Projekt - Szeregi Czasowe

Klaudia Bała, Marta Gacek, Alicja Wiączkowska

Spis treści

1 Wstęp

Celem projektu jest analiza wybranych danych statystycznych, traktowanych jako szereg czasowy. Chcemy zbadać powiązaną z nimi problematykę i zweryfikować prawdziwość swoich założeń, a także przeprowadzić prognozę przyszłych wartości z tej dziedziny.

1.1 Wprowadzenie do danych

Dane, którymi się zajmujemy, pochodzą ze strony Głównego Urzędu Statystycznego i dotyczą polskiej demografii, a dokładniej liczby zawartych małżeństw. Baza podzielona jest na lata i miesiące, dzięki czemu wiadomo ile związków małżeńskich zostało zawartych w konkretnym okresie czasu. Materiał obejmuje zakres od początku roku 2002 do końca 2023.

Omawiane dane zostały wybrane ze względu na ich dobre predyspozycje do analizy. Powszechnie wiadomo, że daty ślubów - zarówno kościelnych, jak i cywilnych - są zazwyczaj wybierane nieprzypadkowo, według pewnych kryteriów. Najważniejszym z nich jest pora roku. Można spodziewać się, że najbardziej popularnym okresem zawierania małżeństw jest sezon letni - ze względu na najlepszą pogodę oraz fakt, że dla wielu ludzi jest to czas wakacji czy urlopów wypoczynkowych.

Istnieją również inne, bardziej nietypowe czynniki. Przykładowo, popularnym przesądem jest zawieranie małżeństwa w miesiącu, który w swej nazwie posiada literę „r", co ma przynosić młodej parze szczęście. Można więc wysnuć teorię, że ma to pewien wpływ na dobór odpowiedniej daty ślubu przez narzeczonych.

Powyższe czynniki - i nie tylko - mogą sugerować, że będziemy mieć do czynienia z wyraźną sezonowością w szeregu czasowym.

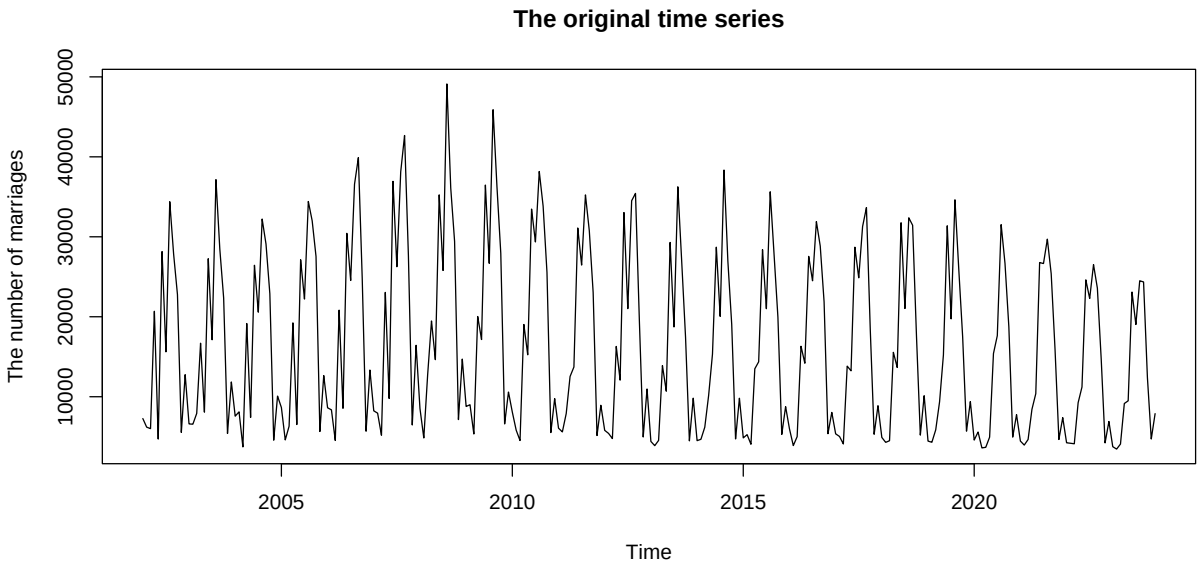
Aby sprawdzić ten i inne elementy procesu, posłużymy się analizą szeregu czasowego, opartego na naszych danych. W tym celu dopasujemy do niego odpowiedni model, a następnie przeprowadzimy prognozy, z których wyciągniemy wnioski.

Ogólnym celem będzie zatem znalezienie realnych zależności pomiędzy okresem w roku a częstotliwością zawierania związków małżeńskich oraz sprawdzenie, czy na podstawie danych uzyskanych do 2012 roku możliwe było przewidzenie zmian, które nastąpiły w kolejnych latach.

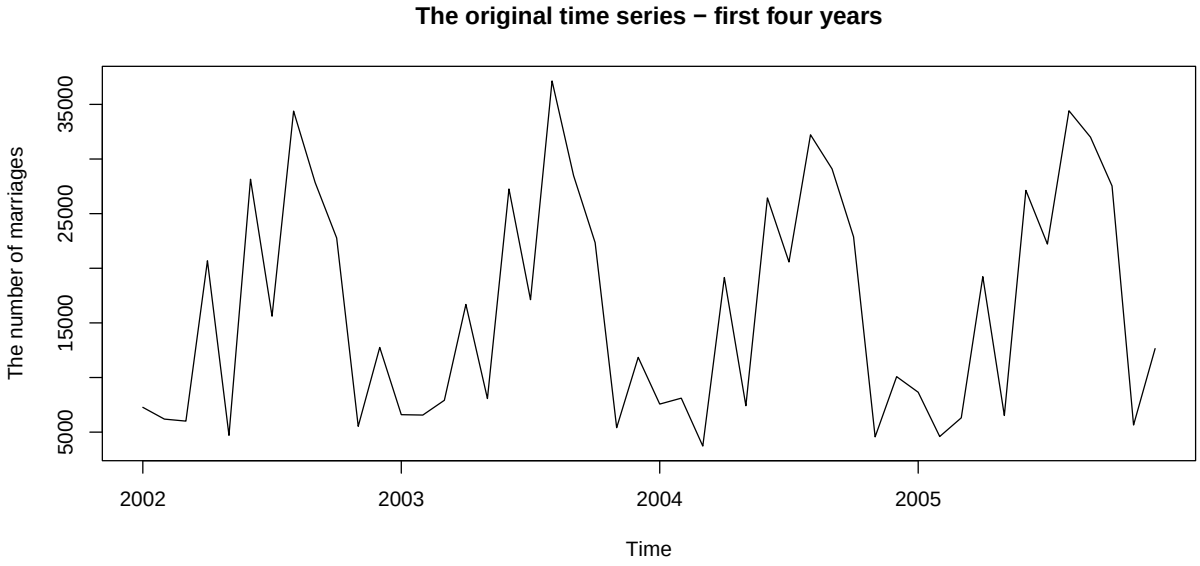
Do przeprowadzenia analizy i operowania na szeregu czasowym posłuży nam program RStudio i język R. Ze względu na istotną rolę, jaką odegrała w dziedzinie demografii pandemia wirusa Covid 19 i obostrzenia z nią związane, można się spodziewać, że dane zebrane po 2020 roku będą w pewien sposób zaburzone. Co będzie powodem, dlaczego większą część analizy przeprowadzimy tylko na części z nich - do roku 2012 - a po uzyskaniu na ich podstawie prognozy porównamy ją z rzeczywistymi wynikami. Dzięki temu dowiemy się jak istotną różnicę w liczbie zawieranych małżeństw wprowadziła obecność koronawirusa.

1.2 Wstępna wizualizacja

Zanim zaczniemy przeprowadzać transformacje na danych, obejrzymy ich oryginalną wersję, aby poczynić pierwsze obserwacje. Poniższy wykres obrazuje omawiany szereg w postaci pierwotnej, przed zastosowaniem na nim jakichkolwiek operacji.

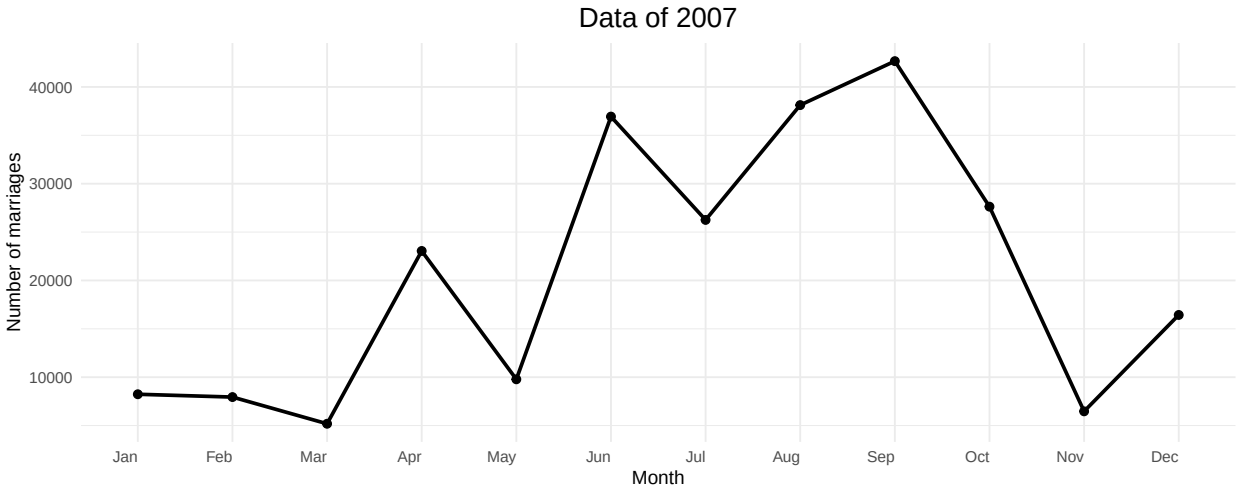


Zgodnie z przypuszczeniami możemy zaobserwować wyraźną sezonowość szeregu. Widać też pewną prawidłowość - co kilka lat następuje lekki wzrost, a następnie spadek liczby zawartych małżeństw, co sprawia, że cały szereg składa się z kilku podobnych okresów, obejmujących po parę lat. Najwięcej ślubów odbyło się pomiędzy 2005 a 2010 rokiem, z kolei po roku 2020 liczba małżeństw wydaje się znacząco spadać, co może być efektem panującej wówczas pandemii koronawirusa. Aby lepiej przyjrzeć się aspektowi sezonowości, ograniczymy teraz wykres do czterech pierwszych lat.



Sezonowość jest teraz dokładnie widoczna. Najwięcej ślubów odbywa się w miesiącach letnich, z kolei w zimie liczba ta drastycznie spada, co zgadza się z intuicją, ponieważ niskie temperatury i pojawiający się śnieg nie są sprzyjającymi dla tego typu wydarzeń czynnikami.

Obejrzyjmy jeszcze wizualizację dla pojedynczego roku, w tym wypadku 2007.



Mało małżeństw zostało zawartych przez pierwsze dwa miesiące, czyli w zimie. W grudniu ślubów było więcej, mimo podobnych warunków atmosferycznych, co może być powiązane ze Świętami Bożego Narodzenia, które bywają chętnie wybierane przez niektóre pary na datę sformalizowania związku. Innym świętem mającym wpływ na częstotliwość zawierania małżeństw jest Wielkanoc, a właściwie poprzedzający ją Wielki Post. Większość ślubów w Polsce to śluby kościelne, które w czasie postu się nie odbywają. Z tego względu po kilkudziesięciodniowym przestoju śluby rozpoczynają się tuż po Wielkanocy, czyli najczęściej w kwietniu. Co ciekawe, w roku 2007 sprawdziła się teoria o literze r - najchętniej wybieranymi miesiącami były czerwiec, sierpień, wrzesień i październik. Ten ostatni okazał się nawet nieco lepszy od lipca, który jest miesiącem wakacyjnym z dobrą pogodą, lecz nie ma w nazwie r. Wynika stąd, że przesądami może kierować się całkiem sporo par.

2 Transformacje i wybór modelu

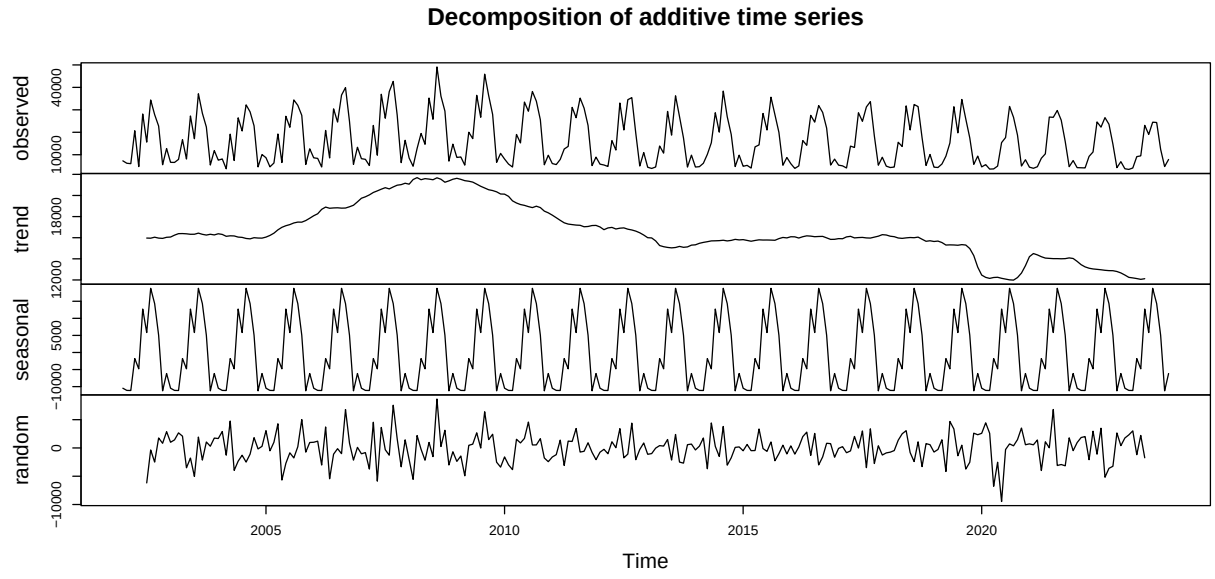
W celu szczegółowej analizy przeprowadzimy odpowiednie operacje na szeregu.

2.1 Dekompozycja

Standardowo zaczniemy od dekompozycji naszego szeregu. Służy ona do rozłożenia szeregu na podstawowe komponenty, co umożliwia dokładną analizę różnych jego aspektów, takich jak sezonowość czy trend. W ten sposób można lepiej zrozumieć strukturę danych i wyciągnąć bardziej precyzyjne wnioski. Możliwe są dwa podejścia do dekompozycji, w zależności od szeregu - metoda addytywna lub multiplikatywna. My wykorzystamy tę pierwszą.

Na początek przeprowadzimy dekompozycję addytywną na danych od 2002 do 2023 roku. Następnie dla obciętych danych od 2002 roku do 2012. Do przeprowadzenia dekompozycji użyjemy funkcji `decompose()`. Następnie, by lepiej zrozumieć nasz szereg za pomocą funkcji `Acf()` i `Pacf()`, narysujemy wykres autokorelacji i cząstkowej autokorelacji dla reszt z dekompozycji.

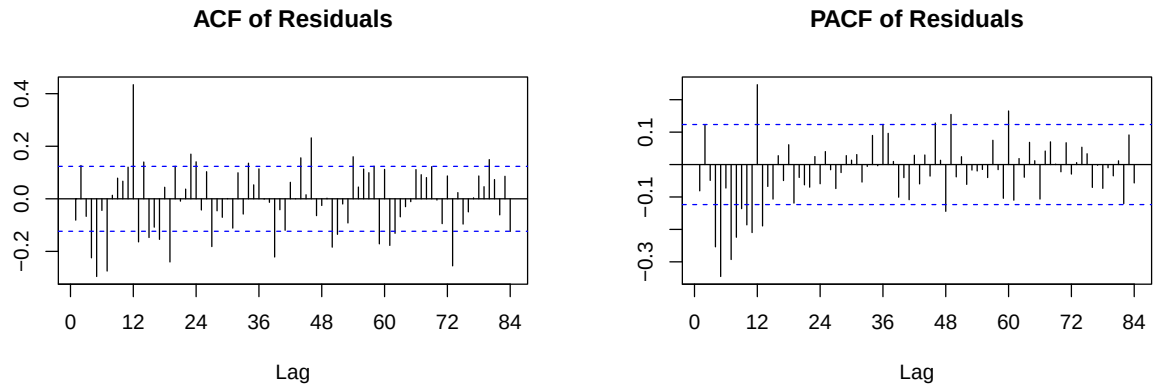
```
library(gridExtra)
decomposition <- decompose(series, type = "additive")
plot(decomposition)
```




```
trend <- decomposition$trend
seasonal <- decomposition$seasonal
random <- decomposition$random
```

Powyżej widzimy nasz oryginalny szereg, wyekstrahowany trend, sezonowość i reszty. Zauważmy, że sezonowość wydaje się mieć dużo istotniejsze znaczenie od trendu na postać szeregu. Trend kawałkami przypomina funkcje liniową, kawałkami wielomian. Ma górkę dla lat 2005-2013 (czego przyczyną mógł być np. wyż demograficzny w latach 80) i dołek dla lat 2020-2021 (co przypada na lata pandemii, kiedy były wprowadzone ostre restrykcje i lockdown), poza tym układa się raczej liniowo. Przy sezonowości widzimy omawiane wcześniej zależności, czyli mamy wyższe wartości dla miesięcy letnich oraz tych zawierających literę „r", wyjątek stanowi marzec. Na pierwszy rzut oka reszty wydają się losowe, oscylują wokół zera. Jednak, by naprawdę sprawdzić, czy są one losowe, przyjrzymy się wykresowi funkcji Acf i Pacf.

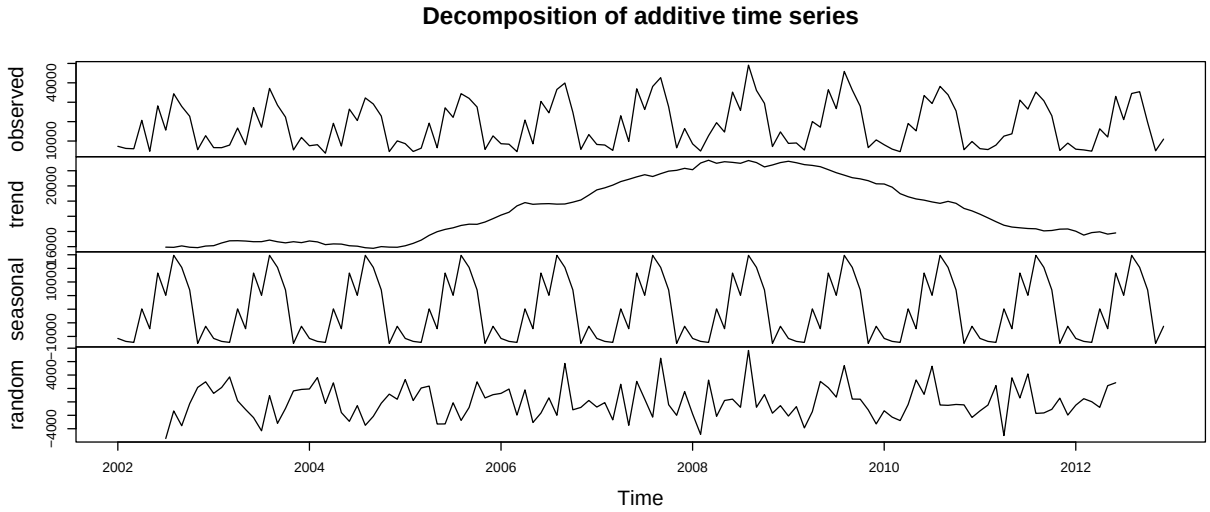
```
par(mfrow=c(1,2))
Acf(random, main="ACF of Residuals", lag.max=7*12, ylab="")
Pacf(random, main="PACF of Residuals",lag.max=7*12, ylab="")
```



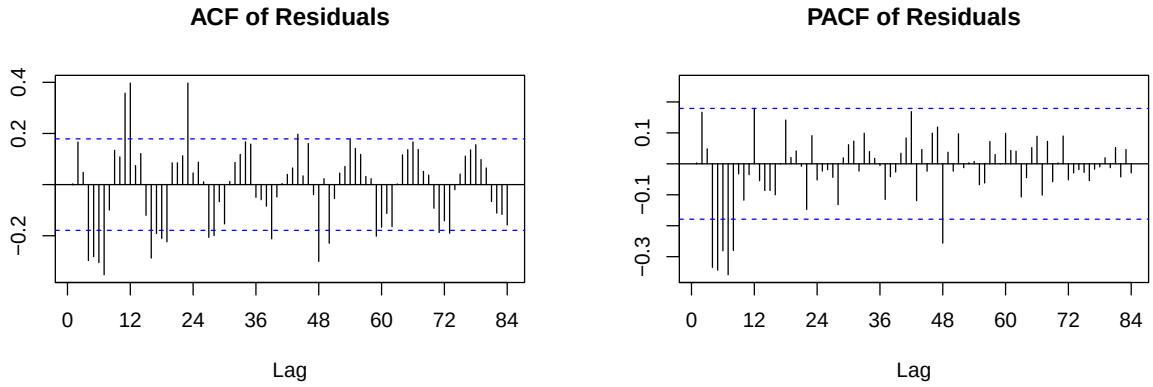
Jak widzimy znaczna część obserwacji nie mieści się w przedziale ufności. Stąd możemy wnioskować, że część obserwacji jest skorelowana. Na poparcie naszych spostrzeżeń wykonamy jeszcze test Boxa-Ljung.

```
library(tseries)
test_result <- Box.test(random, lag = 12, type = "Ljung-Box")
t_p_val <- test_result$p.value
```

P-wartość dla testu Boxa-Ljung wyniosła w przybliżeniu 0, co pozwala nam odrzucić hipotezę zerową, mówiącą, że obserwacje są nieskorelowane. Potwierdza to nasze przypuszczenie, iż obserwacje są skorelowane. Oznacza to, że podstawowa dekompozycja nie jest najlepszym podejściem w naszej sytuacji. Musimy zatem poszukać skuteczniejszej metody radzenia sobie z takim szeregiem. A co w przypadku, gdy dekompozycja nie działa z pewnych powodów na całym szeregu, ale jeśli zastosujemy ją do obciętego szeregu przyniesie odpowiednie rezultaty? Spróbujmy zatem wykonać ponownie dekompozycje tylko tym razem dla szeregu dla danych do 2012 roku.



Ponownie widzimy wyekstrahowany trend, sezonowość i reszty, które nie budzą zastrzeżeń na pierwszy rzut oka. Dlatego ponownie przyjrzyjmy się wykresom Acf i Pacf reszt oraz wynikowi testu Boxa-Ljung.

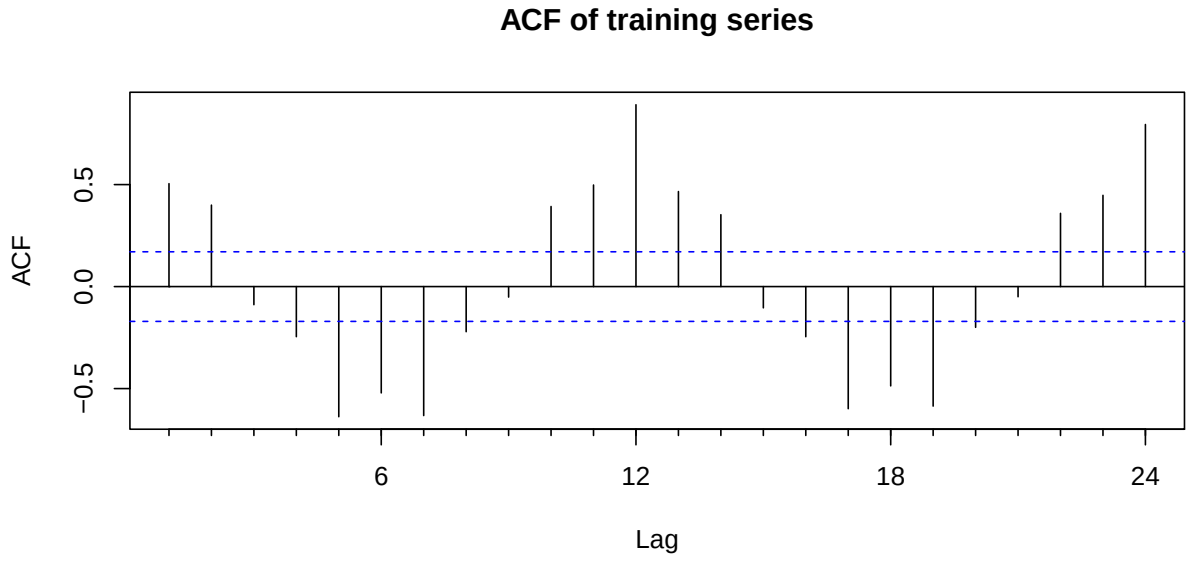


Niestety ponownie duża część obserwacji nie mieści się w przedziale ufności, co razem z wynikiem testu Boxa-Ljung (p-wartość wynosi $2.3314684 \times 10^{-15}$) wskazuje na skorelowanie obserwacji. Dodatkowo obserwujemy sinusoidalny kształt funkcji autokorelacji, który może wskazywać, że sezonowość nie została w pełni usunięta z reszt - co pokazuje zawodność i nieskuteczność tego podejścia. Zatem nie łudząc się już dalej w powodzenie tej metody spróbujemy innego podejścia.

Naszym punktem wyjściowym będzie sprawdzenie czy nasz szereg jest stacjonarny. Do tego posłuży nam test Dickey'a-Fullera. Sprawdzimy stacjonarność zarówno oryginalnego jak i obciętego szeregu.

```
adf_result <- adf.test(series)
p_value_adf <- adf_result$p.value
adf_result12 <- adf.test(series12)
p_value_adf12 <- adf_result12$p.value
```

W obu testach otrzymaliśmy małą p-wartość (dla oryginalnego wyniosła mniej niż 0.01, zaś dla obciętego również mniej niż 0.01), dzięki czemu odrzucamy hipotezę zerową na rzeczy hipotezy alternatywnej, co oznacza, że nasze szeregi są stacjonarne. Nasz obcięty szereg posłuży nam jako dane treningowe, zaś pozostałe dane będą służyły jako dane testowe, stąd od teraz będziemy pracowali na naszym obciętym szeregu. Na początek przyjrzyjmy się wykresowi Acf tego szeregu, by zobaczyć czy powinniśmy zastosować różnicowanie.



Na wykresie funkcji autokorelacji obserwujemy jej sinusoidalne zachowanie, co oznacza, że powinniśmy zastosować różnicowanie sezonowe, tak aby wyeliminować sezonowość.

W celu przeprowadzenia dalszej analizy, dopasujemy odpowiedni model do naszego szeregu. Sensownym wyborem wydaje się model sARIMA (seasonal Autoregressive Integrated Moving Average).

2.2 Opis modelu sARIMA

Sezonowy szereg jest typu $sARIMA(p,d,q)\times(P,D,Q)_s$ z okresem s i rzędami $d \geq 0$ oraz $D \geq 0$, gdy spełnia równanie

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

gdzie $W_t \sim WN(0,\sigma^2)$ jest białym szumem. Model sARIMA jest rodzajem procesu autoregresji ze średnią ruchomą, a dokładniej rozszerzeniem modelu ARIMA o komponenty sezonowe. Wykorzystywane są następujące parametry, będące liczbami naturalnymi:

- p - rząd procesu autoregresji,
- q - rząd średniej ruchomej,
- d - liczba operacji różnicowania,
- P - rząd sezonowej autoregresji,
- Q - rząd sezonowej średniej ruchomej,
- D - rząd sezonowego różnicowania,
- s - długość sezonu.

Rzędy średniej ruchomej i autoregresyjny dotyczą zależności zachodzących między obecną a wcześniejszymi wartościami szeregu czasowego. Z kolei parametr d określa ile różnicowań jest potrzebnych, aby uzyskać stacjonarność szeregu. Skuteczność doboru wartości parametrów ocenia się zazwyczaj na podstawie kryterium AIC.

W przypadku analizowanych przez nas danych model sARIMA może być użyteczny, ponieważ dobrze odnajduje się on w sytuacjach, gdy szereg charakteryzuje się powtarzalnymi wzorcami.

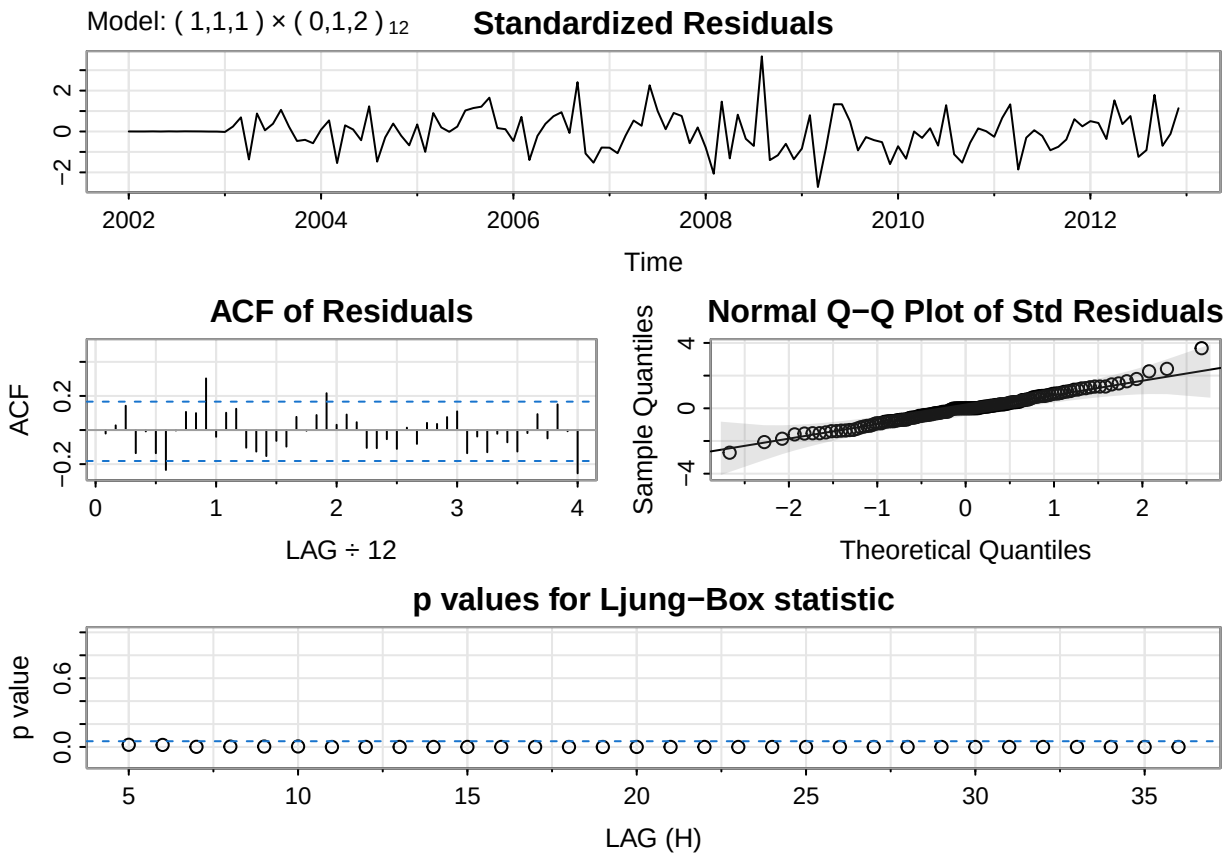
2.3 Automatyczny dobór modelu sARIMA

Aby zamodelować dane o zawieranych małżeństwach w latach 2002-2012 modelem $sARIMA$ potrzebujemy znać parametry: p,d,q,P,D,Q. Aby je wyznaczyć możemy użyć funkcji `auto.arima`.

```
auto.arima(series12)
```

Automatycznie dobrany model to $sARIMA(1,1,1)\times(0,1,2)_{12}$. Sprawdźmy teraz jakość tego dopasowania.

```
library(forecast)
library(astsa)
library(tseries)
fit_series12_auto <- sarima(series12, 1,1,1, 0,1,2, 12, details = T)
```




```
shtest=shapiro.test(resid(fit_series12_auto$fit))
```

Co prawda residua zachowują się jak realizacje zmiennych losowych z rozkładu normalnego - ich kwantyle empiryczne układają się wzdłuż linii teoretycznej na wykresie Q-Q plot, a p-wartość otrzymana w teście Shapiro-Wilka wynosi 0.106, co nie prowadzi do odrzucenia hipotezy zerowej. Test Dickey'a–Fullera potwierdza stacjonarność szeregu residuów. Niestety z testu Ljung-Boxa wynika, że zmienne są ze sobą silnie skorelowane - o czym świadczą zerowe p-wartości umieszczone na wykresie. Residua powinny reprezentować nieskorelowany szum biały, a ich zależność niestety prowadzi do odrzucenia zaproponowanego modelu.

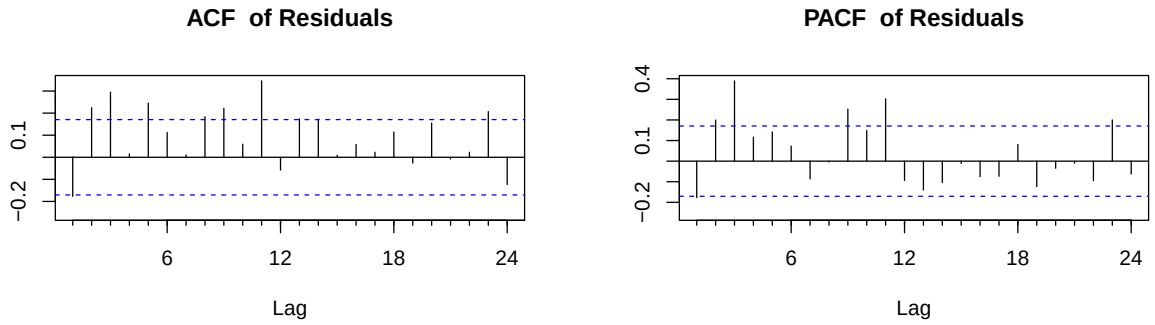
2.4 Dobór parametrów dla modelu sARIMA na podstawie funkcji ACF i PACF

Spróbujmy teraz dobrać model $sARIMA(p,d,q) \times (P,D,Q)_s$ do naszych danych na podstawie analizy funkcji Acf (funkcja autokorelacji) i Pacf (funkcja częściowej autokorelacji). Pierwsza z wymienionych mierzy korelację między bieżącą wartością szeregu czasowego a jego kolejnymi przesunięciami (opóźnieniami). Druga natomiast różni się tym, że usuwa wpływ pośrednich opóźnień, biorąc pod uwagę jedynie zależność między obecną wartością szeregu a konkretnym przesunięciem.

Dobierzemy teraz model $sARIMA(p,d,q) \times (P,D,Q)_s$ do naszych danych. Wykres autokowariancji sugerował znaczącą sezonowość ze względu na swój sinusoidalny kształt. Zróżnicujmy zatem szereg (sezonowo) przyjmując $D = 1$. Oczywiście wiemy, że sezonowość roczna wynosi $s = 12$.

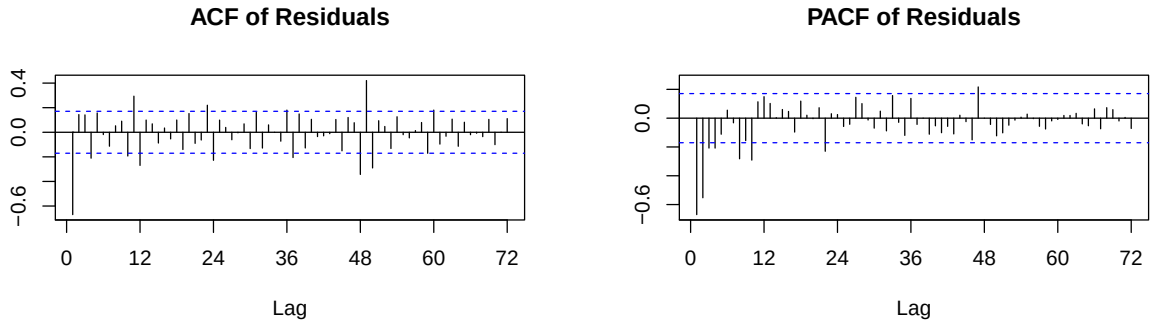
```
library(forecast)
library(astsa)
library(tseries)
```

```
fit_series12_1 <- sarima(series12, 0,0,0, 0,1,0, 12, details=F)
adf.test(resid(fit_series12_1$fit))
```



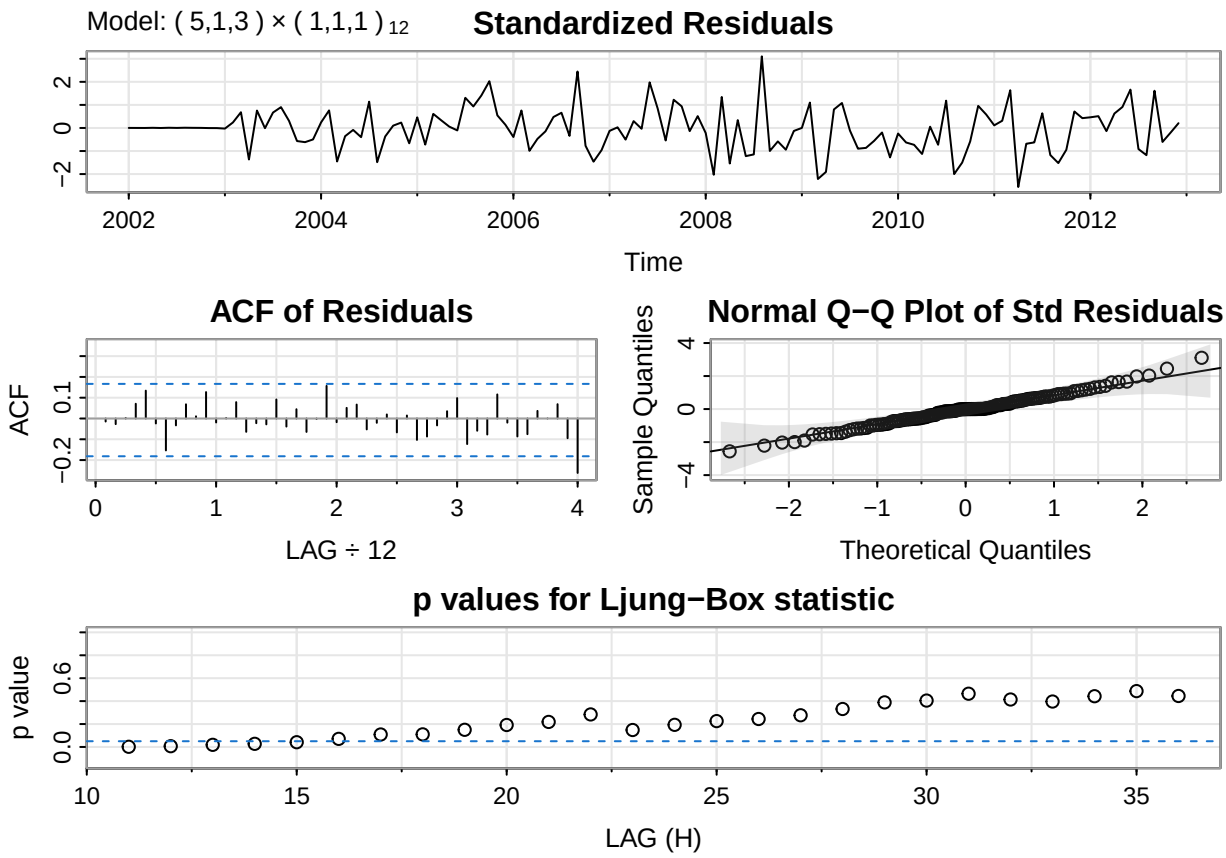
Po usunięciu sezonowości wykres autokowariancji residuów nadal sugeruje dużą zależność między kolejnymi obserwacjami, a z testu Dickey'a–Fullera wynika, że szereg ten nie jest stacjonarny. Aby wyeliminować zależność między obserwacjami dodajmy różnicowanie, przyjmując $d = 1$. Zabieg ten wiąże się również z usunięciem trendu.

```
fit_series12_1 <- sarima((series12), 0,1,0, 0,1,0, 12,details=F)
adf.test(resid(fit_series12_1$fit))
```



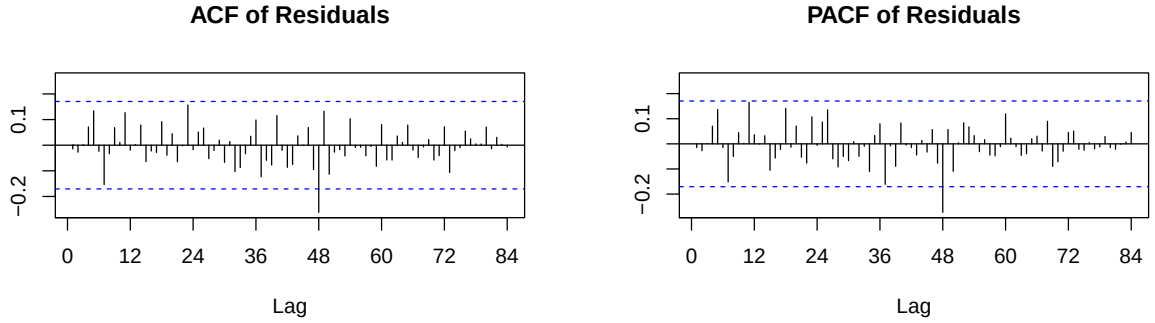
Na obu wykresach dostrzegalne są sezonowe piki występujące co 12 *lag*-ów, zatem niech $P = Q = 1$. Dobierzmy teraz parametry p oraz q. Patrząc na wartości występujące przed pierwszym opóźnieniem sezonowym możemy zaobserwować, że 5 spośród nich wystaje ponad zaznaczony przedział ufności na wykresie Pacf (szósty pik w cyklu długości 12 to właśnie powtarzające się opóźnienie sezonowe), a 3 na wykresie funkcji Acf. Z tego powodu przyjmiemy $q = 3$ oraz $p = 5$. Ostatecznie otrzymujemy model $sARIMA(5,1,3) \times (1,1,1)_{12}$.

```
fit_series12_1 <- sarima((series12), 5,1,3, 1,1,1, 12)
```




```
stest=shapiro.test(residuals(fit_series12_1$fit))
```

Wykresy Acf oraz p-wartości dla testu Ljung-Boxa (związanego z dopasowaniem modelu do danych) wyglądają zadowalająco. Wartości statystyk AIC, AICc, BIC wynoszą kolejno: 18.55, 18.56, 18.8. Test Shapiro-Wilka, osądzający o rozkładzie normalnym residuów, zwraca p-wartość 0.61, co prowadzi do przyjęcia hipotezy zerowej. Podobne wnioski możemy wyciągnąć z wykresu Q-Q plot dla residuów. Poniżej zaprezentowano wykresy Acf i Pacf dla $lag \leq 7 \cdot 12$. Większość obserwacji mieści się w zaznaczonym przedziale ufności. Nie ma powodów do dyskwalifikacji tego modelu.

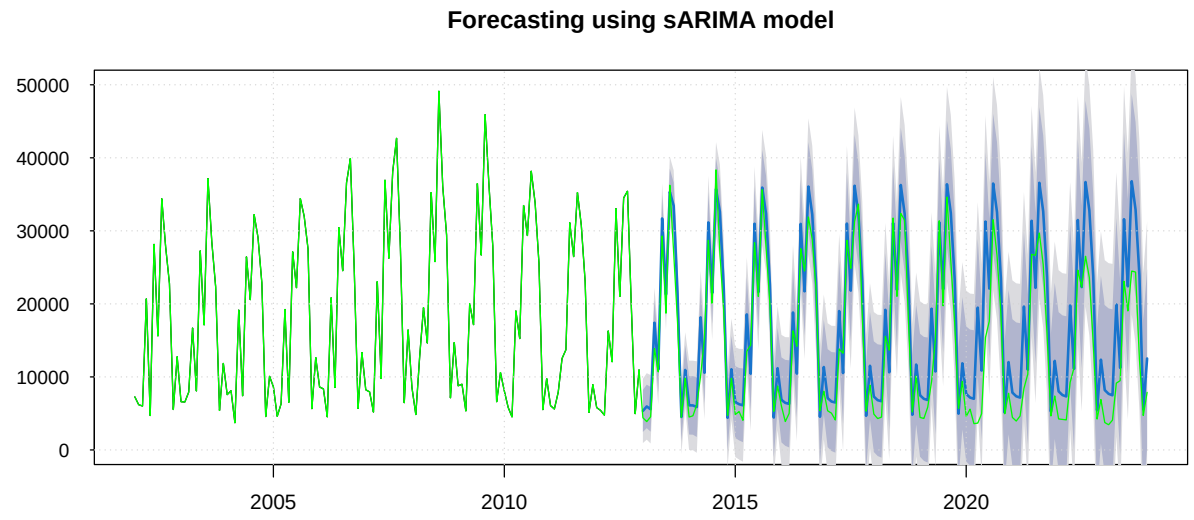


3 Prognozowanie

3.1 Prognozowanie modelem sARIMA$(5,1,3) \times (1,1,1)_{12}$

Do wygenerowania prognozy wykorzystajmy funkcję `forecast`. Przewidywane liczby zawieranych małżeństw w latach 2013-2023 zaznaczono niebieskim kolorem, zielonym rzeczywiste wartości, a szarym przedziały ufności prognozy.

```
forecast_1<-forecast(series12,model=fit_series12_1$fit, h=132)
```



Wartości rzeczywiste mieszczą się w zaznaczonym polu niepewności. Przedziały ufności powyższej prognozy są dość szerokie, a ich szerokość rośnie, przez co w dłuższym horyzoncie czasowym prognoza może się okazać nieskuteczna. Model dość dobrze prognozuje wartości występujące przed rokiem 2020. W późniejszych okresach różnica jest coraz bardziej zauważalna - zwłaszcza w miesiącach letnich.

3.2 Prognozowanie modelem ETS

W celu osiągnięcia szerzej zakrojonego prognozowania wykorzystamy teraz inny model, co umożliwi nam porównanie rezultatów. *ETS*, czyli Exponential Smoothing State Space Model, to metoda prognozowania, która opiera się na wygładzaniu wykładniczym. Model ma szerokie zastosowania ze względu na możliwość licznych modyfikacji parametrów, dzięki czemu precyzyjnie dopasowuje się do badanego szeregu. Istotnymi składnikami są obecny stan szeregu, sposób modelowania trendu oraz sezonowości. Każdy z nich może być zamodelowany przy założeniu addytywności lub multiplikatywności, co prowadzi do wielu możliwych konfiguracji. Sprawia to, że *ETS* jest metodą wyjątkowo elastyczną.

Dodatkową zaletą tego modelu jest jego wygodne użycie w środowisku *R*, w którym pracujemy. Algorytm własnoręcznie dobiera najodpowiedniejsze parametry na podstawie naszych danych, co usprawnia pracę i skutkuje dokładniejszymi wynikami.

Sprawdźmy zatem, jak metoda *ETS* poradzi sobie z naszymi danymi.

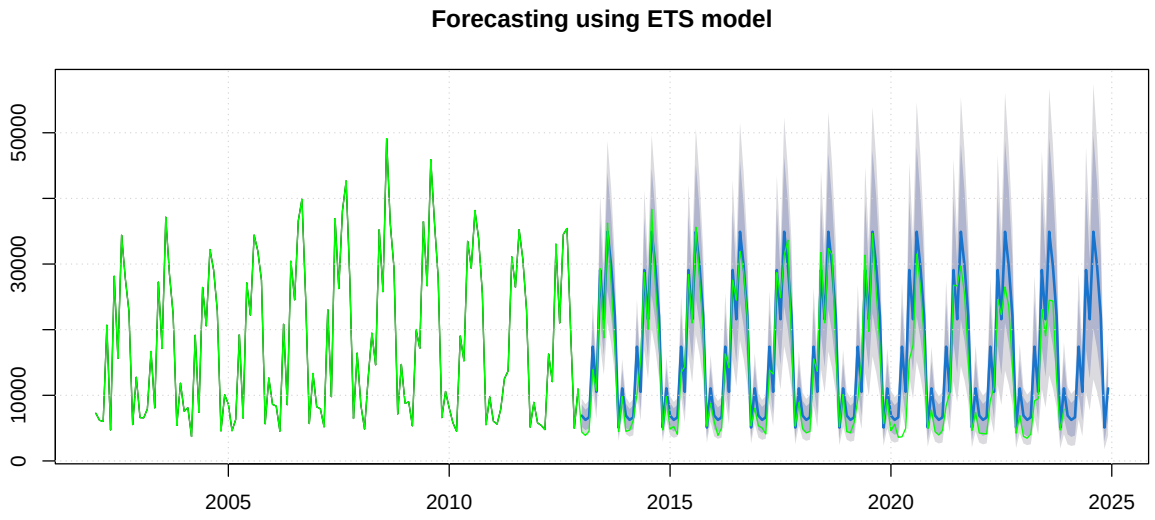
```
ets_model <- ets(series12)
print(ets_model)
```

Widzimy, że funkcja dopasowała do naszego szeregu model $ETS(M, N, M)$. Oznacza to multiplikatywny błąd i sezonowość (wpływ zakłóceń i sezonowości jest proporcjonalny do poziomu szeregu czasowego) oraz brak długoterminowego wzrostu lub spadku wartości (co zgadza się z wcześniejszymi obserwacjami). Algorytm dobrał parametry wygładzania wielkości $alpha = 0.1129$ oraz $gamma = 1e-04$. Pierwszy z nich odpowiada za poziom (obecny stan) szeregu i kontroluje jak szybko model reaguje na zmiany w jego obrębie. Ponieważ w tym przypadku wartość *alpha* jest stosunkowo bliska 0, to model reaguje wolno i mocno wygładza zmiany, co jest odpowiednie dla danych o niskiej zmienności. Parametr *gamma* z kolei odpowiada za sezonowość i sprawdza reakcję modelu na zmiany, które w niej zachodzą. Jest bardzo mały, zatem oznacza to, że sezonowość zmienia się powoli, co zgadza się z naszymi obserwacjami.

Początkowy poziom szeregu wynosi $l = 17325.8899$, natomiast wymienione wielkości s to wartości początkowe sezonowości dla dwunastu okresów (miesięcy). Multiplikatywna sezonowość oznacza, że wartości szeregu są mnożone przez te współczynniki w zależności od miesiąca.

Parametr $sigma = 0.1931$ jest odchyleniem standardowym reszt modelu, a więc miarą nieprzewidywalności danych. Mała wartość sugeruje, że model dobrze dopasowuje dane, ale nie są one do końca przewidywalne. Funkcja wymienia również wyliczone wartości AIC, AICc oraz BIC, wynoszące kolejno $2758.175, 2762.313$ i 2801.417.

Skoro omówiliśmy już działanie metody ETS na naszym szeregu czasowym, wykorzystajmy ją teraz do uzyskania prognozy.



Podobnie jak w poprzednim podrozdziale, przewidywane liczby ślubów w latach 2013-2023 zostały zaznaczone na niebiesko, rzeczywiste wartości na zielono, a szare obszary to przedziały ufności.

Prognoza wygląda dość podobnie do tej uzyskanej za pomocą modelu $sARIMA$. Widać jednak, że przedziały ufności są węższe i bardziej dopasowane do wartości. Z biegiem lat nie zwiększają się również tak bardzo jak przedziały w poprzedniej prognozie, czyli są nieco bardziej precyzyjne. Wszystkie wartości szeregu mieszczą się odpowiednio w wyznaczonych obszarach. Do roku 2020 wartości rzeczywiste i prognozowane są do siebie zbliżone, a więc prognoza jest dokładna. W późniejszych latach pokrywają się coraz słabiej, ponieważ liczba zawieranych małżeństw maleje, co jest skutkiem panującej wówczas pandemii. Widać więc duże różnice między przewidywanymi wartościami, a tymi które rzeczywiście nastąpiły.

3.3 Prognozowanie modelem liniowym

Dla lepszego porównania naszych wyników użyjemy również modeli liniowych do analizy i prognozy naszych danych. Dopasujemy dwa modele liniowe do naszych danych zależne od trendu (który raz będzie wielomianem 1 stopnia, a raz wielomianem 2 stopnia) i sezonowości.

```
library(kableExtra)
model_trend_season <- tslm(series12 ~ trend + season)
m_sum1 <- summary(model_trend_season)
```

Tablica 1: Tabela p-wartości dla zmiennych w modelu

Intercept	Trend	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12
0	0.0053	0.4304	0.3014	0	0.0323	0	0	0	0	0	0.1187	0.0034

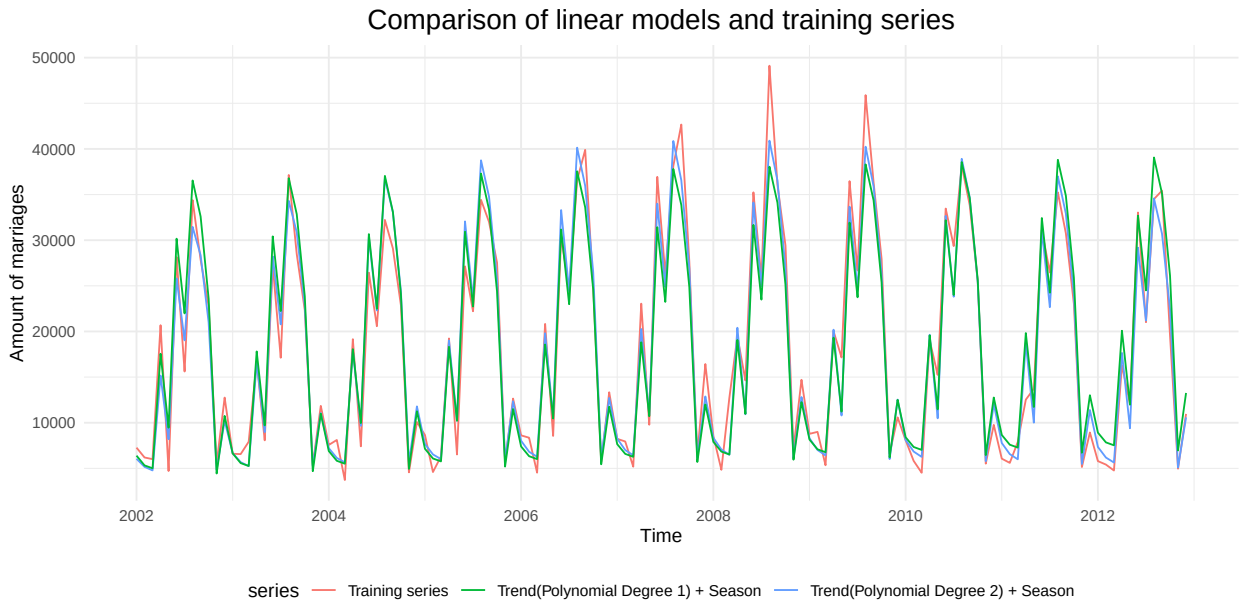
Z summary() dla pierwszego modelu (którego trend jest wielomianem 1 stopnia) widzimy, że p-wartości dla trendu i większości zmiennych sezonowych są bardzo małe, co wskazuje na istotność tych zmiennych. Nieistotnymi okazały się tylko zmienne `S2`, `S3` i `S11`, które odpowiadałyby wpływowi lutego, marca i listopada na model. Statystyka $R^2 = 0.931$, co wskazuje na dobre dopasowanie modelu.

```
model_poly2 <- tslm(series12 ~ poly(trend, 2) + season, lambda = 0)
m_sum2 <- summary(model_poly2)
```

Tablica 2: Tabela p-wartości dla zmiennych w modelu

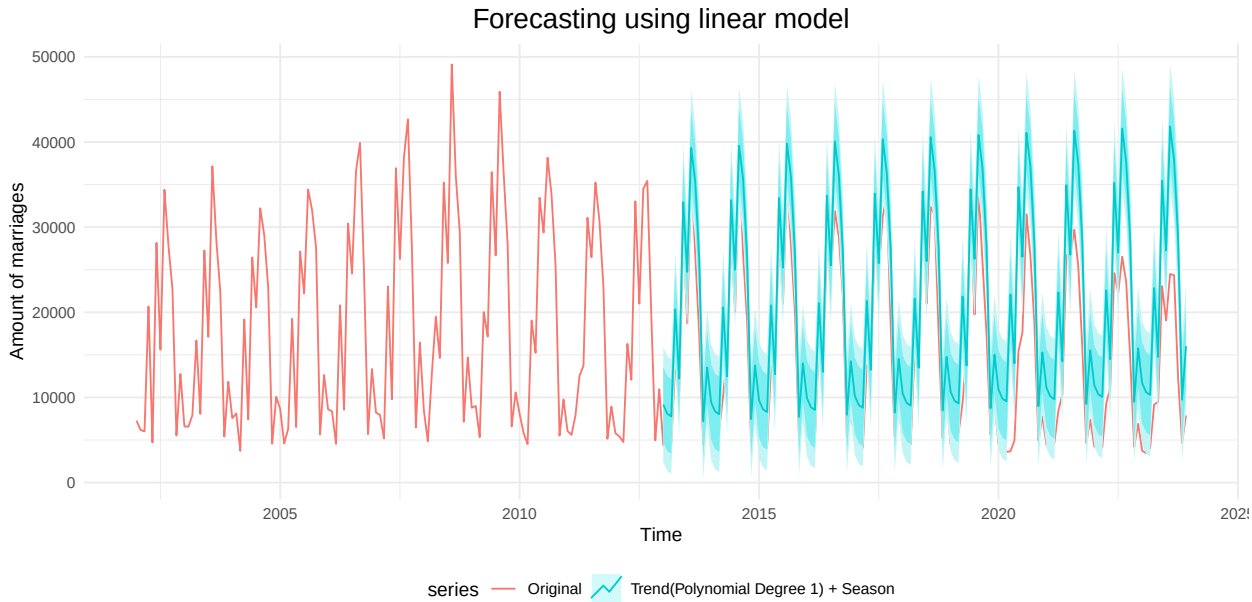
Intercept	T1	T2	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12
0	0.0329	0	0.0477	0.0029	0	0.0015	0	0	0	0	0	4e-04	0

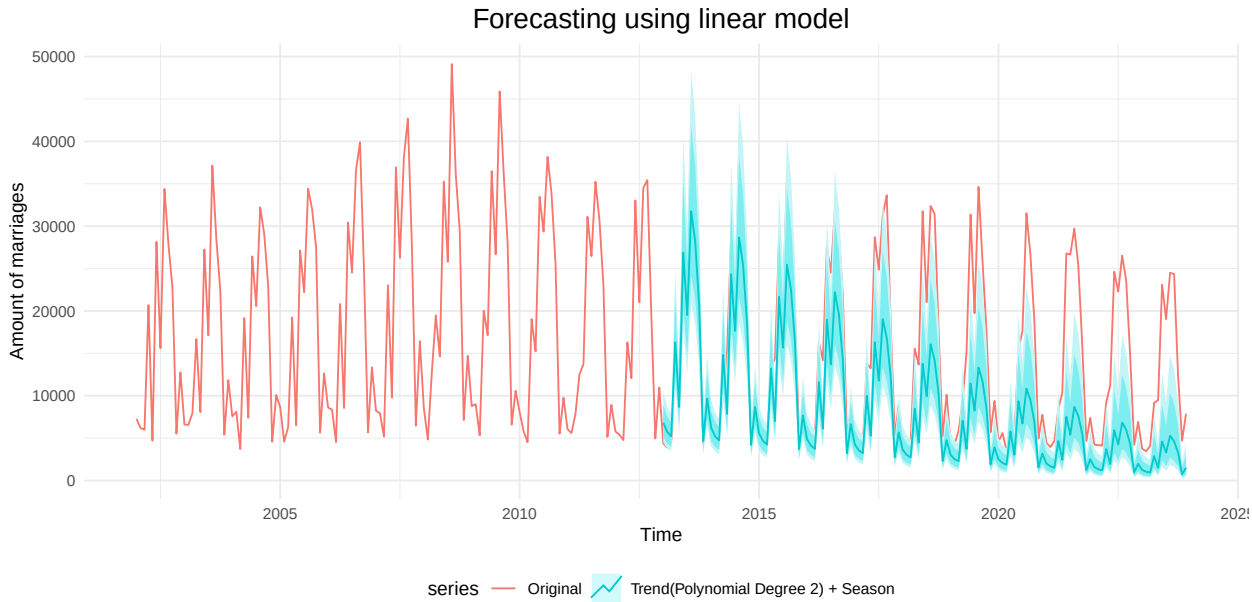
Z kolei dla drugiego modelu widzimy, że wszystkie zmienne są istotne i również $R^2 = 0.934$, co świadczy o dobrym dopasowaniu modelu. Na podstawie obu modeli ponownie możemy dostrzec istotność komponenty trendu i komponenty sezonowości. Zobaczmy jak na wykresie prezentują się dopasowane modele w porównaniu z danymi.



Teraz użyjemy naszych modeli do prognozowania przyszłych wartości z użyciem funkcji `forecast()`, a wyniki przedstawimy poniżej na wykresach.

```
forecasted_values <- forecast(model_trend_season, h = 11*12)
forecasted_values2 <- forecast(model_poly2, h = 11*12)
```





Z powyższych wykresów widzimy, że na początku modele radzą sobie nawet dobrze z prognozowaniem, lecz wraz z prognozą dalszych lat jakość prognozy się pogarsza. I tak dla modelu z trendem stopnia 1 prognozuje on zbyt duże wartości w stosunku do rzeczywistości, zaś model z trendem stopnia 2 prognozuje zbyt małe wartości.

4 Podsumowanie

Porównajmy teraz parametry dla wszystkich wykonanych prognoz. Błąd średniokwadratowy prognozy policzymy dla danych z lat 2013-2018 oraz 2013-2023. Efekty przedstawione są w poniższej tabeli.

Tablica 3: Porównanie różnych kryteriów doboru modelu.

Model	AIC	BIC	MSE do 2018r	MSE do 2023r
sARIMA	18.547	18.564	8731727	23081262
liniowy, trend st. 1	2521.689	2562.048	28313078	56681889
liniowy, trend st. 2	-42.094	1.148	44479891	82072264
ETS	2758.175	2801.417	6251836	14888250

Najlepsze wartości mierników AIC oraz BIC uzyskano dla modelu liniowego z trendem drugiego stopnia. Niewiele gorszy okazał się model $sARIMA(5,1,3) \times (1,1,1)_{12}$. Najmniej optymalną wartość uzyskał model ETS, co jest spowodowane jego dużą złożonością.

Błąd dopasowania prognozy do rzeczywistych danych obrazuje wartość MSE, czyli błąd średniokwadratowy. Pod tym względem najlepszy okazał się model ETS, a zaraz za nim - $sARIMA$.

Przy dobieraniu modelu do prognozowania powinno się brać pod uwagę różne kryteria, zatem model $sARIMA$ wydaje się najoptymalniejszym rozwiązaniem.

Wraz z upływem czasu prognozy dla każdego z rozważanych modeli zaczynają znacznie odbiegać od rzeczywistych wartości, co sprawia, że w dłuższym horyzoncie czasowym są one coraz mniej skuteczne. Nie jesteśmy zatem w stanie przewidywać dalekiej przyszłości, co najwyżej kilka lat do przodu. Ponadto wraz ze zdobywaniem nowych danych modele powinny być aktualizowane i ponownie dopasowywane, co pozwoli skuteczniej przewidywać kolejne okresy czasowe.