Szeregi czasowe – raport z projektu

Turystyka w Japonii w latach 2012-2024

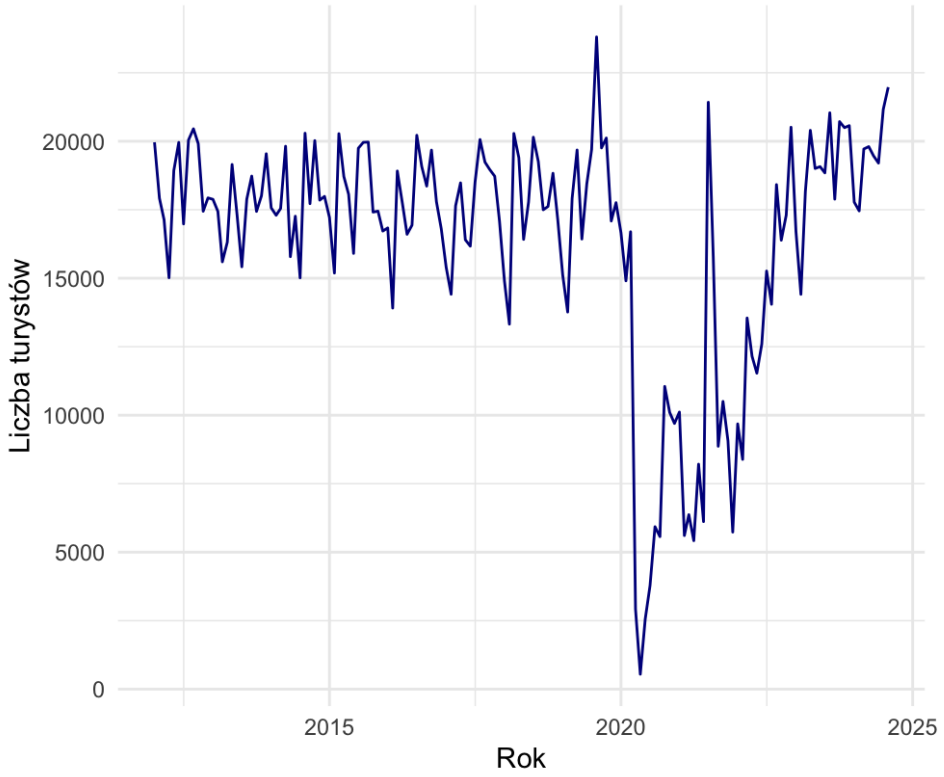
Julia Gąbka, Oktawia Hankus, Magdalena Potok

Wrocław 2025

Spis treści

1 Wstęp

Turystyka odgrywa kluczową rolę w gospodarce Japonii, będąc jednym z istotnych sektorów generujących wpływy finansowe oraz wspierających rozwój lokalnych społeczności. Na przestrzeni lat Japonia przyciągała miliony turystów z całego świata, dzięki swojej unikalnej kulturze, bogatej historii oraz różnorodnym atrakcjom. Jednak globalna pandemia COVID-19, która wybuchła na początku 2020 roku, miała dramatyczny wpływ na przemysł turystyczny na całym świecie, w tym również w Japonii. Okres tego wydarzenia przypada dokładnie od kwietnia 2020 roku do października 2022 roku, wtedy granice Japonii zostały zamknięte i ponownie otwarte.



Rysunek 1: Liczba turystów w Japonii w latach 2012-2024

Wykres przedstawia liczbę turystów odwiedzających Japonię w latach 2012–2024. Dane te pokazują względnie ustabilizowaną liczbę odwiedzających w latach poprzedzających pandemię. Natomiast począwszy od 2020 roku, widoczny jest dramatyczny spadek liczby turystów, związany z wprowadzeniem restrykcji podróżniczych i zamknięciem granic podczas pandemii COVID-19. W kolejnych latach widoczna jest powolna odbudowa.

Celem niniejszego projektu jest szczegółowa analiza szeregu czasowego dotyczącego liczby turystów odwiedzających Japonię w ostatnich latach. W ramach tego badania zostaną przeanalizowane dane historyczne oraz zmiany w liczbie odwiedzających w okresie przed pandemią oraz w jej trakcie.

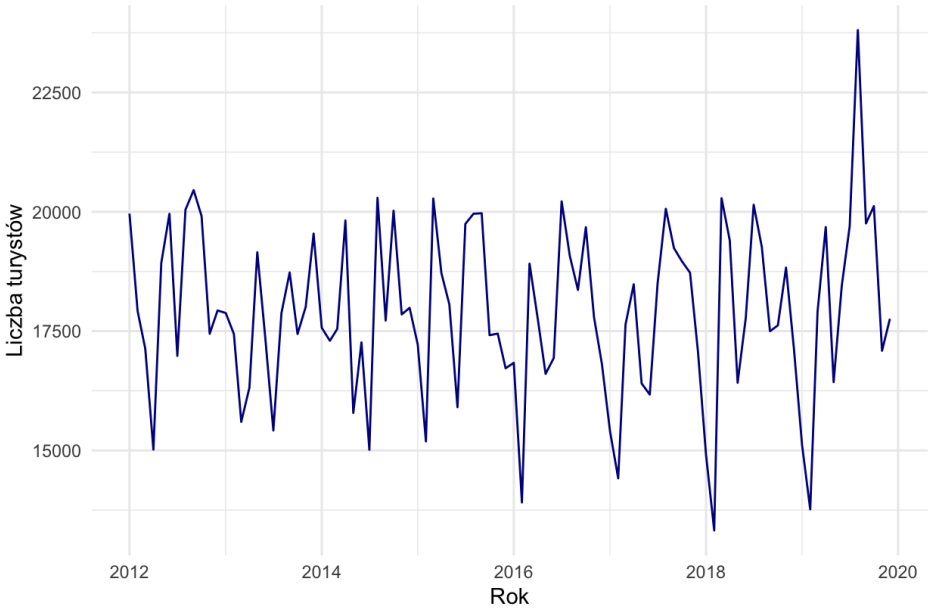
Projekt ten ma na celu nie tylko zrozumienie wpływu pandemii na japońską branżę turystyczną, ale również dostarczenie wartościowych informacji, które mogą zostać wykorzystane w strategiach odbudowy turystyki w przyszłości. Analiza taka może pomóc w identyfikacji kluczowych trendów oraz ocenie potencjalnego tempa odbudowy ruchu turystycznego w Japonii.

2 Analiza danych z lat 2012-2019

W pierwszej części zajmiemy się danymi z lat 2012-2019, czyli czasem sprzed epidemii Covid-19. Dane z tego okresu mają jasną, niezakłóconą nieoczekiwanymi wydarzeniami strukturę, która pozwoli nam na dokładną analizę wszystkich komponentów szeregu.

2.1 Opis danych

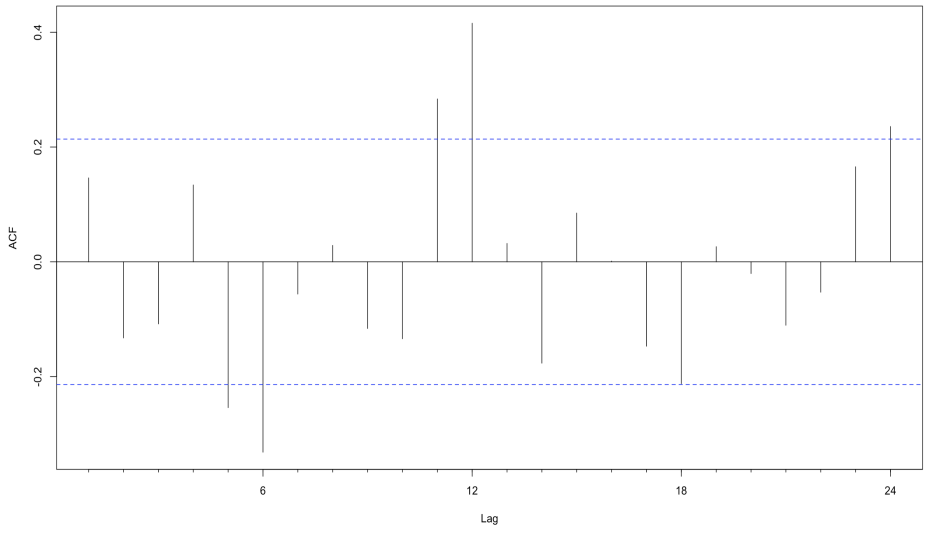
Poniższy wykres przedstawia zaobserwowane dane od stycznia 2012 roku do grudnia 2019.



Rysunek 2: Dane z lat 2012-2019

Na wykresie obserwujemy brak szczególnie widocznego trendu, roczną sezonowość oraz kilka wartości odstających, szczególnie w drugiej połowie 2019 roku. W październiku 2019 roku miało miejsce w Japonii Grand Prix Formuły 1, które odbyło się na torze Suzuka. Wydarzenie to przyciągnęło dużą liczbę turystów z całego świata, co mogło znacząco wpłynąć na wzrost liczby odwiedzających w tym okresie.

Dane podzielono na dwie części: treningowe, na podstawie których chcemy stworzyć model oraz testowe, na których będziemy sprawdzać jego poprawność. Dane treningowe to lata 2012-2018, natomiast testowe to ostatnie 12 miesięcy. Spójrzmy więc na wykres ACF treningowej części danych.



Rysunek 3: ACF danych treningowych

Z powyższego wykresu możemy odczytać, że kolejne obserwacje są ze sobą skorelowane.

2.2 Dekompozycja

Dekompozycja szeregu czasowego polega na podziale szeregu na poszczególne składowe: trend, sezonowość oraz resztę. Proces ten można przeprowadzić w dwóch głównych formach: addytywnej i multiplikatywnej.

W przypadku dekompozycji addytywnej szereg czasowy opisuje równanie:

$$Y_t = T_t + S_t + R_t$$

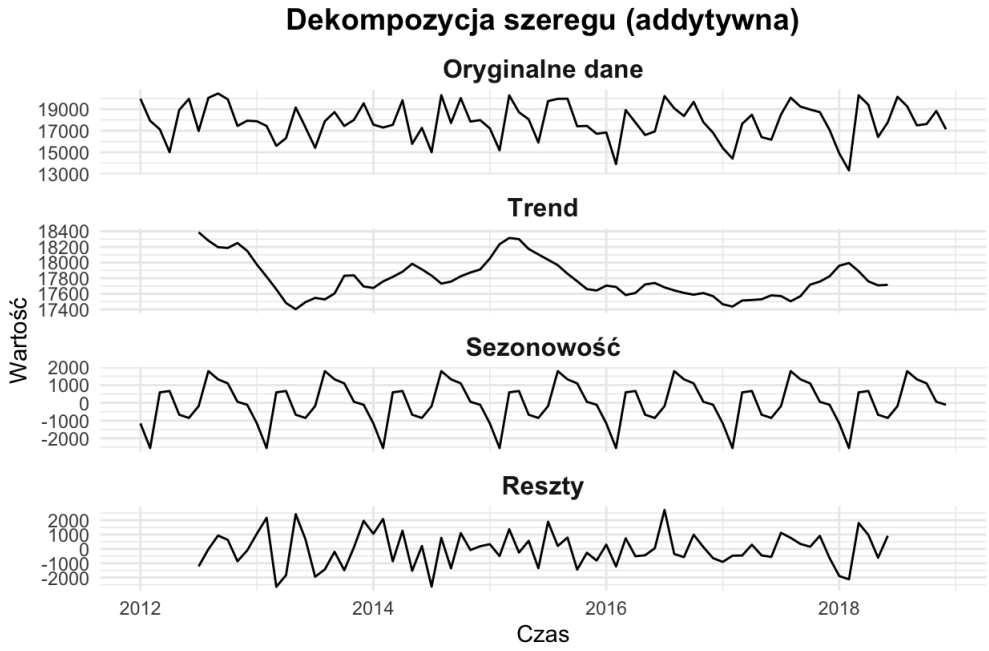
Natomiast w dekompozycji multiplikatywnej stosuje się równanie:

$$Y_t = T_t \times S_t \times R_t$$

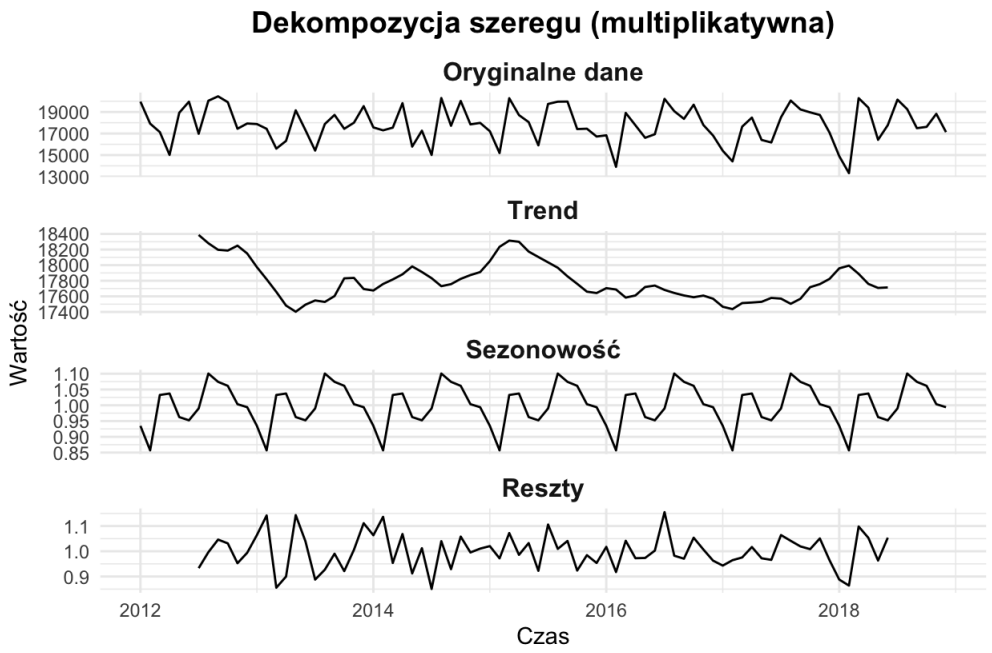
Gdzie:

- Y_t to wartość szeregu czasowego w chwili t,
- T_t to składowa trendu,
- S_t to składowa sezonowości,
- R_t to reszta.

Poniżej zaprezentowano wyniki dekompozycji analizowanego szeregu czasowego.



Rysunek 4: Dekompozycja addytywna

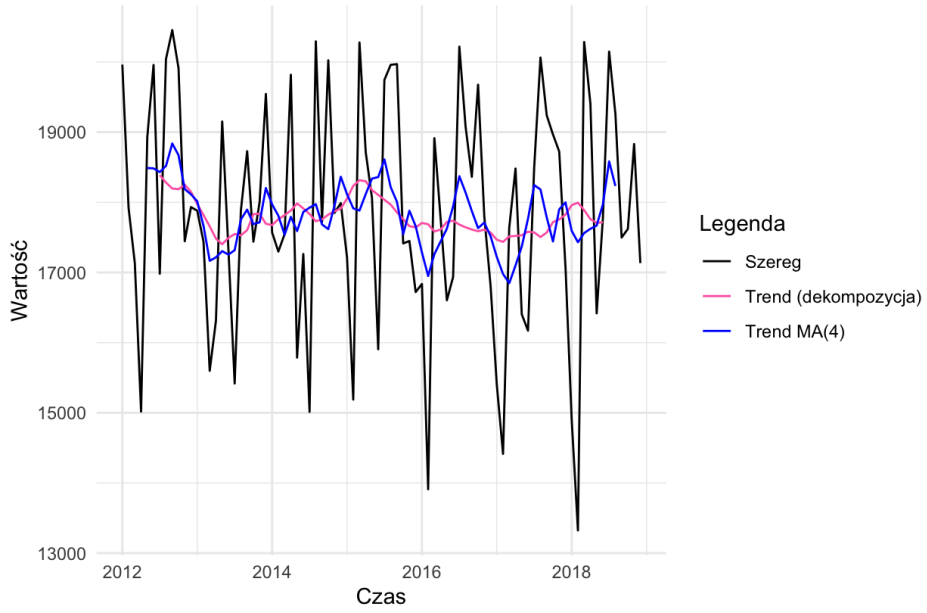


Rysunek 5: Dekompozycja multiplikatywna

Wykresy ukazujące dekompozycje są niemal identyczne. Oba potwierdzają sezonowość danych oraz ukazują stały trend. Największą różnicą pomiędzy metodami jest wykres reszt, a dokładniej jego skala. W przypadku metody addytywnej reszty sięgają 2000, natomiast w drugiej metodzie wahają się w zakresie 0.9-1.1.

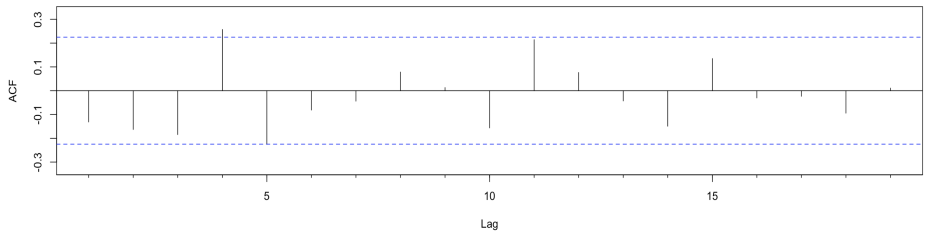
2.3 Metoda średniej ruchomej

Wyestymujemy trend metodą średniej ruchomej (MA(4)) oraz użyjemy go do dekompozycji szeregu, a następnie porównamy reszty z poprzednim przykładem.

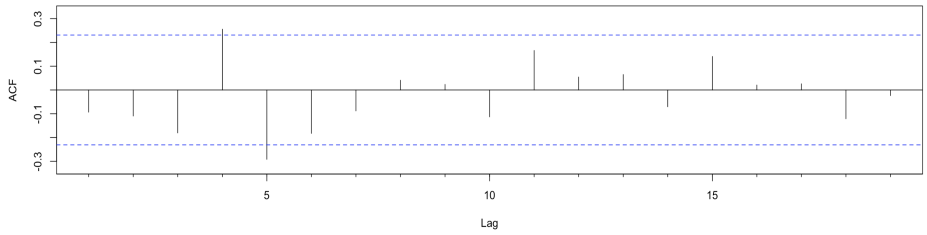


Rysunek 6: Trendy

Wykres prezentuje oryginalne wartości szeregu oraz trend wyestymowany dwiema metodami. Estymacja metodą średniej ruchomej zdaje się być dokładniejsza, przyjrzymy się wynikom porównując wykresy ACF dla reszt z obu modeli.



Rysunek 7: ACF reszt po dekompozycji metodą średniej ruchomej



Rysunek 8: ACF reszt po automatycznej dekompozycji

Wykresy nie różnią się znacznie, jednak sugerują, że metoda średniej ruchomej rzędu 4 może być bardziej adekwatna, w przypadku dekompozycji badanego szeregu.

2.4 Dopasowanie modeli

Modele ARIMA stanowią jedną z metod analizy i prognozowania szeregów czasowych. W ramach tej kategorii modeli wyróżnia się różne typy, takie jak modele autoregresji (AR), modele średniej ruchomej (MA), modele mieszane (ARMA) oraz modele zintegrowane (ARIMA). Modele ARIMA wymagają przeprowadzenia operacji różnicowania pierwszego lub wyższego rzędu, które mają na celu uzyskanie stacjonarności szeregu czasowego poprzez eliminację trendu.

Model ARIMA(p,d,q) opisuje się równaniem:

$$\phi(L)\nabla^d X_t = \Theta(L)W_t, \quad W_t \sim WN(0, \sigma^2),$$

gdzie:

- X_t to wartość szeregu czasowego w chwili t,
- $\phi(L)$ to operator autoregresji AR(p), który można zapisać jako:

  $$\phi(L) = 1 - \phi_1 L - \phi + 2L^2 - \cdots - \phi_p L^p,$$

  gdzie $\phi_1, \phi_2, \ldots, \phi_p$ to współczynniki autoregresji, a L to operator przesunięcia w czasie, czyli $L^k X_t = X_{t-k}$,
- ∇^d to operator różnicowania stopnia d, zdefiniowany jako $\nabla^d = (1 - L)^d$, gdzie d to liczba koniecznych różnicowań, aby szereg stał się stacjonarny,
- $\Theta(L)$ to operator średniej ruchomej MA(q), zapisany jako

  $$\Theta(L) = 1 + \theta_1 L + \theta_2 L^2 + \cdots + \theta_q L^q,$$

  gdzie $\theta_1, \theta_2, \ldots, \theta_q$ to współczynniki średniej ruchomej,

- W_t to niezależne i z identycznym rozkładem zmienne losowe, których wartość oczekiwana to 0, a wariancja σ^2, w modelu interpretujemy je jako błąd w chwili t.

Model SARIMA$(p,d,q) \times (P,D,Q)_s$ opisuje się równaniem:

$$\phi(L)\Phi(L^s)\nabla^d(1-L^s)^D X_t = \theta(L)\Theta(L)W_t, \quad W_t \sim WN(0,\sigma^2),$$

gdzie część (p,d,q) jest taka sama jak dla modelu ARIMA, natomiast

- $\Phi(L^s)$ to operator sezonowej autoregresji SAR(P), zapisany jako:

  $$\Phi(L^s) = 1 - \Phi_1 L^s - \Phi_2 L^{2s} - \cdots - \Phi_P L^{Ps},$$

  gdzie $\Phi_1, \Phi_2, \ldots \Phi_P$ to współczynniki sezonowej autoregresji, a L^s to operator przesunięcia sezonowego,

- $(1-L^s)^D$ to operator sezonowego różnicowania stopnia D, gdzie D to liczba różnicowań sezonowych, a s to długość okresu sezonowego,

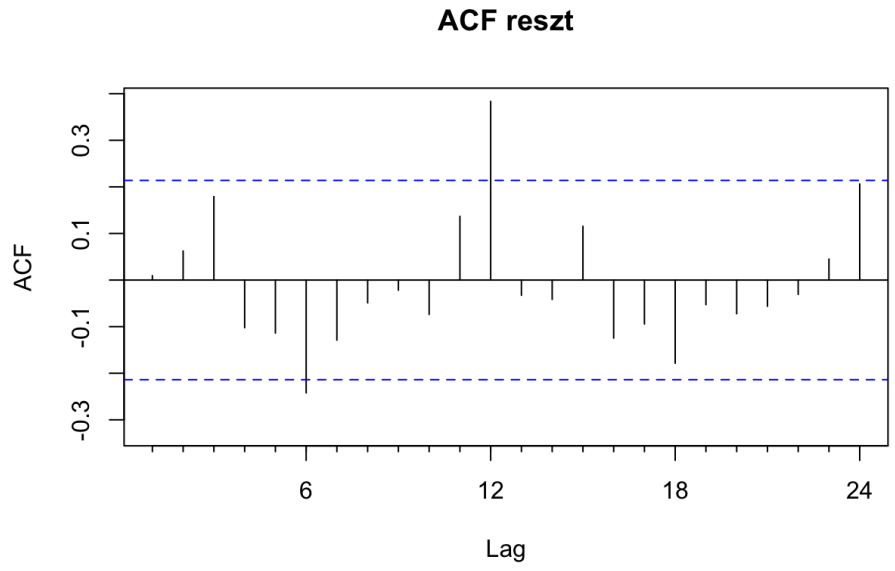
- $\Theta(L^s)$ jest operatorem sezonowej średniej ruchomej SMA(Q)

  $$\Theta(L^s) = 1 + \Theta_1 L^s + \Theta_2 L^{2s} \cdots + \Theta_Q L^{Qs},$$

  gdzie $\Theta_1, \Theta_2, \ldots, \Theta_Q$ to współczynniki sezonowej średniej ruchomej.

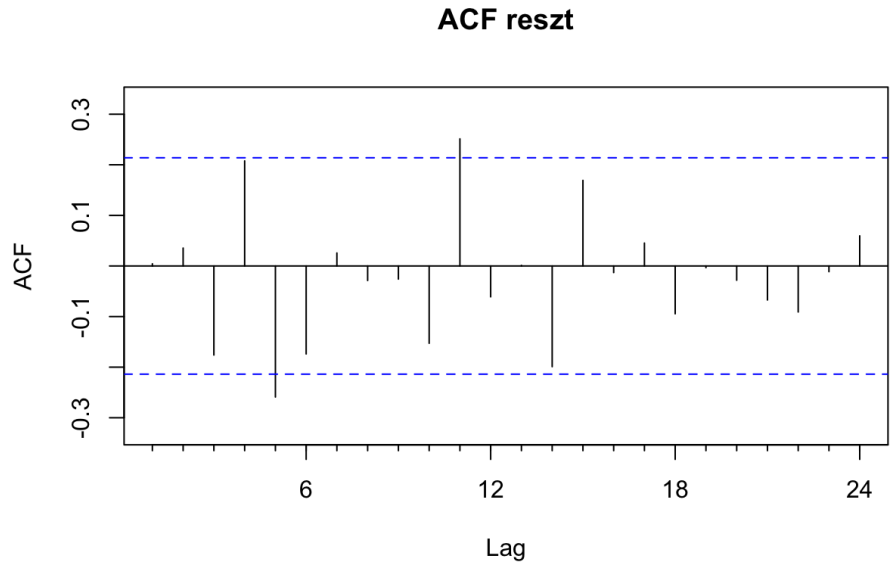
Postaramy się dopasować kilka modeli i sprawdzić ich adekwatność w przewidywaniu przyszłych wartości szeregu.

Pierwszym modelem będzie ręcznie dopasowany model ARIMA, a dokładniej ARMA, ponieważ analiza szeregu wykazała, że różnicowanie nie jest potrzebne. Zatem pierwszym modelem będzie model ARIMA(1,0,4). AIC w tym modelu wynosi 1477.6, BIC 1494.62 a tak wygląda wykres ACF residuów.



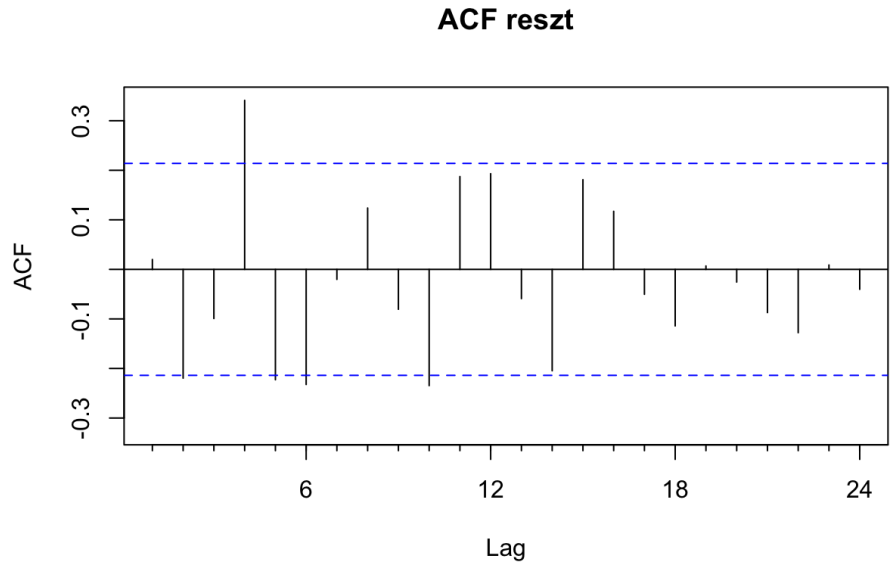
Rysunek 9: ACF reszt modelu ARIMA(1,0,4)

Drugim modelem będzie model dopasowany poprzez funkcję `auto.arima()`, który zaproponował następujące parametyry: SARIMA$(0,0,0) \times (1,0,0)[12]$. AIC w tym modelu wynosi 1466.98, BIC 1474.27, zatem mniej niż w poprzednim modelu, a tak wygląda wykres ACF residuów.



Rysunek 10: ACF reszt modelu SARIMA$(0,0,0) \times (1,0,0)[12]$

Trzecim modelem będzie model stworzony poprzez użycie funkcji `tslm()`, która dopasowuje do szeregu model liniowy. AIC w tym modelu wynosi 1475.88, wynik ten plasuje się pomiędzy dwoma poprzednimi, BIC 1509.91, czyli najwięcej z trzech modeli, a tak wygląda wykres ACF dla tego modelu.

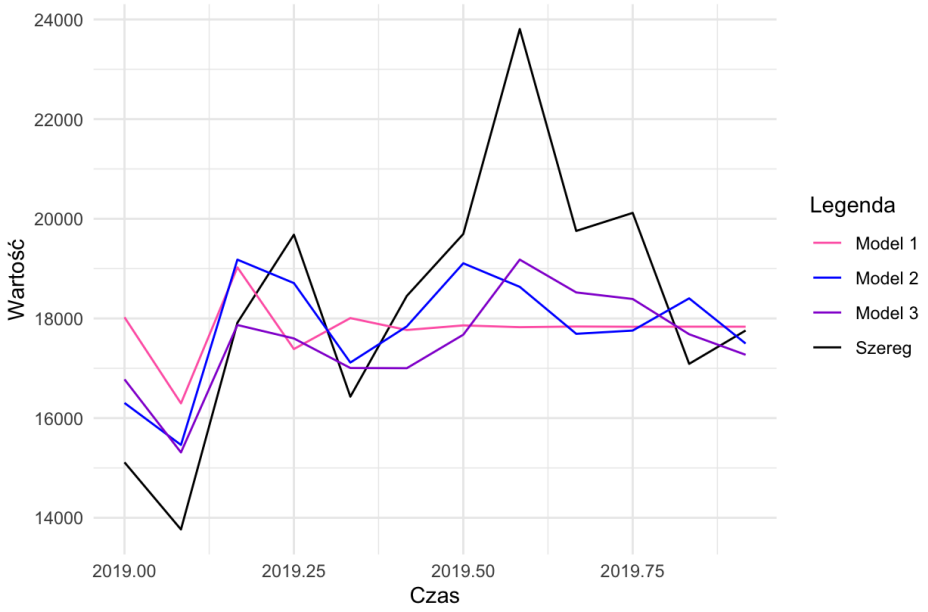


Rysunek 11: ACF reszt modelu liniowego

Tabela 1: Podsumowanie kryteriów informacyjnych

	AIC	BIC	df
Model 1	1477.6	1494.62	7
Model 2	1466.98	1474.27	3
Model 3	1475.88	1509.91	14

Według kryteriów AIC i BIC najlepiej dopasowanym modelem jest model nr 2, używający funkcji `auto.arima()`. Teraz chcemy zobaczyć predykcje każdego z modeli i porównać je z danymi testowymi.



Rysunek 12: Predykcje trzech modeli

Predykcje zostały przygotowane na cały 2019 rok. Jak widać początek roku jest względnie dobrze dopasowany, jednak żaden z modeli nie przewidział nagłego skoku, czego można było się spodziewać, gdyż była to obserwacja odstająca. Ponownie, całościowo najlepiej prezentuje się model nr 2, nr 1 w dalszej perspektywie nie dopasował się w ogóle do rzeczywistych danych, konkurencyjny zdaje się być model nr 3. Obliczymy błąd średniokwadratowy MSE, jego pierwiastek kwadratowy RMSE oraz średni błąd bezwzględny MAE, by porównać wybrane modele.

Tabela 2: Statystyki dla opisanych modeli

	MSE	RMSE	MAE
Model 1	6077812	2465.32	1995.60
Model 2	3867943	1966.70	1514.99
Model 3	3542331	1882.11	1504.34

Wbrew wcześniejszym przypuszczeniom, najmniejszym błędem średniokwadratowym oraz bezwzględnym wykazał się model nr 3. Jednak może być to spowodowane przez najlepszą predykcję obserwacji odstającej, przez co różnica jest mniejsza niż dla pozostałych. Może to skutkować poprawieniem statystyk. Poza poprawą w predykcji końcówki 2019 roku model liniowy gorzej odwzorowywuje zachowanie badanego szeregu na innych przedziałach czasowych. Trzeba zauważyć, że wszystkie błędy dla każdego z modeli są bardzo duże.

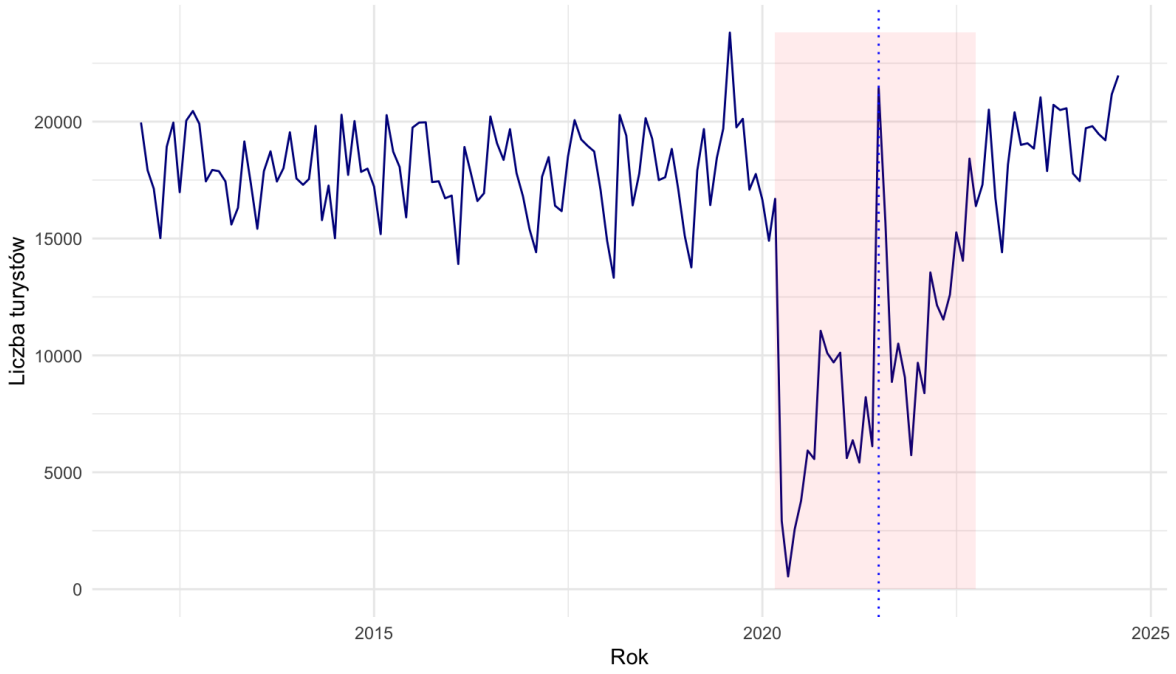
2.5 Wnioski

Biorąc pod uwagę wszystkie poprzednie rozważania, uznajemy że model nr 2, będący modelem zaproponowanym przez funkcję `auto.arima()` jest najlepszym dopasowaniem do badanego szeregu. Nie odzwierciedla on w pełni zmienności prawdziwych danych, jednak jest on najlepszym modelem jaki udało się skonstruować. Tym wnioskiem będziemy się kierować w dalszej analizie, dotyczącej lat dotkniętych COVIDem.

3 Analiza danych z lat 2012-2024

3.1 Wprowadzenie

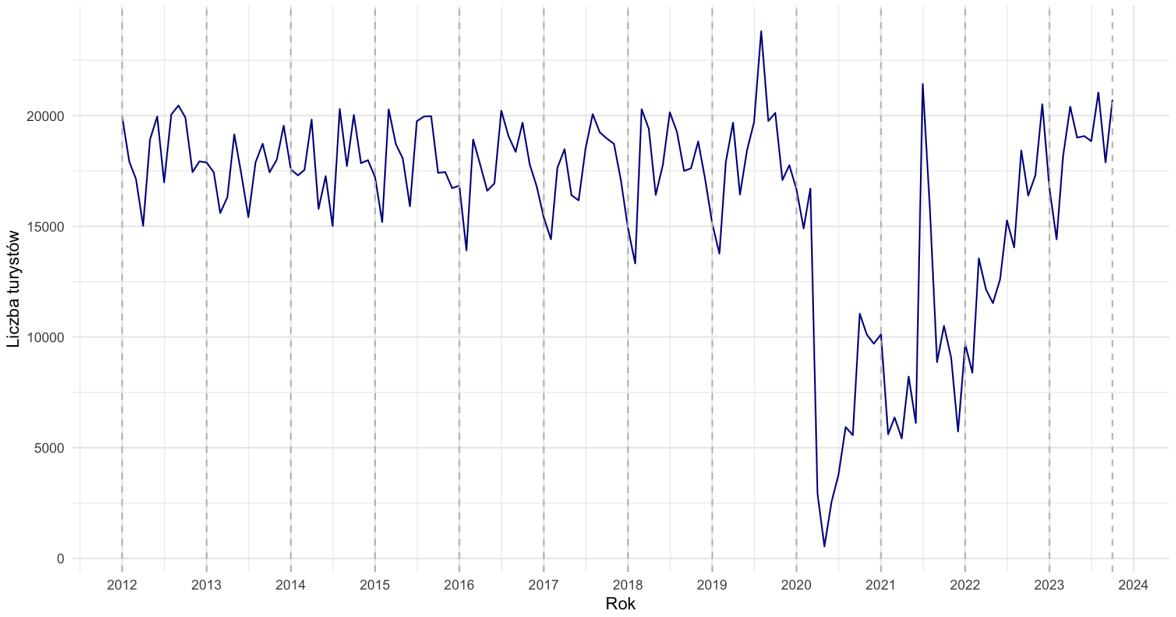
Powrócimy teraz do całości naszych danych, a więc przedmiotem analizy w tym rozdziale jest szereg czasowy przedstawiający dane dotyczące turystyki w Japoniii w latach 2012-2024. Chociaż dane obejmują okres zniekształcony przez pandemię COVID-19, podejmiemy próbę dopasowania modelu SARIMA w celu przewidywania liczby turystów. Z powodu pandemii spadek turystów w Japonii można zauważyć od marca 2020 roku, a otwarcie granic dla wszystkich turystów miało miejsce dopiero w październiku 2022 roku. Dodatkowo, w lipcu 2021 roku w Japonii miały miejsce przełożone o rok igrzyska olimpijskie w Tokio. Japonia w tym okresie była zamknięta dla międzynarodowych turystów, ale kraj otworzył się w ograniczonym stopniu dla osób związanych z organizacją igrzysk, uczestników oraz niektórych oficjalnych gości. Te wydarzenia miały kluczowy wplyw na strukturę danych oraz jakość modelu i prognoz.



Rysunek 13: Liczba turystów w Japonii w latach 2012-2024

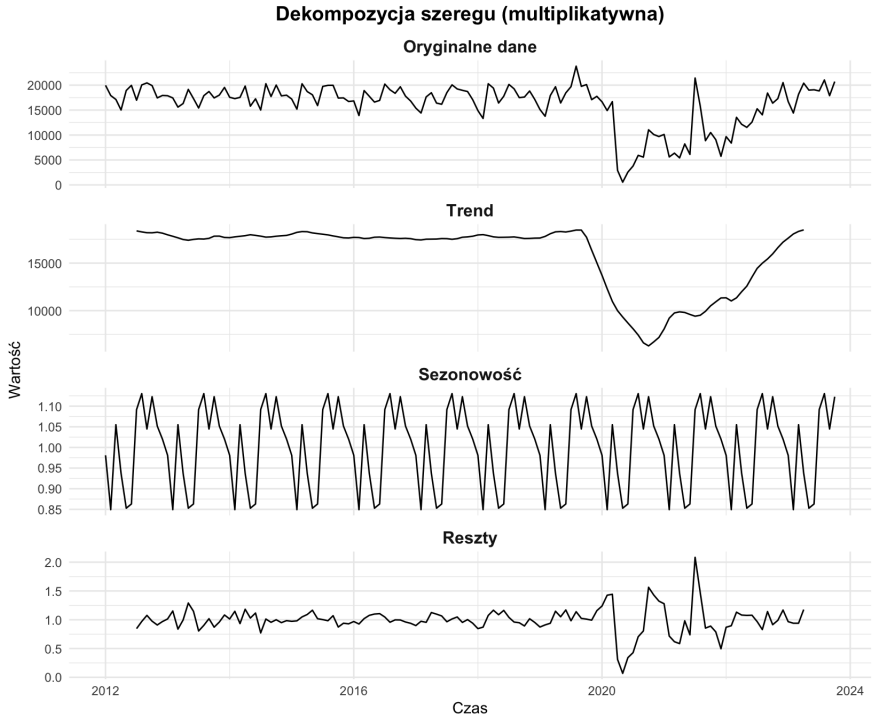
3.2 Analiza danych treningowych

Do modelowania wykorzystano dane treningowe obejmujące okres 2012-2023.10, a dane testowe stanowiły okres od listopada 2023 roku. Spójrzmy teraz na wykres danych treningowych naszego szeregu.



Rysunek 14: Wykres danych treningowych

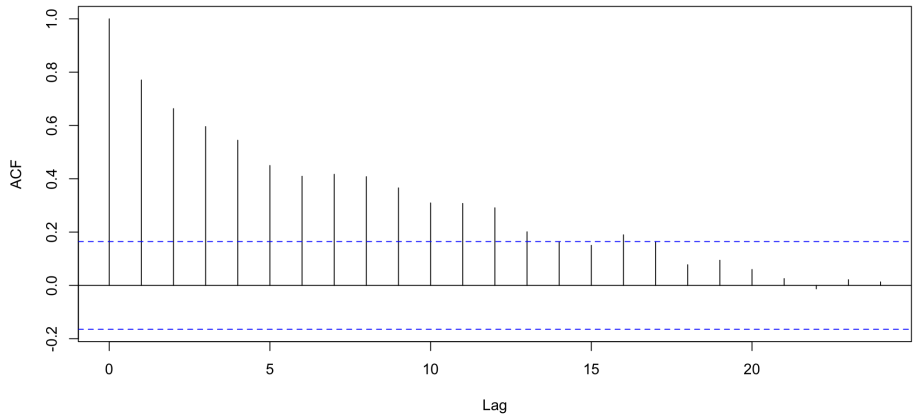
Na wykresie widać znaczące odchylenia w okresie pandemii, to właśnie przez nie stwierdzamy, że nasz szereg nie ma stałej ani średniej, ani wariancji. Ponadto, możemy również zauważyć sezonowy, roczny, wzorzec w naszych danych. Te informacje prowadzą nas do wniosku, że nasz szereg nie jest stacjonarny i powinien być przewidywany przez model uwzględniający sezonowość. Sprawdźmy teraz funkcję `ts_decommpose()` dla naszych danych treningowych.



Rysunek 15: Dekompozycja danych treningowych

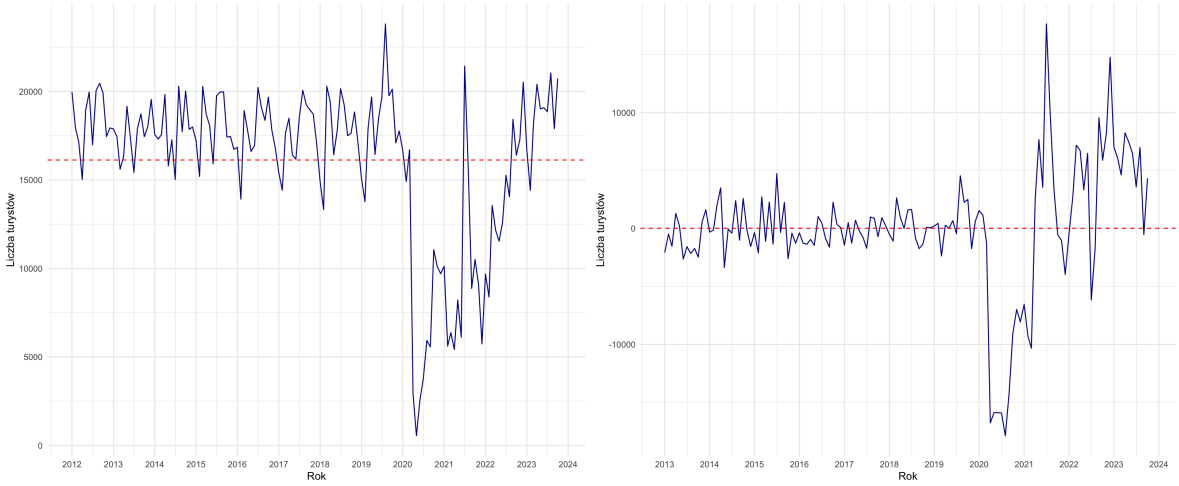
Dekompozycja szeregu czasowego potwierdziła sezonowość oraz ukazała zniekształcenia na wykresach związane z pandemią COVID-19. Trend spadkowy w tym okresie sugeruje zmniejszoną liczbę

turystów, co jest spójne z rzeczywistością. Spodziewamy się, więc, że nasz szereg nie jest stacjonarny, spójrzmy na wykres ACF naszego szeregu.



Rysunek 16: ACF danych treningowych

Na wykresie ACF widać wiele autokorelacji wykraczających poza poziom istotności, co sugeruje brak stacjonarności szeregu. W celu uzyskania stacjonarnego szeregu, można używać funkcji `diff()` i badać jego wygląd. W celu usprawnienia pracy użyłyśmy wbudowanych w R funkcji `ndiffs()` oraz `nsdiffs()`, które wykazały odpowiednio, że liczba potrzebnych różnicowań to: 1, natomiast liczba potrzebnych różnicowań sezonowych to: 0. Porównajmy teraz wykresy szeregu przed i po jednokrotnym różnicowaniu.

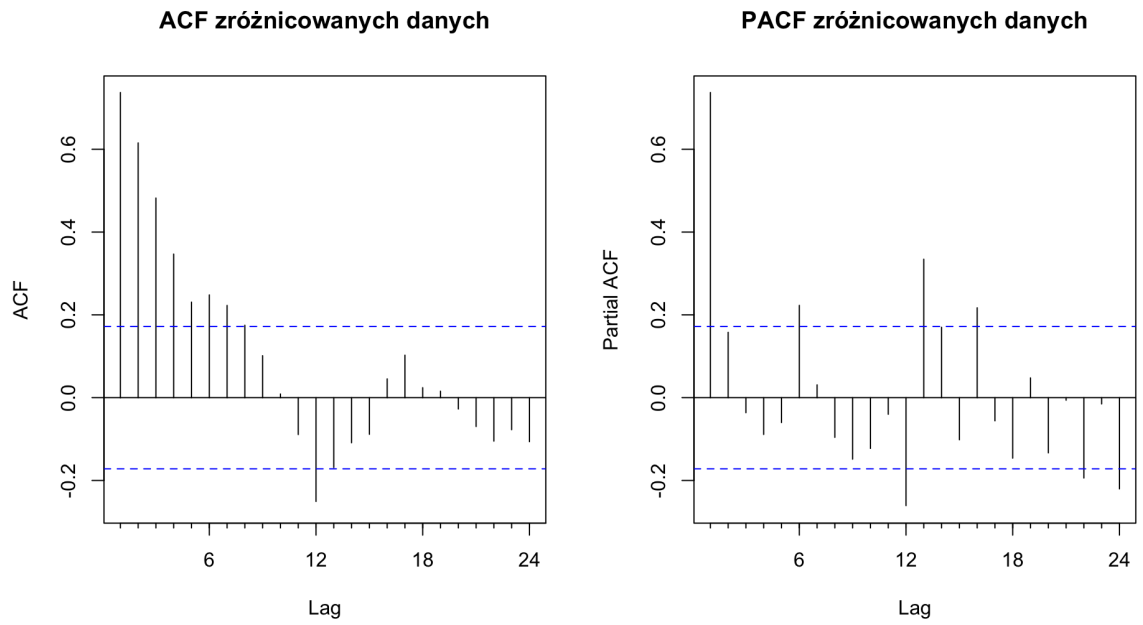


Rysunek 17: Wykres danych treningowych przed i po różnicowaniu

Na wykresie zostały czerwoną, przerywaną linią zaznaczone średnie obu szeregów. Widzimy, że po przeprowadzeniu różnicowania szereg stabilizuje się wokół stałej średniej. Sezonowość również jest mniej wyraźna, jednak w latach pandemicznych nadal obserwuje się odmienną wariancję, co sugeruje, że nasz szereg z latami COVID może być zbyt skomplikowany, aby modele poznane na wykładzie dobrze go naśladowały.

3.3 Dobór parametrów modelu SARIMA

Aby wybrać odpowiednie parametry modelów spójrzmy się na wykresy ACF i PACF dla zróżnicowanego szeregu.



Rysunek 18: ACF i PACF zróżnicowanego szeregu

Biorąc pod uwagę istotne opóźnienia wybrane zostały następujące wartości parametrów: $p = 1, 6,\ q = 1, 4, 6,\ P = 2,\ Q = 1$. Dla sezonowych parametrów wybierane zostały parametry na podstawie istotnych opóźnien na wykresach, które są w wielokrotnościach sezonowych $(12, 24, 36 \dots)$. Na podstawie analizy wykresów oraz złożoności modelu do dalszej analizy wybrano pięć modeli:

- **ARIMA**$(6, 1, 6)$ – ponieważ nie uwzględnia sezonowości wybrane zostały wysokie wartości parametrów p i q, aby zobaczyć, czy ta złożoność wystarczy do przewidywania danych sezonowych,
- **SARIMA**$(6, 1, 6) \times (2, 0, 1)[12]$ – model sezonowy, uwzględniający złożoność danych,
- **SARIMA**$(1, 1, 4) \times (2, 0, 1)[12]$ – prostszy model sezonowy, z mniejszą liczbą parametrów,
- **SARIMA**$(1, 1, 6) \times (2, 0, 1)[12]$ – umiarkowanie złożony model sezonowy,
- **SARIMA**$(1, 1, 1) \times (1, 0, 2)[12]$ – model wybrany automatycznie za pomocą funkcji `auto.arima()`.

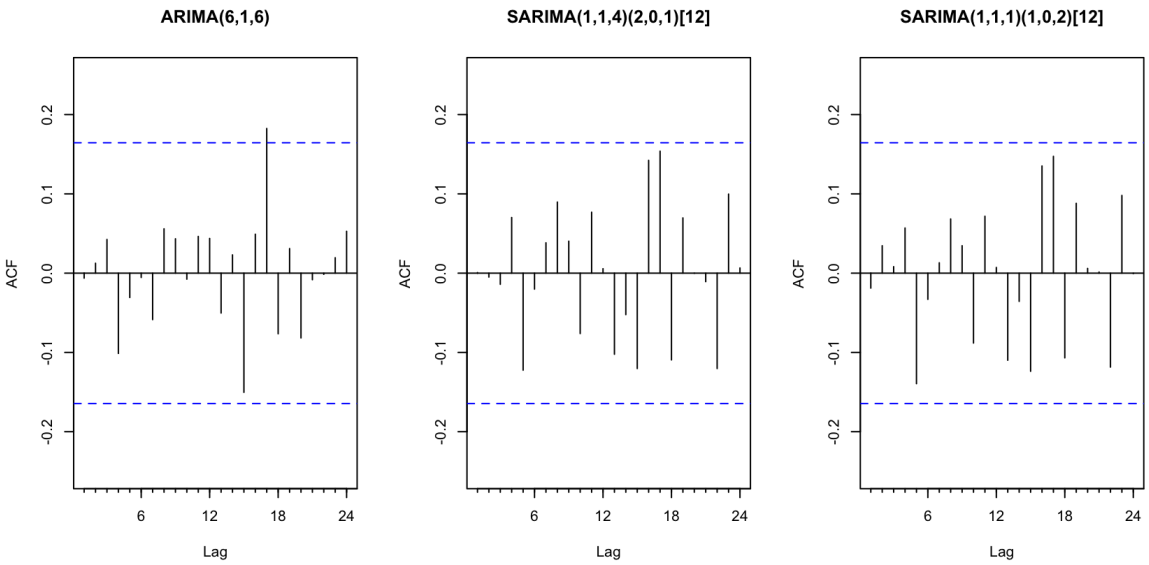
Następnie na podstawie kryteriów informacyjnych AIC i BIC porównamy powyższe modele.

Tabela 3: Kryteria informacyjne dla wybranych modeli

	AIC	BIC	df
ARIMA(6,1,6)	2636.65	2674.99	13
SARIMA(6,1,6)(2,0,1)[12]	2651.66	2698.84	16
SARIMA(1,1,4)(2,0,1)[12]	2645.49	2672.03	9
SARIMA(1,1,6)(2,0,1)[12]	2644.49	2676.93	11
SARIMA(1,1,1)(1,0,2)[12]	2640.11	2657.80	6

Model z najniższą wartością AIC to ARIMA$(6, 1, 6)$, najniższa wartość BIC ma model wybrany automatycznie SARIMA$(1, 1, 1) \times (1, 0, 2)[12]$ – jest to też model o najmniejszej złożoności, tylko 6 stopni swobody. Zaraz za nimi najlepiej w tabelce wypada model SARIMA$(1, 1, 4) \times (2, 0, 1)[12]$, ten model również weźmiemy pod uwagę w dalszej analizie.

W celu oceny jakości dopasowania przeanalizujemy reszty modeli przy użyciu wykresu ACF.

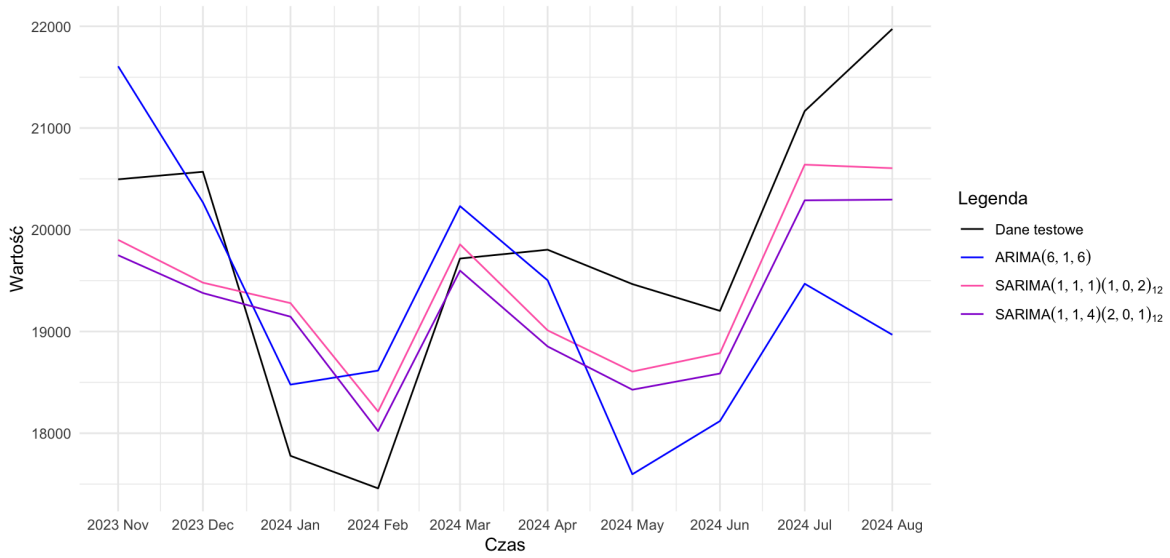


Rysunek 19: Wykresy ACF residuów wybranych modeli

Dla modelu ARIMA jedna z obserwacji wykracza poza przedział istotności, co może wskazywać na niewielkie problemy z modelem. Dla obu modeli typu SARIMA wszystkie obserwacje mieszczą się w przedziałach istotności, co wkazuje na brak autokorelacji reszt.

3.4 Predykcja oraz ocena jakości prognoz

Za pomocą funkcji `forecast()` przewidzimy, dla wcześniej wybranych trzech modeli, miesiące szeregu, które zostały na pocztątku 3. sekcji wybrane jako dane testowe. Sprawdźmy na wykresie jak poradziły sobie wybrane modele.



Rysunek 20: Prognoza wybranych modeli na miesiące szeregu testowego

Najlepiej odwzorowujący dane testowe jest model SARIMA$(1,1,1) \times (1,0,2)[12]$. Chociaż nie odwzoruje ich idealnie, nie przewiduje wzrostu turystów na okres wakacyjny 2024 roku, jednak jego prognozy są najbliższe rzeczywistym wartościom w większości przedziałów czasowych. Model SARIMA$(1,1,4) \times (2,0,1)[12]$ jest mniej dokładny w stosunku do drugiego modelu tego typu, za to ma lepsze wyniki od

ARIMA$(6,1,6)$. Najgorzej dopasowującym się modelem jest model ARIMA$(6,1,6)$, mimo złożoności i dużej ilości współczynników, ten model nie poradził sobie najlepiej z przewidywaniem danych testowych. Ten model ignoruje sezonowość, co może być jedną z głównych przyczyn gorszej wydajności.

Do pełnej oceny wyboru najlepszego modelu policzymy i porównamy statystyki MSE, RMSE oraz MAE.

Tabela 4: Statystyki dla wybranych modeli

	MSE	RMSE	MAE
ARIMA(6,1,6)	2009910	1417.71	1174.28
SARIMA(1,1,1)(1,0,2)[12]	807464.4	898.59	804.35
SARIMA(1,1,4)(2,0,1)[12]	1012417	1006.18	914.77

Najniższe wartości osiąga model wybrany automatycznie przez funkcję `auto.arima()`, czyli SARIMA$(1,1,1) \times (1,0,2)[12]$, oznacza to, że ten model odwozruje najlepiej dane testowe. Drugim na podium jest model SARIMA$(1,1,4) \times (2,0,1)[12]$, osiąga on gorsze wyniki od poprzedniego, wskazują one na średnią jakość dopasowania. Najgorszym, jak się już mogliśmy spodziewać, jest model ARIMA$(6,1,6)$, ten fakt potwierdza, że ignorowanie sezonowości powoduje znaczące problemy z dokładnością predykcji.

3.5 Wnioski

Modelowanie szeregu czasowego zawierającego dane z okresu pandemii COVID-19 stanowi szczególne wyzwanie. Zniekształcenia związane z zamknięciem turystyki, ponownym jej otwarciem oraz w międzyczasie organizowanie igrzysk olimpijskich, prowadzą do nietypowych wzorców w danych i wprowadzają trudności w dostosowaniu modeli statystycznych. W naszej analizie mogliśmy zauważyć, że dobrane modele znacząco odstawały od rzeczywistych danych testowych. Tak naprawdę żaden z modeli nie był w stanie idealnie przewidzieć rzeczywistej ilości turystów, szczególnie w okresie wakacyjnym 2024 roku. Wyniki z ostatniej tabeli, czyli błąd średnio-kwadratowy, pierwiastek kwadratowy błędu średnio-kwadratowego oraz średni błąd bezwzględny są dość duże, co świadczy o trudności modelowania danych z okresu pandemii.

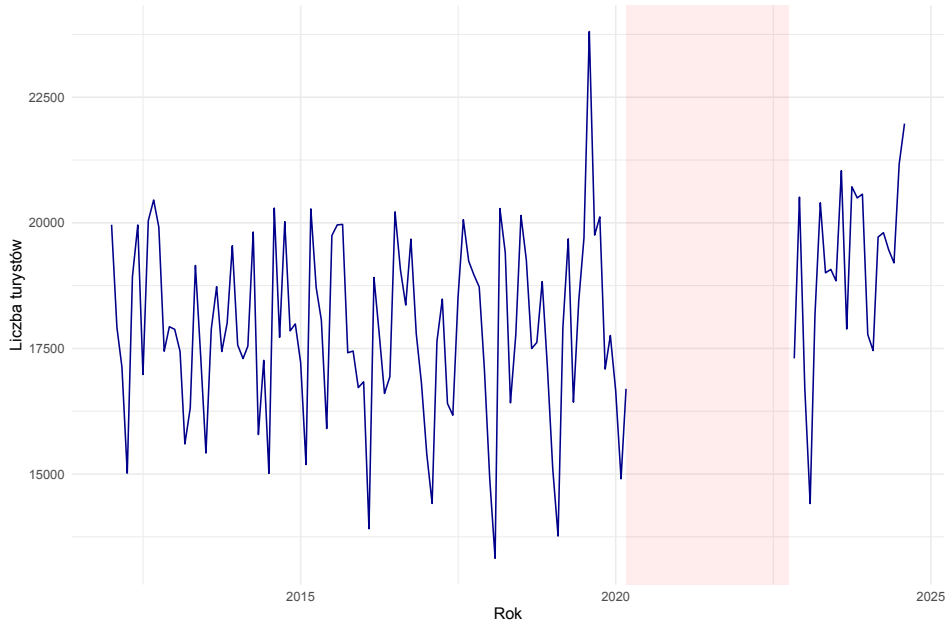
Model z najlepszymi wynikami, czyli SARIMA$(1,1,1) \times (1,0,2)[12]$, nie jest w stanie idealnie przewidzieć rzeczywistych wartości, co wynika z nieprzewidywalności wzorców związanych z pandemią COVID-19.

4 COVID-19

4.1 Wprowadzenie

Jak mogliśmy zauważyć modele stworzone z wykorzystaniem rzeczywistych danych z lat 2012-2024 nie przewidują nowych wartości zbyt dokładnie. Przez wybuch pandemii predykcje były często mniejsze niż rzeczywista liczba turystów od października 2023 roku. Teraz chcielibyśmy rozważyc hipotetyczne pytanie, "A co jeśli pandemii by nie było?". Czy predykcja nowych wartości będzie bardziej dokładna, kiedy najpierw wysetymujemy wartości danych podczas pandemii?

Rozpoczniemy od usunięcia danych covidowych ze zbioru.



Rysunek 21: Liczba turystów w latach 2012-2024 z wyłączeniem danych covidowych

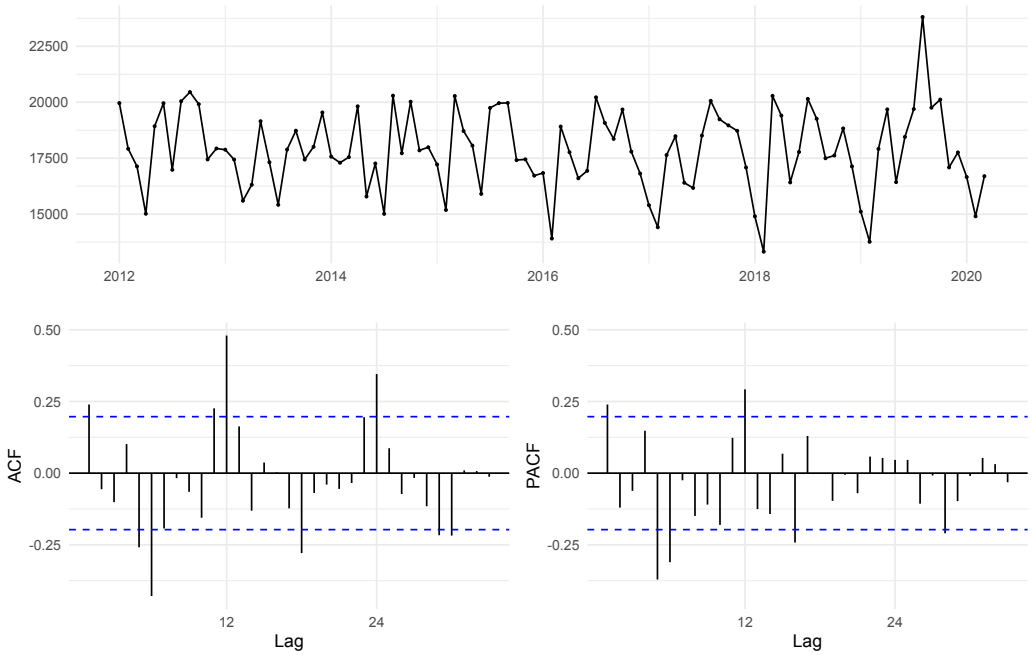
Wciąż pracujemy na danych od początku 2012 roku do sierpnia 2024 roku. Zbiorem treningowym są wartości pomiędzy styczniem 2012 a październikiem 2023, pozostałe to zbiór testowy.

Rozważymy kilka sposobów uzupełniania danych. Jako pierwsze, spróbujemy dokonać predykcji jednocześnie na podstawie przeszłych i przyszłych danych. Zobaczymy, że w naszym przypadku jest to utrudnione ze względu na małą liczbę obserwacji występujących po pandemii. Zatem spróbujemy również ograniczyć się jedynie do wykorzystania danych sprzed COVID-19. Następnie spróbujemy użyć funkcji `R` mających ułatwić nam takie operacje.

Jako, że pierwsza część pracy skupia się na ręcznym dobraniu odpowiednich modeli, po dokonaniu odpowiedniego uzupełnienia, model dla całości dobieramy za pomocą funkcji `auto.arima()`.

4.2 Tworzenie modeli

Rozpocznijmy więc od zobrazowania danych przed wybuchem pandemii.

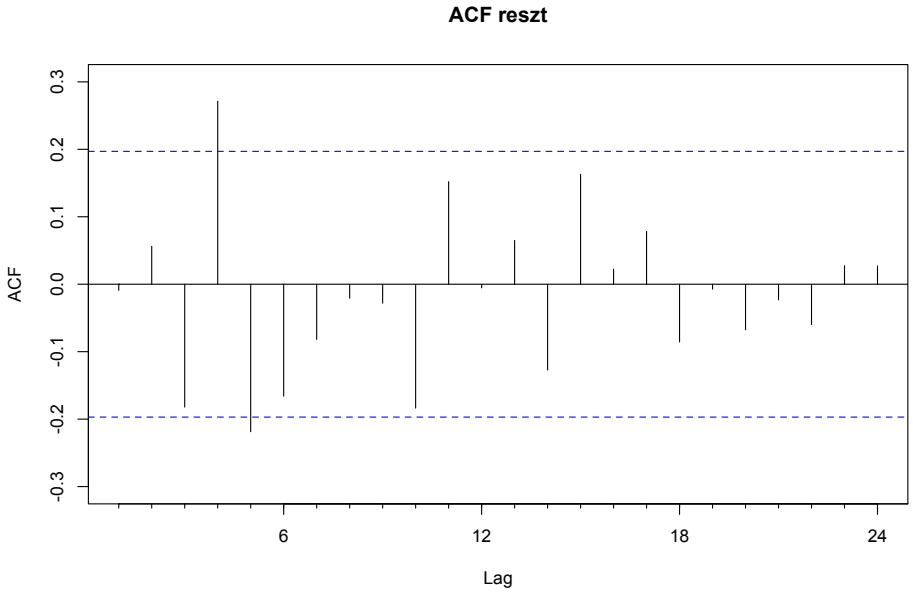


Rysunek 22: Liczba turystów w Japonii przed pandemią COVID-19

Są to dane prawie takie same jak w pierwszej sekcji, uwzględniamy jednak jeszcze część 2020 roku.

Obie funkcje `ndiffs()` oraz `nsdiffs()` wykazują, że różnicowanie szeregu jest niepotrzebne. Z wykresów ACF oraz PACF odczytujemy możliwe wartości parametrów modeli. Wybieramy zatem $p = 1, q = 1, P = 1, Q = 1$ (wykorzystując informację o P i Q jako krotnościach okresu).

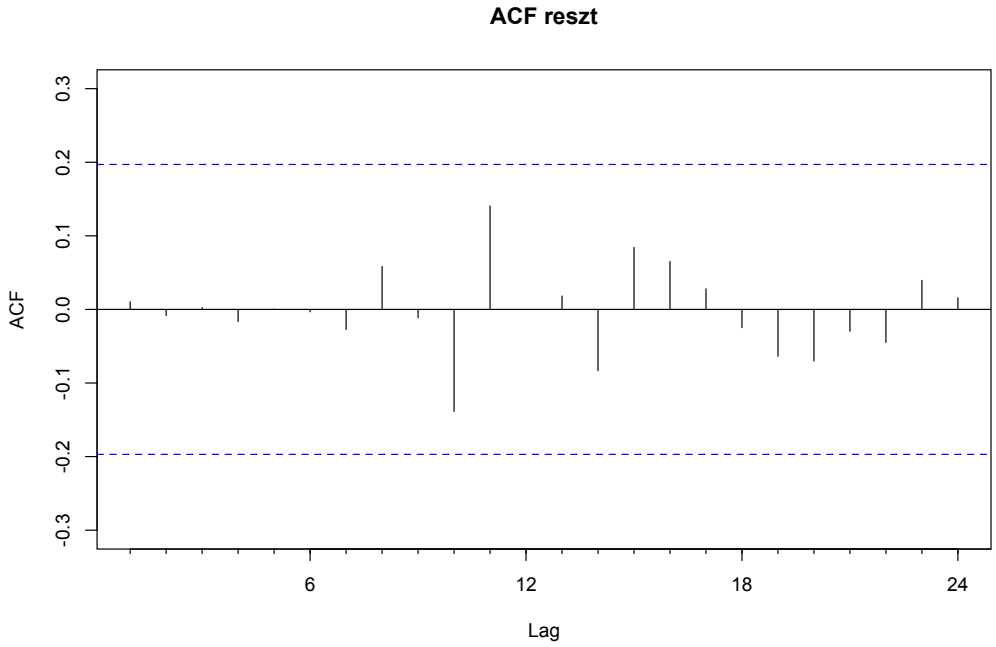
Otrzymujemy następujący wykres ACF dla residuów.



Rysunek 23: Residua - SARIMA$(1, 0, 1) \times (1, 0, 1)[12]$

Widzimy korelację pomiędzy kolejnymi obserwacjami. Oznacza to, że powinniśmy spróbować dobrać lepsze dopasowanie.

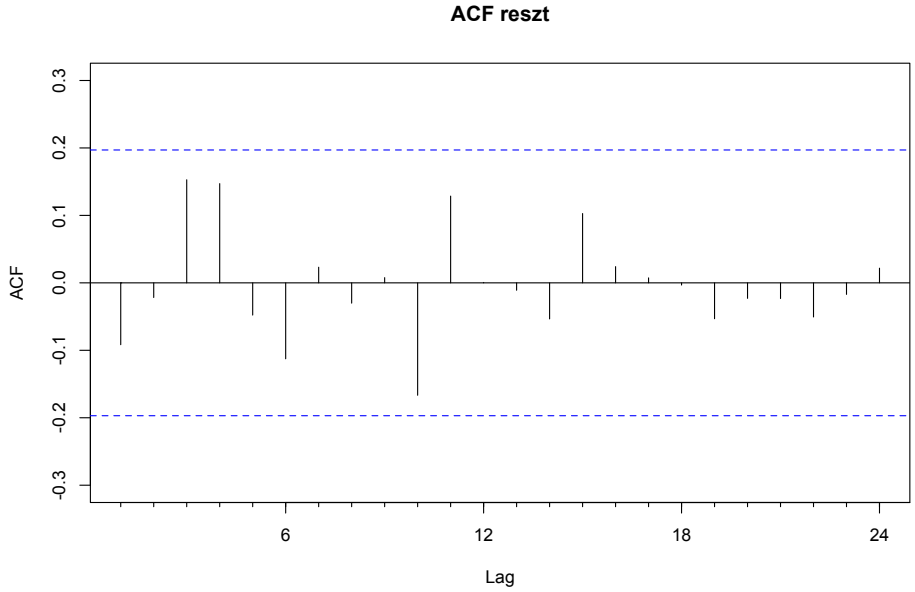
Spróbujmy więc wybierając największe piki na wykresach występujące przed $lag = 12$, stąd $q = 6$, a $p = 5$.



Rysunek 24: Residua - SARIMA$(5, 0, 6) \times (1, 0, 1)[12]$

Tym razem na wykresie ACF residuów wszystkie wartości są w przedziałach ufności. Sugeruje to, że szum jest losowy. Jednak wybrane rzędy są dość duże, spróbujemy je zmniejszyć analizując kolejne modele.

Tak dochodzimy do modelu SARIMA$(2, 0, 3) \times (1, 0, 1)[12]$. Wtedy ACF reszt prezentuje się następująco.

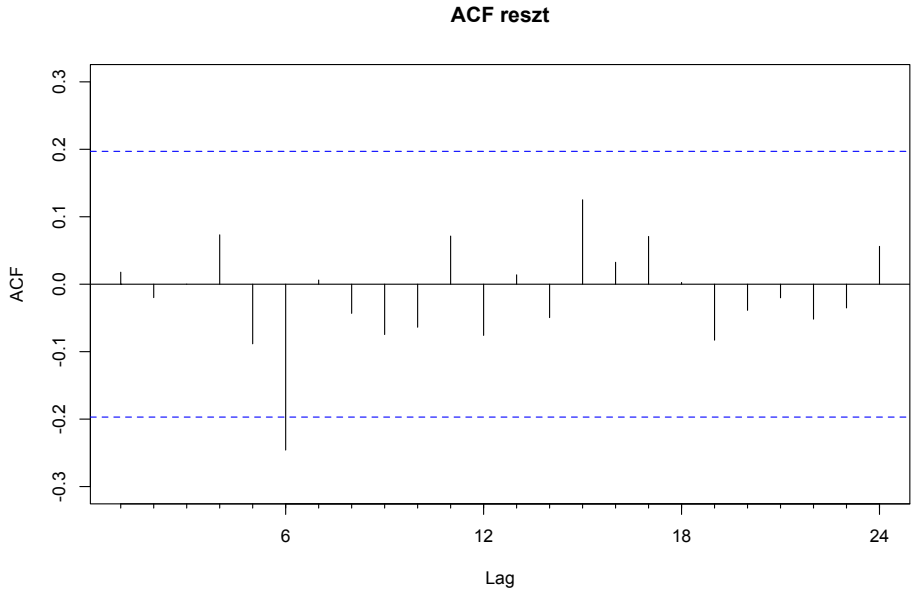


Rysunek 25: Residua - SARIMA$(2, 0, 3) \times (1, 0, 1)[12]$

Kolejne pomniejszenie któregoś z rzędów prowadzi do ukazania korelacji pomiędzy dalszymi wartościami.

Wartość AIC dla takiego modelu to 1720.72, a błąd średnio-kwadratowy wynosi 1582195.

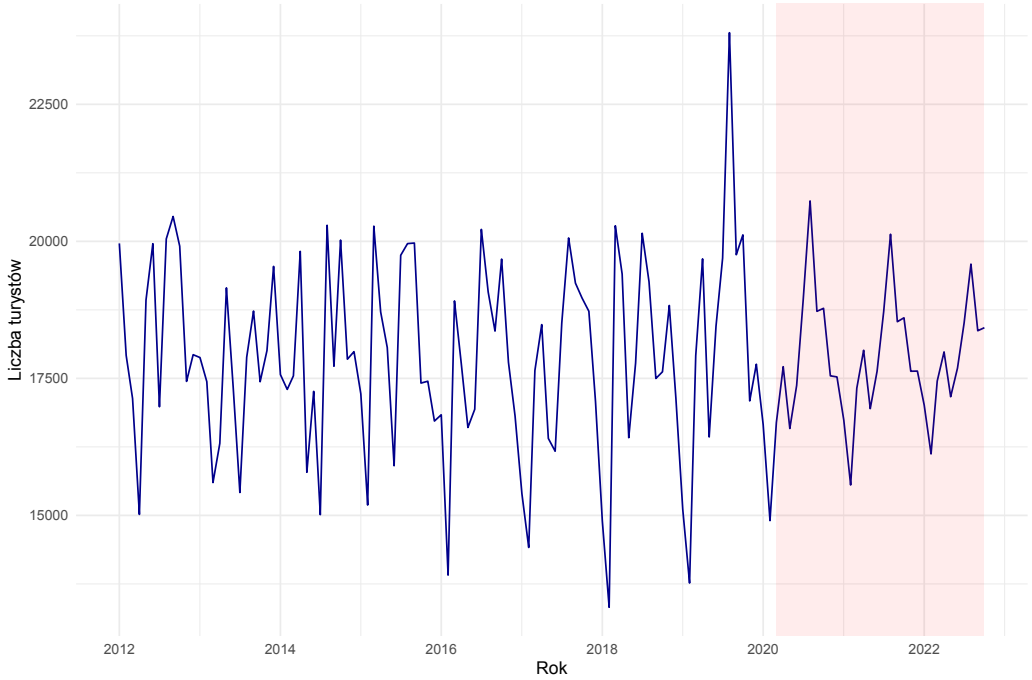
Model wybrany przez funkcję `auto.arima()` to SARIMA$(2,0,2) \times (1,0,0)[12]$. Wykres ACF jego reszt prezentuje się następująco.



Rysunek 26: Residua - SARIMA$(2,0,2) \times (1,0,0)[12]$

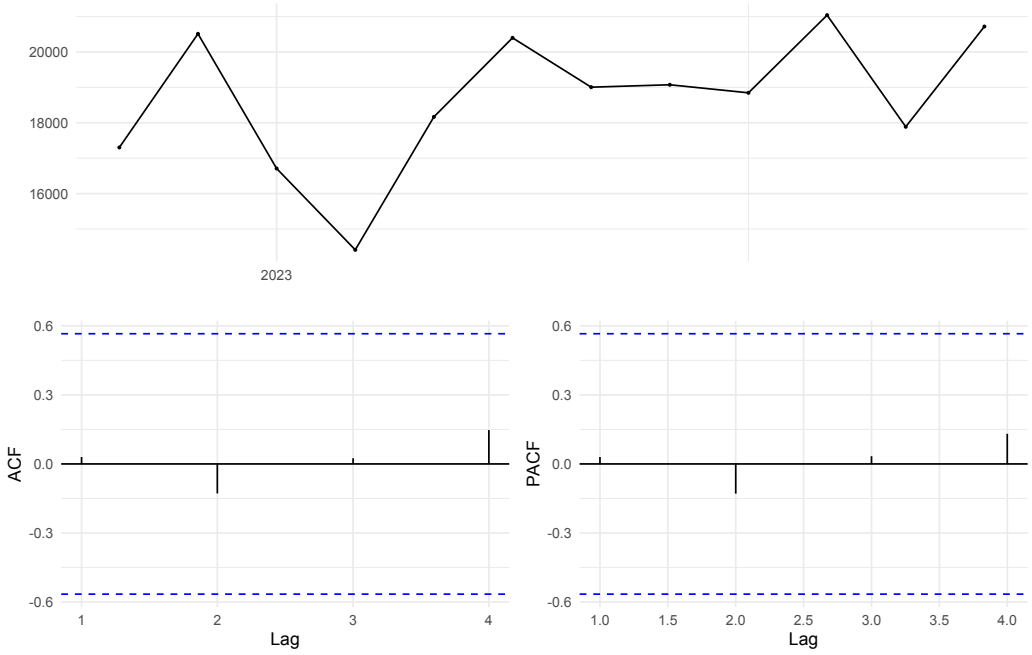
AIC wtedy wynosi 1719.69, widzimy, że nie jest dużo mniejszy od tej wartości dla wybranego przez nas modelu. Błąd średnio-kwadratowy tego modelu jest większy i wynosi 1730275. Przewagą ręcznie dobranego modelu jest również to, że nie występuje autokorelacja dla reszt.

Do dalszej analizy wybieramy pierwszy model - SARIMA$(2,0,3) \times (1,0,1)[12]$. Na jego podstawie estymujemy dane dla przedziału czasowego pandemii.



Rysunek 27: Przed pandemią COVID-19 z predykcją

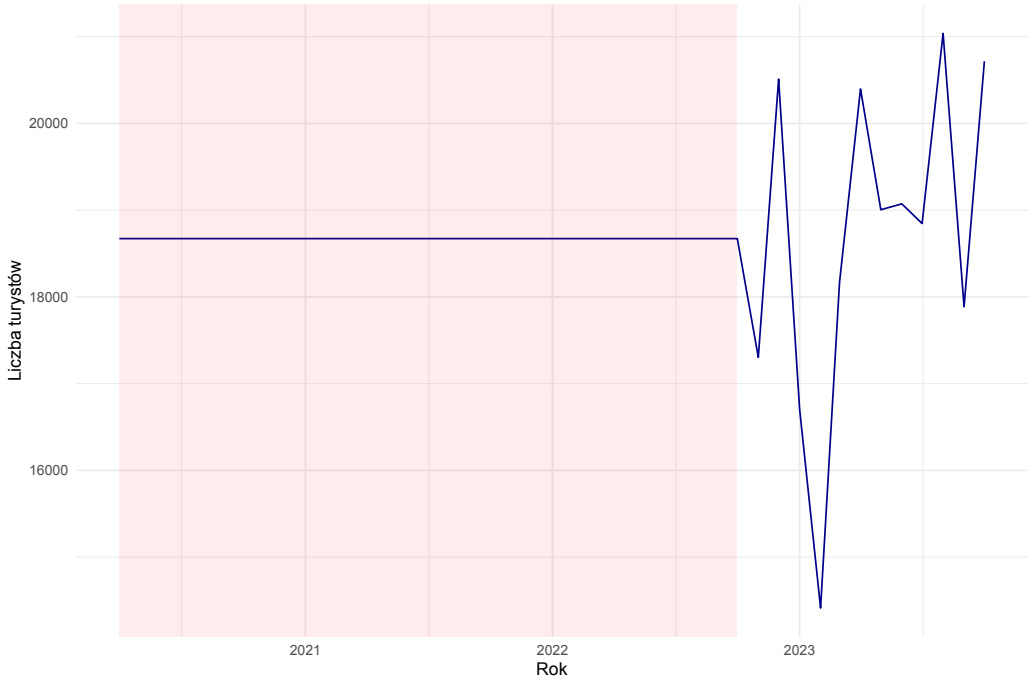
Teraz chcielibyśmy zrobić taką samą analizę na danych po pandemii, a następnie uśrednić uzyskane wyniki.



Rysunek 28: Po pandemii COVID-19

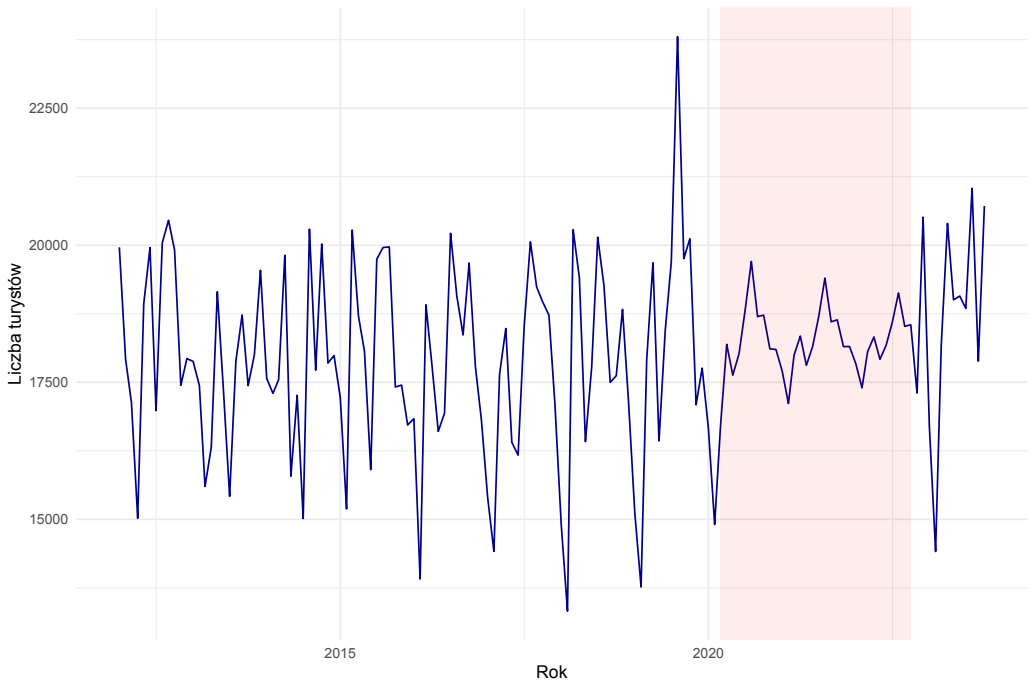
Niestety dysponujemy zbyt małą liczbą takich obserwacji (szczególnie biorąc pod uwagę jedynie zbiór treningowy, jednak uwzględnienie także zbioru testowego nie poprawia sytuacji). Możemy zauważyć, że wykersy ACF oraz PACF nie ukazują żadnych znaczących powiązań między danymi. Nie potrafimy dobrać możliwych rzędów parametrów modeli. Również `auto.arima()` wskazuje, że właść-

ciwym modelem, byłaby ARIMA$(0, 0, 0)$, co oznacza, że policzyliśmy jedynie średnią.



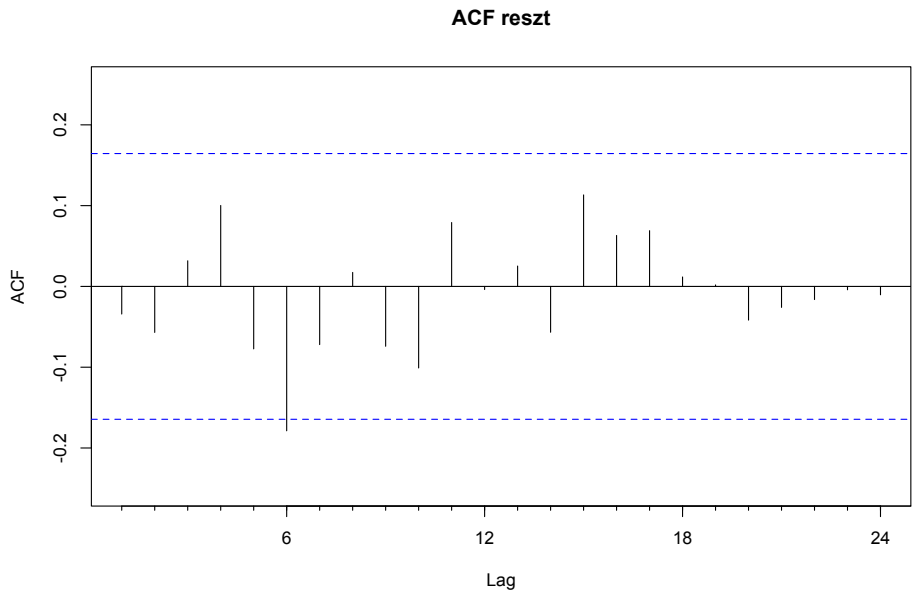
Rysunek 29: Po pandemii COVID-19 z predykcją

Taka predykcja nie mówi za wiele o dokładnych wartościach, ale możemy spróbować uśrednić te wartości z otrzymanymi w poprzednim kroku.



Rysunek 30: Uzupełnienie pandemii COVID-19 z użyciem wartości przed i po

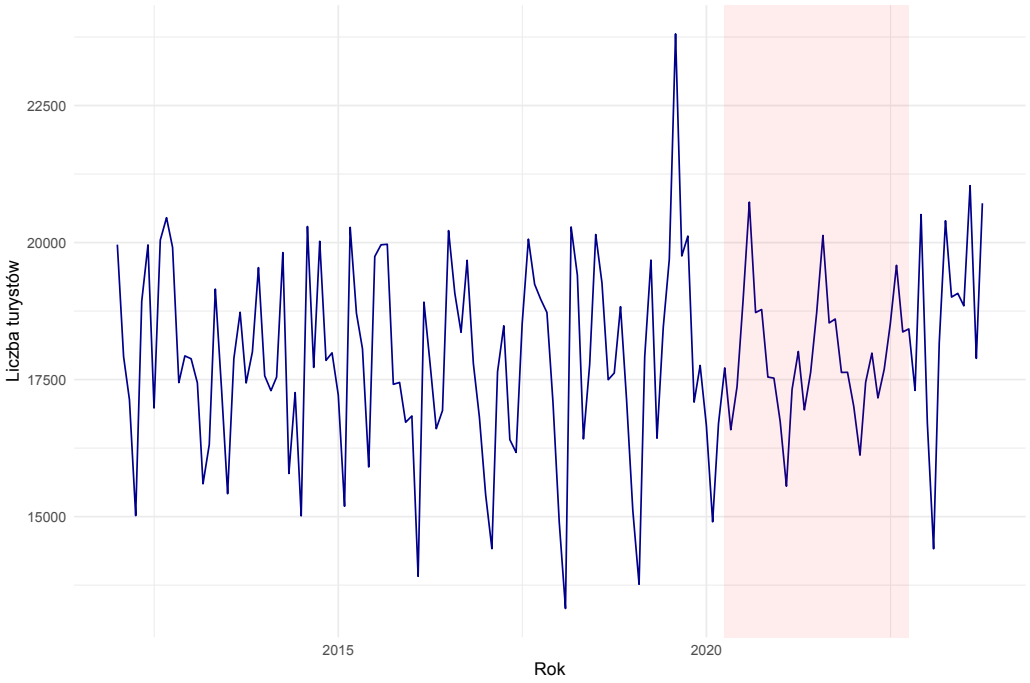
Dla tak otrzymanych danych tworzymy model, aby dokonać predykcji danych testowych.



Rysunek 31: Residua - SARIMA$(2,0,2) \times (1,0,1)[12]$ - z predykcją przed i po

Na wykresie ACF dla residuów tak otrzymanego modelu widzimy lekko odstającą wartość, ale test Ljung-Boxa nie wykazuje zależności pomiędzy resztami.

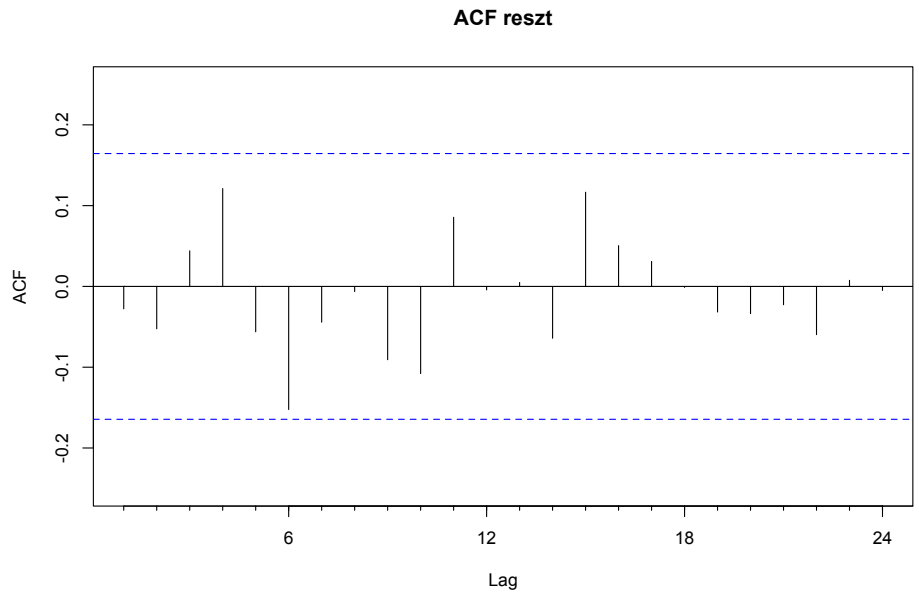
Jako, że dane po pandemii COVID-19 są ograniczone, spróbujmy utworzyć model niebiorący ich pod uwagę. Wtedy uzupełniamy szereg jedynie wartościami predykcji dla danych sprzed pandemii.



Rysunek 32: Uzupełnienie pandemii COVID-19 z użyciem wartości przed

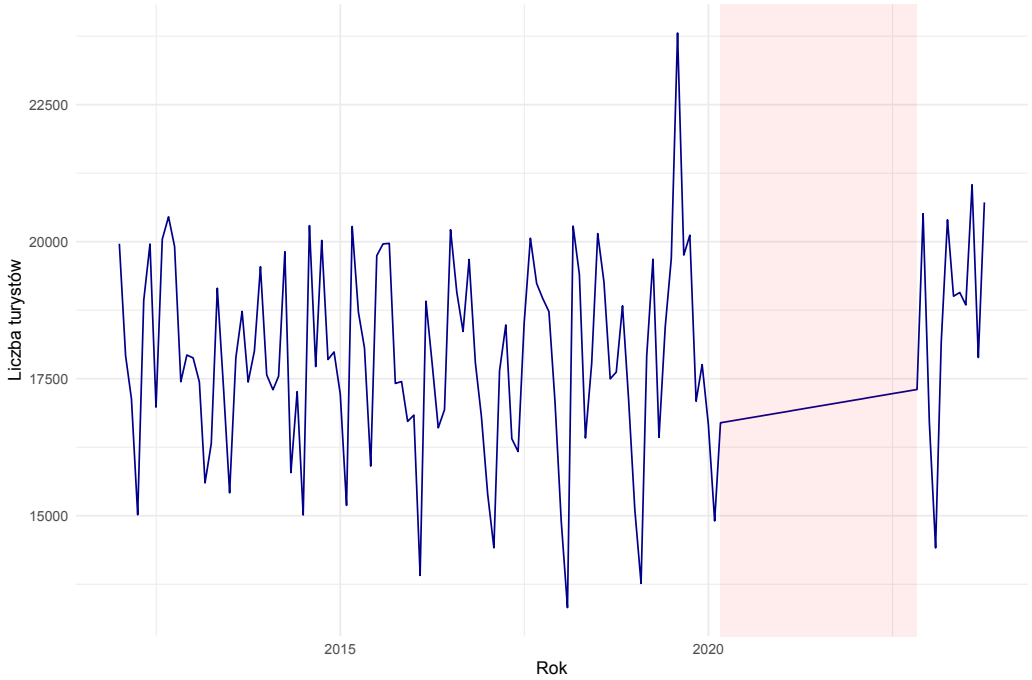
Takie uzupełnienie wydaje się lepsze, ponieważ lepiej zachowuje duże różnice między obserwacjami. Nie "spłaszcza"pików jak poprzednio.

Model utworzony dla tak otrzymanych danych to SARIMA$(2,0,2) \times (1,0,1)[12]$. Wykres ACF dla residuów wygląda nastepująco.



Rysunek 33: Residua - SARIMA$(2,0,2) \times (1,0,1)[12]$ - z predykcją przed

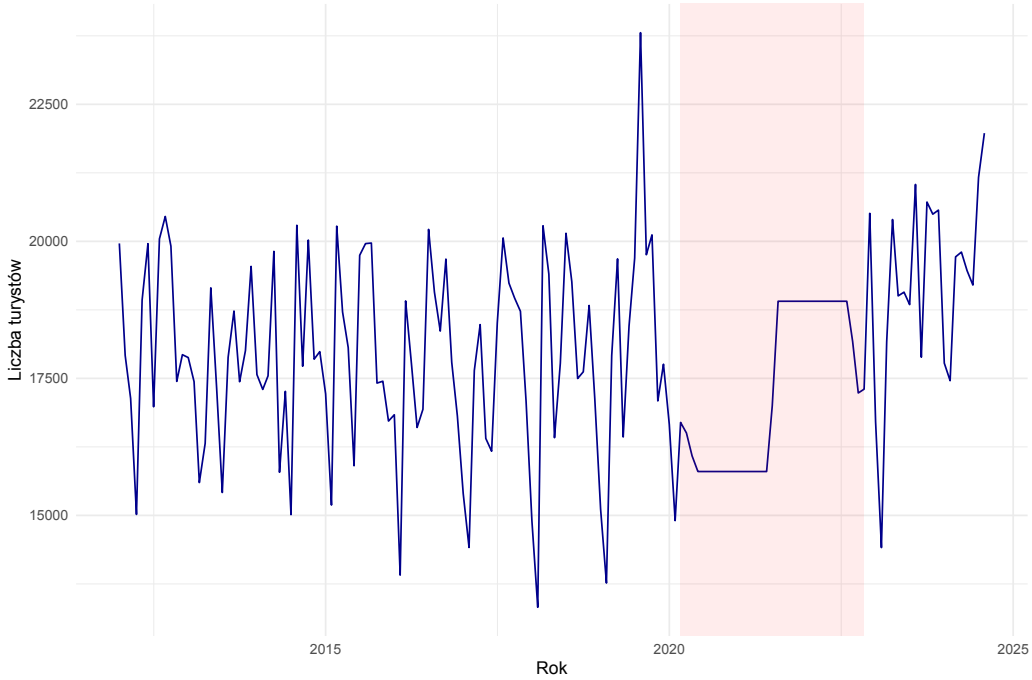
Przejdźmy teraz do wbudowanych metod uzupełniania danych. Jako pierwszą rozważamy metodę liniowej interpolacji. Polega ona na przeprowadzeniu prostej przez punkty ograniczające brakujący przedział. Wyestymowane wartości przedstawiamy na wykresie.



Rysunek 34: Estymacja za pomocą interpolacji

Możemy zobaczyć, że jest to po prostu funkcja liniowa przechodząca przez dwa skrajne punkty. Nie wydaje się właściwe tworzenie modelu dla tak wyglądających danych. Sezonowość i trend widoczne w dostępnych wartościach nie są w żaden sposób obrazowane w tym przypadku.

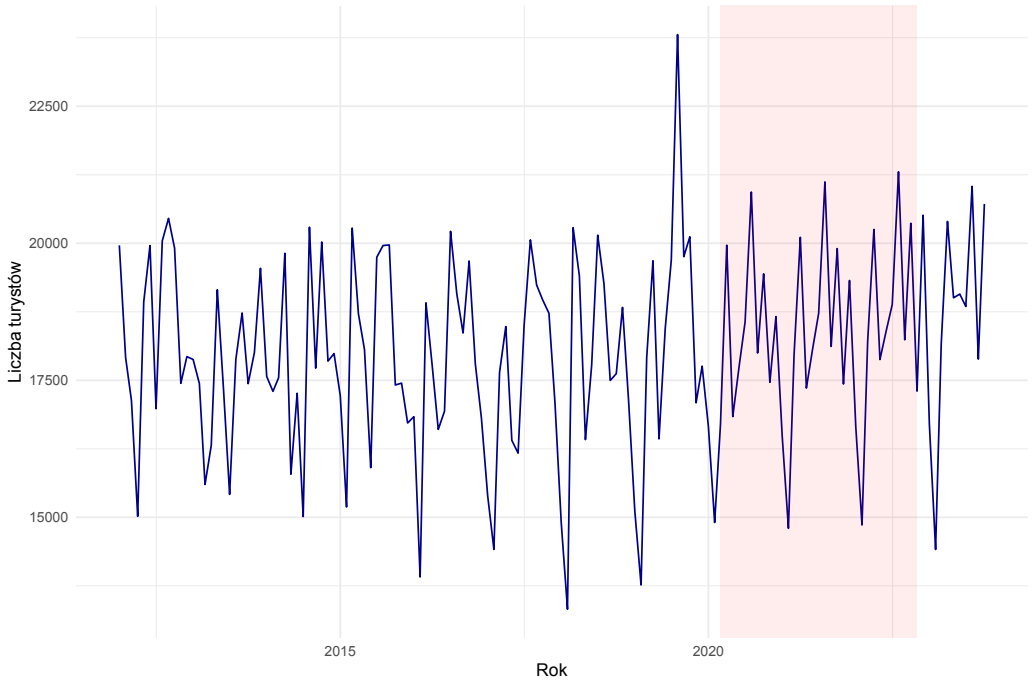
W przypadku estymacji brakujących wartości za pomocą średniej ruchomej również nie otrzymujemy zadowalających wyników - braki w danych są zbyt duże.



Rysunek 35: Estymacja za pomocą średniej ruchomej

Widzimy więc, że proste metody w przypadku braku dużej liczby wartości nie dają estymatorów w jakikolwiek sposób bliskich możliwym prawdziwym obserwacjom.

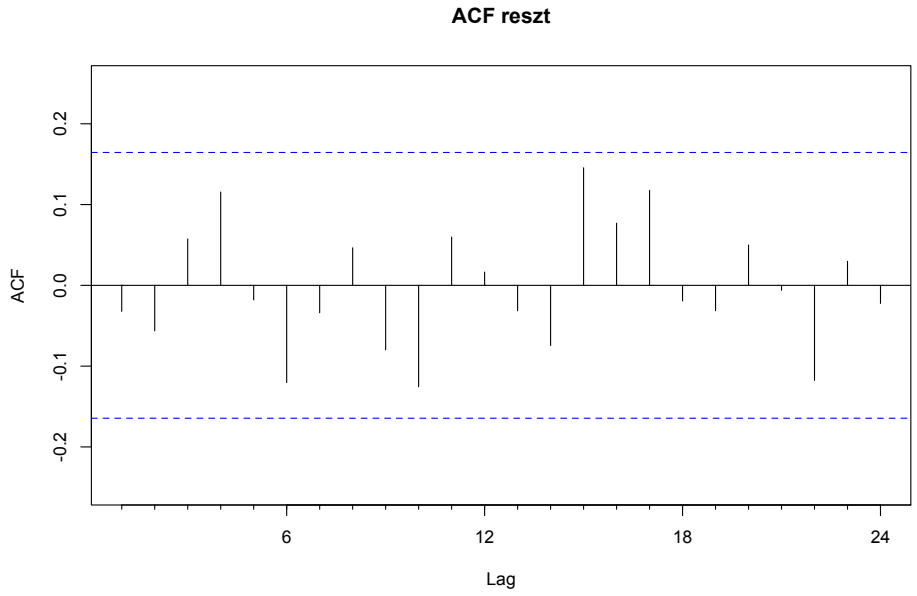
Możemy użyć także bardziej skomplikowanej metody - filtru Kalmana - `na_kalman()`. Tym razem szereg wygląda tak, jakby rzeczywiście reprezentował obserwowane przez nas dane.



Rysunek 36: Estymacja za pomocą filtru Kalmana

Zdecydowanie najlepiej przypomina pozostałe części szeregu. Najlepiej ze wszystkich przedstawia duże różnice w danych.

Na podstawie tych wyników przeprowadzamy dalszą analizę. Widzimy, że dla modelu stworzonego na podstawie takich danych, również nie występuje korelacja pomiędzy kolejnymi resztami.



Rysunek 37: Residua - SARIMA$(2,0,2) \times (0,1,1)[12]$ - filtr Kalmana

4.3 Porównaie modeli

Zobaczmy jak prezentują się predykcje dla wszystkich modeli.

Rysunek 38: Prognoza wybranych modeli na miesiące szeregu testowego

Legenda oznacza odpowiednio model utworzony na podstawie danych uzupełonionych za pomocą filtru Kalmana, następnie model uwzględniający dane przed, jak i po pandemii, a ostatni model to

model utworzony na danych uwzględniający jedynie dane przedpandemiczne.

Możemy zobaczyć, że modele powstałe z wykorzystaniem danych przed i po, oraz jedynie danych sprzed są tego samego typu SARIMA$(2,0,2) \times (1,0,1)[12]$. Są one do siebie bardzo podobne i lepiej odwzorowują kształt danych testowych. Jednak im predykcja jest dalsza tym lepszy wydaje się model SARIMA$(2,0,2) \times (0,1,1)[12]$, utworzony na podstawie danych uzupełnionych filtrem Kalmana. Różnice między wartościami są wtedy mniejsze. Dużo lepiej przewiduje on także wzrost wartości w okresie letnim 2024 roku. Wszystkie modele jednak niedoszacowują danych i zakładają, że liczba turystów będzie mniejsza niż jest w rzeczywistości.

Porównajmy także dokładność modeli na podstawie statystyk liczbowych.

Tabela 5: Statystyki dla wybranych modeli

	MSE	RMSD	MAE
Model 1	3127511	1768.477	1463.678
Model 2	3141142	1772.327	1446.240
Model 3	3289658	1813.741	1534.000

Okazuje się, że model 3 uwzględniający jedynie dane sprzed pandemii przewiduje dane trochę gorzej niż model 2. Jest to spowodowane, pominięciem informacji wnoszonej przez dane po pandemii (mimo, że jest to tylko średnia). Przez to wartości estymowane na podstawie modelu 2 są bliższe rzeczywistym. Najmniejsze wartości statystyk mamy dla modelu 1. Jednak nie różnią się one tak bardzo od pozostałych. Duże odstępstwa na początku rekompensowane są dobrą predykcją od kwietnia 2024 roku.

A na koniec zobaczmy statystyki AIC, BIC oraz wyliczoną dla nich liczbę stopni swobody.

Tabela 6: Statystyki dla wybranych modeli

	AIC	BIC	df
Model 1	2217.554	2234.759	6
Model 2	2440.731	2464.378	8
Model 3	2435.108	2458.754	8

Znów najmniejsze wyniki otrzymujemy dla modelu utworzonego z uzupełnieniem filtrem Kalmana. Jednak różnice nie są bardzo duże. Mniejsza liczba stopni swobody spowodowana jest występowaniem różnicowania w tym modelu, czego nie ma w pozostałych. Tym razem dla modelu drugiego statystyki AIC i BIC są większe niż dla modelu 3. Jest to spowodowane dobraniem trochę innych estymatorów parametrów modeli.

4.4 Wnioski

Dla danych, w których wykonujemy uzupełnienie na podstawie danych poprzedzających i następujących po pandemii, jak i jedynie na podstawie danych poprzedzających, dobierany model jest taki sam. Jednak uwzględnienie danych od listopada 2022 roku pozwala nam na lepsze oszacowanie przyszłych wyników. Oba modele bardzo dobrze odwzorowują monotoniczność zbioru testowego. Model utworzony na podstawie uzupełnienia filtrem Kalmana nie uwzględnia części sezonowej autoregresji, ale występuje w nim sezonowe różnicowanie, czego nie było w dwóch poprzednich. W pierwszych miesiącach zbioru testowego predykcja nie jest najlepsza. Jednak ten model wykazuje nalepszą predykcję bardziej odległych obserwacji. Warto również zauważyć, że wszystkie modele zaczynają od bardzo podobnych wartości, dość odległych od rzeczywistych. Dopiero później przyjmują oczekiwany kształt.

5 Podsumowanie wyników analizy

W tym rozdziale podsumujemy wyniki przeprowadzonej analizy oraz ocenimy skuteczność zastosowanych strategii prognozowania liczby turystów. Dokonamy przeglądu wybranych modeli, zwracając

szczególną uwagę na ich mocne i słabe strony w kontekście różnych podejść do uwzględniania lat pandemicznych.

5.1 Porównananie modeli prognostycznych

W tej sekcji przeanalizujemy, która z wcześniej wybranych strategii prognozowania liczby turystów sprawdziła się najlepiej. W rozdziale drugim zdecydowaliśmy się na stworzenie modelu opartego na danych niezniekształconych przez lata pandemiczne. Model ten dobrze poradził sobie z prognozowaniem na kolejny rok, jednak nie przewidział nagłego wzrostu liczby turystów w drugiej połowie 2019 roku. W kolejnym rozdziale uwzględniliśmy dane z lat pandemii COVID-19, budując model, który również skutecznie prognozował najbliższe miesiące. Niestety, nie przewidział on znaczącego wzrostu liczby turystów w okresie wakacyjnym 2024 roku. Było to zaskakujące, ponieważ można było się spodziewać, że uwzględniając sezonowość, taki wzrost powinien być oczywisty. Niestety, zaburzenia związane z latami pandemicznymi wpłynęły na model, który nie był w stanie poprawnie oszacować tego wzrostu. W ostatnim rozdziale postanowiliśmy przeanalizować alternatywny scenariusz, tj. „co by było, gdyby nie było pandemii?". W tym celu usunęliśmy dane z okresu COVID-19 i zastąpiliśmy je różnego rodzaju estymacjami. Modele stworzone w ten sposób poprawnie przewidywały wzrost liczby turystów w okresie wakacyjnym 2024 roku, co wcześniej okazało się trudnością. Niemniej jednak, wszystkie te modele znacząco niedoszacowały ogólnej liczby turystów. Na wykresach przedstawiających dane testowe krzywa prognoz niemal zawsze znajdowała się poniżej krzywej rzeczywistych danych. Modele te nie uwzględniały, że po ponownym otwarciu Japonii na turystykę średnia liczba turystów znacząco wzrosła w porównaniu do wcześniejszych lat.

Aby wyciągnąć jednoznaczne wnioski, porównaliśmy najlepsze modele z każdej sekcji w tym samym przedziale czasowym: od listopada 2023 roku do sierpnia 2024 roku. Na początek spójrzmy, jak różnice w strategiach wpływają na wyniki prognoz w tym okresie:

Rysunek 39: Prognoza wybranych modeli na okres 2023.11-2024.08

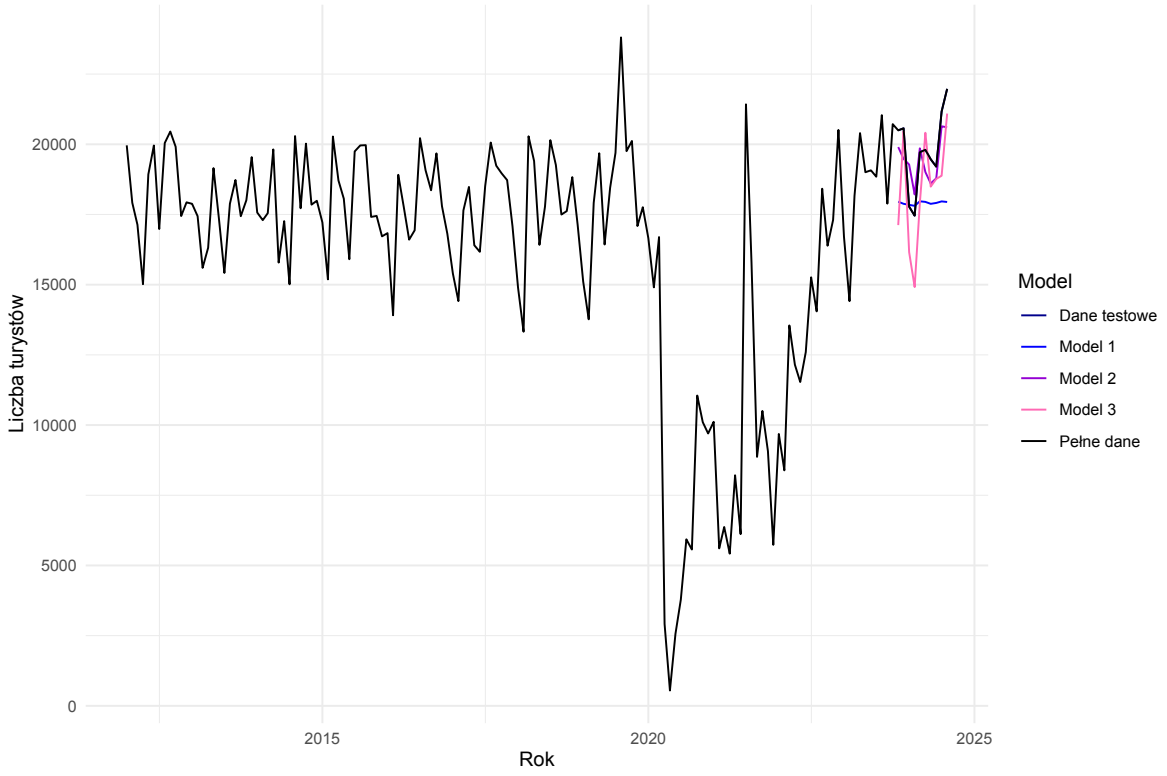
Model 1 w legendzie odpowiada modelowi stworzonemu na danych w latach 2012-2018. Model 2, to model wykorzystujące dane do 2023 roku, uwzględniający dane covidowe, a model 3, to model najpierw dokonujący predykcji danych covidowych, a następnie na tak uzupełnionych danych dobierający

parametry.

Zgodnie z oczekiwaniami, najgorzej poradził sobie model opracowany w rozdziale drugim. Wynika to z faktu, że był on trenowany na danych z lat 2012–2018, gdzie występował jedynie stały trend i sezonowość. Funkcja `auto.arima()` automatycznie przypisała mu jeden parametr sezonowy, co uczyniło model stosunkowo prostym. W efekcie nie był on w stanie dobrze przewidzieć lat bardziej oddalonych od okresu trenowania, zwłaszcza tych zniekształconych przez pandemię.

Model stworzony w rozdziale trzecim, który uwzględniał lata pandemiczne, na niektórych odcinkach wykresu wypada najlepiej – jego krzywa jest najbliższa rzeczywistym danym testowym. Jego największym mankamentem jest jednak spadek liczby turystów na końcu prognozowanego okresu, gdzie w rzeczywistości powinien nastąpić wzrost.

Model stworzony w rozdziale trzecim, w którym dane pandemiczne zostały usunięte i zastąpione przekształceniami, miał problem z niedoszacowaniem liczby turystów. Niemniej jednak poprawnie prognozował zarówno wzrosty, jak i spadki liczby turystów w odpowiednich okresach, w tym wzrost w wakacje 2024 roku. Co ciekawe, spadek liczby turystów między nowym rokiem, a okresem kwitnienia wiśni wydaje się w tym modelu zbyt gwałtowny.



Rysunek 40: Prognoza wybranych modeli na okres 2023.11-2024.08

Na powyższym wykresie widzimy predykcję na tle całego szeregu, czyli w jaki spoób modele dopasowują się do poprzednich danych. Cały czas widzimy, że model 1 jest zbyt daleką predykcją i w żaden sposób nie przypomina dostępnych danych. Możemy zaobserwować, że model 2 o wiele lepiej dostosuwuje się do danych testowych, za to model 3 dość dobrze uwzględnia zachowanie szeregu w latach poprzedzających pandemię.

Aby uzupełnić analizę, porównaliśmy również statystyki testowe dla wybranych modeli:

Tabela 7: Statystyki dla wybranych modeli

	MSE	RMSE	MAE
Model 1	5098379	2257.96	1935.89
Model 2	807464	898.59	804.35
Model 3	3127511	1768.48	1463.678

Statystyki te potwierdzają wcześniejsze obserwacje – model opracowany w rozdziale trzecim, który uwzględniał lata pandemiczne, okazał się najlepiej dopasowany do danych testowych, zwłaszcza do miesięcy następujących zaraz po ponownym otwarciu Japonii na turystykę. Oznacza to, że pandemia miała istotny wpływ na kształtowanie się turystyki, przyczyniając się do jej wzrostu w późniejszych latach.

Ponieważ modele 2 i 3 mają wady i zalety sprawdzimy za pomocą testu Diebolda-Mariano jakość obu zestawów prognoz. Za pomocą funkcji `dm.test()` zbadamy, czy prognozy modelu 2 są bardziej dokładne, niż prognozy modelu 3. P-wartość tego testu wynosi 0.95, co oznacza, że na poziomie istotności 0.05 nie mamy podstaw do odrzucenia hipotezy zerowej. Oznacza to, że model 2 ma statystycznie lepsze prognozy.

Podsumowując, wyniki tej analizy są zaskakujące. Na wczesnym etapie pisania kodu i raportu zakładaliśmy, że usunięcie danych pandemicznych i ich zastąpienie estymacjami będzie najlepszą strategią. Jednak lepszym rozwiązaniem okazało się uwzględnienie lat pandemii, ponieważ modele te lepiej odwzorowują rzeczywiste dane. Pokazuje to, że zamknięcie Japonii w okresie pandemii, paradoksalnie, miało pozytywny wpływ na turystykę w kolejnych latach – liczba turystów wzrosła, przekraczając wartości z lat przed pandemią.

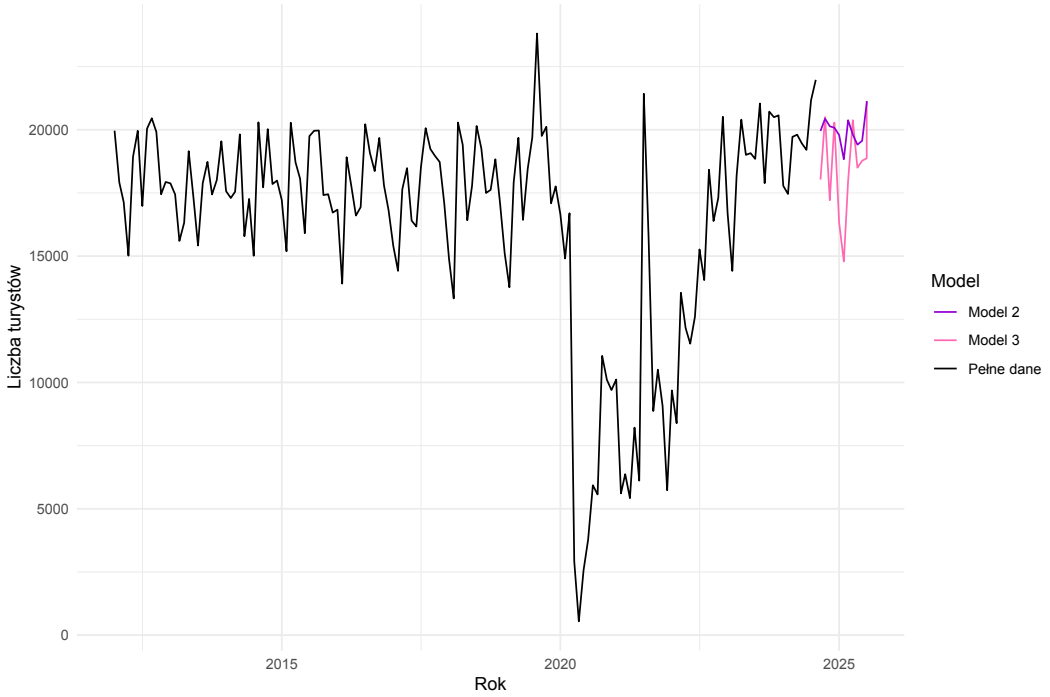
5.2 Wnioski

Analiza wpływu pandemii na turystykę w Japonii pokazuje, że pandemia miała istotny wpływ na dynamikę ruchu turystycznego, powodując początkowe załamanie, a następnie szybki wzrost po otwarciu granic. Modele uwzględniające lata pandemiczne dobrze odzwierciedlały ogólny trend, ale nie przewidziały wzrostu liczby turystów w sezonie wakacyjnym 2024, co wskazuje, że pandemia zakłóciła regularne wzorce sezonowości.

Z kolei modele eliminujące lata pandemiczne i uzupełniające dane odpowiednimi przekształceniami lepiej uchwyciły wzrost liczby turystów w okresach sezonowych, takich jak lato 2024. Niemniej jednak niedoszacowały one całkowitej liczby turystów, co sugeruje, że odrzucenie danych z lat pandemii prowadzi do pominięcia wpływu „efektu odroczonego popytu".

Wnioski z analizy wskazują, że przyszłe strategie odbudowy turystyki powinny uwzględniać nie tylko tradycyjne wzorce sezonowe, ale także zmiany wynikające z nowych trendów po pandemii. Ważne jest przygotowanie na wzrost popytu w najpopularniejszych okresach oraz elastyczne podejście do modelowania potencjalnych zakłóceń w przyszłości.

Jako ciekawostę przedstawiamy również wykres zawierający przewidywania nieistniejących jeszcze, przyszłych danych, od września 2024 roku do sierpnia roku 2025.



Rysunek 41: Prognoza wybranych modeli na okres 2024.09-2025.08

Możemy zauważyć, że predykcje dla modelu 2 mieszczą się w węższym przedziale, przewidując jednocześnie dość dużą liczbę turystów w każdym miesiącu. Obserwowane spadki tej liczby są niewielkie w porównianiu do wahań występujących przed pandemią COVID-19. Model 3 dużo lepiej zachowuje wzorce obserwowane w latach przed 2020 rokiem. Wahania pomiędzy kolejnymi miesiącami są dużo większe niż w modelu 2. Niestety na razie nie jesteśmy w stanie odpowiedzieć na pytanie, który z tych modeli będzie lepszy. Na podstwie zbioru testowego możemy przypuszczać, że będzie to model 2. Jednak, aby wiedzieć, czy w rzeczywistości zachowają się takie wzorce, będziemy musieli poczekać i w przyszłości sprawdzić poprawność naszych prognoz.

Literatura

1. P. J. Brockwell, R. A. Davis *Introduction to time series and forecasting*
2. W. Cygan *Analiza Szeregów Czasowych. Skrypt do wykładu Szeregi czasowe*
3. G. E. P. Box i wsp. *Time Series Analysis*