

Analiza szeregów czasowych - Projekt zaliczeniowy

Alicja Kalwat
Modelowanie Matematyczne i Analiza Danych,
Uniwersytet Gdański

20.01.2024r.

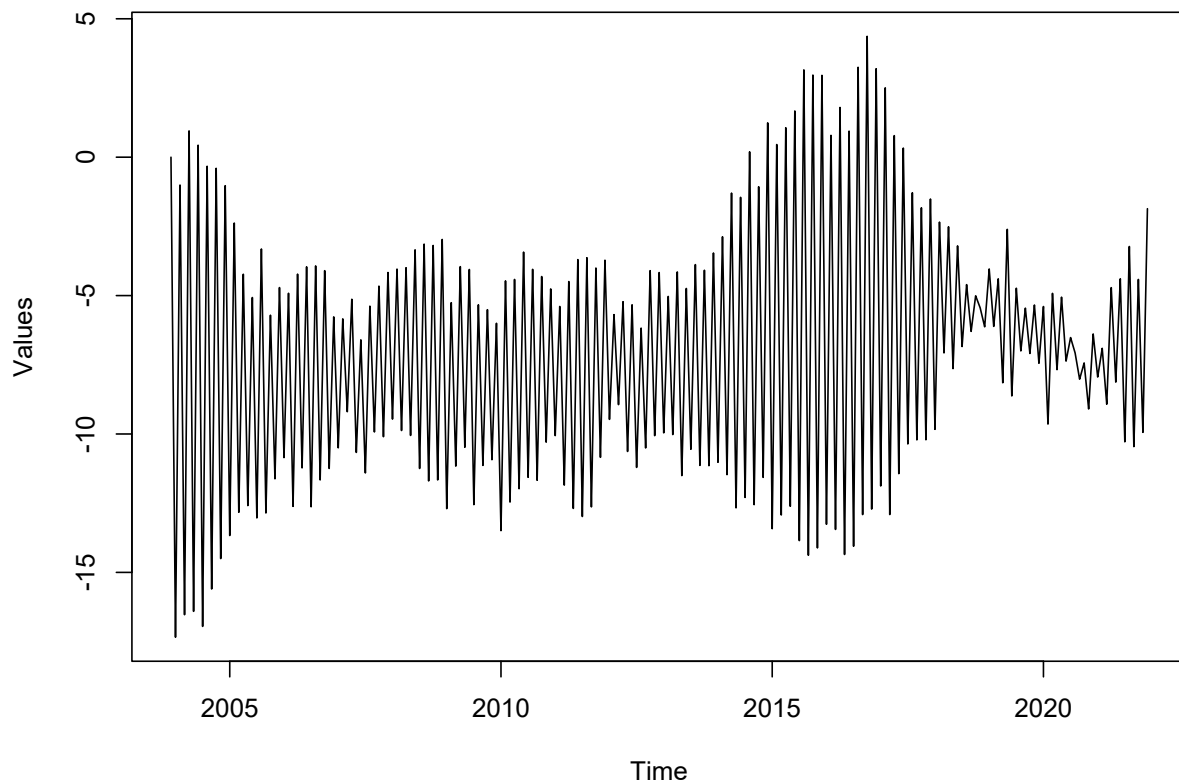
Wstęp

Projekt prezentuje analizę danych miesięcznych (dane nr 12), skupiając się na ocenie stacjonarności szeregu, zastosowaniu przekształceń, modelowaniu $AR(p)$, $MA(q)$ oraz dopasowaniu modelu SARIMA. Diagnostyka obejmuje ocenę reszt oraz dopasowania i precyzji prognoz. Ostatecznie wybrany zostanie model, który najlepiej pasuje do analizowanych danych. Analiza przeprowadzona będzie na zbiorze treningowym, a zbiór testowy (ostatnie 10% danych) wykorzystany zostanie do oceny prognoz.

1. Ocena stacjonarności szeregu

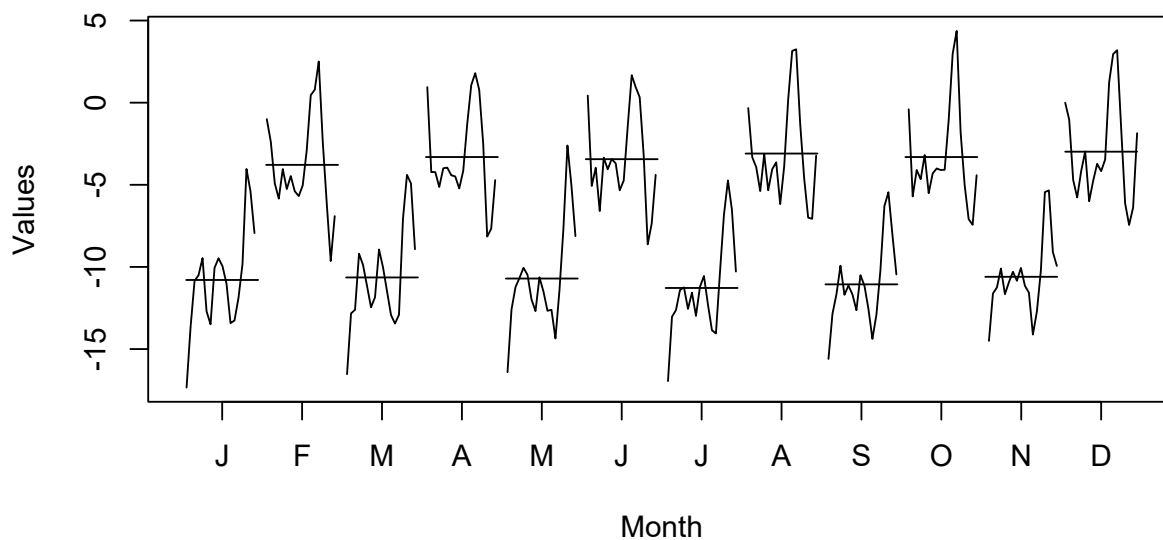
Pierwszym krokiem w analizie szeregu czasowego jest ocena stacjonarności. Cechą, którą możemy zobaczyć już na wykresie szeregu jest sezonowość, która w tym przypadku ma okres dwumiesięczny. Oznacza to, że podobne wartości powtarzają się co dwa miesiące. Świadczy to o braku stacjonarności szeregu.

Wykres szeregu czasowego: grudzień 2003 - grudzień 2021

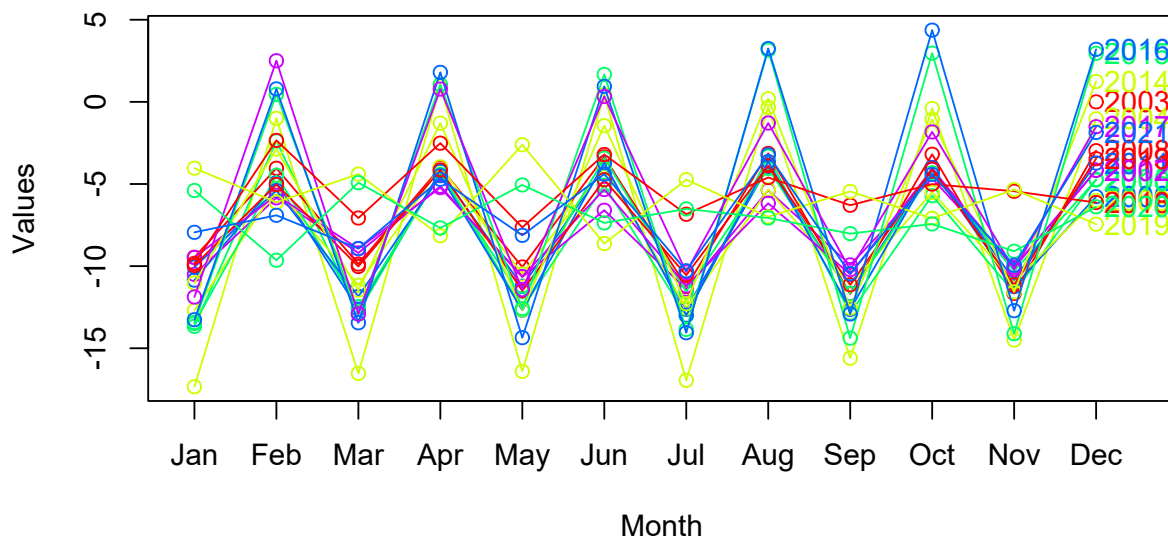


Gdy dane są sezonowe, wtedy warto stworzyć wykresy *seasonplot* oraz *monthplot*. Przyjrzyjmy się w takim razie, jak wyglądają dane podzielone na miesiące - bardzo dobrze widać tutaj opisaną wcześniej sezonowość.

Wykres miesięczny

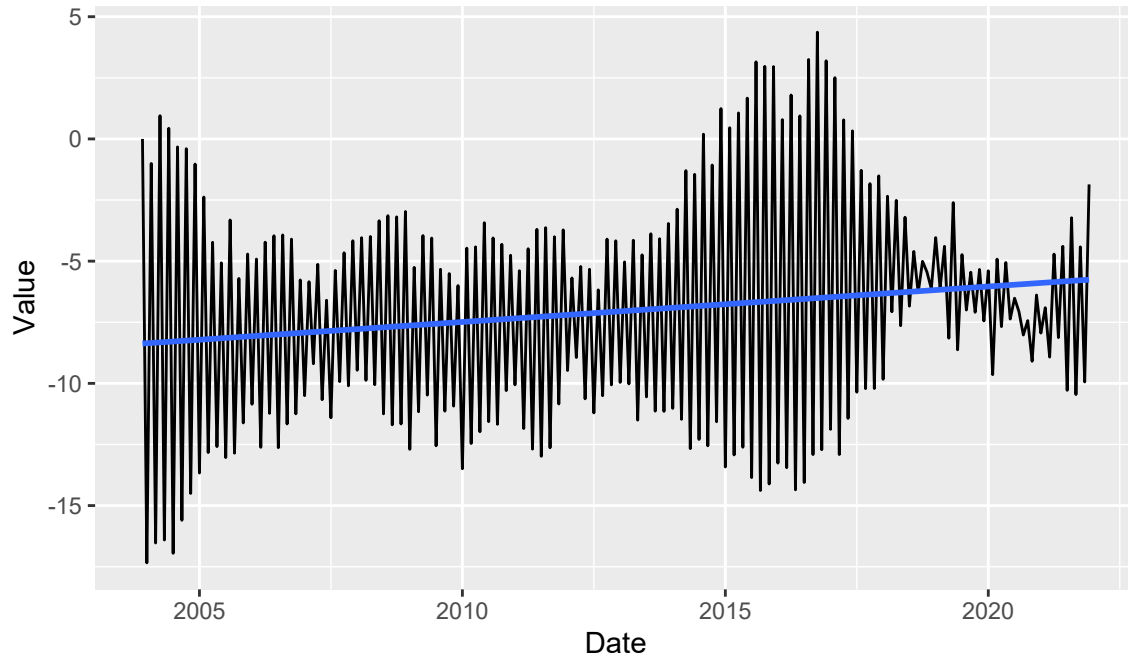


Wykres sezonowy



Drugą cechą, która świadczy o niestacjonarności szeregu, jest obecność trendu. Na pierwszy rzut oka ciężko stwierdzić, czy jest obecny w tym szeregu. W takim razie dodajmy do niego linię regresji.

Szereg czasowy z linią regresji



Patrząc na wykres możemy ocenić, że trend, mimo, że niewielki, to jest obecny. Tworząc model regresji liniowej otrzymujemy w podsumowaniu $p\text{-value} = 0.01845$, co również wskazuje na obecność trendu.

Potwierdza to przypuszczenie, że szereg jest niestacjonarny. Aby upewnić się, że tak faktycznie jest, możemy skorzystać z dwóch testów sprawdzających stacjonarność:

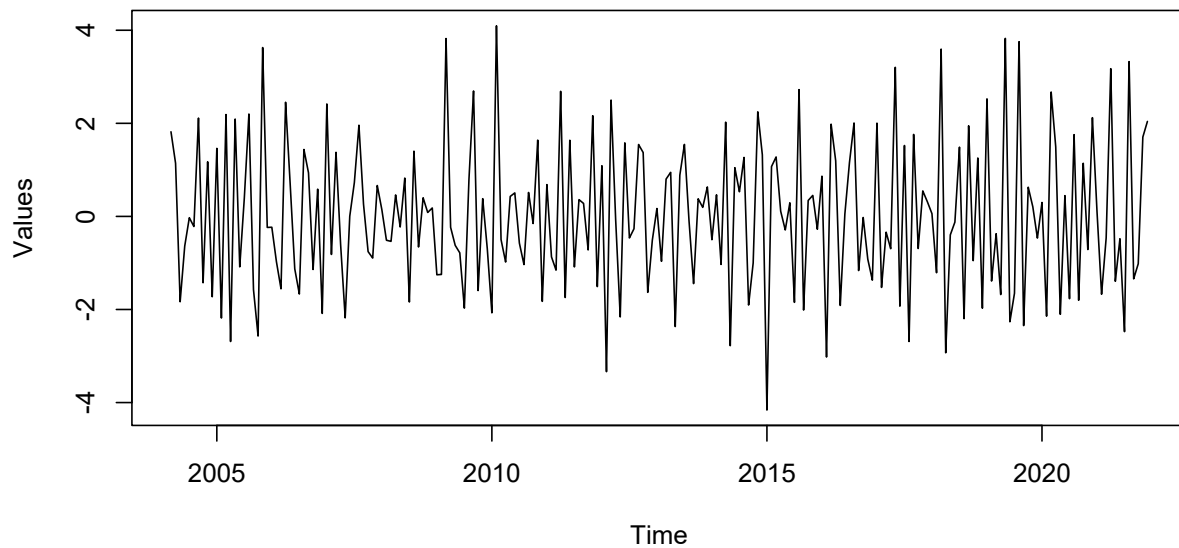
- *adf.test* (Augmented Dickey-Fuller Test) - hipoteza zerowa: szereg niestacjonarny,
- *kpss.test* (Kwiatkowski-Phillips-Schmidt-Shin) - hipoteza zerowa: szereg stacjonarny.

W teście ADF otrzymujemy $p\text{-value} = 0.2512$ (zostawiamy hipotezę zerową), natomiast w KPSS: $p\text{-value} = 0.01$ (odrzucaamy hipotezę zerową). Obie te wartości wskazują na to, że nie jest to szereg stacjonarny.

2. Przekształcanie na szereg stacjonarny

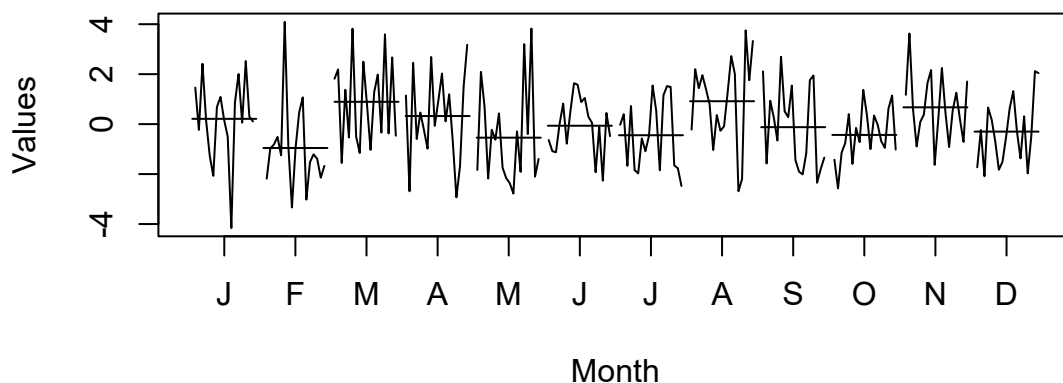
Do dalszej analizy potrzebujemy danych, które są stacjonarne. Zastosujemy więc odpowiednie przekształcenia, aby takie dane uzyskać. Skorzystamy z metody różnicowania. Sprawdzając, jaki jest automatyczny wybór rzędu różnicowania ($ndiffs$, $nsdiffs$) dostajemy w obu przypadkach wynik 1. Oznacza to, że mamy różnicować raz, by pozbyć się trendu i raz, by pozbyć się sezonowości. Zgadza się to z naszymi wcześniejszymi obserwacjami. Tak więc robimy. W przypadku sezonowości ustawimy opóźnienie = 2.

Ostateczny szereg czasowy po zastosowaniu różnicowań

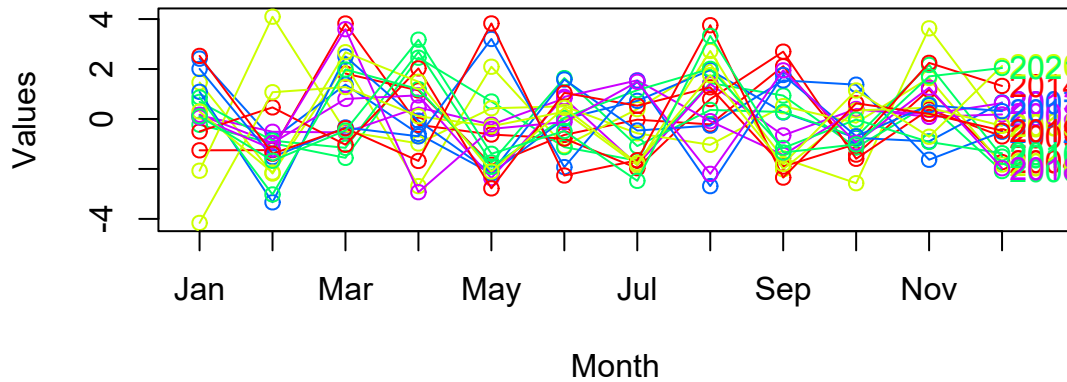


Teraz szereg wygląda na stacjonarny. Upewnijmy się, że nie ma już sezonowości, patrząc po raz kolejny na wykresy *seasonplot* oraz *monthplot*:

Wykres miesięczny

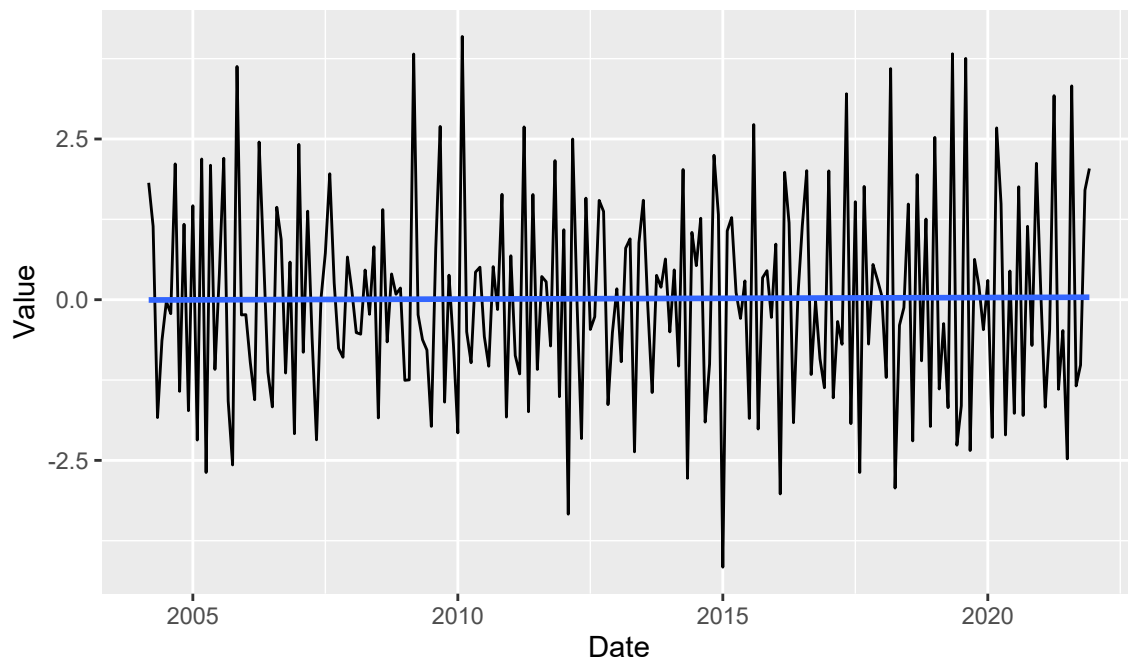


Wykres sezonowy



Nie widać sezonowości. Co więcej, na wykresie miesięcznym widać, że średnie wartości w każdym miesiącu zbliżyły się do zera. Zobaczmy teraz, czy zniknął również trend.

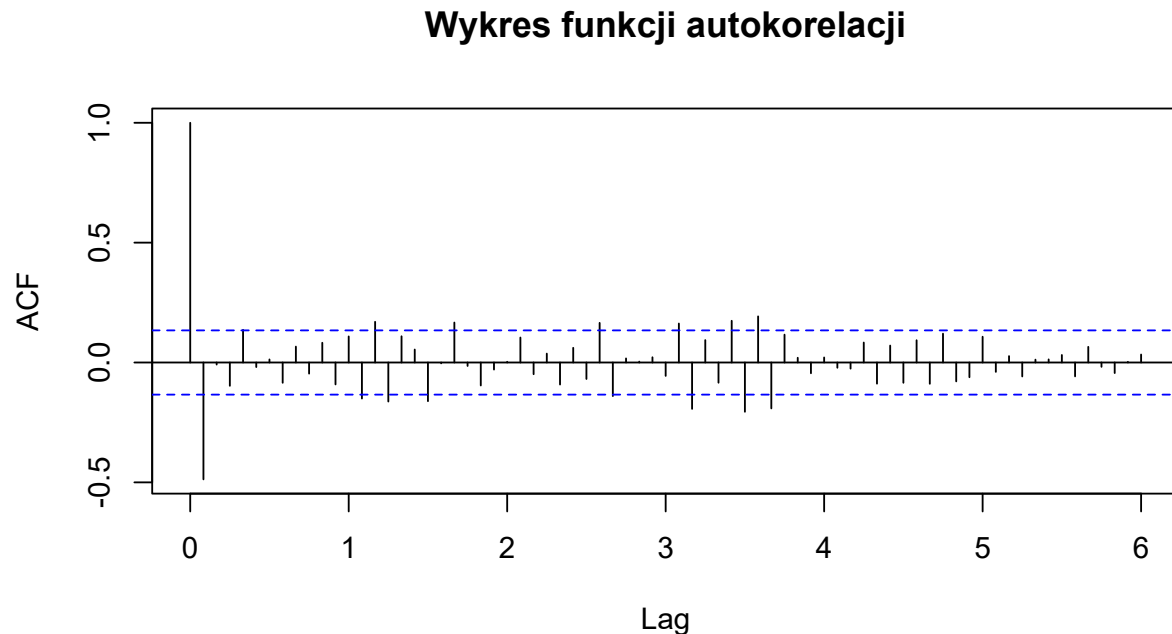
Szereg czasowy z linią regresji (po różnicowaniu)



Trend zniknął (teraz w modelu regresji liniowej $p\text{-value} = 0.9123$). W takim razie nie ma już ani trendu, ani sezonowości. Wykonując dla sprawdzenia test ADF (H_0 : szereg niestacjonarny) dostajemy $p\text{-value} = 0.01$. Zatem możemy uznać te dane za stacjonarne.

3. Dopasowanie modelu szeregu AR(p) oraz szeregu MA(q).

Do danych stacjonarnych możemy dopasować modele szeregu autoregresji oraz szeregu średniej kroczącej. Zajmijmy się najpierw średnią kroczącą - MA(q). Aby dobrać odpowiednie q spójrzmy na wykres autokorelacji.



Potrzebujemy znaleźć takie q , żeby silna korelacja była obecna tylko na opóźnieniach do q . Następne wartości powinny być bliskie zeru (albo przynajmniej znajdować się wewnątrz akceptowalnego przedziału).

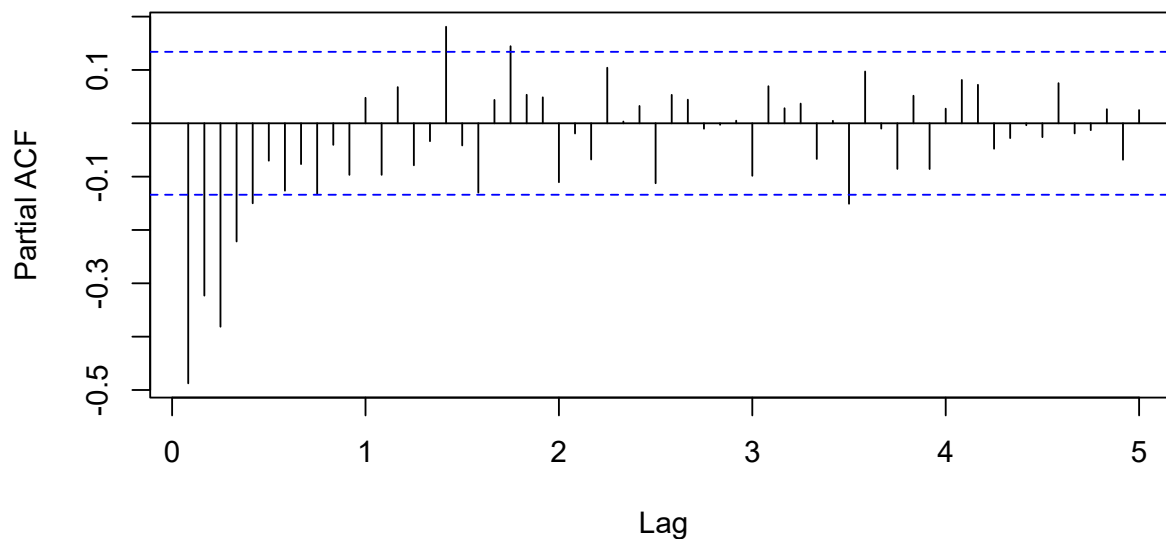
Tutaj najbezpieczniejszym wyborem wydaje się być $q = 44$, bo w tym miejscu jest ostatnia znacząca korelacja. Jednak jest to bardzo duża liczba, więc warte zastanowienia może być też, trochę mniejsze, $q = 20$.

Jest to ostatnia znacząca wartość, po której długo nie ma żadnych wystających poza dopuszczalny przedział słupków. Jest ich potem jeszcze kilka, jednak mieszczą się one w 5%, które mogą poza ten przedział wychodzić. W naszym przypadku 5% długości całego szeregu czasowego to około 11 wartości. Co prawda nie są to duże przekroczenia i kusiłoby wzięcie mniejszego q , jednak należy pamiętać, że nasz model musi spełnić wspomniane wcześniej założenie.

Mając to na uwadze, do dalszej analizy użyjemy wcześniej wybraną wartość $q = 20$, czyli będzie to model średniej kroczącej rzędu 20.

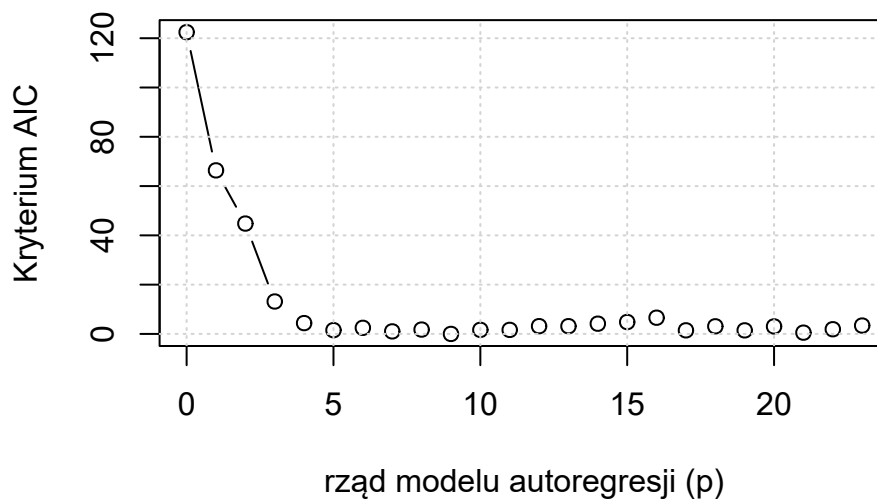
Aby dobrać model szeregu autoregresji - $AR(p)$, przyjrzyjmy się wykresowi częściowej autokorelacji.

Wykres funkcji częściowej autokorelacji



W tym przypadku parametr p również ma być ostatnim, dla którego obecna będzie znaczna korelacja. Dobrym kandydatem wydaje się tu być $p = 21$. Jest to bezpieczny wybór, bo jest to ostatni słupek wystający poza przedział (nie licząc późniejszych pojedynczych wartości mieszczących się w 5%).

Jako pomoc w doborze p możemy zastosować funkcję *ar.aic*, która za pomocą kryterium AIC wybiera najlepsze p , oraz wykres wartości kryterium AIC dla różnych p . Funkcja *ar.aic* w naszym przypadku wybrała $p = 9$. Na wykresie widać, że ma ona najniższą wartość AIC, jednak jest ona bardzo zbliżona do tej dla $p = 21$. Po ocenie wykresu PACF decydujemy się więc na pozostanie przy $AR(21)$.



Podsumowując: dla naszego szeregu stacjonarnego dopasowaliśmy dwa modele:

- szereg autoregresji AR(21),
- szereg średniej kroczącej MA(20).

4. Dopasowanie modeli SARIMA.

Modele można dopasowywać także do danych niestacjonarnych. Spróbujmy więc dopasować model SARIMA do naszych oryginalnych danych (treningowych). Będzie to model złożony z części autoregresyjnej, kroczącej średniej, a także uwzględniający trend i sezonowość. Na początek zastosujemy funkcję *auto.arima*, która automatycznie wybiera model SARIMA za pomocą kryteriów AIC, AICC lub BIC.

Wszystkie kryteria proponują wybór ARIMA(2,0,1)(2,1,0)[12].

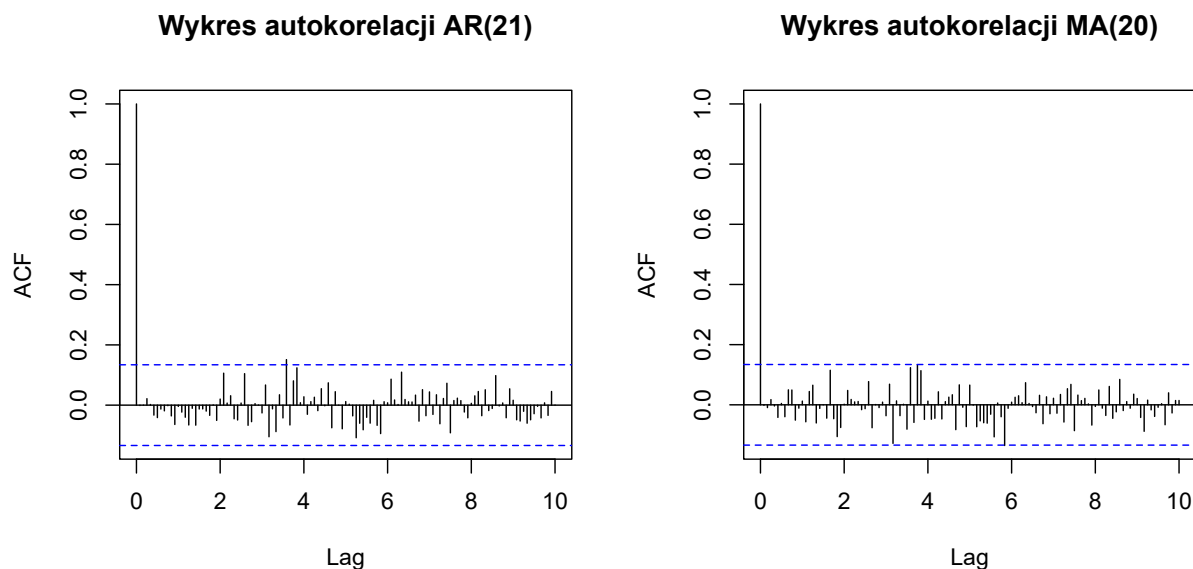
Dziwny wydaje się jednak być okres sezonowości równy 12, skoro na wykresie było wyraźnie widać okres co dwa miesiące. Kolejną zastanawiającą rzeczą jest brak różnicowania w celu niwelacji trendu. Stworzymy więc drugi “ulepszony” model: ARIMA(2,1,1)(2,1,0)[2] - w którym zmienimy okres sezonowości na 2, oraz dodamy różnicowanie $d = 1$.

5. Diagnostyka modeli: ocena jakości dopasowania i prognoz.

Mamy już przygotowane modele dla danych stacjonarnych oraz dla danych początkowych. Wykonamy teraz ich diagnostykę, tj: sprawdzimy, czy reszty są losowe i normalne, wyliczymy wartości kryteriów informacyjnych, a na koniec wyliczymy miary błędu prognoz, by ocenić ich jakość.

Dane stacjonarne

Jako pierwszymi zajmiemy się danymi stacjonarnymi. Na wykresach widać funkcję autokorelacji dla obu modeli - wyglądają poprawnie, nie widać żadnych znaczących korelacji między opóźnieniami.



Losowość reszt można też sprawdzić za pomocą testu Boxa - Ljunga. W tym teście hipoteza zerowa mówi, że reszty są losowe. Patrząc poniżej na p-value reszt obydwu modeli zdecydowanie zostawiamy tę hipotezę.

- p-value testu Boxa-Ljunga dla AR(21): 0.99
 - p-value testu Boxa-Ljunga dla MA(20): 0.98
-

Następnie możemy ocenić normalność reszt używając testu Shapiro-Wilka. Hipoteza zerowa mówi o pochodzeniu danych z rozkładu normalnego, więc stwierdzamy, że nasze reszty są normalne.

- p-value testu Shapiro - Wilka dla AR(21): 0.3645
 - p-value testu Shapiro - Wilka dla MA(20): 0.2144
-

Teraz porównajmy wartości kryteriów informacyjnych: AIC, AICc oraz BIC dla tych modeli. Im mniejsze wartości kryteriów, tym lepiej. Patrząc na wartości są one porównywalne dla obydwu z nich.

- AR(21): AIC = 637.06 AICc = 642.87 BIC = 714.48
 - MA(20): AIC = 637.83 AICc = 643.13 BIC = 711.88
-

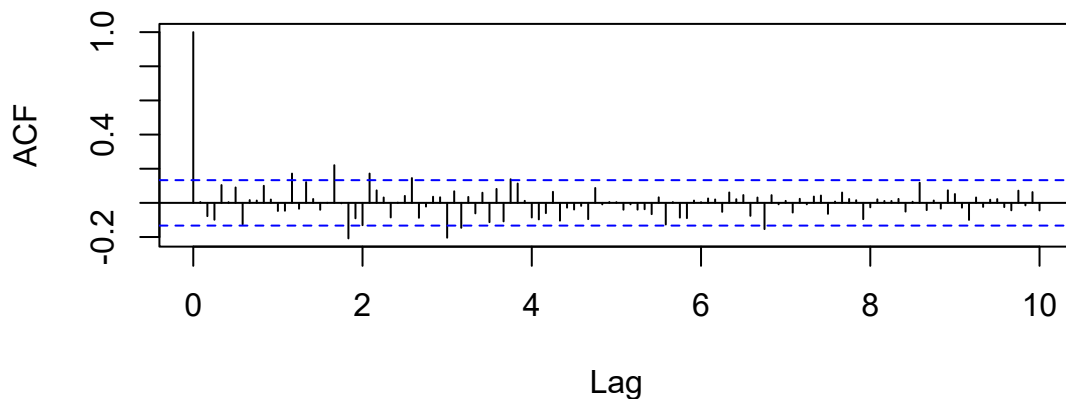
Następnym krokiem w diagnostyce jest sprawdzenie dokładności prognoz, porównamy wartości miar błędów: RMSE (Root Mean Squared Error), MAE (Mean Absolute Error), MAPE (Mean Absolute Percentage Error) oraz MASE (Mean Absolute Scaled Error) - im będą mniejsze, tym lepiej. Tutaj również wartości są porównywalne, jednak niepokojący wydaje się być wysoki wynik MAPE. Wyższa wartość MAPE oznacza większy procentowy błąd prognozy, co może oznaczać, że prognozy są niestabilne.

- AR(21): RMSE= 0.95, MAE = 0.78, MAPE = 281.46, MASE = 0.46
 - MA(20): RMSE = 0.94, MAE = 0.74, MAPE = 334.52, MASE = 0.44
-

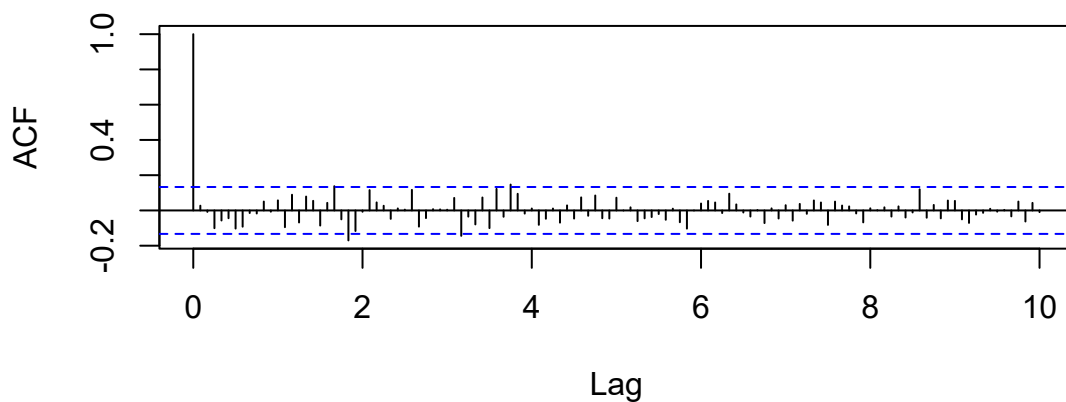
Dane oryginalne

Teraz zajmiemy się danymi oryginalnymi, czyli niestacjonarnymi. Na poniższych wykresach widać funkcję autokorelacji dla dwóch modeli SARIMA. Dla pierwszego modelu możemy zaobserwować znacznie więcej wartości wystających poza dozwolony przedział.

ARIMA(2,0,1)(2,1,0)[12]



ARIMA(2,1,1)(2,1,0)[2]



Sprawdźmy teraz losowość reszt testem Boxa-Ljunga - okazuje się, że pierwszy model nie ma losowych reszt, ale drugi już tak. Losowość reszt jest bardzo ważnym założeniem w trakcie tworzenia modelu i powinna zostać spełniona, więc model ARIMA(2,0,1)(2,1,0)[12] najprawdopodobniej będzie musiał zostać odrzucony.

-
- p-value testu Boxa-Ljunga dla ARIMA(2,0,1)(2,1,0)[12]: 0.0095
 - p-value testu Boxa-Ljunga dla ARIMA(2,1,1)(2,1,0)[2]: 0.2709
-

Oceńmy normalność reszt. Normalność reszt jest założeniem, które warto by było, aby było spełnione przez model. Tutaj znacznie lepiej wygląda drugi model. P-value dla pierwszego modelu jest mniejszy niż 5%, przez co możemy stwierdzić, że nie ma on reszt o rozkładzie normalnym.

- p-value testu Shapiro - Wilka dla ARIMA(2,0,1)(2,1,0)[12]: 0.0229
 - p-value testu Shapiro - Wilka dla ARIMA(2,1,1)(2,1,0)[2]: 0.4304
-

Sprawdźmy wartości kryteriów informacyjnych dla tych model - W tym zestawieniu lepszy wydaje się model ARIMA(2,1,1)(2,1,0)[2], bo ma mniejsze wartości kryteriów.

- ARIMA(2,0,1)(2,1,0)[12]: AIC=652.29 AICc=652.71 BIC=672.22
 - ARIMA(2,1,1)(2,1,0)[2]: AIC=632.23 AICc=632.64 BIC=652.43
-

Ostatnim krokiem będzie porównywanie wartości miar błędów prognoz. Różnią się one nieznacznie, nieco mniejsze są dla modelu ARIMA(2,1,1)(2,1,0)[2]. Tutaj MAPE jest dużo mniejsze niż w przypadku modeli dla danych stacjonarnych, co wskazuje na lepsze dopasowanie.

- ARIMA(2,0,1)(2,1,0)[12]: RMSE = 1.10, MAE = 0.82, MAPE = 24.32, MASE = 0.43
 - ARIMA(2,1,1)(2,1,0)[2]: RMSE = 1.01, MAE = 0.81, MAPE = 26.20, MASE = 0.42
-

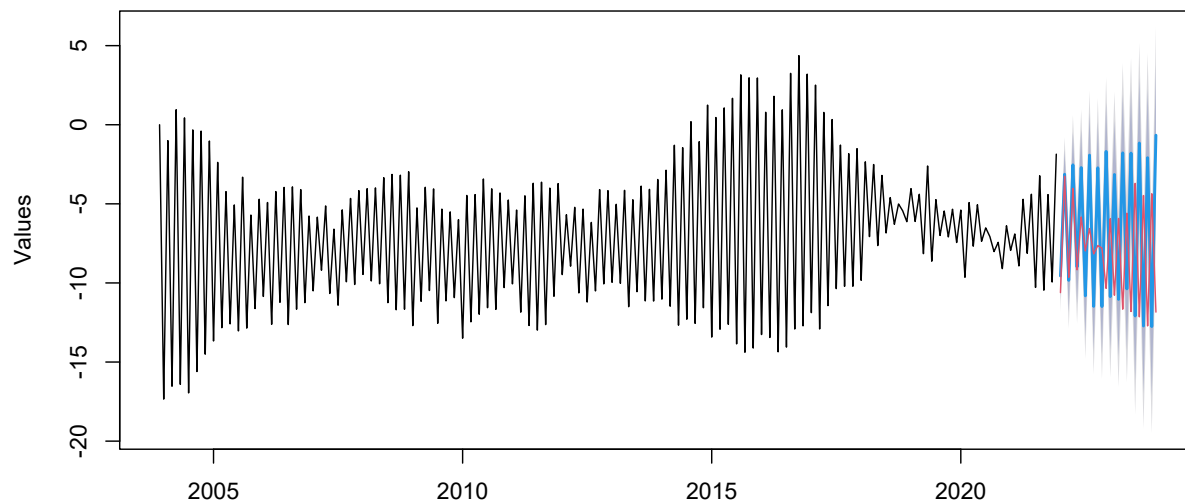
6. Wyznaczanie oraz ocena prognoz

Możemy teraz przystąpić do tworzenia prognoz. Wykonamy je za pomocą funkcji *forecast*. Zbiór testowy ma długość 24, więc na tyle miesięcy wyznaczmy prognozę - od stycznia 2022, do grudnia 2023 roku. Zaczniemy od modeli SARIMA i danych niestacjonarnych.

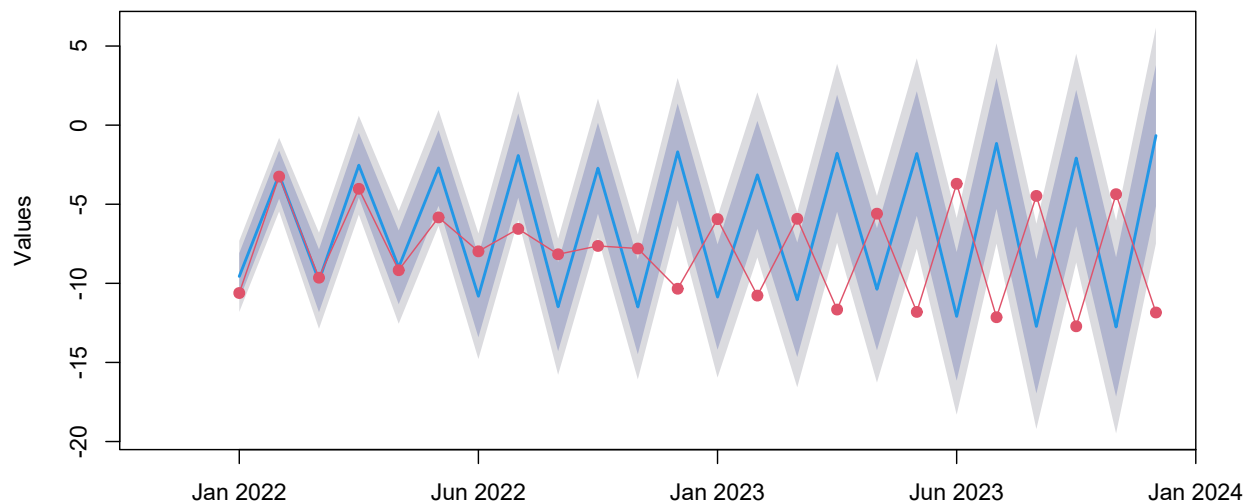
Model $ARIMA(2,0,1)(2,1,0)[12]$

Model dobrze poradził sobie z prognozowaniem w pierwszych miesiącach, potem jednak wartości zaczęły się trochę “rozjeżdżać”. Niemniej jednak została zachowana obecna sezonowość.

Forecasts from $ARIMA(2,0,1)(2,1,0)[12]$



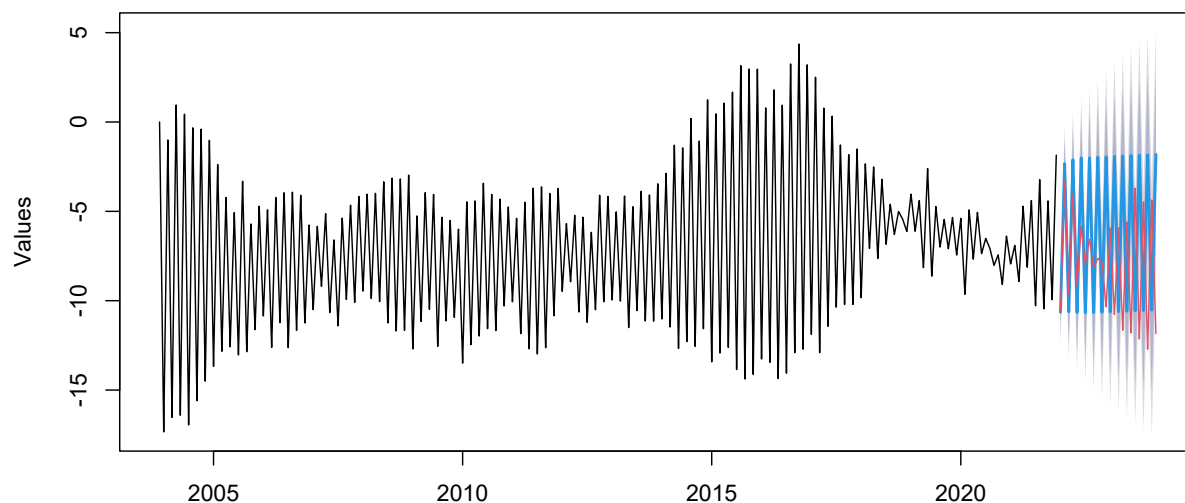
Forecasts from $ARIMA(2,0,1)(2,1,0)[12]$



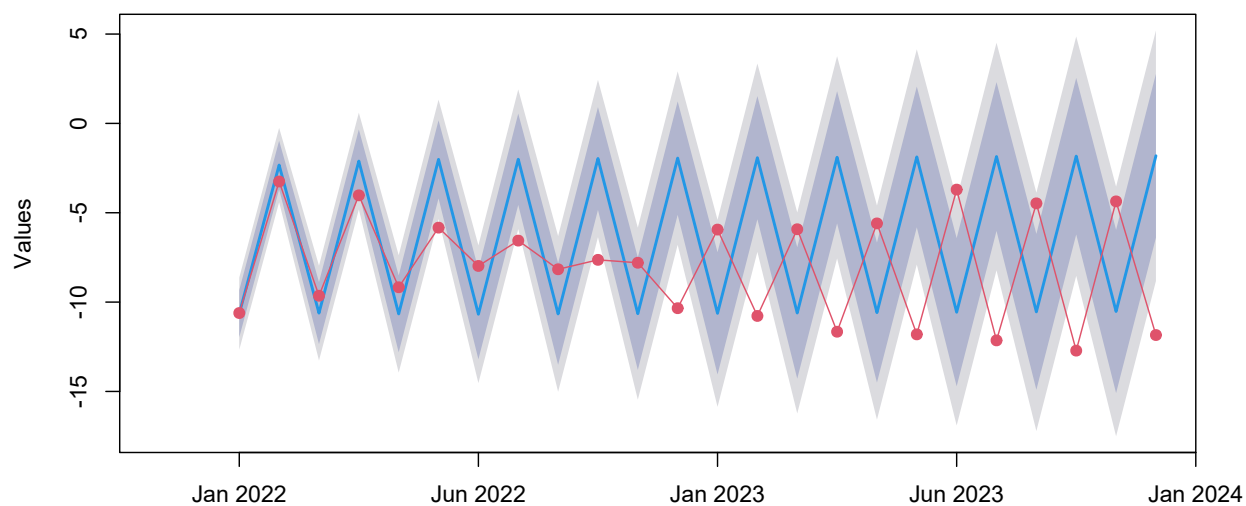
Model $\text{ARIMA}(2,1,1)(2,1,0)[2]$

Ta prognoza jest bardzo podobna, pierwsze miesiące przewidziane są niemal idealnie, dopiero później sytuacja się pogarsza. Mimo, że prognozy wyglądają podobnie, należy pamiętać o wynikach diagnostyki obu modeli. Model drugi ($\text{ARIMA}(2,1,1)(2,1,0)[2]$) miał lepsze wyniki, więc on będzie lepszym wyborem.

Forecasts from $\text{ARIMA}(2,1,1)(2,1,0)[2]$



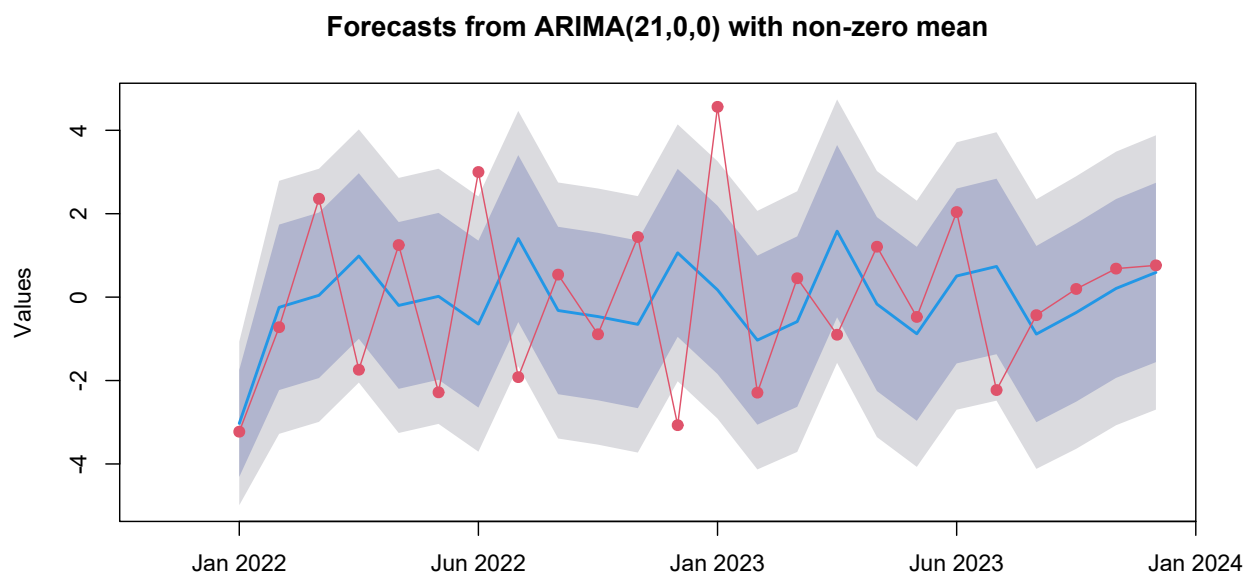
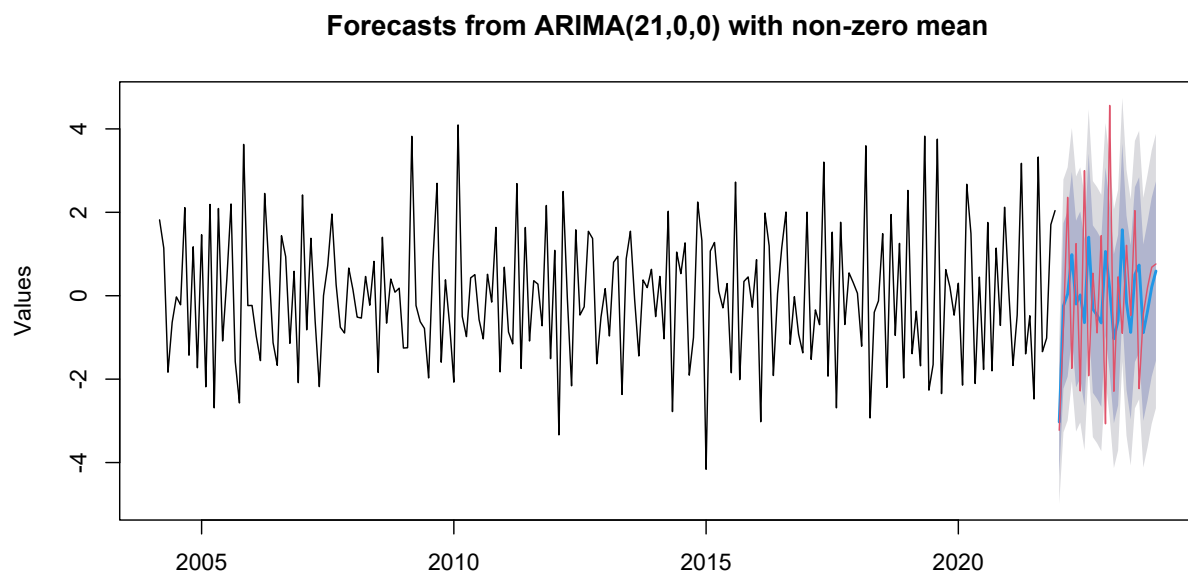
Forecasts from $\text{ARIMA}(2,1,1)(2,1,0)[2]$



Wyznamy teraz prognozy na danych stacjonarnych.

Model AR(21)

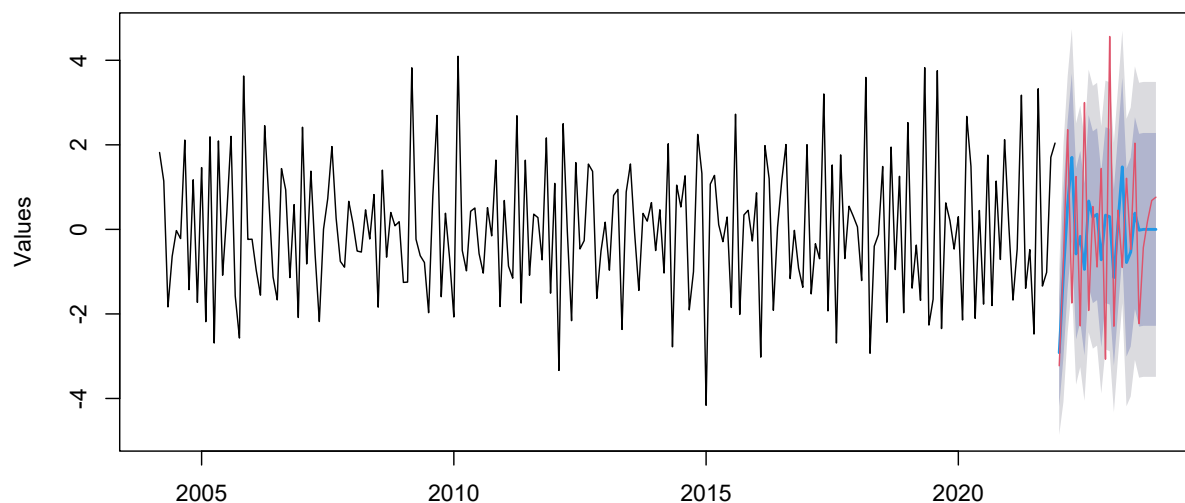
Tutaj, z racji stacjonarności szeregu wartości są bardziej chaotyczne. Mimo to, model poradził sobie z przewidzeniem niektórych wartości, a znaczna większość znajduje się w przedziale ufności owych prognoz.



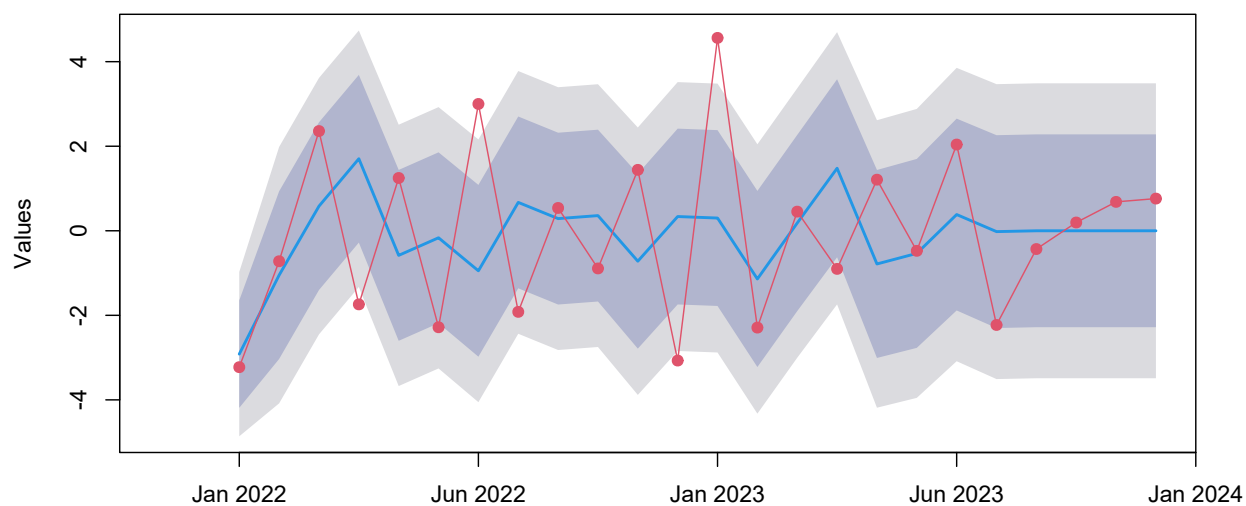
Model MA(20)

W tym przypadku prognozy również są do siebie zbliżone. W tym modelu, podobnie jak w AR(21) niektóre wartości udało się przewidzieć, a niektóre wystają lekko poza przedział ufności. Patrząc na ostatnie miesiące lepiej poradził sobie AR(21), więc on może okazać się być nieco lepszy niż MA(20).

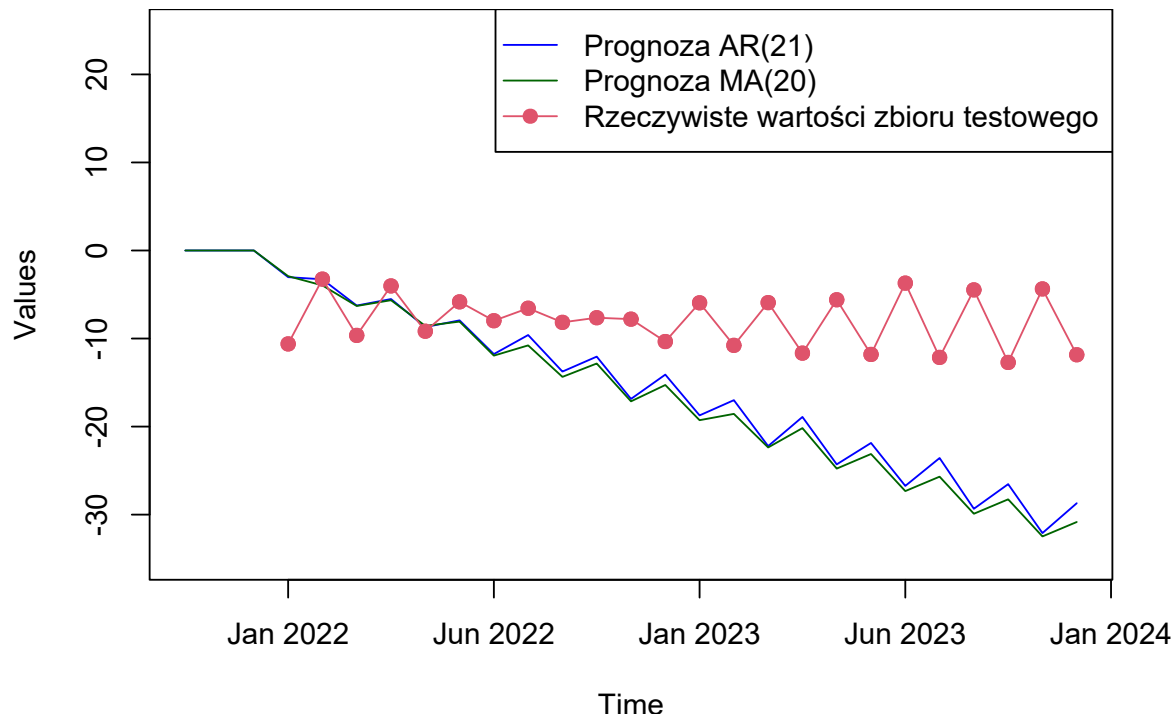
Forecasts from ARIMA(0,0,20) with non-zero mean



Forecasts from ARIMA(0,0,20) with non-zero mean



Możemy spróbować teraz odwrócić różnicowanie, aby zastosować modele AR(21) i MA(20) bezpośrednio na danych oryginalnych. Zastosujemy funkcję *diffinv* i najpierw odwrócimy różnicowanie z lagiem 2 (dodanie sezonowości), a potem z lagiem 1 (dodanie trendu). Prognoza zostanie wyrysowana tylko dla wartości średnich.



Widzimy, że tak, jak niektóre początkowe wartości w miarę się zgadzają, to później zaczynają one zbyt szybko maleć w porównaniu do faktycznych wartości ze zbioru testowego.

Porównując wszystkie modele w kwestii danych oryginalnych można stwierdzić, że SARIMA poradziła sobie lepiej. Natomiast jeśli chodzi o dane stacjonarne, to AR(21) i MA(20) osiągnęły bardzo podobne wyniki.

7. Podsumowanie

W analizie szeregu czasowego skoncentrowaliśmy się na ocenie stacjonarności, przekształceniach danych, dopasowaniu modeli AR(p), MA(q), a także SARIMA. Po przeprowadzeniu diagnostyki wybieramy model $ARIMA(2,1,1)(2,1,0)[2]$ jako najlepiej dopasowany do naszych danych.

Model ten charakteryzuje się najniższymi wartościami kryteriów AIC, AICc i BIC, a reszty modelu spełniają założenia o losowości i normalności. Natomiast w porównaniu miar błędu prognozy, model $ARIMA(2,1,1)(2,1,0)[2]$ uzyskuje niższe wartości RMSE, MAE, MAPE i MASE, co świadczy o lepszej jakości prognoz. Co prawda przy ocenie długoterminowej prognozy widać pewne wyzwania związane z jej dokładnością. Mimo wszystko model skutecznie radzi sobie w pierwszych miesiącach, co może być korzystne w kontekście krótkoterminowej prognozy.

Warto jest jednak pamiętać, że ważne jest ciągle sprawdzanie modelu, oraz zdolność dostosowywania go do zmieniających się warunków. To kluczowe, aby nasze prognozy były nadal skuteczne, zwłaszcza w obliczu nietypowych i nieprzewidywalnych zdarzeń.