Uniwersytet Warszawski

Analiza szeregów czasowych – projekt

Mateusz Tomczak

Nr albumu: 432561

Projekt zaliczeniowy na zajęcia

z Analizy Szeregów Czasowych, gr. nr 303

prowadzone przez dr Natalię Nehrebecką

Semestr letni 2022/2023

Warszawa, czerwiec 2023

*Oświadczenie autora pracy*

Świadom odpowiedzialności prawnej oświadczam, że niniejsza praca została napisana samodzielnie i nie zawiera treści uzyskanych w sposób niezgodny z obowiązującymi przepisami.

Data Podpis autora pracy

18.06.2023r.

**WSTĘP**

W pracy przedstawiono analizę dwóch szeregów czasowych. Pierwszy z nich dotyczy łącznej liczby mil przebytych przez pasażerów kolei w Stanach Zjednoczonych w każdym miesiącu od stycznia 2000 roku do grudnia 2022 roku i jest on szeregiem sezonowym. Drugi szereg, niesezonowy, przedstawia zmiany światowej ceny kawy wyrażonej w centach amerykańskich za funt (U.S. cents per pound) w przedziale miesięcznym od stycznia 2000 roku do lutego 2023 roku.

Dla każdego z przedstawionych szeregów dokonana została dekompozycja, następnie dopasowany został model z klasy ARIMA (w przypadku szeregu niesezonowego) lub SARIMA (w przypadku szeregu sezonowego) i model ekstrapolacyjny, następnie porównane zostały jakości prognoz z tych modeli. Każdy z szeregów został podzielony na część in-sample oraz out-of-sample w celu zbadania jakości prognozy. W przypadku szeregu sezonowego próbka out-of-sample wyniosła 12 najnowszych obserwacji (pełny cykl sezonowy), natomiast dla szeregu niesezonowego próbka ta wyniosła 3 najnowsze obserwacje. W badaniach założony został poziom istotności na poziomie 5%. Analiza przeprowadzona została przy wykorzystaniu środowiska R oraz dostępnych w nim bibliotek.

**ROZDZIAŁ I**

**SZEREG NIESEZONOWY**

* 1. **Opis szeregu**

Wybrany szereg niesezonowy przedstawia wartości światowej ceny kawy. Źródłem tych danych jest baza danych Federal Reserve Economic Data (FRED). Badany szereg obejmuje łącznie 277 obserwacji. Pierwsza obserwacja występuje w styczniu 2000 roku, natomiast ostatnie przypada na luty 2023 roku. Wartości obserwacji wyrażone są w centach amerykańskich na funt kawy. Na Rys. 1 przedstawiony został wykres obserwacji w czasie.

Obraz zawierający tekst, diagram

Opis wygenerowany automatycznieRys. 1. Wykres szeregu niesezonowego

*Źródło:* Opracowanie własne.

* 1. **Dekompozycja szeregu czasowego**

Na podstawie Rys. 1 nie można określić jednorodnego trendu, co potwierdza dekompozycja szeregu przedstawiona na Rys. 2. Nie można także na bazie obserwacji stwierdzić, że szereg posiada sezonowość, co potwierdziła analiza szeregu w programie JDemetra+, która wykazała brak sezonowości. Analiza wykazała również jedną obserwację odstającą o charakterze długotrwałym (level shift), która wystąpiła w październiku 2008 roku.

Rys. 2. Dekompozycja szeregu niesezonowego

Obraz zawierający tekst, Czcionka, diagram, Wykres

Opis wygenerowany automatycznie

*Źródło:* Opracowanie własne.

* 1. **Model ARIMA**

W celu budowy modelu, który pozwoli na prognozowanie analizowanego szeregu niesezonowego posłużono się modelem ARIMA (ang. Autoregressive Integrated Moving Average), zbudowanego przy wykorzystaniu procedury Boxa-Jenkinsa (identyfikacja – estymacja – diagnostyka – prognozowanie). Model ten składa się z procesów: autoregresyjnego AR(p), średniej ruchomej MA(q) oraz stopnia integracji I(d). Do przeprowadzenia prognozy obserwacje w szeregu niesezonowym muszą zostać podzielone na okresy in-sample oraz out-of-sample. Ponieważ model ARIMA może być wykorzystywany jedynie do prognozowanie krótkoterminowego, okres out-of-sample wyniesie 3 miesiące. Dodatkowo wymagane jest, aby prognozowany przy pomocy tego modelu szereg wykazywał stacjonarność, co należy sprawdzić w pierwszej kolejności.

* + 1. **Stacjonarność szeregu**

Do weryfikacji stacjonarności szeregu posłużono się w pierwszej kolejności testem Dickey-Fullera (DF). Hipoteza zerowa tego testu stwierdza, że weryfikowana zmienna jest błądzeniem przypadkowym, czyli nie wykazuje stacjonarności. Dodatkowo, aby wyeliminować ewentualną autokorelację reszt posłużono się rozszerzoną wersją testu Dickey-Fullera (ADF). Natomiast do weryfikacji występowania autokorelacji reszt wykorzystano test Breuscha-Godfreya. Wyniki wspomnianych testów zostały przedstawione w Tabeli 1. Zastosowany został wariant testu ze stałą. W przypadku testu DF (ADF z liczbą opóźnień k=0) występuje autokorelacja reszt, zatem wymagane jest wykorzystanie testu ADF. Na podstawie wyników zauważyć można, że problem autokorelacji wyeliminowany został dla rozszerzenia w wysokości 1. Również w każdym z przypadków p-value dla testu ADF było wyższe od założonego poziomu istotności 5%, zatem nie ma podstaw do odrzucenia hipotezy zerowej stwierdzającej, że szereg jest niestacjonarny. Wymagane jest zatem różnicowanie zmiennej. Wyniki tych samych testów dla pierwszych różnic zmiennej przedstawione zostały w Tabeli 2.

Tabela 1. Wyniki przeprowadzonych testów dla szeregu niezesonowego

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Liczba opóźnień k** | **Test DF/ADF** | | **p-value dla danego rzędu korelacji w teście B-G** | | | |
| **Statystyka testowa** | **p-value** | **1** | **2** | **3** | **4** |
| 0 | -1.291 | 0.576 | 1.935e-05 | 1.085e-04 | 3.423e-04 | 8.891e-04 |
| 1 | -1.756 | 0.402 | 0.953 | 0.985 | 0.972 | 0.976 |
| 2 | -1.772 | 0.368 | 0.975 | 0.998 | 0.997 | 0.997 |

*Źródło*: Opracowanie własne.

Tabela 2. Wyniki przeprowadzonych testów dla pierwszych różnic zmiennej

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Liczba opóźnień k** | **Test DF/ADF** | | **p-value dla danego rzędu korelacji w teście B-G** | | | |
| **Statystyka testowa** | **p-value** | **1** | **2** | **3** | **4** |
| 0 | -12.532 | <0.01 | 0.984 | 0.961 | 0.982 | 0.990 |
| 1 | -9.868 | <0.01 | 0.981 | 0.984 | 0.987 | 0.992 |
| 2 | -8.121 | <0.01 | 0.972 | 0.998 | 0.999 | 0.999 |

*Źródło:* Opracowanie własne.

Wykres przedstawiający szereg po pierwszym różnicowaniu przedstawiony został na Rys. 3. Można na nim zauważyć, że pierwsze różnice zmiennej oscylują wokół wartości 0, zatem możliwe jest zastosowanie testu ADF bez stałej. Jak widać z wyników pierwsze różnicowanie zmiennej niesezonowej wystarczyło, aby otrzymać szereg stacjonarny. Potwierdza to także test KPSS, dla którego wartość statystyki testowej wyniosła 0.0595 natomiast wartość krytyczna dla poziomu istotności 5% wyniosła 0.463. Nie ma więc podstaw do odrzucenia hipotezy zerowej stwierdzającej stacjonarność szeregu. Zatem zmienna ta jest zintegrowana rzędu 1. Również w takim przypadku autokorelacja reszt nie występuje już od liczby opóźnień na poziomie 0.

Rys. 3. Wykres zmiennej niesezonowej po pierwszym różnicowaniu

Obraz zawierający tekst, Czcionka, diagram, zrzut ekranu

Opis wygenerowany automatycznie

*Źródło:* Opracowanie własne.

* + 1. **Identyfikacja, estymacja oraz diagnostyka**

Po analizie przeprowadzonej w poprzednim podrozdziale wiemy, że parametr d budowanego modelu ARIMA wynosi 1, ponieważ analizowany szereg jest zintegrowany rzędu 1. Jako kolejny etap budowy modelu należy znaleźć parametry p, czyli rząd procesu AR, oraz q, czyli rząd procesu MA. Do tego celu wykorzystywane są wykresy funkcji ACF (przedstawiony na Rys. 4) oraz PACF (przedstawiony na Rys. 5) dla różnicowanej zmiennej. Do ustalenia parametru p wykorzystany został wykres PACF. Na bazie tego wykresu można przypisać wspomnianemu parametrowi wartość 1. Z kolei do ustalenia wartości parametru q posłużono się wykresem ACF, który wskazuje, że także może on przyjąć wartość 1.

Rys. 4. Wykres funkcji ACF

Obraz zawierający tekst, zrzut ekranu, linia, diagram

Opis wygenerowany automatycznie

*Źródło:* Opracowanie własne.

Rys. 5. Wykres funkcji PACF

Obraz zawierający tekst, linia, diagram, zrzut ekranu

Opis wygenerowany automatycznie

*Źródło:* Opracowanie własne.

Zatem wartości przedstawione na wykresach ACF oraz PACF sugerują model ARIMA(1,1,1). Natomiast w celu znalezienia optymalnego modelu przetestowane zostały wszystkie kombinacje dla wartości p = {0,1} oraz q = {0,1}. Jedynie modele ARIMA(0,1,1) oraz ARIMA(1,1,0) posiadały istotne statystycznie oszacowania. Jednak mimo nieistotnych oszacowań postanowiono, że model ARIMA(1,1,1) pozostanie w badaniu do dalszej analizy w celu lepszego porównania modeli. Dane dotyczące ostatecznie analizowanych modeli przedstawione zostały w Tabeli 3.

Tabela 3. Wybrane modele ARIMA

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **AIC** | **BIC** | **Test Ljunga-Boxa** | |
| **Statystyka testowa** | **p-value** |
| ARIMA(1,1,0) | 1524.487 | 1531.713 | 21.873 | 0.587 |
| ARIMA(0,1,1) | 1525.217 | 1532.443 | 21.330 | 0.619 |
| ARIMA(1,1,1) | 1528.322 | 1542.774 | 21.628 | 0.601 |

*Źródło:* Opracowanie własne.

W przypadku testu Ljunga-Boxa ma on za zadanie weryfikację założenia, że reszty w analizowanych modelach są białym szumem. Jak można zauważyć reszty każdego z badanych modeli miały charakter białego szumu, co także potwierdziła graficzna analiza wykresów funkcji ACF oraz PACF. Kierując się kryterium informacyjnym Akaike (AIC) oraz Bayesowskim kryterium Schwarza (BIC) można stwierdzić, że model ARIMA(1,1,0) jest najlepszy, ponieważ w jego przypadku oba kryteria posiadają najniższą wartość. Natomiast w celu porównania wyników, do dalszej analizy wykorzystano również pozostałe analizowane modele, czyli ARIMA(1,1,1) oraz ARIMA(0,1,1).

* + 1. **Prognozowanie**

Po znalezieniu optymalnych wartości parametrów p oraz q w modelu ARIMA można przejść do etapu prognozowania szeregu. Prognozowane będą trzy obserwacje, które zostały usunięte z danych w celu uzyskania próbki in-sample. Należy nadmienić, iż wykorzystany model ARIMA zawiera maksymalny rząd p oraz q na poziomie 1, więc jego predykcje będą najlepsze w przypadku prognozowania na jeden okres naprzód, jednak mimo tego zdecydowano się na pozostanie przy prognozowaniu trzech okresów w przeprowadzonym badaniu. Wykres obrazujący prognozę modelu ARIMA(1,1,0) przedstawiony został na Rys. 6. Wspomniany wykres przedstawia jedynie grupę najbardziej aktualnych obserwacji w celu łatwiejszego odczytania otrzymanych prognoz. Następnie w celu zweryfikowania jakości oszacowań policzone zostały wartości MAE (ang. Mean Absolute Error), MSE (ang. Mean Square Error), MAPE (ang. Mean Absolute Percentage Error) oraz AMAPE (ang. Adjusted Mean Absolute Percentage Error), których wyniki przedstawione zostały w Tabeli 4. Zdecydowano się na weryfikację wszystkich błędów, mimo iż w przypadku analizowanego szeregu teoretycznie najlepszy byłby MSE (wartości obserwacji są dodatnie oraz co do modułu większe od 1). Jak widać na przedstawionej tabeli, mimo iż model ARIMA(1,1,0) okazał się najlepszy pod względem kryteriów AIC oraz BIC, to model ARIMA(0,1,1) okazał się być najlepszy pod względem jakości prognozy.

Rys. 6. Prognoza szeregu niesezonowego

Obraz zawierający tekst, diagram, linia, Wykres

Opis wygenerowany automatycznie

*Źródło:* Opracowanie własne.

Tabela 4. Wartości błędów prognozy ex post analizowanych modeli ARIMA

|  |  |  |  |
| --- | --- | --- | --- |
|  | **ARIMA(1,1,0)** | **ARIMA(0,1,1)** | **ARIMA(1,1,1)** |
| MAE | 8.088 | 6.959 | 7.968 |
| MSE | 91.933 | 71.669 | 89.485 |
| MAPE | 0.081 | 0.069 | 0.079 |
| AMAPE | 0.043 | 0.036 | 0.042 |

*Źródło:* Opracowanie własne.

* 1. **Model ekstrapolacyjny**

Jako kolejną metodę prognozowania szeregu czasowego wykorzystano model ekstrapolacyjny. W tym przypadku do analizy wykorzystany został model liniowy Holta. Konieczne w tym przypadku jest określenie parametrów α oraz β, które dobierane są w taki sposób, aby minimalizować błąd średniokwadratowy. Na Rys. 7 przedstawiony został wygładzony przy użyciu liniowej metody Holta analizowany szereg z wygenerowanymi automatycznie wartościami α=1 i β=0.152. Natomiast na Rys. 8 pokazana została prognoza analizowanego szeregu na 3 najbardziej aktualne okresy przy użyciu wspomnianej metody. Ponownie w celu weryfikacji jakości prognozy obliczone zostały wartości jej błędów, przedstawione w Tabeli 5.

Rys. 7. Szereg niesezonowy wygładzony metodą liniową Holta

Obraz zawierający tekst, diagram, Wykres, linia

Opis wygenerowany automatycznie

*Źródło:* Opracowanie własne.

Rys. 8. Prognoza szeregu przy użyciu metody liniowej Holta.

Obraz zawierający tekst, diagram, linia, Wykres

Opis wygenerowany automatycznie

*Źródło:* Opracowanie własne.

Tabela 5. Błędy prognozy dla modelu liniowego Holta

|  |  |
| --- | --- |
|  | **Model liniowy Holta** |
| MAE | 8.178 |
| MSE | 104.684 |
| MAPE | 0.081 |
| AMAPE | 0.043 |

*Źródło:* Opracowanie własne.

* 1. **Porównanie otrzymanych modeli**

Jako podsumowanie przedstawionej analizy szeregu niesezonowego porównane zostały otrzymane prognozy modelu ARIMA oraz liniowego Holta. Ponieważ model ARIMA(0,1,1), bazując na wartościach błędów prognozy ex post, był dokładniejszy w swoich prognozach to właśnie on został porównany do wyników otrzymanych przy wykorzystaniu modelu ekstrapolacyjnego. Prognozy porównywanych modeli przedstawione zostały na Rys. 9. Jak można zauważyć na wykresie oraz wcześniejszych tabelach przedstawiających błędy prognoz dla porównywanych modeli, model ARIMA(0,1,1) okazał się lepszy w prognozowaniu analizowanego szeregu.

Obraz zawierający tekst, diagram, linia, Wykres

Opis wygenerowany automatycznieRys. 9. Prognozy modeli ARIMA oraz liniowego Holta

*Źródło:* Opracowanie własne.

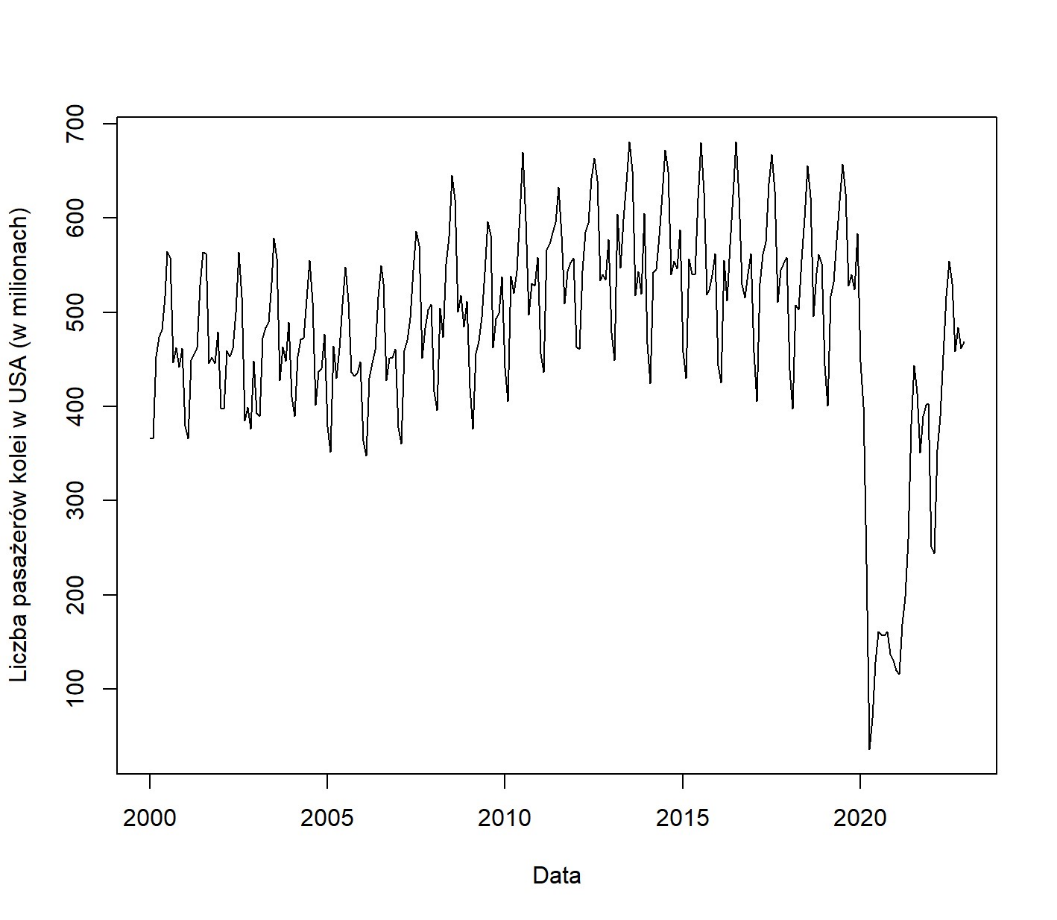
**ROZDZIAŁ II**

**SZEREG SEZONOWY**

* 1. **Opis szeregu**

Analizowany szereg sezonowy przedstawia łączną ilość przebytych mil przez pasażerów kolei w Stanach Zjednoczonych. Dane pochodzą z bazy danych Federal Reserve Economic Data (FRED). Szereg obejmuje łącznie 263 obserwacje. Pierwsza obserwacja występuje w styczniu 2000 roku, natomiast ostatnia przypada na grudzień 2022 roku. Na Rys. 10 przedstawiony został wykres obserwacji w czasie.

Rys. 10. Szereg sezonowy



*Źródło:* Opracowanie własne.

* 1. **Dekompozycja szeregu**

Na bazie Rys. 10 można stwierdzić sezonowość szeregu, co potwierdza program JDemetra+. Natomiast na bazie obserwacji sezonowość można uznać za addytywną, ze względu na brak jasno widocznej zmiany amplitudy wahań sezonowych wraz z trendem. Dodatkowo program JDemetra+ wskazał pięć obserwacji odstających. Obserwacje o charakterze długotrwałym (level shift) pochodzą z marca oraz kwietnia 2020 roku i stycznia oraz czerwca 2021 roku. Z kolei obserwacja odstająca z września 2020 roku ma charakter przejściowy (temporary change). Na Rys. 11 przedstawiona została dekompozycja analizowanego szeregu sezonowego na jego poszczególne składowe. Zauważyć można, że trend w szeregu był stały lub lekko rosnący z widocznym spadkiem i następującym po nim wzroście wynikającym z pandemii COVID-19 w 2020 roku.

Rys. 11. Dekompozycja szeregu sezonowego

Obraz zawierający tekst, diagram, Czcionka, linia

Opis wygenerowany automatycznie

*Źródło:* Opracowanie własne.

* 1. **Model SARIMA**

Do budowy modelu pozwalającego na prognozowanie szeregu wykorzystano model SARIMA, czyli sezonową wersję modelu ARIMA. Ponownie do jego budowy wykorzystana została procedura Boxa-Jenkinsa. Model ten można podzielić na procesy: AR(p), MA(q), I(d) oraz sezonowego autoregresyjnego SAR(P), sezonowego średniej ruchomej SMA(Q) i sezonowego stopnia integracji SI(D). W celu budowy modelu i przeprowadzenia prognozy obserwacje zostały podzielone na okresy in-sample oraz out-of-sample. Ponieważ analizowany szereg składa się z obserwacji miesięcznych, okres out-of-sample wyniesie 12 obserwacji, które będą prognozowane. Ponownie, w pierwszej kolejności należy zweryfikować stacjonarność szeregu.

* + 1. **Stacjonarność szeregu**

Do przeprowadzenia prognozy szeregu przy wykorzystaniu modelu SARIMA niezbędne jest, aby wykazywał on stacjonarność oraz brak sezonowości. W pierwszym kroku stacjonarność oraz sezonowość sprawdzone będą na bazie weryfikacji graficznej. W kolejnym kroku stacjonarność zostanie zweryfikowana na podstawie testów Dickey-Haszy-Fullera oraz KPSS. Na Rys. 12 oraz Rys. 13 przedstawione zostały, w kolejności, wykresy ACF oraz PACF dla analizowanego szeregu sezonowego.

Rys. 12. Wykres ACF szeregu sezonowego

Obraz zawierający tekst, diagram, linia, Wykres

Opis wygenerowany automatycznie

*Źródło:* Opracowanie własne.

Jak można wywnioskować ze wspomnianych rysunków szereg prawdopodobnie cechuje się niestacjonarnością (wypustki nie znikają w szybkim tempie i nie są bliskie zeru) oraz na wykresie ACF widać silną sezonowość poprzez istotne wypustki w pozycjach 12, 24 oraz 36. Można zatem przypuszczać, że szereg jest niestacjonarny regularnie i sezonowo. Zastosowane zostało zatem pierwsze różnicowanie szeregu sezonowego. Korelogramy dla szeregu zróżnicowanego w stopniu pierwszym widoczne są na Rys. 14 oraz Rys. 15. Jak można na nich zaobserwować, występuje silna sezonowość (wypustki statystycznie istotne dla opóźnień równych 12, 24, 36). Wymagane jest zatem wykorzystania różnicowania sezonowego. Na Rys. 16 oraz Rys. 17 przedstawione zostały wykresy funkcji ACF oraz PACF dla szeregu zróżnicowanego regularnie w stopniu pierwszym oraz sezonowo w stopniu pierwszym (d = 1 oraz D = 1).

Obraz zawierający tekst, diagram, linia, Wykres

Opis wygenerowany automatycznieRys. 13. Wykres PACF szeregu sezonowego

*Źródło:* Opracowanie własne.

Rys. 14. Korelogram zróżnicowanego szeregu (ACF)

Obraz zawierający diagram, linia, tekst, Wykres

Opis wygenerowany automatycznie

*Źródło:* Opracowanie własne.

Rys. 15. Korelogram zróżnicowanego szeregu (PACF)

Obraz zawierający tekst, diagram, linia, Wykres

Opis wygenerowany automatycznie

*Źródło:* Opracowanie własne.

Rys. 16. Korelogram szeregu zróżnicowanego regularnie i sezonowo (ACF)

Obraz zawierający tekst, linia, diagram, Wykres

Opis wygenerowany automatycznie

*Źródło:* Opracowanie własne.

Obraz zawierający diagram, tekst, linia, Rysunek techniczny

Opis wygenerowany automatycznieRys. 17. Korelogram szeregu zróżnicowanego regularnie i sezonowo (PACF)

*Źródło:* Opracowanie własne.

W celu sprawdzenia zasadności wykorzystania różnicowania sezonowego posłużono się testem Dickey-Haszy-Fullera. W przypadku zerowego opóźnienia reszty wykazywały silną autokorelację. Zastosowana została zatem rozszerzona wersja testu. W celu usunięcia autokorelacji wymagane było zastosowanie 3 opóźnień w teście. W takim przypadku autokorelacja reszt nie występowała dla badanych pięciu pierwszych rzędów. Ostateczna wartość statystyki testowej wyniosła -0.840, natomiast wartość statystyki krytycznej była na poziomie około -5.83. Nie ma zatem podstaw do odrzucenia hipotezy zerowej stwierdzającej, że występuje sezonowy pierwiastek jednostkowy. Wykorzystanie różnicowania sezonowego jest zatem zasadne.

W kolejnym kroku sprawdzona została regularna stacjonarność szeregu przy wykorzystaniu testu ADF. Co ciekawe, już w przypadku szeregu jedynie różnicowanego sezonowo test ten wykazywał na stacjonarność szeregu (wartość p-value równa 0.01 po wyeliminowaniu autokorelacji reszt w 1 opóźnieniu). Podobnie w przypadku testu KPSS, który dla poziomu istotności 5% (wartość krytyczna równa 0.463) także stwierdzał stacjonarność (statystyka testowa równa 0.224). Zatem sprawdzony został wykres korelogramu dla funkcji ACF, przedstawiony na Rys. 18, w celu weryfikacji, czy również na nim widoczna jest stacjonarność szeregu.

Rys. 18. Korelogram szeregu różnicowanego sezonowo (ACF)

Obraz zawierający tekst, diagram, linia, Wykres

Opis wygenerowany automatycznie

*Źródło:* Opracowanie własne.

Jak można zauważyć na Rys. 18 wykres funkcji ACF nie wykazuje jasno cech wskazujących na stacjonarność szeregu – wypustki nie spadają w znacząco szybkim tempie do zera, natomiast zbliżają się do zera stopniowo, pozostając relatywnie długo w przedziale istotności. Podobnie widoczne „odbicie” w 10 opóźnieniu tworzące stopniowy szybki wzrost oraz wolny spadek do zera budzą pewne podejrzenia co do wniosków, na które wskazują formalne testy na stacjonarność. Zatem na bazie obserwacji graficznej i sprzecznie z wnioskami wynikającymi z testów ADF oraz KPSS zdecydowano się na wykorzystanie szeregu różnicowanego regularnie oraz sezonowo.

Wyniki testu ADF i B-G dla szeregu różnicowanego regularnie i sezonowo przedstawione zostały w Tabeli 6, natomiast jej wykres widoczny jest na Rys. 19. Jak można zauważyć na wspomnianym wykresie jasno widoczny jest wpływ obserwacji odstających wynikających z pandemii COVID-19. Bazując na fakcie, że zmienna oscyluje wokół zera wykorzystany został test ADF bez stałej. Jak można zauważyć już w przypadku zerowego opóźnienia szereg wykazuje stacjonarność. Na podobne wnioski wskazuje test KPSS, który także wykazuje stacjonarność szeregu (wartość statystyki testowej równa 0.069 dla wartości krytycznej 0.463).

Tabela 8. Wyniki przeprowadzonych testów dla pierwszych różnic regularnych i sezonowych zmiennej

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Liczba opóźnień k** | **Test DF/ADF** | | **p-value dla danego rzędu korelacji w teście B-G** | | | |
| **Statystyka testowa** | **p-value** | **1** | **2** | **3** | **4** |
| 0 | -13.384 | <0.01 | 0.705 | 0.062 | 0.123 | 0.216 |
| 1 | -8.791 | <0.01 | 0.908 | 0.982 | 0.921 | 0.974 |
| 2 | -8.066 | <0.01 | 0.986 | 0.999 | 0.999 | 0.998 |

*Źródło:* Opracowanie własne.

Rys. 19. Wykres zmiennej różnicowanej regularnie i sezonowo

Obraz zawierający tekst, diagram, pismo odręczne, Czcionka

Opis wygenerowany automatycznie

*Źródło:* Opracowanie własne.

* + 1. **Identyfikacja, estymacja oraz diagnostyka**

Kolejnym krokiem w budowie modelu SARIMA jest identyfikacja jego parametrów (p,d,q) oraz (P,D,Q) s. Parametr d przyjmuje wartość 1 ze względu na to, że wykonane zostało pierwsze różnicowanie szeregu. Także parametr D przyjmuje wartość 1 ze względu na różnicowanie sezonowe szeregu. Widoczna zmienna s przyjmuje wartość 12, ponieważ analizowane dane bazują na miesięcznych obserwacjach. Z kolei do identyfikacji parametrów p i q oraz P i Q należy wykorzystać wykresy funkcji ACF i PACF dla szeregu różnicowanego regularnie i sezonowo, widoczne na Rys. 16 oraz Rys. 17.

Wnioskując z wykresu funkcji ACF można założyć, że parametr q przyjmie wartość 2. W przypadku parametru Q, jak można zauważyć, istotna jest jedynie wypustka na 12 pozycji, zatem przyjmie on wartość 1. Analizując wykres funkcji PACF można założyć, że parametr p przyjmie wartość 2, natomiast w przypadku szacowania wartości parametru P zauważyć można, że każda wypustka na pozycji 12, 24 oraz 36 jest istotna, zatem przyjmie on wartość 3.

Ostatecznie w celu znalezienia optymalnego modelu przeanalizowane zostaną modele z różnymi wartościami parametrów. Dla każdego z nich zastosowany został następujący możliwy przedział wartości: p = {0, 1, 2}, q = {0, 1, 2}, P = {0, 1, 2, 3}, Q = {0, 1}. W pierwszej kolejności sprawdzone zostały wszystkie kombinacje dla parametrów sezonowych. Po usunięciu modeli posiadających nieistotne zmienne, parametry sezonowe przyjęły ostatecznie wartość P = 0, D = 1, Q = 1. W innych przypadkach stałe przy najwyższej wartości parametrów były nieistotne zatem nie zdecydowano się na wykorzystanie do analizy modeli z innymi wartościami oszacowań parametrów sezonowych. Dla takiego modelu wypustki dla wartości 12, 24 oraz 36 w wykresach dla reszt modelu funkcji ACF oraz PACF były nieistotne, zatem sezonowość została wyeliminowana. Również kryterium informacyjne AIC oraz BIC wskazały model ze wspomnianymi parametrami sezonowymi jako najlepszy.

W kolejnym etapie szacowane były wartości parametrów regularnych modelu SARIMA. Po usunięciu modeli, dla których zmienne najwyższej wartości parametru były nieistotne, oraz dla których test LR (Likelihood-ratio), czyli test ilorazu wiarygodności, nie dawał podstaw do odrzucenia hipotezy zerowej mówiącej, że model ograniczony jest prawdziwy, czyli jego dopasowanie do danych nie różni się istotnie od modelu bez ograniczeń, w analizie pozostały cztery modele. Wartości ich kryteriów informacyjnych AIC i BIC oraz informacji na temat testu Ljunga-Boxa zostały przedstawione w Tabeli 9.

Tabela 9. Wybrane modele SARIMA

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **AIC** | **BIC** | **Test Ljunga-Boxa** | |
| **Statystyka testowa** | **p-value** |
| SARIMA(2,1,2)×(0,1,1)12 | 2458.700 | 2479.852 | 30.493 | 0.728 |
| SARIMA(1,1,2)×(0,1,1)12 | 2458.252 | 2475.879 | 33.519 | 0.587 |
| SARIMA(0,1,1)×(0,1,1)12 | 2455.566 | 2466.142 | 35.657 | 0.485 |
| SARIMA(1,1,0)×(0,1,1)12 | 2455.262 | 2465.838 | 35.189 | 0.507 |

*Źródło:* Opracowanie własne.

Jak można zauważyć w Tabeli 9, kierując się kryteriami informacyjnymi AIC oraz BIC, najlepszym modelem okazał się model SARIMA(1,1,0)×(0,1,1)12, ponieważ w jego przypadku przyjmują one najniższą wartość. Jednak w celach porównawczych, pod względem błędów predykcji przeanalizowane zostaną wszystkie zaprezentowane modele.

* + 1. **Prognozowanie**

Mając już gotowe wartości zmiennych do modeli SARIMA możliwe jest przejście do prognozowania szeregu. Tak jak zostało wcześniej wspomniane prognozowane będzie 12 najnowszych obserwacji w szeregu. Na Rys. 20 przedstawiona została prognoza najlepszego, pod względem kryteriów informacyjnych, z modeli, czyli SARIMA(1,1,0)×(0,1,1)12, natomiast w Tabeli 10 przedstawione zostały wartości błędów prognozy dla każdego z analizowanych modeli. Również w przypadku szeregu sezonowego zdecydowano się przeanalizować wartości wszystkich błędów prognozy ex post.

Rys. 20. Wykres prognozy modelu SARIMA(1,1,0)×(0,1,1)12

Obraz zawierający tekst, diagram, linia, Wykres

Opis wygenerowany automatycznie

*Źródło:* Opracowanie własne.

Tabela 10. Wartości błędów prognozy analizowanych modeli SARIMA

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **SARIMA**  **(2,1,2)×(0,1,1)12** | **SARIMA**  **(1,1,2)×(0,1,1)12** | **SARIMA**  **(0,1,1)×(0,1,1)12** | **SARIMA**  **(1,1,0)×(0,1,1)12** |
| MAE | 77.978 | 81.564 | 80.284 | 80.929 |
| MSE | 7103.167 | 7618.838 | 7391.415 | 7501.012 |
| MAPE | 0.176 | 0.185 | 0.182 | 0.184 |
| AMAPE | 0.095 | 0.100 | 0.099 | 0.100 |

*Źródło:* Opracowanie własne.

Jak można zauważyć, w przypadku porównania jakości predykcji najlepszym z modeli jest model SARIMA(2,1,2)×(0,1,1)12 ze znaczącą przewagą w wartościach błędów prognozy w porównaniu do pozostałych modeli. Zatem zostanie on wykorzystany do porównania jakości prognozy z modelem ekstrapolacyjnym, zbudowanym w następnym podrozdziale.

* 1. **Model ekstrapolacyjny**

Inną opcją przeprowadzania prognoz szeregów czasowych jest wykorzystanie modeli ekstrapolacyjnych. Ponieważ analizowana zmienna zawiera wahania sezonowe, wykorzystany został model Holta-Wintersa. Jak zostało wcześniej ustalone efekty sezonowe w analizowanym szeregu wydają się mieć charakter addytywny, ponieważ są stałe w czasie. W związku z tym wykorzystany zostanie addytywny wariant modelu Holta-Wintersa. Parametry modelu dobrane w celu minimalizacji błędu średniokwadratowego przyjęły następujące wartości: α=1, β=0 oraz γ=1.149e-13. Na Rys. 21 przedstawiony został szereg z wygładzonymi wartościami analizowanego modelu. Natomiast na Rys. 22 przedstawiona została prognoza szeregu przy wykorzystaniu metody Holta-Wintersa. W Tabeli 11 pokazane zostały wartości błędów predykcji modelu.

Rys. 21. Szereg wygładzony przy wykorzystaniu modelu Holta-Wintersa

Obraz zawierający tekst, zrzut ekranu, Czcionka, diagram

Opis wygenerowany automatycznie

*Źródło:* Opracowanie własne.

Obraz zawierający tekst, diagram, linia, Wykres

Opis wygenerowany automatycznieRys. 22. Prognoza wykonana przy użyciu metody Holta-Wintersa

*Źródło:* Opracowanie własne.

Tabela 5. Błędy prognozy dla addytywnego modelu Holta-Wintersa

|  |  |
| --- | --- |
|  | **Addytywny model HW** |
| MAE | 55.884 |
| MSE | 3586.861 |
| MAPE | 0.140 |
| AMAPE | 0.070 |

*Źródło:* Opracowanie własne.

* 1. **Porównanie modeli**

Jak można zauważyć na bazie wartości błędów prognozy w przypadku szeregu sezonowego model ekstrapolacyjny okazał się lepszym predyktorem zachowań analizowanego szeregu. Model Holta-Wintersa osiągnął znaczące lepsze wyniki niż najlepszy znaleziony model sezonowy ARIMA, czyli SARIMA(2,1,2)×(0,1,1)12. Porównanie prognoz wspomnianych modeli przedstawione zostało na Rys. 23.

Rys. 23. Porównanie prognoz szeregu sezonowego

Obraz zawierający tekst, diagram, linia, Wykres

Opis wygenerowany automatycznie

*Źródło:* Opracowanie własne.

**ZAKOŃCZENIE**

Przedstawione w pracy badanie miało na celu pokazanie procesu analizy i prognozowania dwóch szeregów czasowych – nieposiadającego efektów sezonowych oraz posiadającego efekty sezonowe. Szereg niesezonowy został poddany krótkookresowej prognozie, wynoszącej 3 najnowsze obserwacje. W tym przypadku najlepszy okazał się model ARIMA(0,1,1), którego prognozy były dokładniejsze niż modelu liniowego Holta. Z kolei szereg sezonowy poddany został prognozie wynoszącej 12 miesięcy. W jego przypadku model ekstrapolacyjny okazał się lepszym predyktorem. Addytywny model Holta-Wintersa był w stanie uzyskać lepsze wyniki prognozowania w porównaniu do modelu SARIMA.