ANALIZA LICZBY ODWIEDZAJĄCYCH DOLNOŚLĄSKIE TURYSTYCZNE OBIEKTY NOCLEGOWE W LATACH 2009−2023

Projekt z szeregów czasowych 2024/25

Kinga Kępa, Dominika Ochalik

1. Wprowadzenie

Celem projektu jest zbadanie z wykorzystaniem szeregów czasowych, jak w ostatnich latach zmieniała się liczba osób odwiedzających dolnośląskie obiekty noclegowe[1] − za istotne zadania wyznaczamy sobie przede wszystkim: dostrzeżenie prawidłowości, dobór odpowiedniego modelu wraz z uzasadnieniem wyboru, a także próbę przewidzenia, jak wyglądałyby dane, gdyby nie było pandemii SARS-Cov-2. W rozważaniach wstępnych porównamy ponadto dane dla województwa dolnośląskiego z danymi dla innych województw, omówimy widoczne tendencje i zastanowimy się, czy przyjęty przez nas sposób działania byłby adekwatny również w przypadku analizy liczby turystów w innych obszarach Polski.

Na początku przyjrzymy się strukturze danych. Zanim jednak to zrobimy, powiedzmy kilka słów o tym, jakie dokładnie informacje przekazują. W projekcie zajmiemy się łączną liczbą turystów, którzy w latach 2009−2023 (z podziałem na miesiące) skorzystali z usług noclegowych obiektów turystycznych. Podczas zbierania danych za noclegowy obiekt uznawano obiekt, który posiada 10 lub więcej miejsc noclegowych[2] (łącznie z pokojami gościnnymi i kwaterami agroturystycznymi).

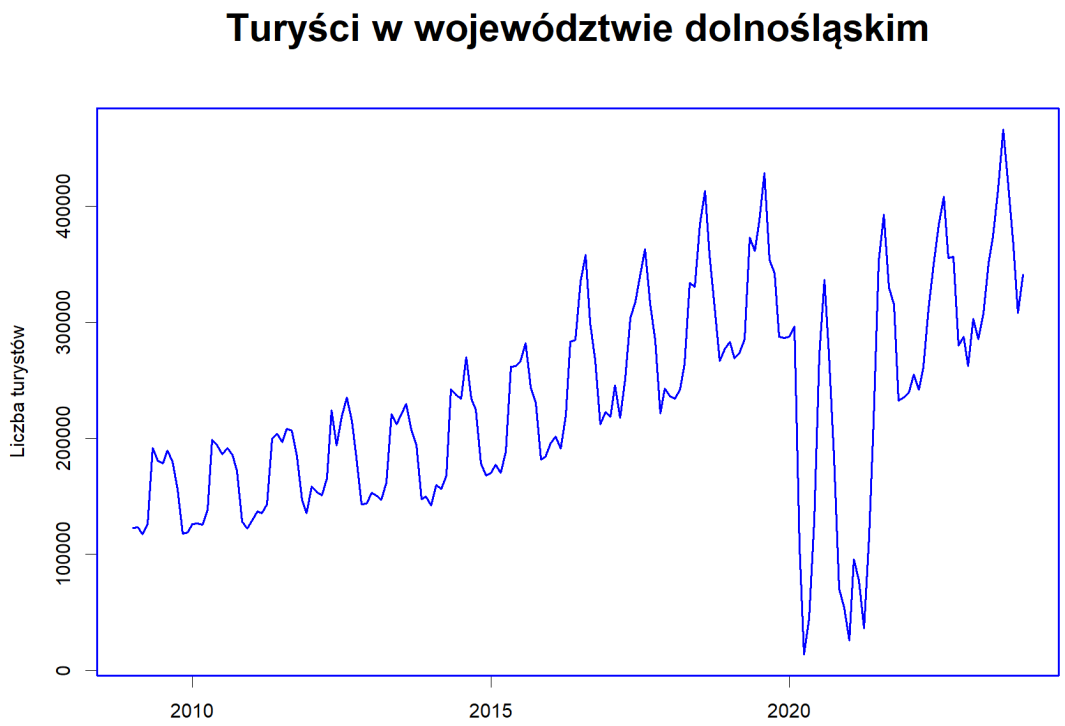
Poniżej zamieszczono tabelę zestawiającą dane o liczbie turystów odwiedzających województwo dolnośląskie, a także obrazujący je wykres.

[1]Dane wykorzystane w projekcie pochodzą z Banku Danych Lokalnych Głównego Urzędu Statystycznego i są dostępne pod adresem internetowym: `https://bdl.stat.gov.pl/bdl/dane/podgrup/wymiary`[dostęp: 12.01.2025]

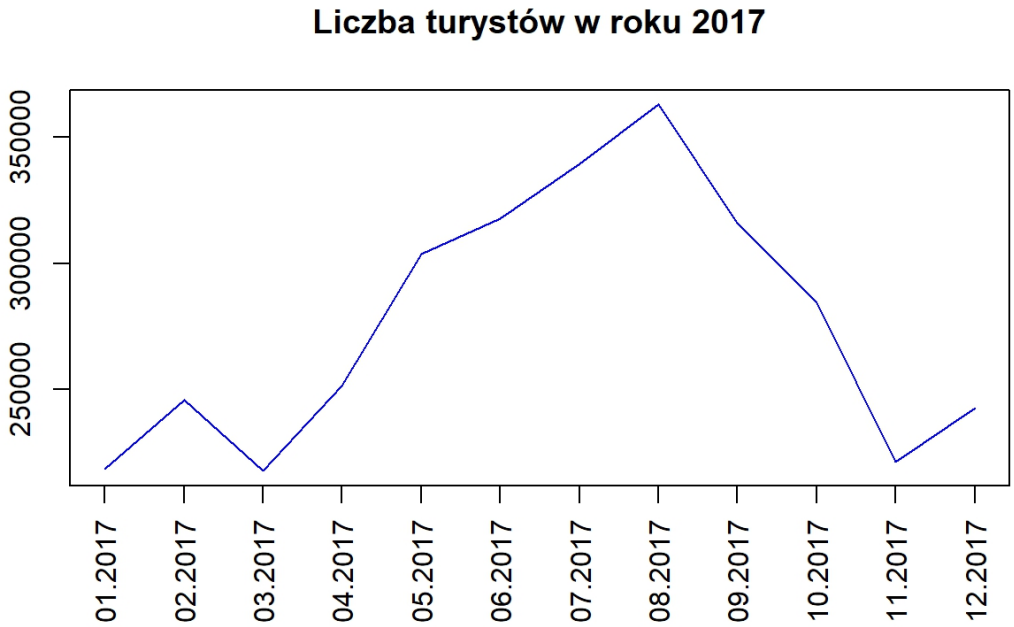
[2]Dla porządku zaznaczmy, że specyfikacja ta odpowiada danym zbieranym od stycznia 2012 r., wcześniejsze odnosiły się do obiektów zbiorowego zakwaterowania. Nie jest to jednak czynnik, który w jakikolwiek znaczący sposób oddziaływałby na strukturę badanych przez nas danych.

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2009	122.8	123.2	117.1	126.3	191.5	180.8	178.6	189.4	180.0	155.9	117.7	119.0
2010	126.0	126.4	125.6	138.1	198.9	193.9	186.4	192.0	185.8	172.1	128.2	122.0
2011	129.7	137.1	135.3	143.2	199.7	203.8	196.4	208.5	206.7	184.0	146.8	135.6
2012	158.7	153.6	150.9	165.7	224.3	193.8	218.6	235.3	214.8	183.3	143.3	143.9
2013	153.3	150.4	147.3	161.9	220.8	212.0	220.5	230.0	207.5	194.5	147.4	149.6
2014	142.0	160.0	156.5	168.3	242.6	237.3	234.2	270.0	234.3	224.7	178.4	168.1
2015	170.0	177.6	170.1	188.5	261.9	262.5	266.8	282.2	244.3	230.7	182.0	183.7
2016	195.8	201.6	191.2	220.0	283.2	284.7	335.2	358.0	299.6	267.1	212.2	222.8
2017	218.5	245.9	217.9	251.5	303.8	317.7	339.5	362.9	316.0	284.5	221.3	243.0
2018	236.5	233.9	242.2	264.3	334.1	330.4	384.8	413.5	358.4	313.0	266.7	276.9
2019	283.2	268.7	274.1	285.2	373.0	361.6	387.2	428.7	353.9	342.3	287.7	286.8
2020	287.9	296.5	106.8	13.7	44.6	128.4	272.1	336.7	272.7	184.1	70.3	54.1
2021	25.6	95.6	77.8	36.2	115.9	215.0	352.9	393.0	330.1	314.9	232.6	235.2
2022	239.3	255.4	241.8	262.0	312.1	349.9	384.4	408.6	355.7	356.6	279.7	287.5
2023	262.2	303.2	285.2	307.2	351.1	374.4	415.0	466.3	417.3	365.6	307.9	341.2

Tabela 1: Liczba turystów w województwie dolnośląskim

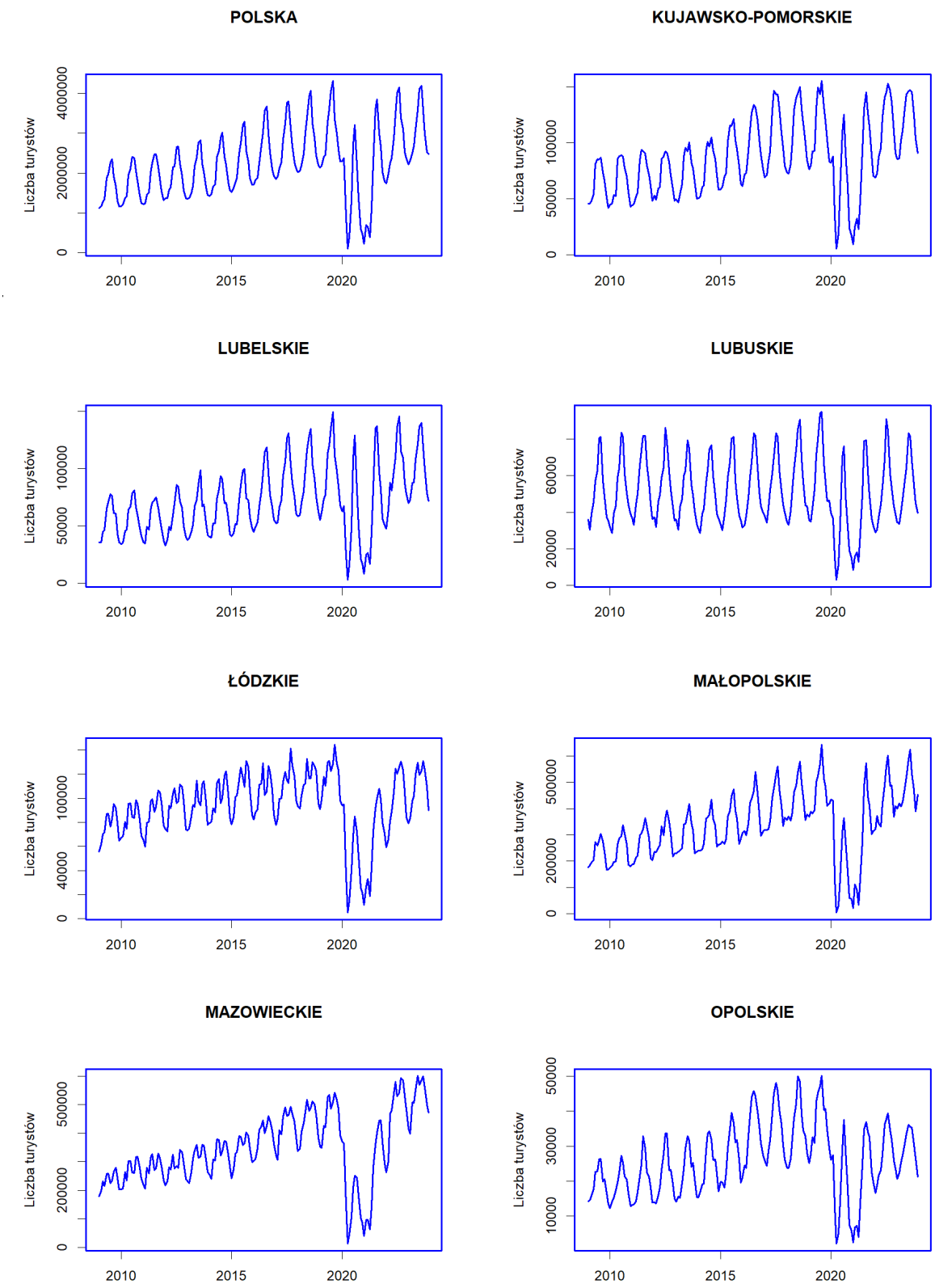


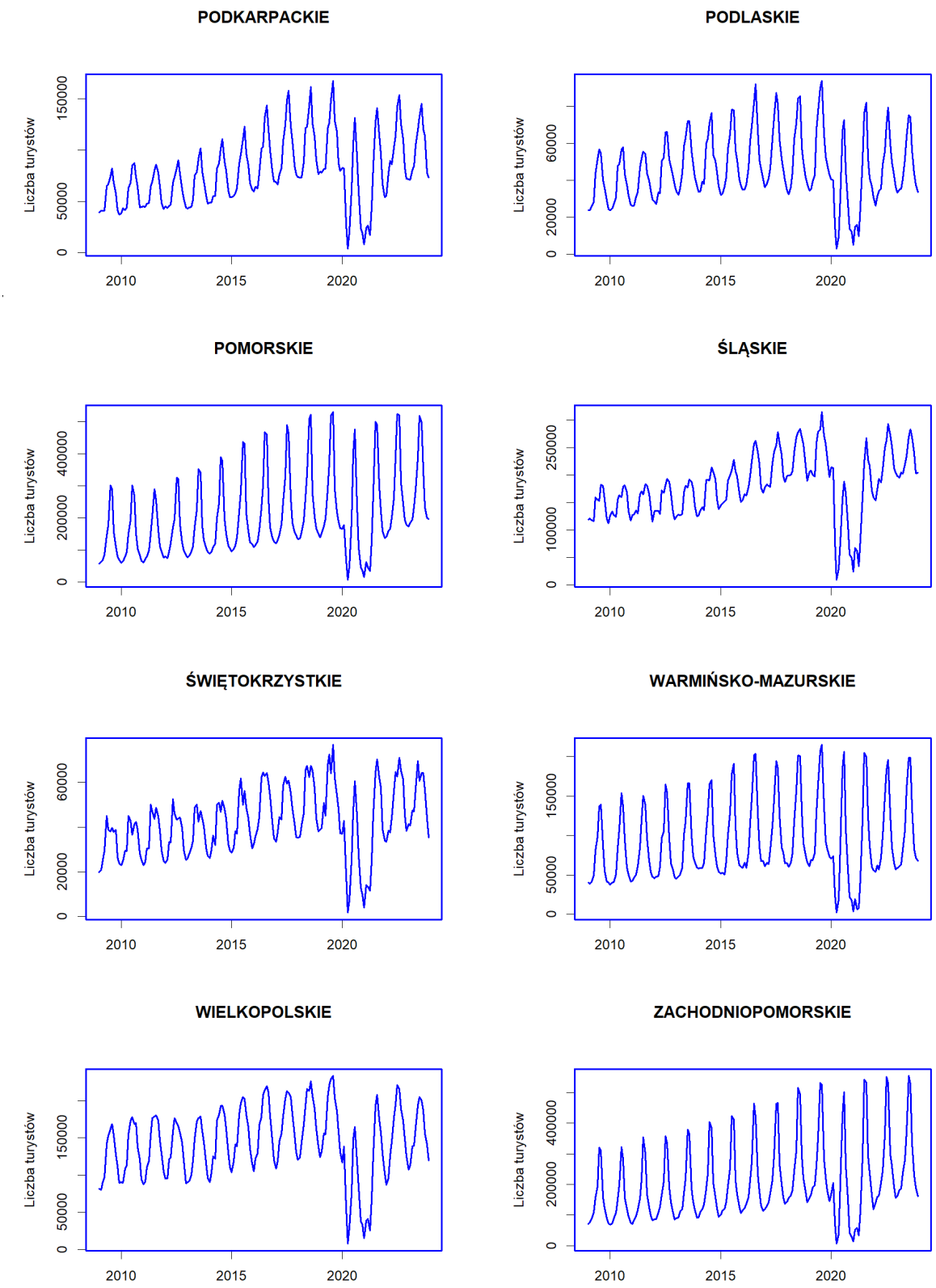
Spójrzmy jeszcze, jak zmienia się liczba odwiedzających Dolny Śląsk w trakcie przykładowego roku. W tym celu prześledzimy wykres dla roku 2017.



Już bardzo wstępna analiza powyższych wykresów i tabeli wystarczy, by zauważyć, że zdecydowanie większą popularnością cieszy się Dolny Śląsk w miesiącach letnich niż w zimowych. Dla pierwszej połowy rozpatrywanego czasu można wyodrębnić trwający od maja do października sezon letni, w którym liczba turystów utrzymywała się na podobnym, wysokim poziomie. W przypadku kolejnych lat największą wartość, znacznie różniącą się od pozostałych, odnotowuje się w sierpniu. Obserwowana struktura wykazuje wyraźną sezonowość, co nie może budzić zdziwienia w przypadku danych odnoszących się do turystyki. Bardzo charakterystyczną i łatwą do rozpoznania cechą jest wyraźny spadek liczby turystów w latach 2020−2021 w porównaniu do lat poprzednich, spowodowany panującą wówczas pandemią SARS-CoV-2, która znacznie ograniczyła ruch turystyczny. Wstępna, graficzna analiza danych pozwala też stwierdzić, że województwo dolnośląskie staje się dla turystów coraz bardziej atrakcyjne − dostrzegamy wyraźny trend wzrostowy, który jest widoczny także w latach po pandemii.

Na dwóch kolejnych stronach zamieszczone zostały wykresy liczby turystów dla pozostałych województw, a także dla całej Polski:


POLSKA
Liczba turystów
0
2000000
4000000
2010
2015
2020
KUJAWSKO-POMORSKIE
Liczba turystów
0
50000
100000
2010
2015
2020
LUBELSKIE
Liczba turystów
0
50000
100000
2010
2015
2020
LUBUSKIE
Liczba turystów
0
20000
60000
2010
2015
2020
ŁÓDZKIE
Liczba turystów
0
40000
100000
2010
2015
2020
MAŁOPOLSKIE
Liczba turystów
0
200000
500000
2010
2015
2020
MAZOWIECKIE
Liczba turystów
0
200000
500000
2010
2015
2020
OPOLSKIE
Liczba turystów
10000
30000
50000
2010
2015
2020



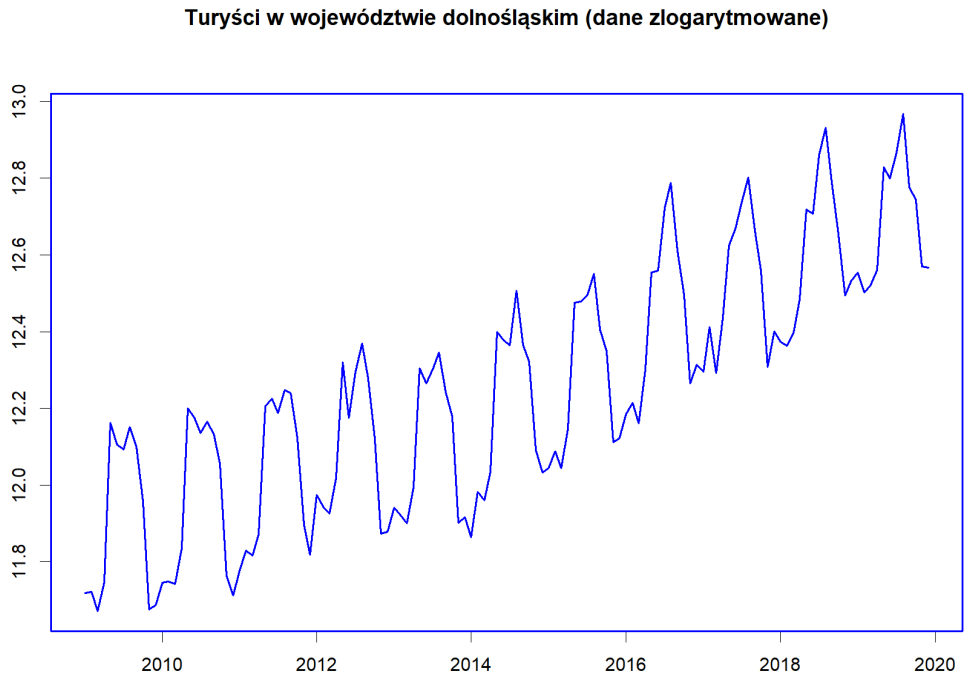
Widzimy, że województwa można podzielić na pewne grupy pod względem zmiany liczby turystów. Jedną z takich grup będą stanowiły województwa bez trendu albo z bardzo małym trendem (województwo lubuskie, warmińsko-mazurskie). Niektóre województwa mają wyraźny jeden miesiąc, w którym są odwiedzane najczęściej (pomorskie, zachodniopomorskie, podlaskie). Najmniej wpłynął wirus na województwa warmińsko-mazurskie,

zachodniopomorskie i pomorskie. Województwo dolnośląskie dobrze odzwierciedla to, co się dzieje w całej Polsce. Jedyną różnicą jest obserwacja, że województwo dolnośląskie ma kilka miesięcy, w których liczba turystów jest wyraźnie większa, a Polska jeden taki miesiąc.

2. Dekompozycja modelu

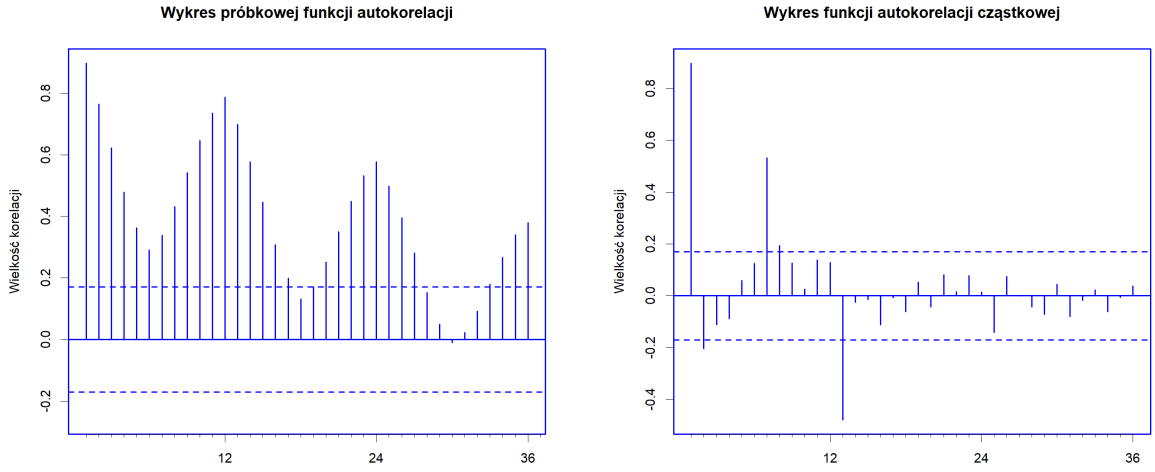
Ze względu na to, że wystąpienie pandemii można właściwie traktować jako anomalię, która wyraźnie zaburzyła prawidłowości obserwowane w ruchu turystycznym w województwie dolnośląskim, przy próbie dopasowania modelu nie będziemy uwzględniać wyników z lat 2020−2023. W pierwszym kroku dobierzemy model odpowiadający latom 2009−2019 i w oparciu o niego wykonamy prognozę na lata 2020−2023, której wyniki będziemy traktować w dalszych rozważaniach jako hipotetyczne dane, odpowiadające sytuacji, w której pandemia nie nastąpiła.

Nietrudno zauważyć, że dane, które analizujemy, charakteryzują się dużą wariancją, dlatego na początku nałożymy na nie logarytm[3], co pozwoli zmniejszyć ich rozrzut i ułatwi dopasowanie do nich modelu. Po usunięciu informacji o latach 2020−2023 i zlogarytmowaniu interesujący nas szereg przedstawia się następująco:



Oczywiście otrzymany szereg jest niestacjonarny. Wykonamy więc dekompozycję, która pozwoli sprowadzić go do postaci stacjonarnej. W analizie sezonowości i trendu pomogą też funkcje ACF i PACF zamieszczone poniżej.

[3]Podczas dopasowywania modelu (co zostanie omówione w dalszej części raportu) przeprowadzono też transformację Boxa-Coxa, której logarytmowanie jest szczególnym przypadkiem. Transformaja ta nie pozwoliła jednak dopasować modelu lepszego niż dla danych zlogarytmowanych.



Wykres próbkowej funkcji autokorelacji (sACF) wskazuje na roczną sezonowość szeregu: wartość korelacji osiąga szczyty dla opóźnień będących wielokrotnościami 12. Ponadto widzimy, że wartości korelacji powoli zmniejszają się (szczyty są coraz mniejsze), co wskazuje na obecność trendu.

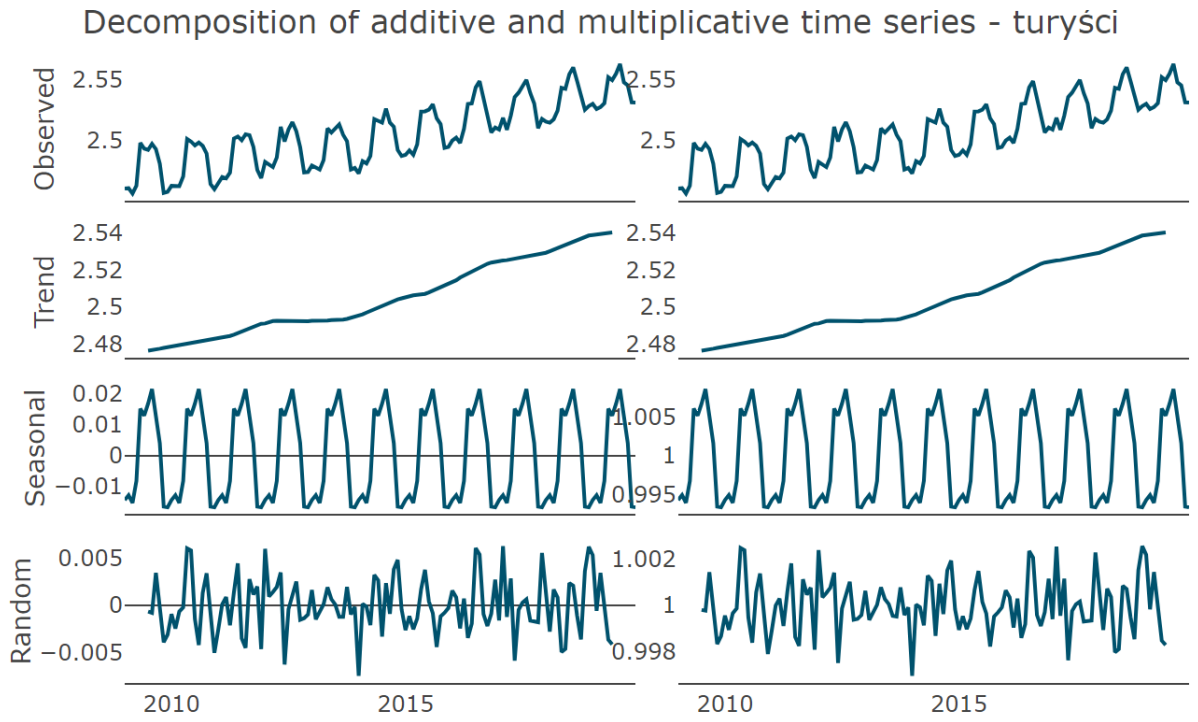
Wykres funkcji autokorelacji cząstkowej (PACF) wskazuje na silną korelację z danych z poprzednimi wartościami, a także na sezonowość, ponieważ widzimy znaczące piki dla h=1 i h=13.

Stacjonarność rozumiemy w sensie słabym, tzn. za stacjonarny uznajemy taki szereg, który spełnia następujące warunki:

- $\mathrm{Var}(X_t) < \infty$ dla każdego t,
- $\mathbb{E} = m = \text{const}$ dla każdego t,
- $\gamma_X(t, s) = \gamma_X(t + r, s + r)$ dla dowolnych t, s, r.

Zauważmy ponadto, że warunki te pociągają stałą wariancję dla wszystkich X_t.

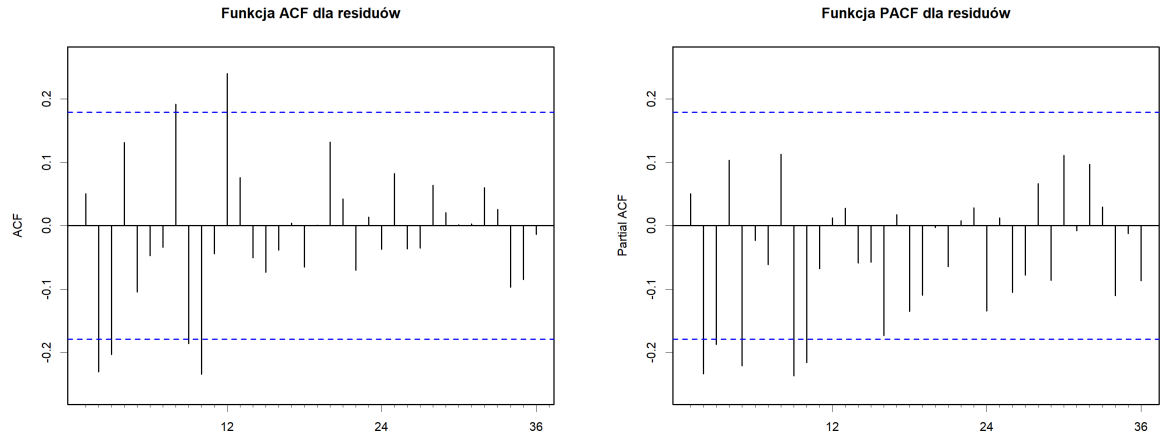
Patrząc na wykres oryginalnych wartości szeregu dla województwa dolnośląskiego, widzimy, że wariancja dla danych z lat 2009-2019 delikatnie rośnie. Po zlogarytmowaniu wartości szeregu wzrost wariancji staje się niewidoczny. Łącząc oba wnioski, przeprowadzimy dekompozycję szeregu zarówno dla modelu addytywnego, który charakteryzuje się stałą wariancją, jak i multiplikatywnego, w którym wariancja jest zmienna w czasie.



Dekompozycja szeregu. Po lewej stronie model addytywny, po prawej model multiplikatywny

W przypadku obu modeli widzimy wyraźny trend wzrostowy oraz sezonowość.

Przyjrzyjmy się teraz residuom otrzymanym po dekompozycji (dla modelu addytywnego). Aby zbadać relacje między nimi, posłużymy się się funkcjami ACF i PACF dla reszt.



Otrzymane wyniki są na ogół satysfakcjonujące, choć zarówno dla funkcji ACF, jak i PACF pewna część obserwacji mieści się poza przedziałem ufności. Sama dekompozycja nie wystarcza zatem, by móc z satysfakcjonującą dokładnością modelować dane dotyczące liczby turystów.

3. Dobór modelu dla województwa dolnośląskiego

Do analizowanych danych postaramy się dobrać model SARIMA. Jest on jednym z podstawowych modeli, które sprawdzają się w przypadku niestacjonarnych szeregów, w których dostrzega się trend i sezonowość.

Model taki, oznaczany przez:

SARIMA$(p, d, q) \times (P, D, Q)_s$,

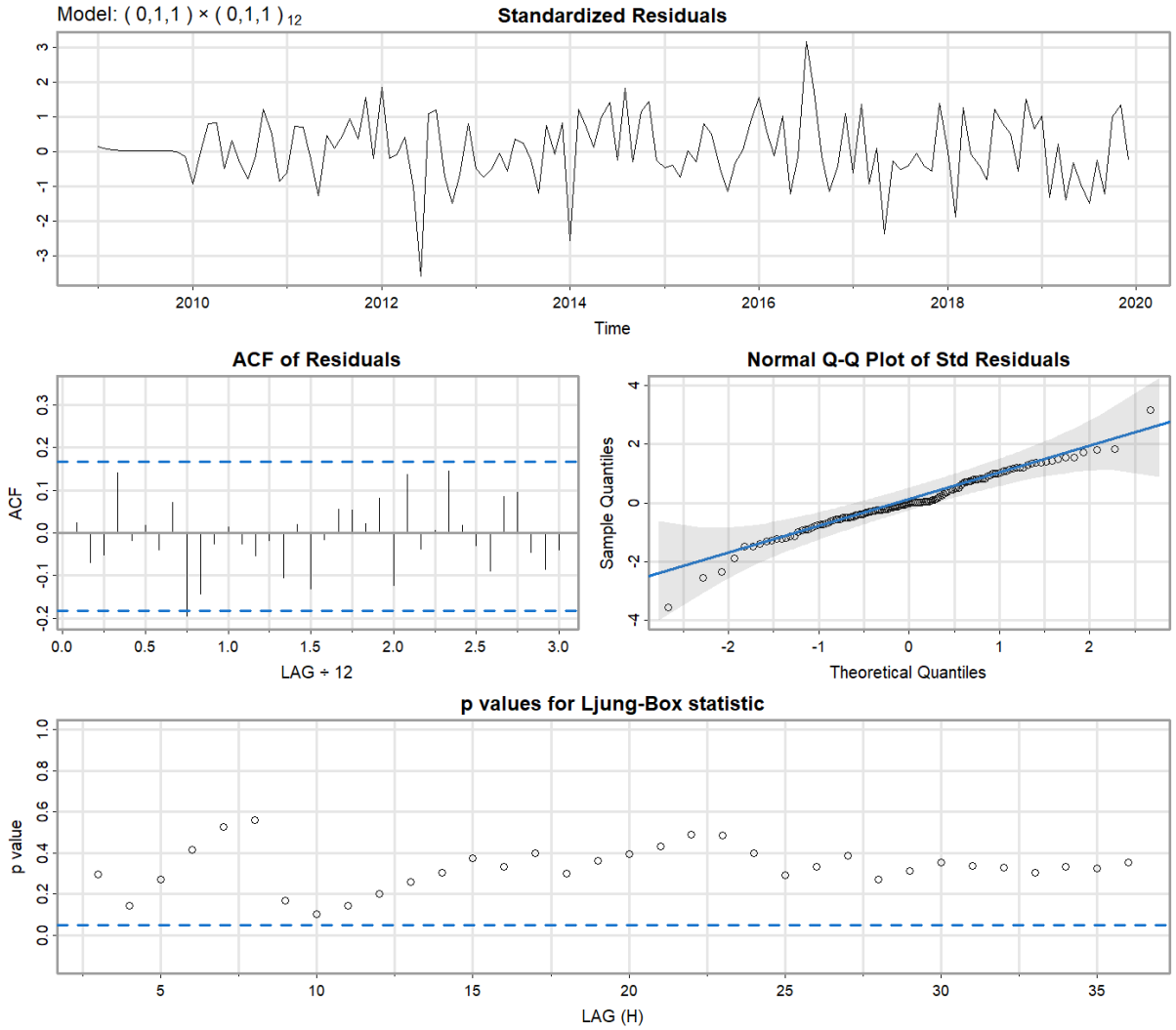
odpowiada szeregowi X_t spełniającemu równanie:

$$\phi(L)\Phi(L^s)\Delta^d(1 - L^s)^D X_t = \theta(L)\Theta(L^s)W_t$$

i składa się z siedmiu parametrów. Parametr s oznacza długość sezonu (w naszym przypadku, jako że analizujemy dane miesięczne, będzie on równy 12), parametry p, d, q są związane z niesezonową częścią ARIMA i odpowiadają kolejno rzędowi autoregresji, liczbie różnicowań i rzędowi średniej ruchomej, a P, D, Q są ich odpowiednikami sezonowymi.

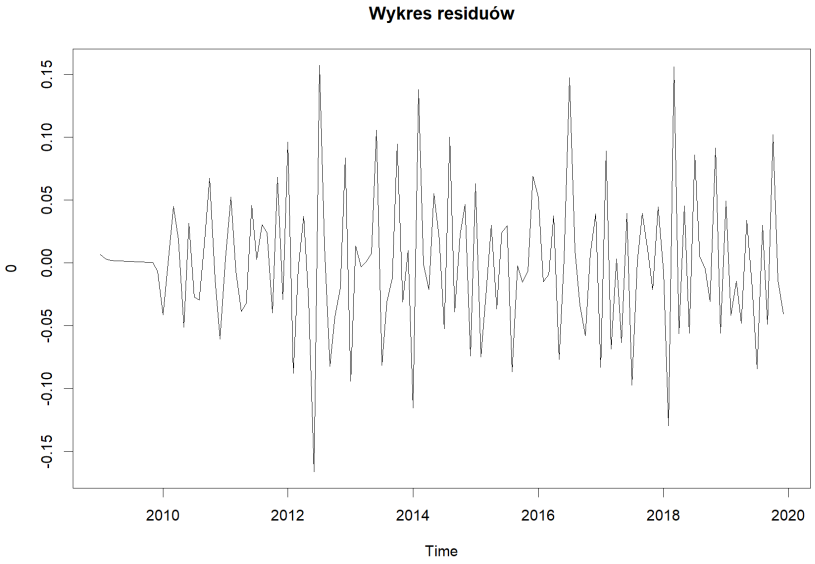
Przy wyborze modelu zastosujemy dwa podejścia. Na początku użyjemy funkcji **auto.arima** z pakietu **forecast**, która dobierze model automatycznie. Następnie postaramy się dobrać model samodzielnie, na podstawie odpowiednich funkcji ACF oraz PACF i porównamy otrzymane wyniki. Dla wybranych modeli dokonamy analizy reszt, w szczególności za pomocą testu Shapiro-Wilka sprawdzimy, czy pochodzą one z rozkładu normalnego. Porównamy też wartości statystyk AIC, BIC oraz AICc.

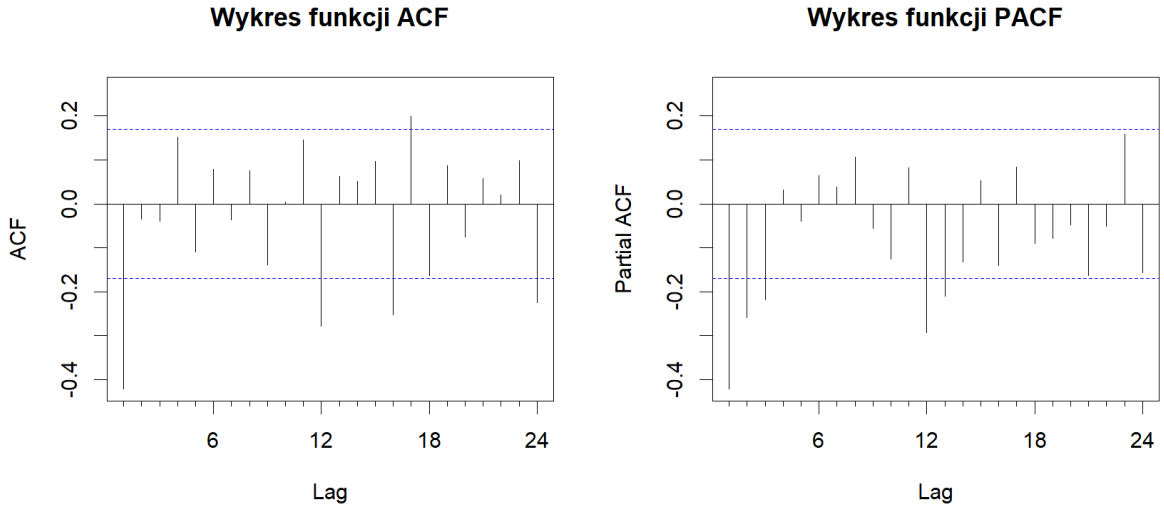
Funkcja **auto.arima** dla zlogarytmowanego szeregu wskazała model SARIMA$(0,1,1) \times (0,1,1)_{12}$. Spróbujmy więc ocenić jakość dopasowania tego modelu do danych.



Otrzymane wyniki dotyczące diagnostyki reszt wydają się satysfakcjonujące. Wszystkie p-wartości w teście Ljunga-Boxa, badającym związek między resztami, są większe od przyjętego poziomu istotności i wskazują na to, że nie występuje zależność między residuami. Również wartości funkcji autokorelacji (poza jedną) mieszczą się w 95% przedziale ufności, wykres kwantylowo-kwantylowy, zgodnie z oczekiwaniami, wskazuje też, że rozkład residuów jest bliski normalnemu o średniej bliskiej 0 (średnia reszt z dokładnością do czterech miejsc po przecinku to $\bar{x} = 0.017$. Aby upewnić się, czy reszty mogą pochodzić z rozkładu normalnego, wykonamy jednak jeszcze test Shapiro-Wilka (użyjemy funkcji **shapiro.test**). Hipoteza zerowa tego testu mówi, że dane pochodzą z rozkładu normalnego, za poziom istotności standardowo przyjmiemy 0.05. Otrzymana p-wartość wynosi $p = 0.01971$, co każe nam jednak odrzucić hipotezę zerową, a w dalszym kroku poszukać alternatywnego modelu.

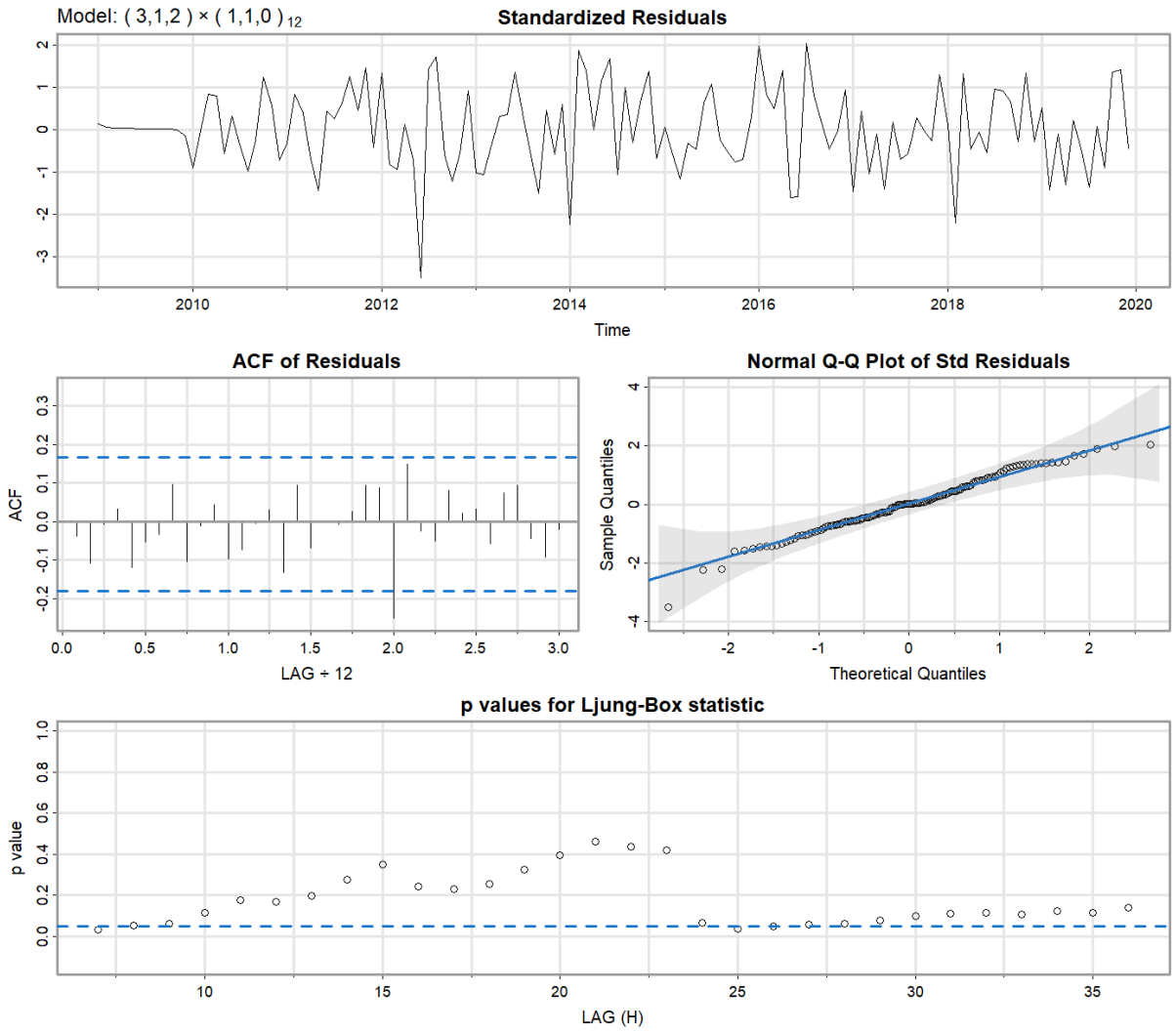
Wybór parametrów $d = 1$ i $D = 1$ pozwala otrzymać stacjonarne reszty. Wykres residuów oraz funkcji ACF i PACF dla modelu SARIMA$(0,1,0) \times (0,1,0)_{12}$ przedstawiają się następująco:





Na wykresie PACF widzimy, że przed pierwszym opóźnieniem sezonowym 3 obserwacje wykraczają poza 95% przedział ufności. Stąd mamy p=3, przyjmiemy też P=1, ponieważ widzimy, że trzynasta obserwacja również wykracza poza przedział ufności. Na wykresie ACF widzimy dwie obserwacje wystające poza przedział ufności przed pierwszym opóźnieniem sezonowym, co nam daje q=2.

Spójrzmy, jak w przypadku tego modelu zachowują się reszty.



Reszty w modelu dobranym ręcznie zachowują się w podobny sposób jak te w modelu wyznaczonym przez funkcję **auto.arima**. Średnia residuów jest tym razem niższa, w przybliżeniu to $\bar{x} = 0.0007$. Wynik testu Ljunga-Boxa jest zasadniczo ten sam, choć warto odnotować, że otrzymane w nim p-wartości często były niższe niż poprzednio – może to wskazywać na pewne mankamenty modelu. W teście Shapiro-Wilka otrzymujemy tym razem p-wartość równą $p = 0.1274$, a zatem większą niż poziom istotności. Nie mamy zatem podstaw, by odrzucić hipotezę zerową, co jest w tym przypadku optymistycznym wnioskiem.

Na koniec porównajmy statystyki AIC, AICc i BIC obliczone dla kilku wybranych modeli. W przypadku tych statystyk mniejsza wartość świadczy o lepszym dopasowaniu modelu. Uwzględnimy dwa omówione wyżej modele. Oprócz nich sprawdzimy te same modele dopasowane do szeregu, wobec którego w celu zmniejszenia rozrzutu danych zastosowano transformację Boxa-Coxa, oznaczoną w tabeli jako BC (funkcja **auto.arima** dla tych danych również zwraca model SARIMA$(0,1,1)\times(0,1,1)_{12}$) oraz jeden „kontrolny” model, który podczas ręcznego dobierania modelu w testach został wzięty pod uwagę, a teraz posłuży za punkt odniesienia.

	AIC	AICc	BIC
SARIMA(0,1,1)x(0,1,1)12	-3.254331	-3.253461	-3.184269
SARIMA(3,1,2)x(1,1,0)12	-3.147800	-3.141497	-2.984322
SARIMA(0,1,1)x(0,1,1)12 [BC]	-2.020563	-2.019693	-1.950501
SARIMA(3,1,2)x(1,1,0)12 [BC]	-1.922917	-1.916615	-1.759439
SARIMA(2,1,2)x(1,1,0)12	-3.126864	-3.122402	-2.986740

Tabela 2: Statystyki AIC, AICc oraz BIC dla kilku modeli

Otrzymane wyniki, z wyłączeniem tych (nieco słabszych), które odpowiadają modelom dopasowanym do danych po transformacji Boxa-Coxa, są zbliżone. Najlepsze z nich uzyskano dla modeli SARIMA$(0,1,1)\times(0,1,1)_{12}$ i SARIMA$(3,1,2)\times(1,1,0)_{12}$ i to właśnie te modele posłużą w dalszych rozważaniach do prognozowania liczby turystów.

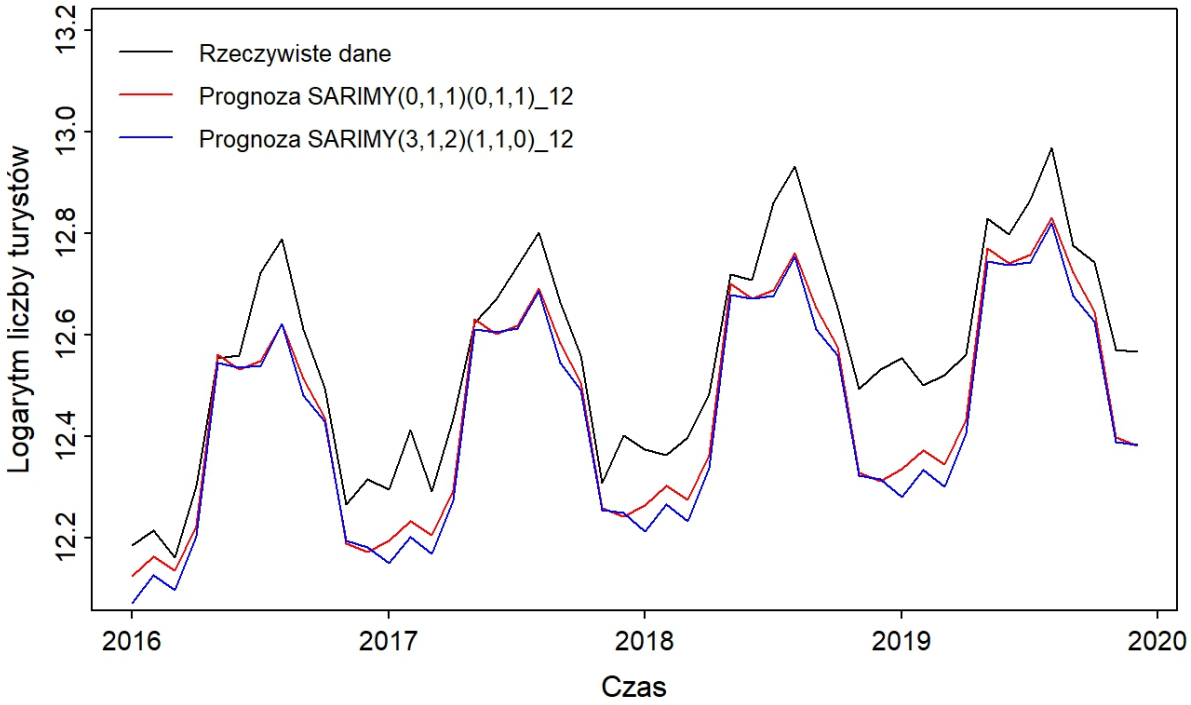
4. Prognozowanie

4.1. Testowanie jakości prognoz dla obu modeli

W poprzednim rozdziale udało się dopasować 2 modele do naszego szeregu: $\text{SARIMA}(0,1,1)\times(0,1,1)_{12}$ oraz $\text{SARIMA}(3,1,2)\times(1,1,0)_{12}$. Teraz zajmiemy się testowaniem jakości predykcji dla obu tych modeli, aby później móc wybrać jeden z nich do prognozowania.

Testując, przeprowadzono prognozę szeregu na 4 lata na okres 01.2016-12.2019. Zbiorem treningowym były dane od 01.2009 do 12.2015.

Prognozę przeprowadzono z użyciem funkcji **forecast** w RStudio. Poniższy wykres przedstawia porównanie faktycznych wartości szeregu z prognozowanymi przez oba modele:



Widzimy, że prognozy obu modeli podobnie się prezentują. Największe różnice między nimi możemy dostrzec w miesiącach zimowych, kiedy liczba turystów jest mniejsza - wtedy predykcje dla szeregu $\text{SARIMA}(0,1,1)\times(0,1,1)_{12}$ osiągają większe wartości od predykcji drugiego modelu. Tym samym nieco mniej różnią się od danych faktycznych. Możemy zauważyć, że prognozowane wartości są niższe niż dane faktyczne, zwłaszcza w miesiącach letnich i zimowych Obie prognozy zachowują podstawowe własności badanego szeregu: sezonowość oraz trend wzrostowy.

Aby zobaczyć, jak zmienia się dokładność prognoz wraz z upływem czasu, obliczymy średni błąd prognozy dla każdego roku (2016, 2017, 2018 i 2019). Z racji, że prognozy obu modeli nie różnią się znacząco, obliczymy średnie błędy dla modelu $\text{SARIMA}(0,1,1)\times(0,1,1)_{12}$. Otrzymujemy wyniki:

- dla roku 2016: 0.07955,
- dla 2017: 0.09491,
- dla 2018: 0.11779,
- dla 2019: 0.12688.

Rezultaty są zgodne z intuicją: średni błąd prognozy rośnie wraz z upływem czasu.

Porównamy jeszcze wartości paru statystyk dla obu prognoz:

- **Mean Squared Error (MSE)**, czyli błąd średniokwadratowy, który jest liczony według wzoru:
$$MSE = \frac{1}{n}\sum_{t=1}^{n}\left(Y_t - \hat{Y}_t\right)^2$$
- **Mean Absolute Percentage Error (MAPE)**, czyli średni procentowy błąd bezwzględny, który wyraża się wzorem:
$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{Y_t - \hat{Y}_t}{Y_t}\right|$$

Wartości wyżej wymienionych statystyk dla obu modeli znajdziemy w poniższej tabeli:

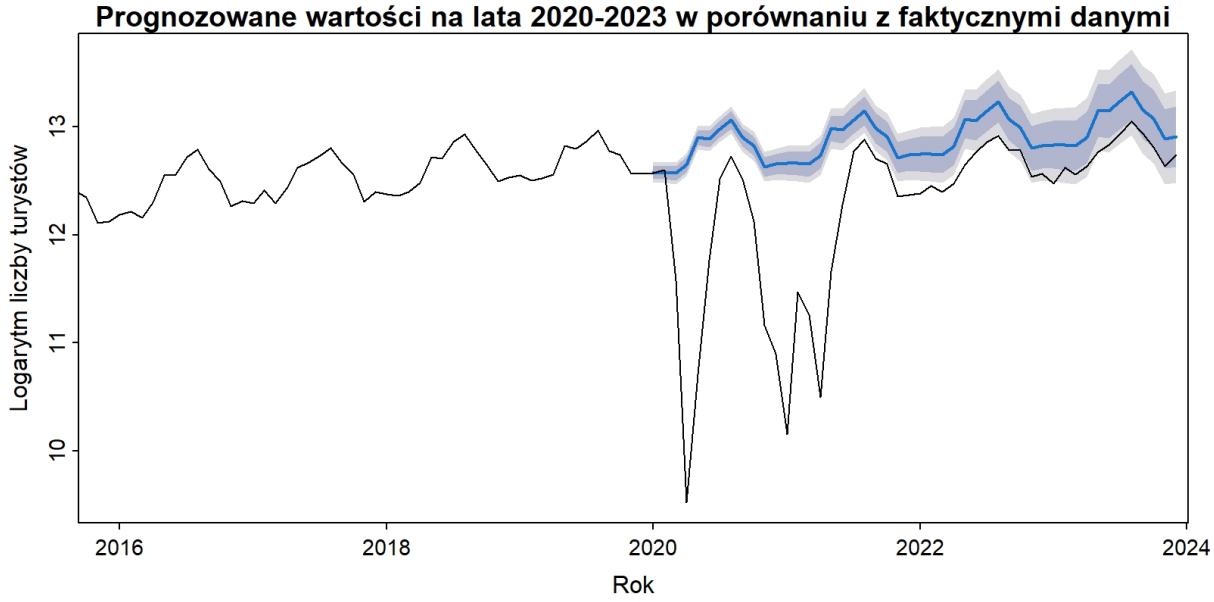
Model	MSE	MAPE
SARIMA$(0,1,1)\times(0,1,1)_{12}$	0.01423	0.00838
SARIMA$(3,1,2)\times(1,1,0)_{12}$	0.01897	0.00993

Możemy zauważyć, że model SARIMA$(0,1,1)\times(0,1,1)_{12}$ ma mniejsze wartości MSE oraz MAPE od modelu SARIMA$(3,1,2)\times(1,1,0)_{12}$.

Podsumowując, lepsze predykcje ma model SARIMA$(0,1,1)\times(0,1,1)_{12}$. Użyjemy go zatem do przeprowadzenia prognozy na lata 2020-2023.

4.2. Prognozowanie na lata 2020-2023

Teraz zajmiemy się prognozą liczby turystów na lata 2020-2023, aby przekonać się, jakie wartości moglibyśmy obserwować, gdyby nie pandemia.



Niebieskim kolorem zaznaczono prognozowane wartości, natomiast szare obszary to przedziały ufności (ciemnoszary - 80% przedział ufności, jasnoszary - 95% przedział ufności).

Widzimy, że w 2023 roku rzeczywiste wartości szeregu zawierają się w 95% przedziale ufności, co oznacza, że szereg zbliża się do wartości prognozowanych przez model. Można wyciągnąć wniosek, że pandemia znacząco wpłynęła na liczbę turystów na Dolnym Śląsku jedynie krótkotrwale.