej obserwacji odstających. Wyraźnie widoczne są także punkty maksimum i minimum. Skrajne wartości mogą wskazywać na występowanie anomalii pogodowych, lecz bardziej uzasadnioną interpretacją wydaje się być błąd pomiarowy. Sprzęty używane przez XVIII-wiecznych meteorologów były stosunkowo prymitywne i charakteryzowały się znacznie mniejszą dokładnością niż te używane współcześnie. Większe ryzyko wystąpienia błędu pomiarowego jest zatem dobrze uzasadnione w tym kontekście.

2 Analiza danych

Jak już wcześniej wspomnieliśmy, nasze dane zamodelujemy za pomocą modelu ARMA. Ta część pracy poświęcona jest dokonywaniu odpowiednich transformacji danych, celem przygotowania ich do procesu modelowania. Zanim wykonamy dalsze kroki, przypomnjmy, jak zdefiniowany jest model, który będziemy wykorzystywać.

Definicja 1. *Proces stochastyczny* $\{X_t\}_{t\in\mathbb{N}}$ *jest modelem* $ARMA(p,q)$ *jeśli* $\{X_t\}$ *jest stacjonarny i dla każdego* t *spełnia następujące równanie*

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

gdzie $\{Z_t\} \sim WN(0,\sigma^2)$ *jest białym szumem, a* $\phi_{(.)}$ *i* $\theta_{(.)}$ *parametrami modelu.*

Dodatkowo, do zastosowania modelu ARMA, wymagane jest spełnienie następujących założeń:

- Rozważany szereg czasowy jest stacjonarny - jego średnia, wariancja i autokorelacja są stałe w czasie
- Residua pochodzą z białego szumu - są nieskorelowane, mają stałą średnią i wariancję

Teraz wykonamy na naszym szeregu przekształcenia, które zapewnią spełnienie powyższych założeń, co pozwoli nam przejść do procesu modelowania

2.1 Dekompozycja szeregu czasowego

Wizualizacja danych oraz znajomość kontekstu (pochodzenia danych) pozwala nam spodziewać się, że na rozważany przez nasz szereg czasowy oprócz procesu losowego, składają się także pewne składowe deterministyczne - trend i sezonowość. Możemy zatem zapisać go w postaci

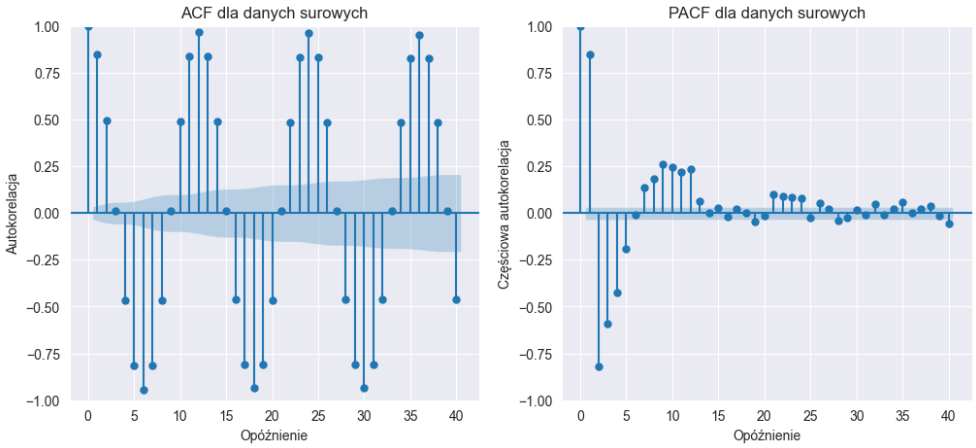
$$X_t = m_t + s_t + Y_t$$

gdzie

- X_t to surowe dane,
- m_t to wolno zmienna funkcja (trend),
- s_t to periodycznie zmienna funkcja (sezonowość),
- Y_t to proces losowy (szum).

Proces dekompozycji polega na usunięciu z szeregu składowych deterministycznych m_t i s_t, celem otrzymania stacjonarnego procesu losowego Y_t.

Najpierw wygenerujemy wykresy funkcji autkorelacji (ACF) i cząstkowej autokorelacji (PACF) dla surowego szeregu.



Rysunek 2: Wykresy ACF i PACF

Wykres funkcji autokorelacji pokazuje wyraźne skoki w regularnych odstępach czasu, co sugeruje, że szereg posiada znaczący komponent sezonowy. Z tego wykresu możemy także odczytać okres funkcji sezonowej, który wynosi około 12. Co oczywiście ma swoje uzasadnienie w kontekście tematyki naszej pracy, ponieważ rokrocznie spodziewamy się podobnych wartości pomiarów temperatury. Wartości funkcji autokorelacji istotnie odbiegają od zera i nie wykazują znaczącej tendencji malejącej wraz ze wzrostem opóźnienia, co sugeruje występowanie silnego, długoterminowego trendu w danych.

Proces dekompozycji rozpoczniemy od wykonania rozszerzonego testu Dickeya-Fullera (Augmented Dickey-Fuller Test) na stacjonarność szeregu czasowego. Szereg czasowy nazywamy stacjonarnym, gdy ma niezależne od czasu: wartość

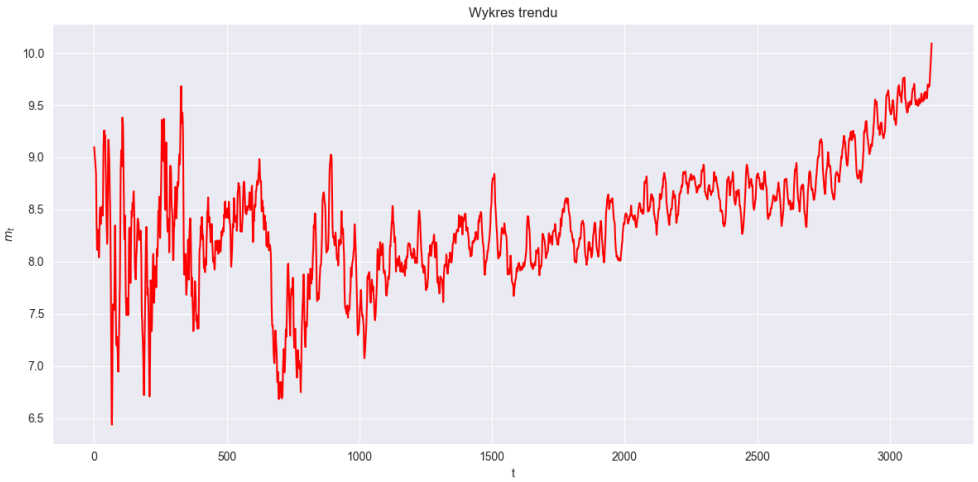
średnią, wariancję i autokorelację. Test ADF sprawdza, czy model autoregresyjny posiada pierwiastek jednostkowy, którego obecność świadczy o niestacjonarności szeregu. Nie jest jednak wyczulony na inne formy niestacjonarności. Z tego powodu, jeśli szereg czasowy zawiera silne komponenty sezonowe, ale jest stacjonarny w sensie średniej i wariancji w krótkim okresie, wynik testu może wskazywać na stacjonarność, mimo obecności sezonowości. Wyniki tego testu dla surowych danych prezentują się następująco:

- Wartość statystyki testowej: $\tau = -3.653$
- p-wartość: 0.005
- Wynik: Odrzucamy hipotezę o niestacjonarności szeregu.

Test ADF wykazał zatem, że rozważany szereg czasowy jest stacjonarny, pomimo że spodziewamy się obecności sezonowości i pewnego wpływu długoterminowego trendu na nasze dane. Z tego powodu konieczne jest przeprowadzenie dalszych etapów dekompozycji, by mieć pewność, że pozbędziemy się tych składowych.

Właściwy proces dekompozycji wykonujemy za pomocą metody `seasonal_decompose` z pakietu `statsmodels.tsa.seasonal`. Wyodrębnione w ten sposób składowe zwizualizujemy na wykresach.

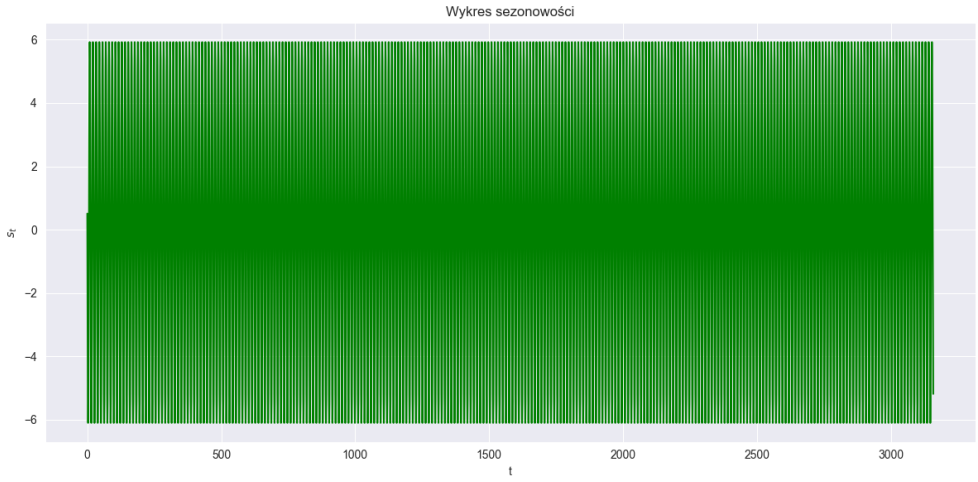
Zależność składowej trendu m_t od czasu prezentuje się następująco:



Rysunek 3: Wykres trendu

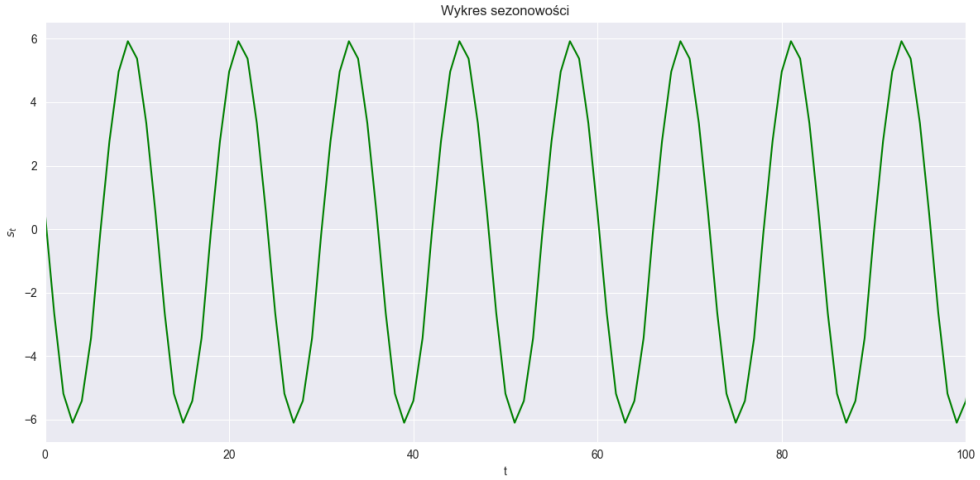
Na przedstawionym wykresie trendu widzimy ogólną tendencję wzrostową w badanym szeregu czasowym. Na początku trend charakteryzuje się dużą zmiennością, która stopniowo staje się coraz bardziej uporządkowana. Widoczne na wykresie spadki wskazują, że chociaż trend ogólnie ma charakter rosnący, to występują okresowe fluktuacje wynikające z czynników sezonowych. Zależność ta jest bardzo istotna w kontekście tematyki naszej pracy. Można bowiem swtierdzić, że na przestrzeni ostatnich trzystu lat występuje tendencja wzrostowa w średniej miesięcznej temperaturze na powierzchni Ziemi. Efekt ten nazywany jest globalnym ociepleniem klimatu. Wpływ człowieka na to zjawisko oraz dalszy kierunek jego rozwoju wciąż stanowi kwestię dyskusyjną w świecie naukowym, niemniej jego istnienie jest niepodważalne.

Składowa sezonowa s_t zmienia się w czasie w następujący sposób:



Rysunek 4: Wykres sezonowości dla pełnego zakresu czasu

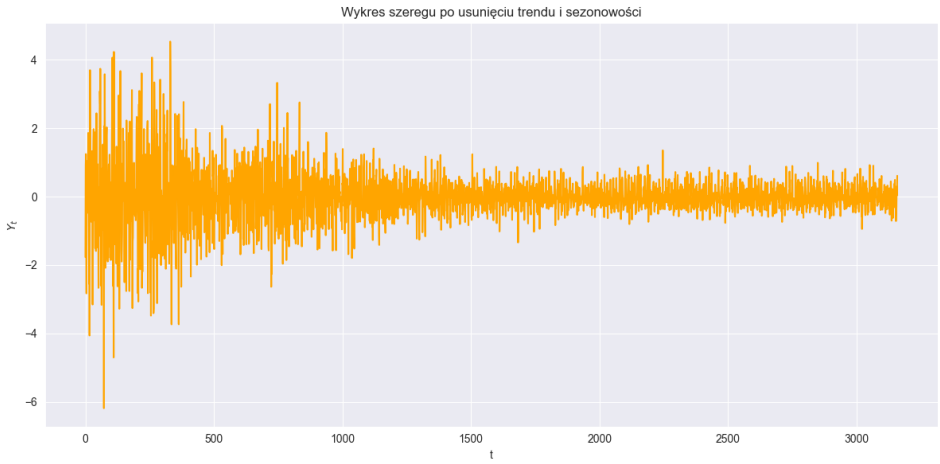
Wykres sezonowości przy pełnym horyzoncie czasowym nie mówi nam zbyt dużo, ponieważ jest bardzo gęsty i nakładające się na siebie linie utrudniają wykrycie szczegółów. Dobrze widoczna jest jednak stała amplituda zmian - punkty oscylują w przedziale $(-6, 6)$. Celem lepszej analizy skróciliśmy horyzont czasowy wykresu, by częstotliwość zmian była lepiej widoczna.



Rysunek 5: Wykres sezonowości dla skróconego zakresu czasu

Skrócenie horyzontu czasowego pozwala nam dokładniej przyjrzeć się oscylacjom temperatury. Powyższy wykres ujawnia wyraźny wzorzec sinusoidalny o okresie 12 miesięcy. Ma to oczywiście związek z cyklicznym występowaniem pór roku - spadki obserwowane są podczas zimy, a szczyty podczas lata.

Usunięcie składowych deterministycznych m_t i s_t z oryginalnego sygnału miało skutkować otrzymaniem stacjonarnego szeregu czasowego Y_t. Sprawdźmy jak prezentują się pozostałe dane, pozbawione efektu trendu i sezonowości.

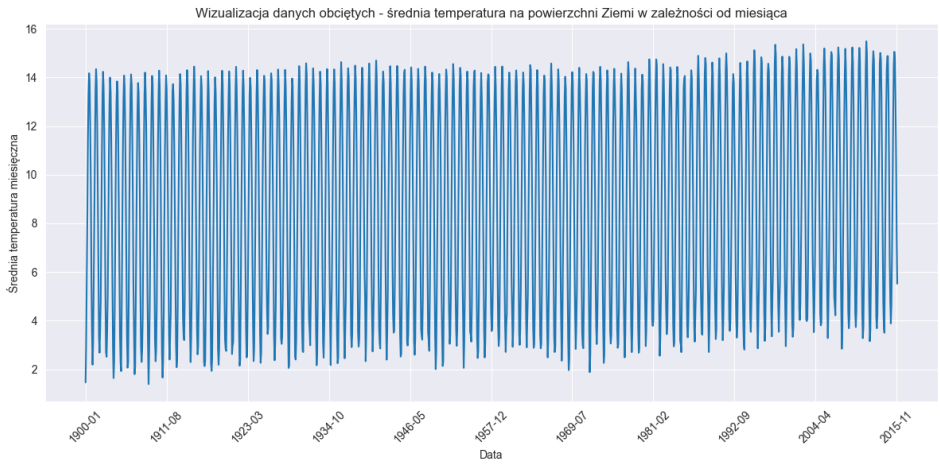


Rysunek 6: Szereg po dekompozycji

Po dekompozycji szereg zdaje się mieć charakter stacjonarny, gdyż jest pozbawiony wyraźnego trendu oraz sezonowych fluktuacji. Oscylacje wydają się być bardziej losowe i mają średnią w okolicach zera. Wariancja residuów jest znacznie większa na początku wykresu, niż w dalszej jego części. Wskazuje to na mocną zmienność wariancji w czasie, czyli heteroskedastyczność. Wykryliśmy zatem naruszenie jednego z założeń, o których wspomnieliśmy wyżej.

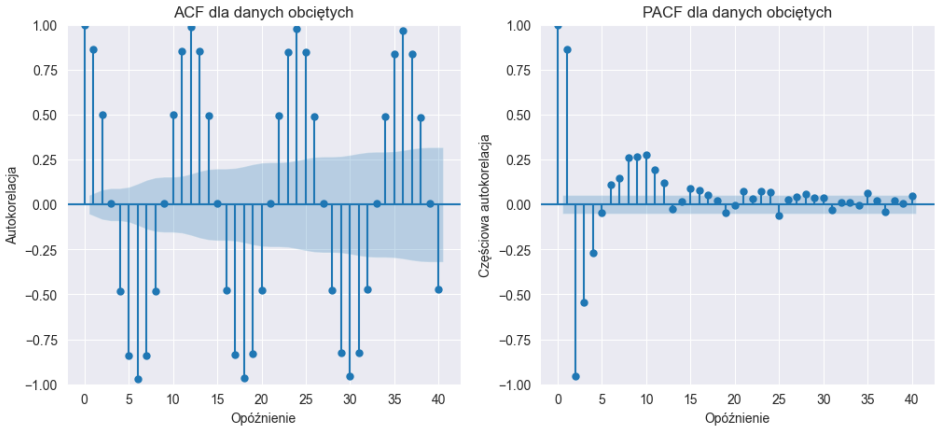
Pomimo naruszonego założenia o homoskedastyczności, zdecydowaliśmy się nie porzucać naszej analizy, a jedynie zmienić lekko rozważany problem. Spójrzmy ponownie na wykres szeregu po dekompozycji i zwróćmy uwagę, że wariancja po czasie wydaje się stabilizować. Dobrym rozwiązaniem może okazać się obcięcie zakresu danych. Rozważmy zatem nowy szereg - opisujący średnią miesięczną temperaturę na powierzchni Ziemi, począwszy od stycznia 1900 roku (obserwacja nr 1767), aż do grudnia 2015 roku. Zmienia się więc okres czasu, na którym będziemy operować, ale sam przedmiot i cel analizy pozostaje ten sam - za pomocą trendu i sezonowości opiszemy proces zmian średnich temperatur na naszej planecie i nakreślimy dalszy kierunek tego zjawiska, uwzględniając efekt globalnego ocieplenia.

Spójrzmy na wizualizację nowego szeregu - obciętych danych surowych.



Rysunek 7: Wizualizacja szeregu obciętego

Następnie wygenerujemy dla niego wykresy funkcji autokorelacji i częściowej autokorelacji.



Rysunek 8: Wykresy ACF i PACF dla szeregu obciętego

Funkcja autokorelacji nie zmieniła swojej postaci w porównaniu do rozważanego wcześniej szeregu. Częściowa autokorelacja znacznie odbiega od zera dla wielu opóźnień, co może sugerować niestacjonarność szeregu.

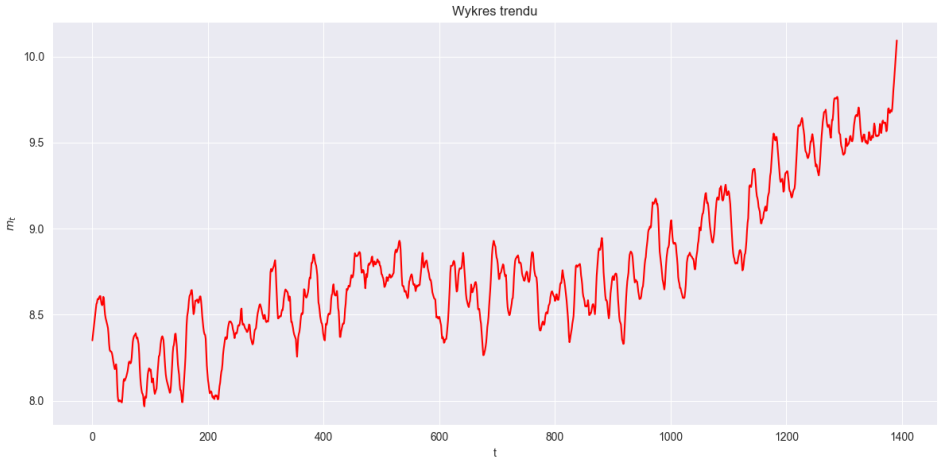
Na nowym szeregu, podobnie jak wcześniej, wykonamy test stacjonarności ADF. Wyniki prezentują się w następujący sposób:

- Wartość statystyki testowej: $\tau = -0.691$
- p-wartość: 0.849
- Wynik: Nie ma podstaw do odrzucenia hiptozey zerowej. Zakładamy, że szereg jest niestacjonarny.

Wynik rozszerzonego testu Dickeya-Fullera wykazał, że szereg po obcięciu danych charakteryzuje się niestacjonarnością. Rezultat ten pozornie może wydawać się sprzeczny z poprzednimi rozważaniami -test ADF był bowiem niewrażliwy na niestacjonarność szeregu oryginalnego. Za najbardziej pradopodobną przyczynę występowania tego zjawiska można przyjąć ustabilizowanie wariancji. Mniejszy rozrzut danych sprawił, że występujący w danych trend stał się lepiej wykrywalny dla testu, dzięki czemu wskazał on na niestacjonarność nowego szeregu.

Żeby móc dopasować odpowiedni model ARMA musimy doprowadzić nasz szereg do postaci stacjonarnej. W tym celu wykonamy wcześniej omawiane już kroki dekompozycji.

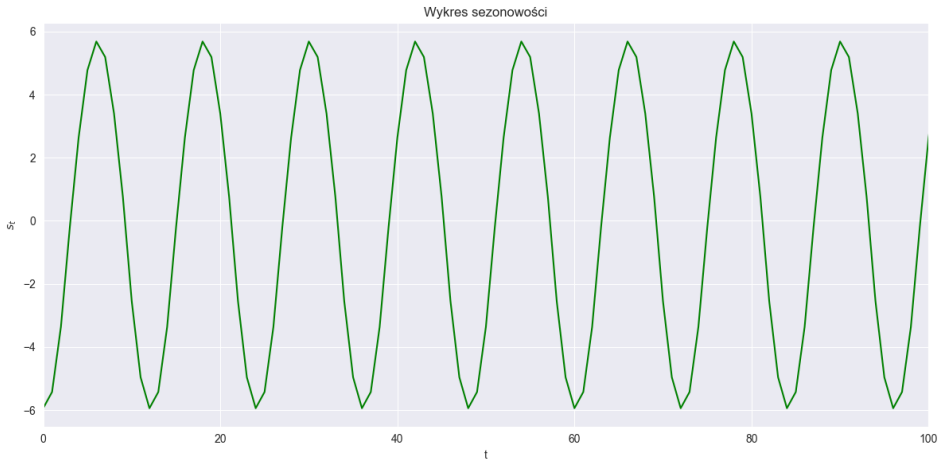
Wykryta składowa trendu m_t zmienia się w czasie w następujący sposób



Rysunek 9: Wykres trendu

Powyższy wykres dobrze pokazuje ogólny trend rosnący w rozważanym szeregu. Nowy trend jest pozbawiony początkowych nieregularnych fluktuacji, które występowały w szeregu uwzględniającym wszystkie dane, a które wynikały najprawdopodobniej z błędów pomiarowych. Wzrostowy charakter trendu potwierdza wysunięte wcześniej wnioski o stopniowym ocieplaniu klimatu Ziemi.

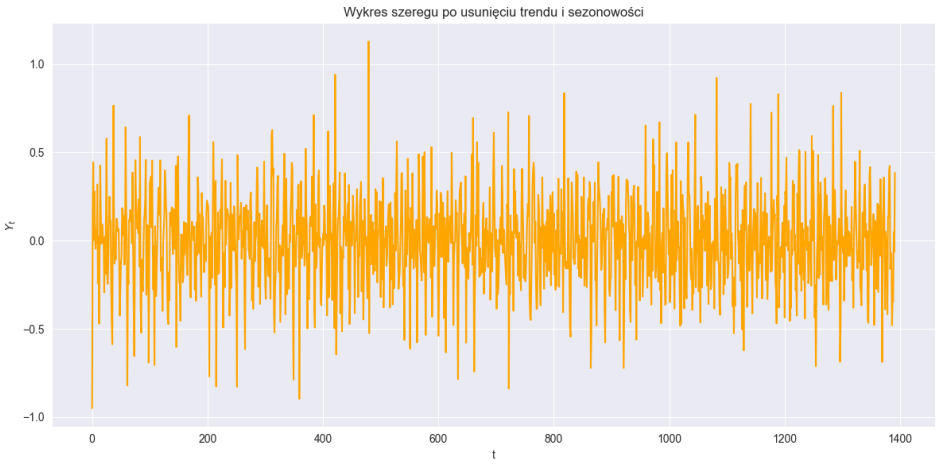
Przyjrzyjmy się teraz wykresowi składowej sezonowej s_t, przy skróconym horyzoncie czasowym.



Rysunek 10: Wykres sezonowości

Zgodnie z naszymi oczekiwaniami, wykres składowej sezonowej nowego szeregu różni się od analogicznego wykresu dla oryginalnego szeregu jedynie przesunięciem w fazie. Wcześniejsza analiza wykazała, że składowa s_t wykazuje cykliczne oscylacje o okresie 12 miesięcy i stałej w czasie amplitudzie. Obcięcie zakresu danych zmienia jedynie przesunięcie funkcji, ale jej ogólny charakter pozostaje taki sam.

Pozbyliśmy się składowych deterministycznych z nowego szeregu i tym razem spodziewamy się otrzymać stacjonarny szereg czasowy Y_t, który będzie spełniał założenia modelu ARMA. Wizualizacja danych "oczyszczonych" znajduje się poniżej.



Rysunek 11: Wykres szeregu po usunięciu trendu i sezonowości

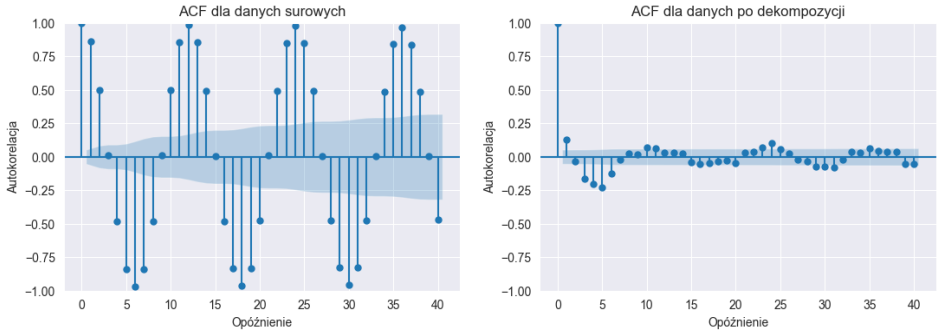
Szereg oczyszczony ze składowych deterministycznych wygląda w tym przypadku dużo lepiej niż w jego początkowej formie. Jego struktura zdecydowanie bardziej przypomina realizację stacjonarnego procesu losowego, niż to miało miejsce wcześniej. Na wykresie nie znajdujemy żadnych cyklicznych wzorców, a zatem możemy uznać, że składowa sezonowa została w większości skutecznie usunięta. Ponadto szereg wykazuje oscylacje wokół zera o względnie stałej amplitudzie, co nie tylko udowadnia pozbycie się składowej trendu, ale także sugeruje, że mamy do czynienia z szeregiem homoskedastycznym.

Żeby ostatecznie ocenić wpływ składowych deterministycznych na nasz szereg, policzyliśmy średni błąd bezwzględny między wartościami surowymi a otrzymanymi w wyniku dekompozycji szeregu. Wyniósł on około

$$MAE \approx 8.763$$

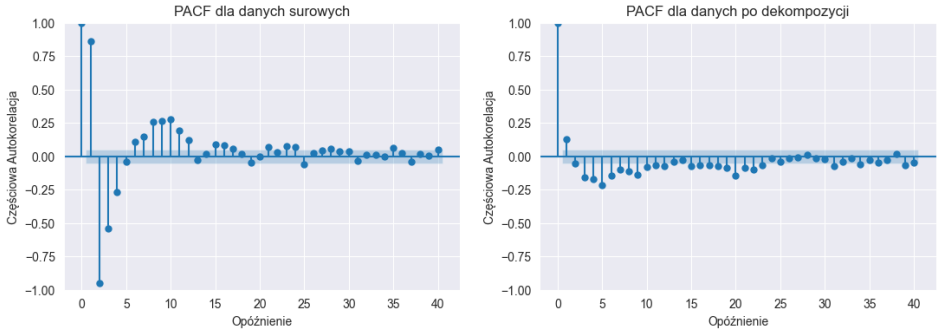
Zwróćmy uwagę, że wyżej wyliczony średni błąd bezwzględny stanowi bardzo dużą część całego zakresu analizowanych przez nas danych. Można zatem stwierdzić, że wpływ składowych deterministycznych na rozważany szereg był znaczący.

Dla otrzymanego w wyniku dekompozycji szeregu wygenerowaliśmy również wykresy funkcji autokorelacji i częściowej autokorelacji. Porównaliśmy je z analogicznymi wykresami wygenerowanymi wcześniej.



Rysunek 12: Porównanie ACF dla danych sprzed i po dekompozycji

Po dekompozycji widoczny jest znaczny spadek autokowariancji już dla pierwszych wartości opóźnień. Dalej funkcja zgodnie z oczekiwaniami zaczyna oscylować wokół zera. Nawet jeśli drobne wzorce sezonowości dalej są widoczne, to zdecydowana większość tej składowej została usunięta, a wartości ACF w dużej części zawierają się w przedziale ufności.



Rysunek 13: Porównanie PACF dla danych sprzed i po dekompozycji

Wartości częściowej autokorelacji szeregu "oczyszczonego" gwałtownie spadają już po pierwszym opóźnieniu. Wartości PACF odpowiadające kolejnym lagom stopniowo zbliżają się do zera i większości wewnątrz przedziału ufności. Oba wykresy wskazują, że po dekompozycji dane stały się słabiej skorelowane, dzięki czemu będziemy w stanie otrzymać bardziej wiarygodny model ARMA.

Dekompozycja szeregu miała doprowadzić nas do postaci stacjonarnej, którą następnie będziemy mogli łatwo analizować oraz modelować. Sprawdzimy, czy otrzymany szereg faktycznie zachowuje tę własność, ponownie z wykorzystaniem

rozszerzone testu Dickeya-Fullera. Wyniki testu prezentują się następująco:

- Wartość statystyki testowej: $\tau = -16.207$
- p-wartość: 0.0
- Wynik: Odrzucamy hipotezę o niestacjonarności szeregu. Zakładamy, że szereg jest stacjonarny.

Zgodnie z naszymi oczekiwaniami, szereg po dekompozycji zachowuje własność stacjonarności. Możemy zatem uznać, że nasze dane są gotowe do procesu modelowania.

3 Modelowanie danych przy pomocy ARMA

W wyniku dekompozycji otrzymaliśmy stacjonarny szereg, więc do stworzenia modelu ARMA potrzebna nam jest jeszcze znajomość rzędu modelu (p oraz q) oraz jego parametrów ($\phi_{(.)}$ i $\theta_{(.)}$). W dalszej części pracy, do oceny poprawności naszego modelu, będziemy potrzebowali zbioru danych rzeczywistych, z którym porównamy naszą predykcję. W tym celu podzieliliśmy nasze dane na zbiór treningowy oraz testowy, w proporcji 4:1. Zbiór treningowy pozwoli nam dopasować odpowiedni model, za pomocą którego wyestymujemy przyszłe wartości szeregu i porównamy je ze zbiorem testowym. Od tej pory będziemy operowali jedynie na części treningowej.

Rząd modelu możemy znaleźć poprzez zastosowanie kryteriów informacyjnych. My wybraliśmy następujące trzy kryteria:

Definicja 2 (Kryterium informacyjne Akkaikego)**.** *Najlepiej dopasowany jest model $ARMA(p, q)$, dla którego najmniejsza jest wartość*

$$AIC(p, q) = 2(p + q) - 2 \ln L(p, q, X_1, \ldots, X_n)$$

gdzie L jest funkcją wiarygodności modelu.

Definicja 3 (Kryterium informacyjne Bayesowskie)**.** *Najlepiej dopasowany jest model $ARMA(p, q)$, dla którego najmniejsza jest wartość*

$$BIC(p, q) = (p + q) \ln n - 2 \ln L(p, q, X_1, \ldots, X_n)$$

gdzie L jest funkcją wiarygodności modelu.

Definicja 4 (Kryterium informacyjne Hannana-Quinna)**.** *Najlepiej dopasowany jest model $ARMA(p,q)$, dla którego najmniejsza jest wartość*

$$HQIC(p,q) = 2(p+q)\ln(\ln n) - 2\ln L(p,q,X_1,\ldots,X_n)$$

gdzie L jest funkcją wiarygodności modelu.

Wszystkie te kryteria zawierają w sobie funkcję największej wiarygodności L. Wybór tej funkcji zależy oczywiście od rozkładu szumu rozważanego modelu, a tej informacji narazie nie posiadamy. Przy procesie modelowania założymy, że reszty modelu pochodzą z rozkładu normalnego, a w ostatniej części pracy zweryfikujemy tę hipotezę. Oczywiście założenie to naraża nas na potencjalne błędy w predykcji, ale jej poprawność również sprawdzimy.

Za pomocą metod `aic`, `bic` i `hqic` z pakietu `statsmodels.tsa.arima.model` policzymy wartości kryteriów informacyjnych dla różnych parametrów p i q. Wartości, które będą najlepiej minimalizować wszystkie trzy kryteria zostaną wybrane jako rząd naszego modelu. Wyniki prezentują się następująco:

- $AIC = -106.27$ - wybrany model $ARMA(4,1)$
- $BIC = -69.86$ - wybrany model $ARMA(4,1)$
- $HQIC = -91.69$ - wybrany model $ARMA(4,1)$

Wszystkie trzy kryteria wykazały, że najlepiej dopasowany do naszych danych będzie model $ARMA(4,1)$.

Skoro znamy już postać naszego modelu, możemy przejść do wyestymowania jego parametrów. Wykorzystamy w tym celu metodę `fit` z pakietu `statsmodels.tsa.arima.model`, podając jako metodę estymacji `innovation_mle`, czyli metodę największej wiarygodności. Metoda ta, podobnie jak wcześniej wymienione kryteria informacyjne, zakłada normalność szumu modelu. Podsumowanie modelu podaje nam wiele istotnych informacji, takich jak: wartości kryteriów informacyjnych, wyestymowane parametry modelu, estymator wariancji szumu oraz testy jego heteroskedastyczności i normalności. Wyniki estymacji parametrów zawarte są w poniższej tabeli.

Parametr	Wartość wyestymowana
ϕ_1	0.8736
ϕ_2	-0.1545
ϕ_3	-0.0851
ϕ_4	-0.1137
θ_1	-0.9999
σ^2	0.0521

Tabela 2: Wyestymowane wartości parametrów

P-wartości testów istotności wszystkich wyestymowanych parametrów są równe 0, zatem wszystkie z nich możemy uznać za istotne statystycznie.

Otrzymawszy estymowane wartości parametrów, możemy zapisać ostateczną postać naszego modelu $ARMA$:

$$X_t - 0.8736X_{t-1} + 0.1545X_{t-2} + 0.0851X_{t-3} + 0.1137X_{t-4} = Z_t - 0.9999Z_{t-1}$$

gdzie $Z_t \sim WN(0, 0.0521)$

Nasz model jest teraz gotowy do przeprowadzenia procesu prognozy i oceny jego skuteczności.

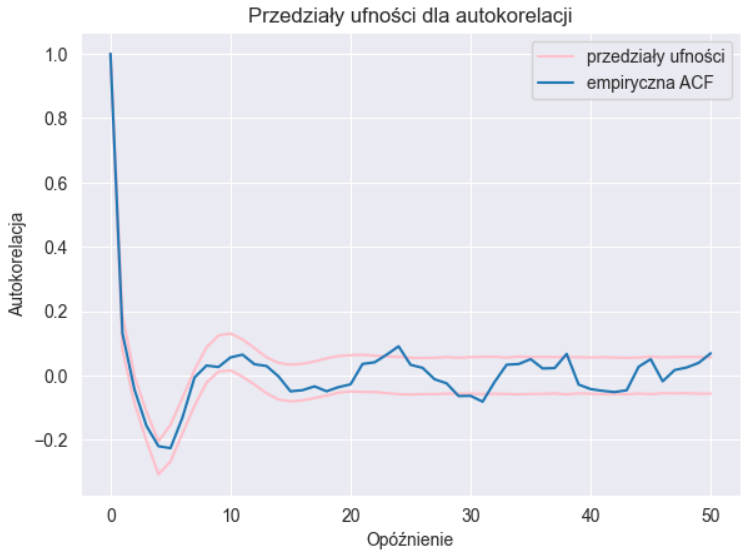
4 Ocena dopasowania modelu

Przejdziemy teraz do segmentu oceny dopasowania modelu. Zrobimy ją bazując na analizie przedziałów ufności dla autokorelacji ACF i częściowej autokorelacji PACF, porównania linii kwantylowych z trajektorią danych treningowych oraz prognozy przyszłych obserwacji (czyli u nas danych z próbki testowej) oraz skonfrontowania jakości dopasowania dobranego modelu ARMA(4,1) z oryginalnymi danymi, po uwzględnieniu z powrotem sezonowości i trendu.

4.1 Przedziały ufności dla ACF i PACF

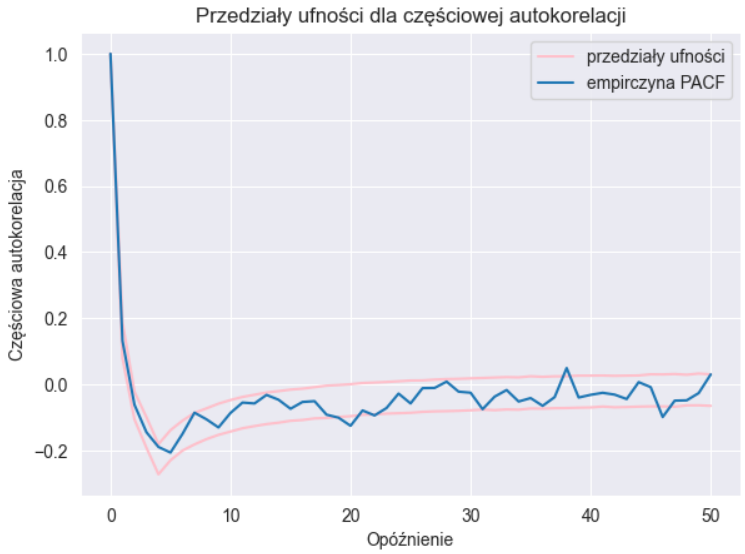
Na początku na tapet weźmiemy funkcje ACF i PACF dla naszych danych po dekompozycji i porównamy je z granicami przedziałów ufności skonstruowanych dla 90% (5% od dołu i 95% od góry).

Procedura wyznaczenia ich będzie polegała na wysymulowaniu 5000 próbek Monte Carlo, o długości takiej samej jak analizowane przez nas ostatecznie dane, metodą `generate_sample` z odpowiednio wyznaczonymi wcześniej dzięki pakietowi `statsmodels.tsa.arima.model.ARIMA` parametrami, a następnie analizie funkcji ACF i PACF wszystkich trajektorii w odpowiadających sobie punktach. W ten sposób dla uzyskanych wartości sporządzamy dla każdego opóźnienia rzędu od 0 do 50 kwantyle 5% i 95%, a następnie wizualizujemy je razem z empirycznymi wartościami tych funkcji dla naszych danych po dekompozycji.



Rysunek 14: Przedziały ufności dla ACF

Na powyższym wykresie widzimy, że większość empirycznych danych utrzymuje sie w ramach przedziałów ufności 90%. Spełnia to założenie owych granic albowiem dla takiej wartości poza nimi nie powinno wychodzić nie więcej niż 10%, czyli w tym przypadku dla liczby 50 różnych opóźnień – 5. Widoczny jest znaczny spadek około 6 wartości opóźnienia oraz odbicie się do wzniosu w okolicach „lagu" 10, aby po przekroczeniu punktu 18 ustabilizować się wokół autokorelacji równej 0. Znajduje to odzwierciedlenie w istocie naszych danych- miesiące 6 i 18 są oddalone od siebie o dokładnie rok, a w tym czasie miesiąc 10 jest przeciwną porą roku niż oba. Stąd te początkowe wahania dla małych wartości opóźnienia.

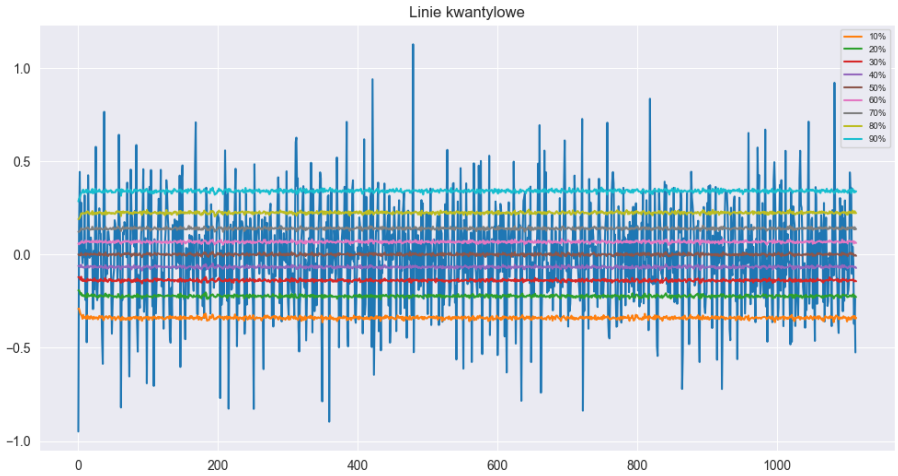


Rysunek 15: Przedziały ufności dla PACF

PACF mierzy bezpośrednią zależność między obserwacją a jej przeszłymi wartościami, usuwając efekty innych opóźnień. Na wykresie widzimy jak częściowa autokorelacja spada do wartości ujemnych po zerowym „lagu", by później powoli stabilizować się wokół 0. Można dostrzec, że linia prezentująca empiryczne wartości zaczyna rosnąć dla około 4 opóźnienia, co może sugerować ostatni istotną wartość, dla której powinniśmy rozważać autoregresję AR, z kolei długie wygaszanie (do momentu oscylacji obu przedziałów ufności wokół 0) wskazuje na niewystarczające niskie wartości dla niej, rzędu 1 lub 2, stąd $ARMA(4,1)$ wpasowuje się w te ideę.

4.2 Linie kwantylowe a trajektoria

Do konstrukcji linii kwantylowych posłużyliśmy się podobną metodą jak przy okazji liczenia ACF i PACF. Tym razem jednak liczyliśmy odpowiednie kwantyle dla każdego punktu na długości naszych danych z samych wygenerowanych 5000 trajektorii, uzyskując linie porównane z danymi jak poniżej.



Rysunek 16: Linie kwantylowe a trajektoria

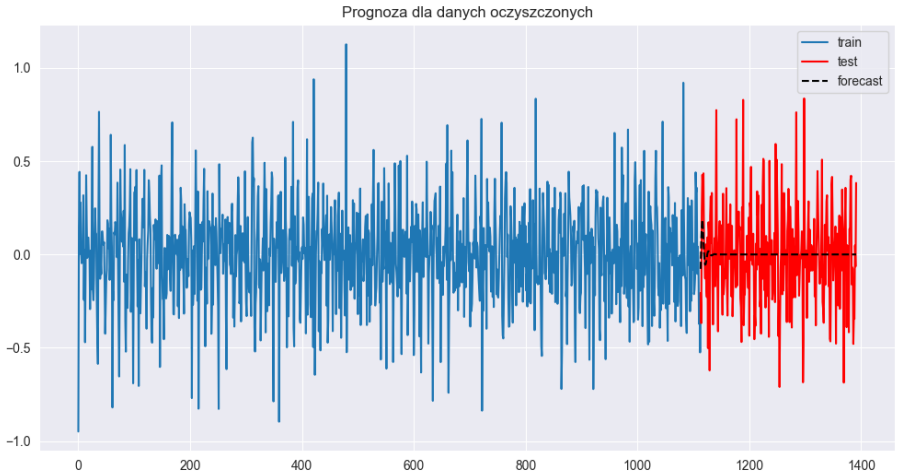
Linie kwantylowe nie są bardzo od siebie wzajemnie oddalone, co sugeruje że wahania na całym przedziale nie są znaczne. Granice na końcach mają największy rozstęp względem sąsiadujących, a będąca w środku symbolizująca 50% mieści się wokół 0 i ma najmniejszą odległość od sąsiadujących z nią bezpośrednio. Dodatkowo żadna linia nie doświadcza większych anomalii na całym przedziale, co każe sądzi, że żadne znaczne wahania nie występują. Jak można zauważyć od granic lini kwantylowych odbiega sporo danych, jednakże wobec ponad 1000 rekordów jest to zrozumiałe, a czy faktycznie jest ich dużo, teraz się przekonamy.

Rysunek 17: Rozkład przedziałów kwantylowych

Zobrazowaliśmy liczbę procentową punktów trajektorii, które wpadają do poszczególnych przedziałów wyznaczonych przez kwantyle. Wobec tego, że jest ich 10, po 10% każdy, należałoby się spodziewać, że liczba wszystkich będzie równa około 10%. Jak możemy wywnioskować z wykresu, tak też się dzieje. Pomimo zauważalnie większej liczby danych w środkowych przedziałach w stosunku do tych na krańcach, wartości są z przedziału $[8, 13]$, czyli nie odbiegają znacząco od tych zakładanych, czyli nasz rozkład kwantyli wyznaczonych symulacyjnie Monte Carlo ma jak najbardziej zastosowanie.

4.3 Prognoza przyszłych obserwacji na bazie danych treningowych i testowych

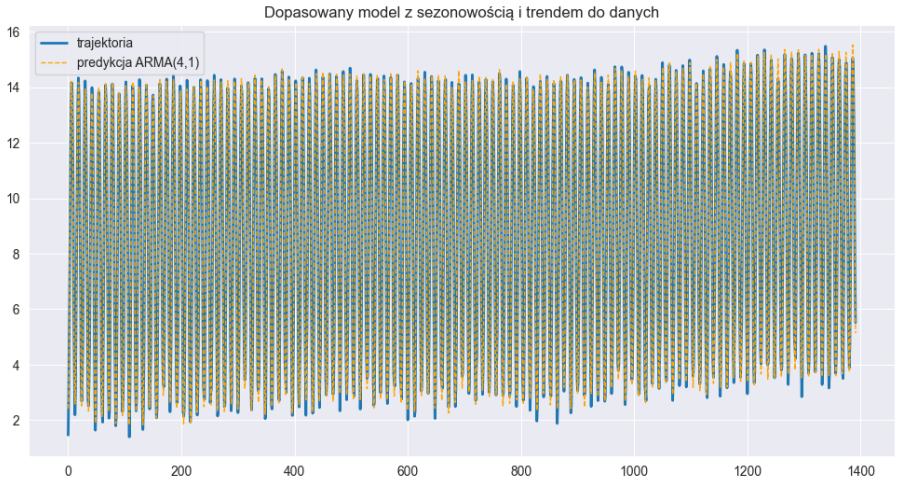
Teraz przejdziemy chyba do najciekawszej części projektu, mianowicie predykcji wartości. Będzie ona dla nas wyznacznikiem, jak dobrze dobrany model $ARMA(4, 1)$ dopasowuje się do danych, które przygotowaliśmy po obcięciu. W tym celu dla naszego modelu, stworzonego na bazie próbki treningowej, stosujemy metodę `ARIMA.forecast()` dla danych z przedziału testowego, uzyskując prognozę dla tzw. przyszłych obserwacji (u nas są one również znane, nazywamy je tak, gdyż $ARMA(4, 1)$ wybrany został na bazie danych sprzed nich, więc z tej perspektywy faktycznie dopiero nadejdą).



Rysunek 18: Prognoza dla danych oczyszczonych

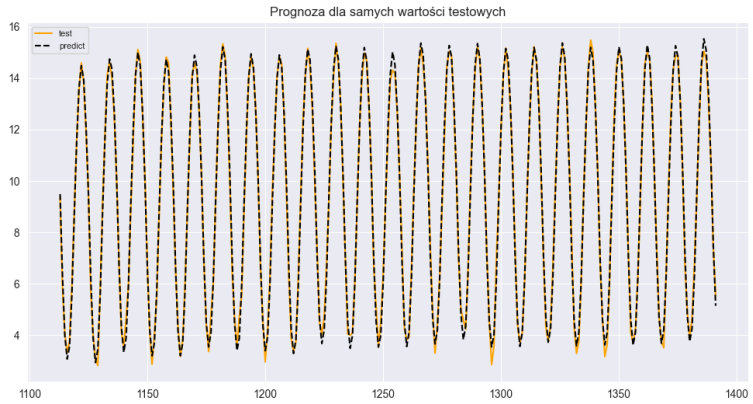
Jak widać predykcja poczyniona na samych danych oczyszczonych nie zdaje się z nimi samymi zgadzać i zbiega do 0 bardzo szybko. To jednak nie koniec tego etapu. Otrzymaliśmy właśnie model odpowiadający danym oczyszczonym- residuuom po dekompozycji. To nie do niego powinien dopasować się, a do danych oryginalnych przed dekompozycją.

Aby niejako „wrócić" na sam początek łączymy dane wyczyszczone dla przedziału treningowego (`ARIMA.predict()`) oraz tych z testowego (`ARIMA.forecast()`) oraz dodajemy do uzyskanej w ten sposób listy sezonowość i trend, który wydzieliliśmy przy okazji procesu dekompozycji. Po dodaniu tych wszystkich komponentów i wizualizacji wraz z rzeczywistymi wartościami, otrzymujemy następujące rezultaty.



Rysunek 19: Dopasowany model z sezonowością i trendem

Spoglądając na wykres powyżej można być lekko skonfundowanym, ale wbrew pozorom to bardzo dobrze, albowiem dane zaznaczone przerywaną pomarańczową linią prezentują model dopasowany wraz z uwzględnieniem sezonowości i trendu, który otrzymaliśmy w punkcie z dekompozycją. Można stwierdzić, że jedynie niektóre wartości na krańcach nie są dość dobrze predykowane, jednakże sama podstawa i środek są znakomicie odwzorowane. W szczególności zobrazowane jest to przez lekki trend wzrostowy, który następuje po tysięcznej danie. Model także dla tego zbioru dobrał dobrze wartości, symbolizujące zwiększanie się regularnie średniej temperatury na Ziemi w ostatnich latach, co nazywamy efektem globalnego ocieplenia. Przyjrzyjmy się teraz samemu zbiorowi testowemu, który do tego przedziału z wzrostem w trendzie należy.



Rysunek 20: Prognoza samych danych testowych

Wykres ponownie potwierdza w jak dobry sposób modelowanie przedstawia faktyczne zachowanie dla przyszłych obserwacji. Jedynie kilka wartości pozostaje niedoszacowanymi, a wzrost średniej temperatury ponownie widoczny. Czy model na pewno jest dobrze oszacowany ukaże nam kolejna sekcja, w której zweryfikujemy warunki szumu naszego modelu $ARMA(4,1)$, dzięki czemu będziemy wiedzieli, czy ma on wobec naszych danych zastosowanie.

5 Weryfikacja założeń dotyczących szumu

Do przystąpienia do tego segmentu musimy wydzielić residua modelu komendą `ARIMA.resid`, otrzymując następujące dane reszt modelu.

Rysunek 21: Residua

Na początku naszej analizy, wypisaliśmy warunki, które muszą być spełnione, by możliwe było zastosowanie modelu ARMA. W poprzednich sekcjach pracy dokonaliśmy pewnych założeń, żeby możliwe było przeprowadzenie procesu modelowania i predykcji. Teraz zajmiemy się weryfikacją tych założeń.

5.1 Normalność rozkładu

Na etapie modelowania potrzebne było nam założenie dotyczące normalności rozkładu szumu. Wymagało tego obliczenie wartości kryteriów informacyjnych oraz przeprowadzenie estymacji parametrów modelu.

Będziemy testować hipotezę

- H_0: Szum pochodzi z rozkładu normalnego

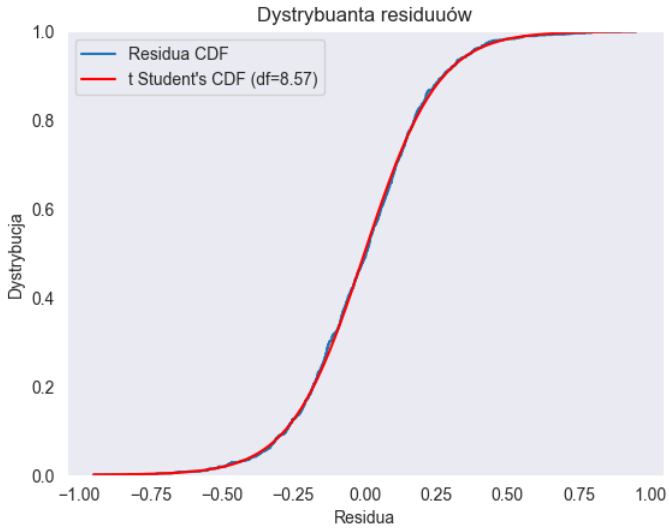
przeciwko

- H_1: Szum nie pochodzi z rozkładu normalnego

Przetestujemy to testem Shapiro-Wilka. Jeśli p-wartość dla niego wyniesie mniej niż 0.05, będziemy zmuszeni odrzucić hipotezę zerową o normalności szumu.

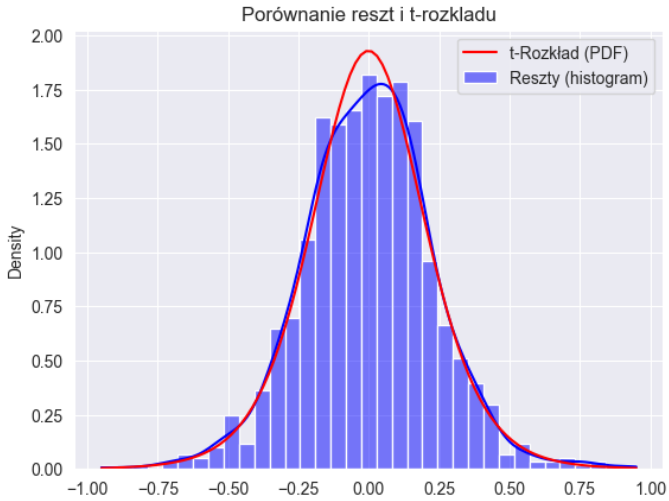
- Statystyka Shapiro-Wilka = 0.9915878176689148
- P-wartość = $5.588492058450356e - 06$
- Wynik: Odrzucamy hipotezę o normalności rozkładu szumu

Jak pokazuje test Shapiro-Wilka residua modelu nie spełniają założenia normalności, aczkolwiek po wykonaniu wykresu dystrybuanty i gęstości zauważyliśmy inną własność.



Rysunek 22: Dystrybuanta

Po dopasowaniu parametrów do rozkładu t-Studenta od naszego modelu widzimy, że dystrybuanta w dużym stopniu pokrywa się z tą zobrazowaną dla naszych danych. Możemy założyć na razie, że jest to prawdopodbne, że szum ten pochodzi właśnie nie z rozkładu normalnego, a z rozkładu t-Studenta, bardzo do niego zbliżonego, jednak charakteryzującego się cięższymi ogonami. Teraz pokażemy gęstość.



Rysunek 23: Gęstość w porównaniu z t-Studentem

Trzeba przyznać, że po środku linie gęstości różnią się od siebie, z kolei na końcach całkiem dobrze się pokrywają. Wykonaliśmy także test Kołomogorowa Smirnowa, dla którego przyjmowane jest u nas:

- H_0: Szum pochodzi z t-rozkładu

przeciwko

- H_1: Szum nie pochodzi z t-rozkładu

Wartości dla niego prezentują się następująco:

- Statystyka KS: 0.0225
- P-wartość: 0.6197
- Wynik: Brak podstaw do odrzucenia hipotezy zerowej

Test Kołmogorowa-Smirnowa nie dał nam podstaw do odrzucenia hipotezy o t-rozkładzie residuuów, więć możemy spekulować, że to właśnie z niego pochodzą residua dla naszego modelu.

5.2 Średnia

Podczas modelowania założyliśmy także, że szum posiada stałą w czasie średnią równą zero.

Przetestujemy hipotezę

- H_0: $\mu = 0$

przeciwko

- H_1: $\mu \neq 0$

Użyjemy do tego celu t-testu z modelu `scipy.stats`. Oto co otrzymaliśmy:

- Statystyka: -0.07809733919000025
- P-wartość: 0.9377646738311822
- Wynik: Średnia równa 0

Jak się okazuje dla naszych residuuów spełnione jest założenie stałej średniej równej 0, dzięki czemu możemy dalej zakładać podstawy do modelowania naszych danych modelem typu $ARMA$.

5.3 Wariancja

Stałość wariancji szumu, czyli homoskedastyczność, jest założeniem, które musi być spełnione, by model ARMA działał poprawnie. Bez niego bowiem szereg nie jest stacjonarny.

Teraz będziemy testować

- H_0: Wariancja szumu nie zmienia się w czasie (homoskedastyczność)

przeciwko

- H_1: Wariancja szumu zmienia się w czasie (heteroskedastyczność)

Przetestujemy to zmodyfikowanym testem Levene. Jest on przeprowadzany dla 2 grup, dla których sprawdzamy stałość wariancji względem obu. W celu

uzyskania ich podzielimy nasze residua (1 grupę) na dwie, dobierając je na podstawie średniej przeciętnej 10% (trimmed mean 10%). Sam test Levene ma zastosowanie przy okazji statystyk nie należących do rozkładu normalnego, natomiast ten konkretny wariant przeprowadza się dla rozkładów ciężkoogonowych (takich jak na przykład w rozkładzie t-Studenta), których kurtoza jest większa niż 3. U nas wartość kurtozy dla reszt to 4.080562073147987, więc jak najbardziej metoda ta ma zastosowanie. Liczymy średnią przeciętną `scipy.stats.trim_mean` i dzielimy dane - większe do jednej grupy, mniejsze lub równe do drugiej. Następnie wykonujemy sam test.

- Levene Statistic: 0.0045
- P-wartość: 0.9463
- Wynik: Nie ma podstaw do odrzucenia hipotezy zerowej (reszty są homoskedastyczne)

Jak pokazują wyniki testu Levene'a udowodniliśmy homoskedastyczność residuuów, czyli kolejne założenie dla modelu $ARMA$.

5.4 Niezależność

Ostatnie założenie, które zweryfikujemy dotyczy niezależności szumu, czyli braku korelacji między wartościami reszt modelu w różnych momentach czasu.

Przeprowadzimy test na

- H_0: Reszty modelu są niezależne

przeciwko

- H_1: Reszty modelu są zależne

Niezależność wypróbujemy testem Ljunga-Boxa, dla którego hipoteza zerowa zakłada, że reszty są niezależne i są białym szumem. Wyniki tego testu dla residuuów dobranego modelu:

- lb stat: 49.373687
- lb p-wartość: 0.08392

- Wynik: Nie ma podstaw do odrzucenia hipotezy zerowej (residua są niezależne)

Test Ljunga-Boxa potwierdza, że reszty modelu są niezależne, co powinniśmy również uzyskać przy okazji tej próby.

6 Podsumowanie

W naszej pracy przedstawiliśmy kompleksową analizę danych meteorologicznych z użyciem modelu ARMA. Jako przedmiot badania obraliśmy średnią miesięczną temperaturę na powierzchni Ziemi. Oryginalne dane zawierały okres czasu od października 1752 roku do grudnia 2015 roku. Brak odpowiednich sprzętów pomiarowych w XVIII i XIX wieku był źródłem dużych błędów pomiarowych, które zakłamywały wyniki naszej analizy. Problemu tego udało nam się pozbyć poprzez obcięcie zbioru danych i przeanalizowanie jedynie wartości od roku 1900 w górę. Dekompozycja szeregu czasowego pozwoliła nam wyodrębnić wyraźną tendencję wzrostowa w danych, która wskazuje na występowanie efektu globalnego ocieplenia klimatu. Składowa sezonowa dobrze pokazała jak średnia temperatura na Ziemi kształtuje się w zależności od panującej pory roku. Dopasowanie odpowiedniego modelu pozwoliło nam przeprowadzić skuteczną predykcję.

Z przeprowadzonej analizy możemy wyciągnąć następujące wnioski:

- Odpowiednio dopasowany model ARMA jest skutecznym narzędziem do predykcji zjawisk zmieniających się w czasie, np. danych meteorologicznych
- Dobranie odpowiednich danych stanowi klucz do zbudowania skutecznego modelu
- Zastosowanie modelu ARMA wymaga spełnienia odpowiednich założeń, których zignorowanie może prowadzić do błędnej analizy
- Czasami warto jest ograniczyć zbiór danych roboczych, celem uzyskania skutecznego modelu, nawet jeśli ogranicza to obszerność analizy