# Szeregi czasowe - Projekt

Analiza konsumpcji i wydatków na komunikację w UK

Patryk Gronkiewicz 164157

**Streszczenie**

Analiza danych dotyczących komunikacji i ogólnych wydatków gospodarstw domowych w UK pozwala na przewidzenie zachowań rynku oraz zaplanowania wydatków zarówno dla osób prywatnych, jak i firm pocztowych i telekomunikacyjnych.

## Spis treści



## 1 Użyte dane

W projekcie użyto danych na temat:

1. Całkowitego kosztu konsumpcji w gospodarstwach domowych UK jako szereg z sezonowością (zawiera także trend).
2. Wydatków na komunikację w UK jako szereg z trendem.

Dane te pochodzą z ONS (odpowiednik GUS-u). W analizie zostanie pominięty okres od 2019Q4 jako anomalia ze względu na pandemię.

Już na oficjalnej stronie można zauważyć, że dane te w obu przypadkach zawierają wyraźny trend wzrostowy, natomiast jedynie całkowity koszt konsumpcji ma wyraźną sezonowość z peakiem w czwartym kwartale każdego roku.

Szereg zawierający dane nt. komunikacji odnosi się do wydatków na usługi pocztowe oraz telefon i fax (z uwzględnieniem sprzętu, jak i usług).

W danych dotyczących wydatków Brytyjczyków uwzględnione zostały wydatki w gospodarstwach zarówno rezydentów i nierezydentów (osób posiadających brytyjski paszport lub nie - jest to koncept inny od obywateli państwa)

W obu przypadkach dane opublikowane zostały 31.03.2021 roku z danymi za 2020Q4, więc można zauważyć, że dostępne są z kwartalnym opóźnieniem.

Analiza tych szeregów pozwala na lepsze planowanie wydatków, nawet na poziomie pojedynczego gospodarstwa ze względu na możliwość uwzględnienia wzrostu cen czy inflacji stylu życia. Analiza wydatków na komunikację jest także dobrym wskaźnikiem do przekazania jak bardzo "zdalne" społeczeństwo jest.

W społeczeństwie, w którym małe grupy ludzi dzielą znaczne odległości wydatki na takie usługi będą wyższe ze względu na częstość wykorzystania takich możliwości.

Do ich obróbki zostały użyte biblioteki zaimportowane poniżej

```
library(forecast)
```

# 2 Analiza

## 2.1 Główne cechy danych

Na początku dane zostały załadowane z plików CSV. W nie interesują nas niektóre z linii widocznych w pliku (linie 1-44 ze względu na metadane i dane roczne, a nie kwartalne).

```
wydatki <- ts(read.csv('wydatki.csv',
                       skip=43,
                       col.names = c("q", "v"))$v,
              start = c(1985, 01),
              frequency = 4)
komunikacja <- ts(read.csv('komunikacja.csv',
                       skip=43,
                       col.names = c("q", "v"))$v,
              start = c(1985, 01),
              frequency = 4)
wydatki <- window(wydatki, start = start(wydatki), end = c(2019, 04))
komunikacja <- window(komunikacja, start = start(komunikacja), end = c(2019, 04))
```

Na początku zostały przedstawione dane na kilku wykresach.

```
par(mfrow=c(2,1), mar=c(2,4,2,2))
plot(wydatki)
plot(komunikacja)
```

Na wykresie 1 bardzo wyraźnie widać sezonowość w postaci "ząbków" dla wydatków ogólnych, czego na pierwszy rzut oka nie można stwierdzić o wydatkach na komunikację. Oba szeregi zawierają wyraźny trend.

```
par(mfrow = c(1,2), mar=c(5,3,4,1))
monthplot(wydatki, ylab = NA, xlab = "wydatki")
monthplot(komunikacja, ylab = NA, xlab = "komunikacja")
```

Na wykresie 2 widać wyraźnie trendy wzrostowe między odpowiadającymi kwartałami, więc zależność została zachowana w przypadku obu szeregów. Jak łatwo zauważyć dla wydatków nie występuje "ząbkowanie" na poszczególnych wykresach, dlatego można wnioskować, że ich sezonowość to pewna wielokrotność 4. Może to wynikać z wyższych kosztów w kwartale 4 ze względu na ogrzewanie i droższą żywność ze względu na zwiększony import w miesiącach jesienno-zimowych. Załamanie w wydatkach wynika z kryzysu w 2007-2009 roku spowodowanym załamaniem rynku kredytów hipotecznych wysokiego ryzyka.

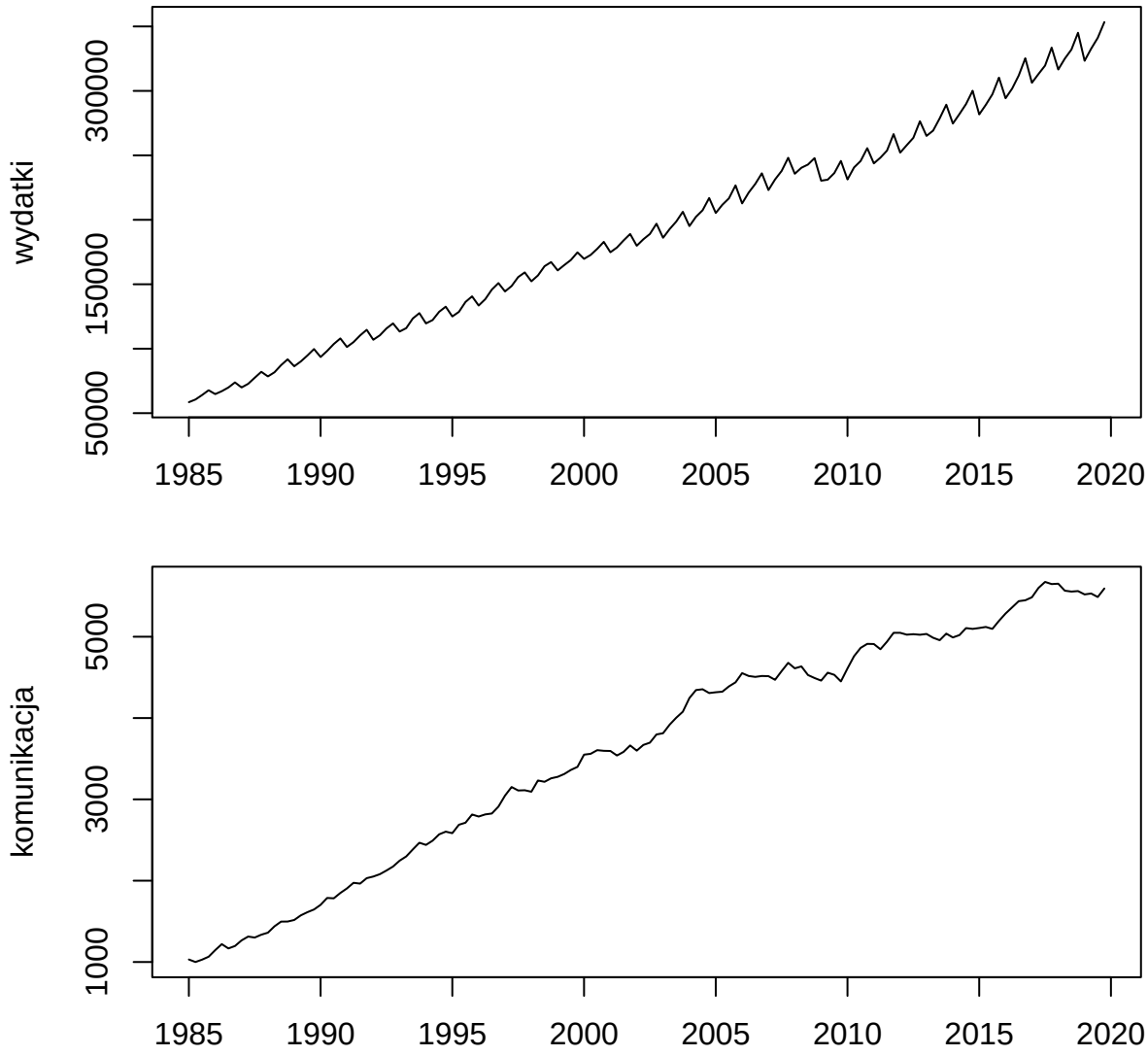
```
par(mfrow = c(1,2))
boxplot(wydatki, xlab="wydatki")
boxplot(komunikacja, xlab="komunikacja")
```

Wykresy na rysunku 3 nie zostały pokazane na jednej osi ze względu na bardzo rozbieżne wartości między szeregami, przez co dane nt. komunikacji nie były czytelne. Jak można zauważyć dużo dłuższe linie błędu są w górę w przypadku wydatków i w dół dla komunikacji.

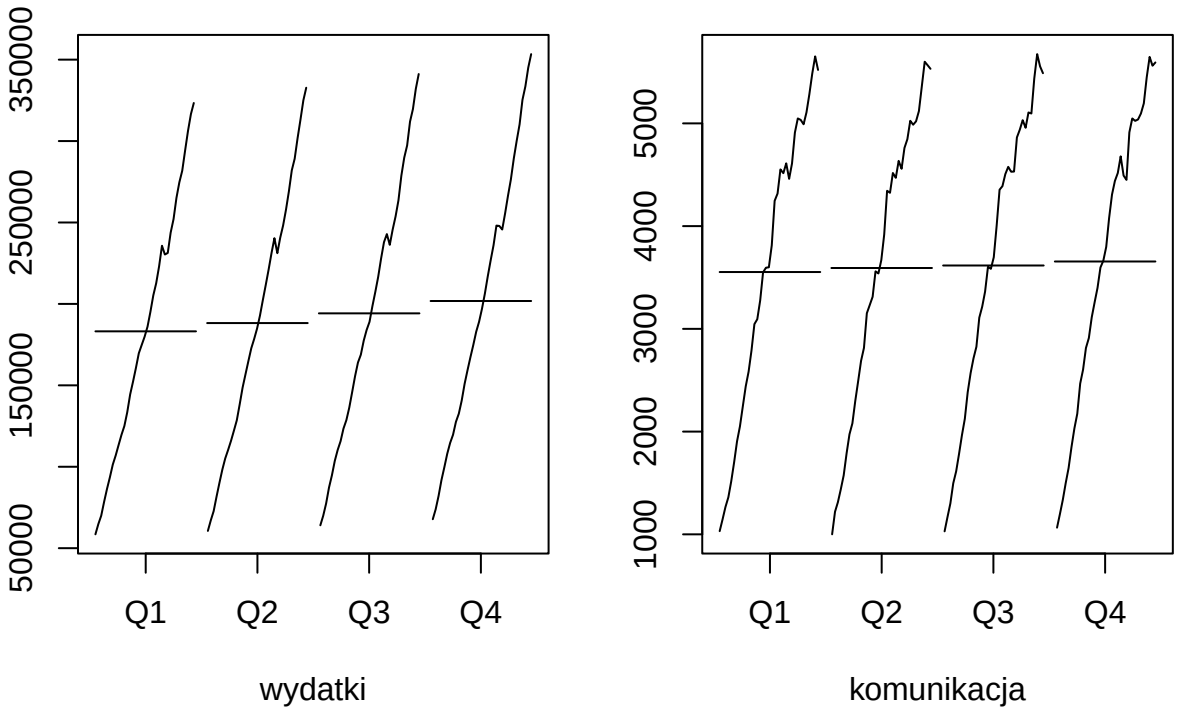
```
lag.plot(wydatki, lags = 4, labels = F)
```

```
lag.plot(komunikacja, lags = 4, labels = F)
```

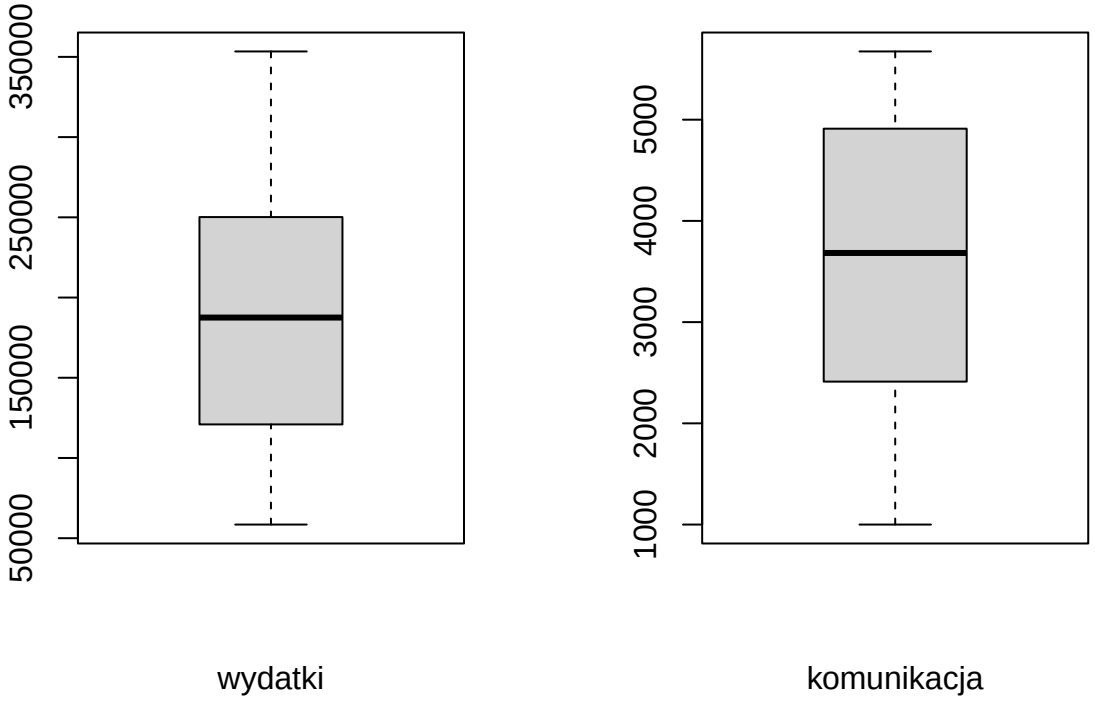
Jak widać na wykresach z rysunku 4 najbardziej skoncentrowane wartości są dla `lag=4`, czego można się



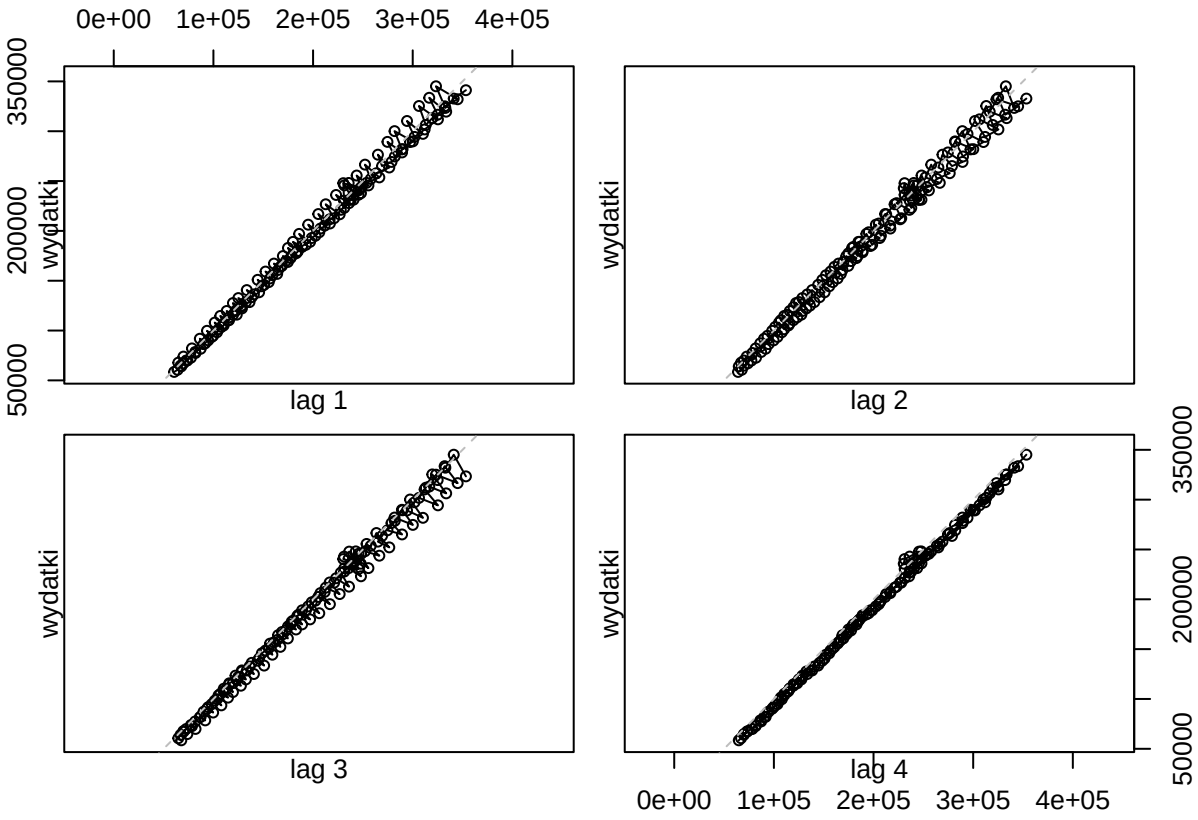

Rysunek 1: Wydatki w czasie przed jakąkolwiek obróbką



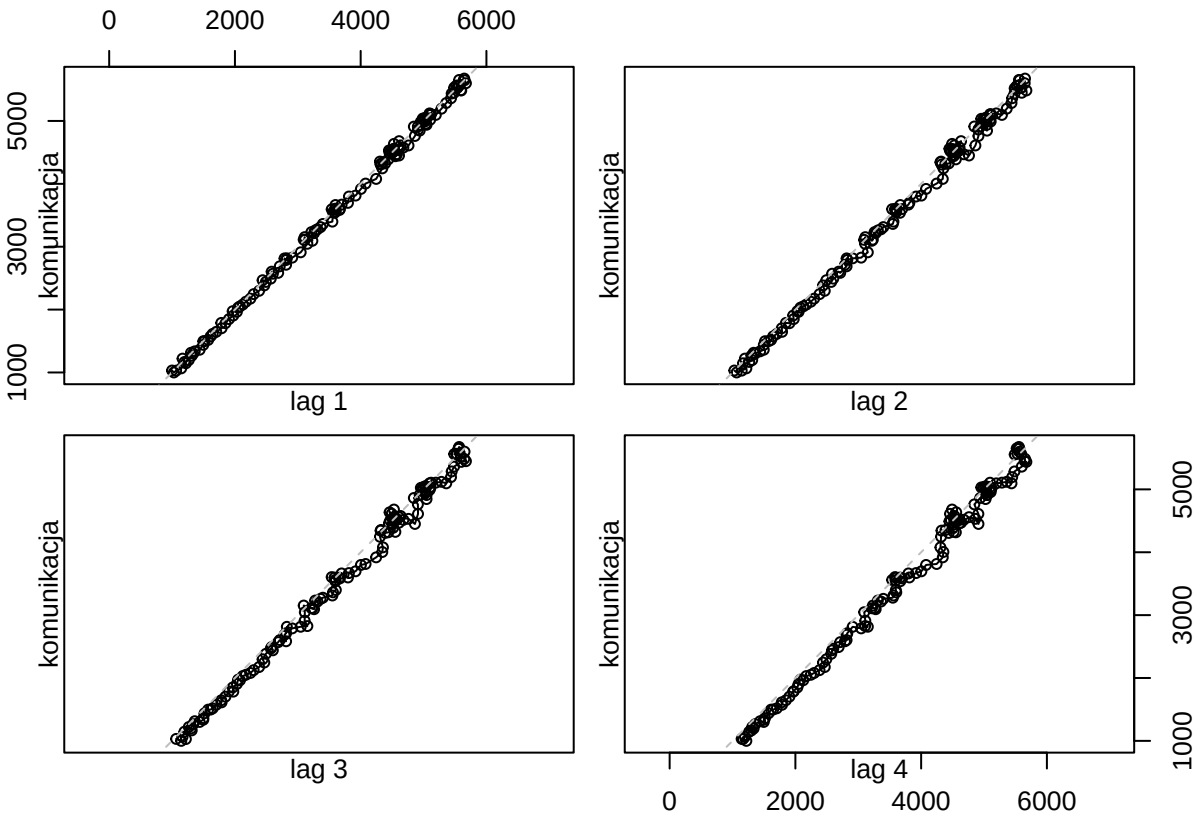

Rysunek 2: Wykresy monthplot jak widać mogą dotyczyć nie tylko miesięcy, ale innych okresów w roku, takich jak kwartały. Dla czytelności wykresy zostały podpisane na osi poziomej, jednak kwota wydatków jest na osi pionowej




Rysunek 3: Wydatki w gospodarstwach domowych ogółem oraz na komunikację.



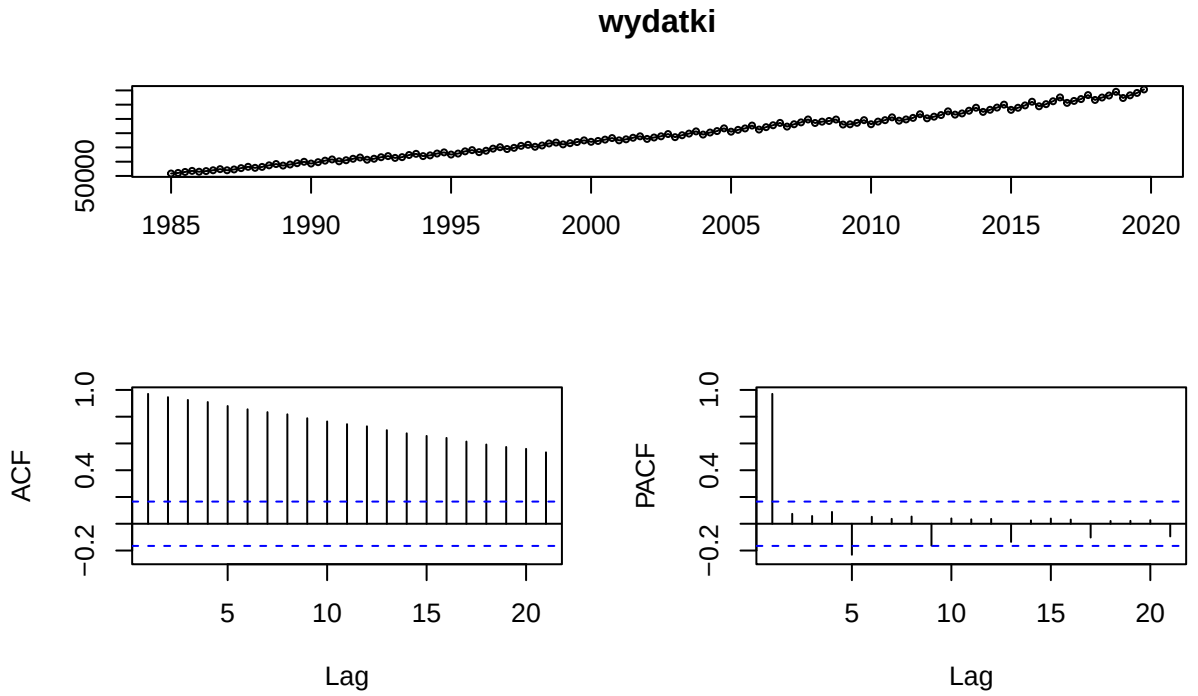

Rysunek 4: Lag plot wydatków




Rysunek 5: Lag plot komunikacji

było spodziewać przy analizie rysunku 2. Dla komunikacji natomiast najbardziej skoncentrowane wartości są dla `lag=1`.

```
tsdisplay(wydatki)
```




Rysunek 6: Wykresy z autokorelacją dla wydatków

```
tsdisplay(komunikacja)
```

Na wykresach generowanych przez funkcję `tsdisplay` (rysunki 6 i 7) bardzo dobrze widać, że dla ogólnych wydatków największa korelacja jest z rocznym opóźnieniem. Korelacja z opóźnieniem dwuletnim jest na granicy istotności, więc w naszych analizach ją pominiemy. Zgodnie z przewidywaniami dla wydatków na komunikację nie istnieje żadna istotna sezonowość.

## 2.2 Dekompozycja

Wykorzystana została dekompozycja na podstawie modelu regresji liniowej - na pierwszy rzut oka wygląda na adekwatny dla tych szeregów.

```
wydatkiDM <- decompose(wydatki, type = "multiplicative")
tsdisplay(wydatkiDM$random)
```

Jak widać na rysunku 8 czysta dekompozycja nie była w stanie sobie poradzić z tym szeregiem. Nadal bardzo widoczna jest autokorelacja dla `lag` $= 4n, n \in \mathbb{N}$

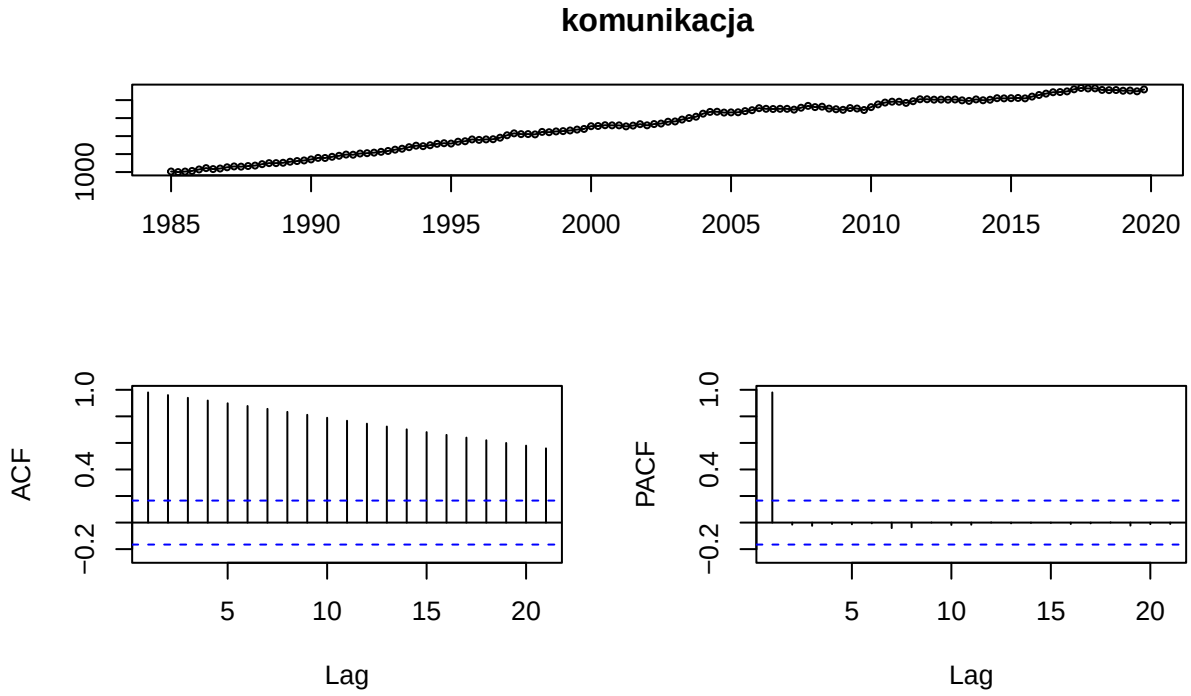
```
wydatkiDA <- decompose(wydatki, type = "additive")
tsdisplay(wydatkiDA$random)
```

W tym wypadku model multiplikatywny z rysunku 8 wydaje się lepszym rozwiązaniem ze względu na charakter danych - zmiany takie jak inflacja nakładają się mnożąc zmiany (np. coś przy inflacji na poziomie 5% po dwóch latach będzie kosztować $1,05^2x$, a nie $(1,05+1,05)x$).

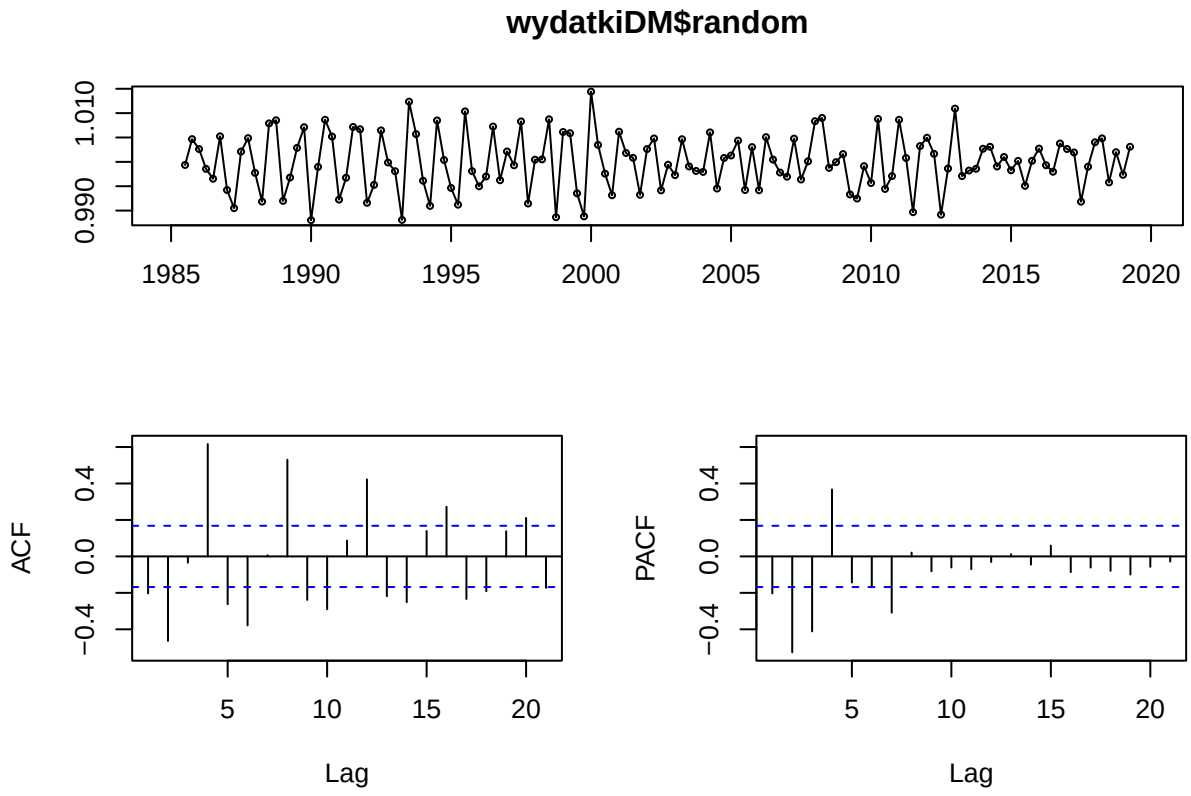
Analogicznie postąpiono dla danych nt. wydatków na komunikację. Wykorzystano dekompozycję multiplikatywną.

```
komunikacjaDM <- decompose(komunikacja, type = "multiplicative")
tsdisplay(komunikacjaDM$random)
```

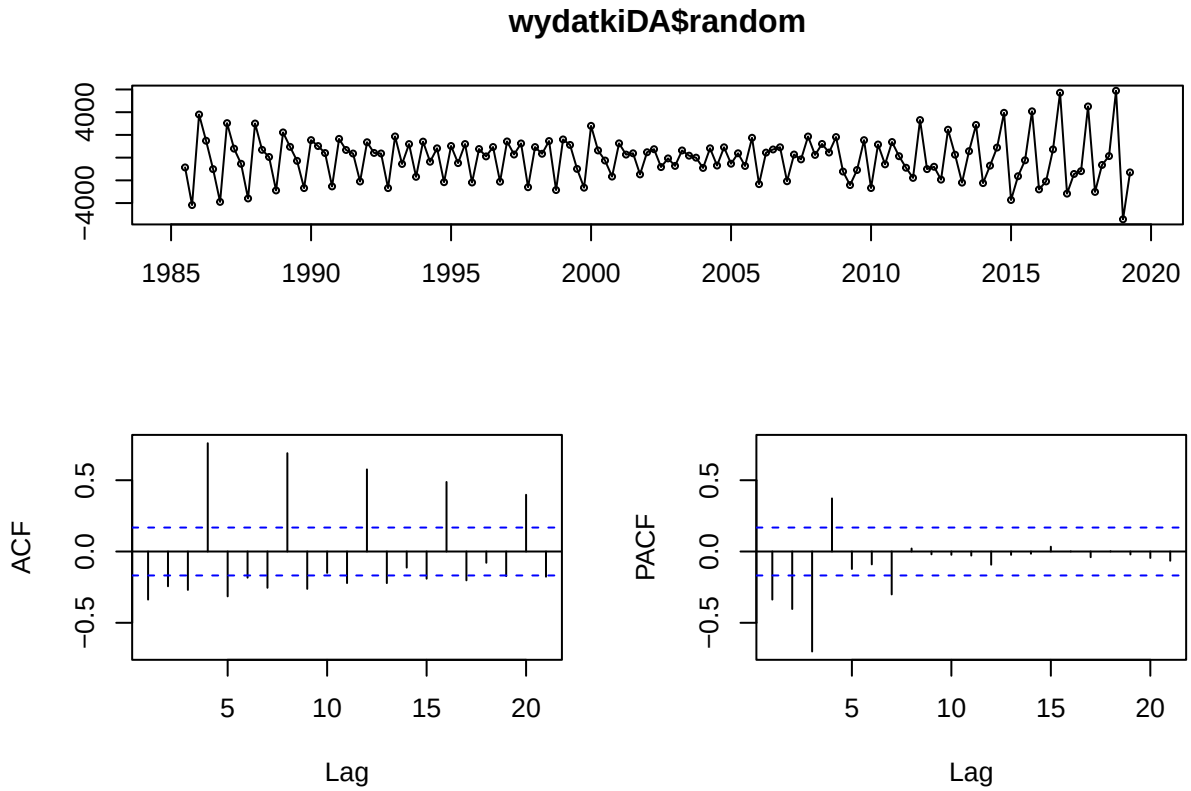
A następnie addytywną (rys. 11).



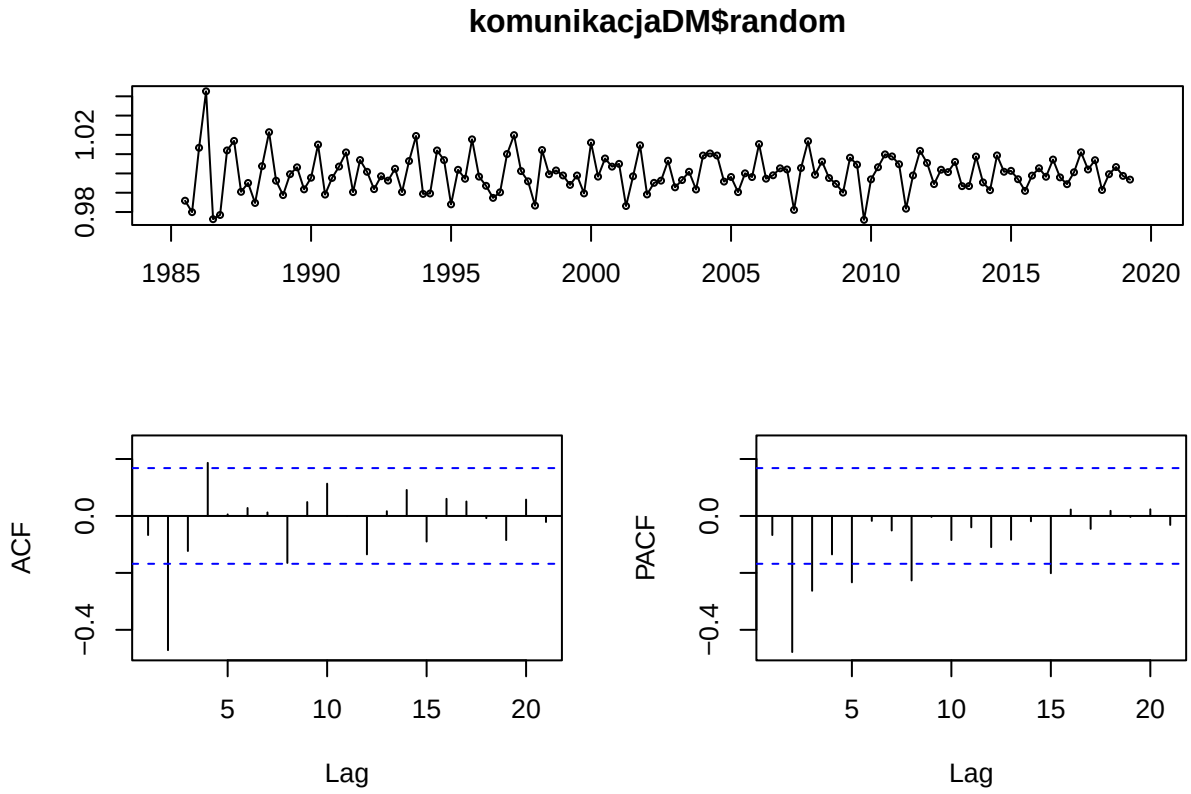

Rysunek 7: Wykresy z autokorelacją dla komunikacji




Rysunek 8: Reszty dla szeregu po dekompozycji multiplikatywnej trendu z wydatków

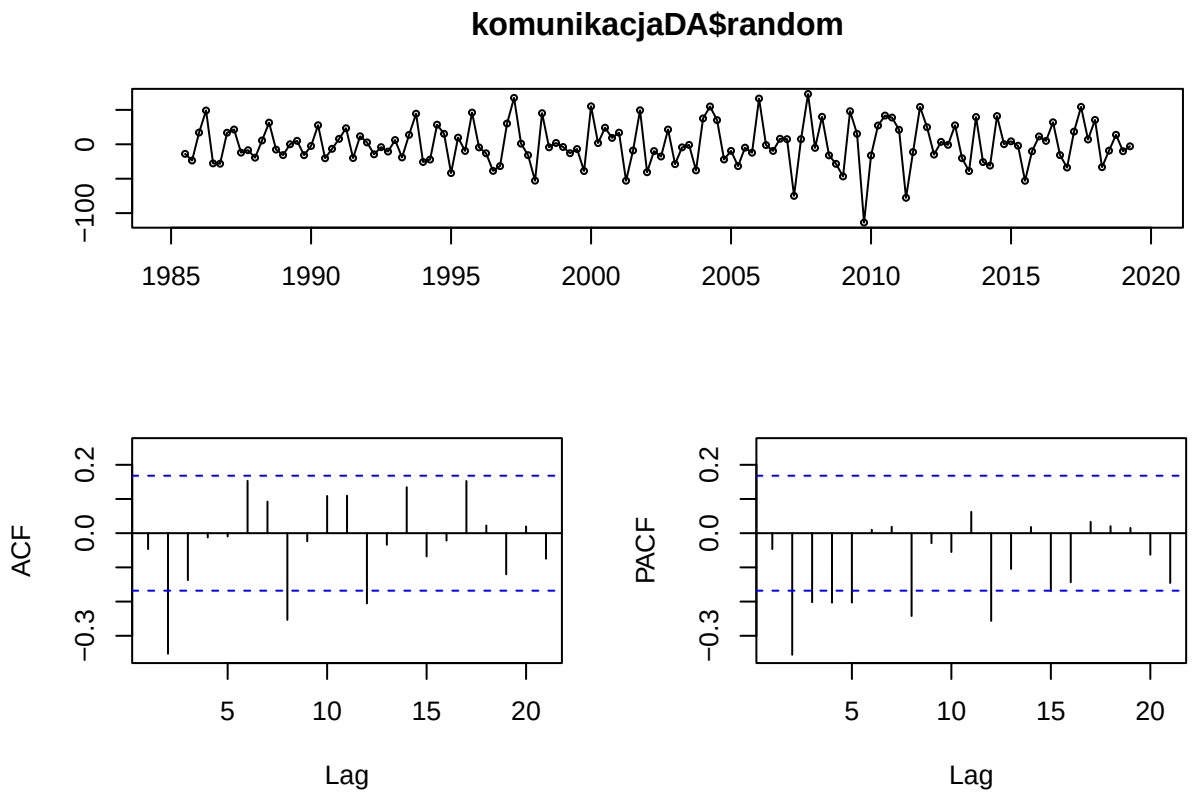



Rysunek 9: Reszty dla szeregu po dekompozycji addytywnej trendu z wydatków




Rysunek 10: Reszty dla szeregu po dekompozycji multiplikatywnej trendu z komunikacji

```
komunikacjaDA <- decompose(komunikacja, type = "additive")
tsdisplay(komunikacjaDA$random)
```




Rysunek 11: Reszty dla szeregu po dekompozycji addytywnej trendu z komunikacji

Jak widać na rysunkach 10 i 11 dekompozycja dla komunikacji dała od razu dużo lepsze efekty niż w przypadku wydatków ogólnych. Wynika to z faktu, że ten szereg nie zawiera żadnych istotnych autokorelacji. Nie zmieniło się natomiast nic w kwestii doboru metody - lepszym rozwiązaniem w tym wypadku jest dekompozycja multiplikatywna.

## 2.3 Eliminacja trendu i sezonowości

Do wyznaczenia współczynników użyta została metoda graficzna - po każdej z transformacji na danych tworzymy wykres, z którego możemy odczytać potencjalne wartości.

```
wydatkiL <- BoxCox(wydatki, BoxCox.lambda(wydatki))

tsdisplay(wydatkiL)
```

Po transformacji Boxa-Coxa jedyną szpilką wyróżniającą się na tle pozostałych jest lag równy 5 dla wydatków (12), pozostałe są nieistotne statystycznie.

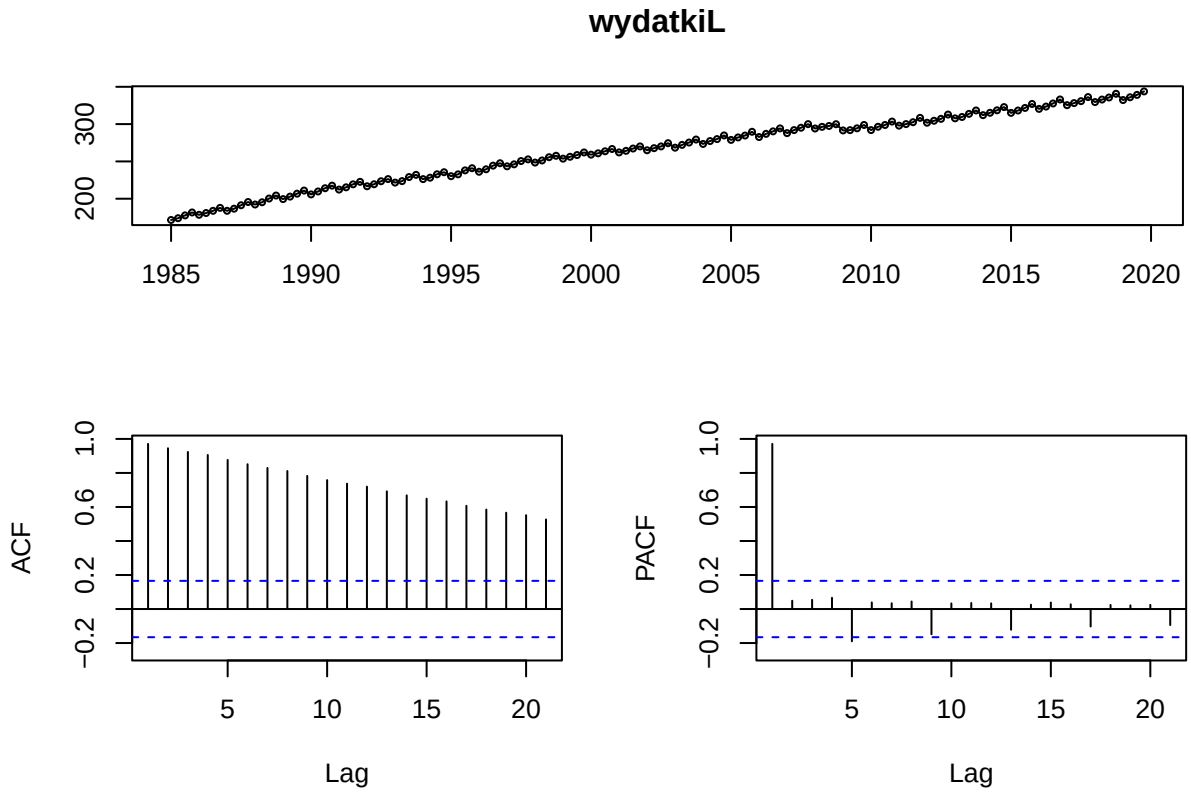
```
komunikacjaL <- BoxCox(komunikacja, BoxCox.lambda(komunikacja))
tsdisplay(komunikacjaL)
```

Natomiast dla wydatków na komunikację nie istnieje żadna wyróżniająca się wartość (rys. 13) - znaczy to, że na ten moment nie będziemy używać różnicowania na tym szeregu.

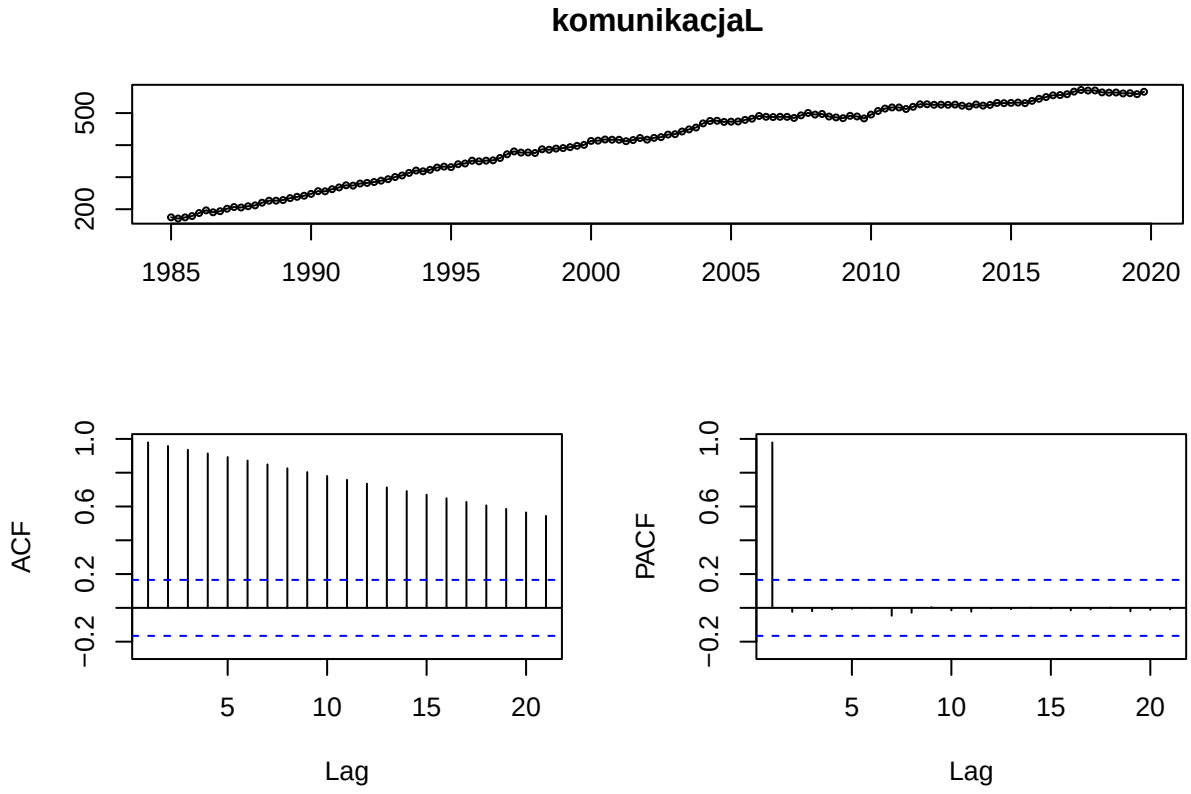
```
wydatkiL.1 <-diff(wydatkiL, lag=1)
tsdisplay(wydatkiL.1)
```

Bardzo widoczną szpilką jest teraz `lag=4`, dlatego ponowiono działanie.

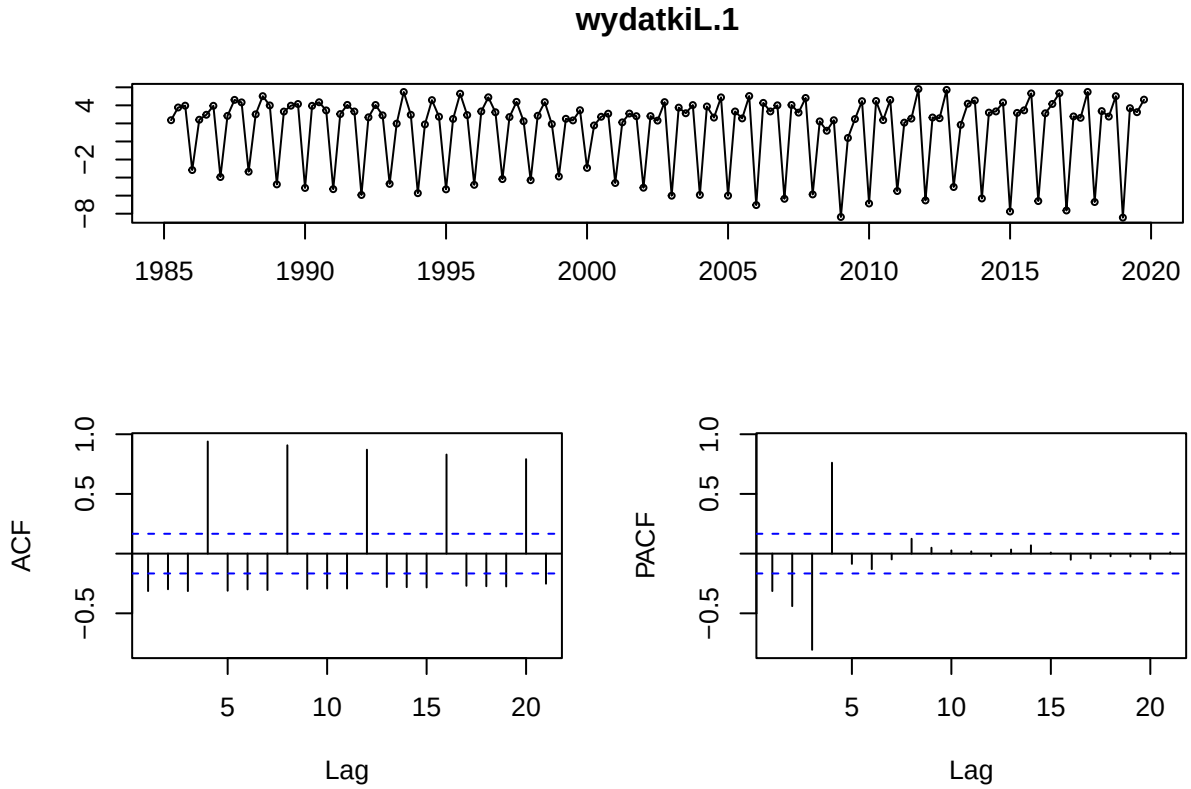
```
wydatkiL.1.4 <-diff(wydatkiL.1, lag=4)
tsdisplay(wydatkiL.1.4)
```



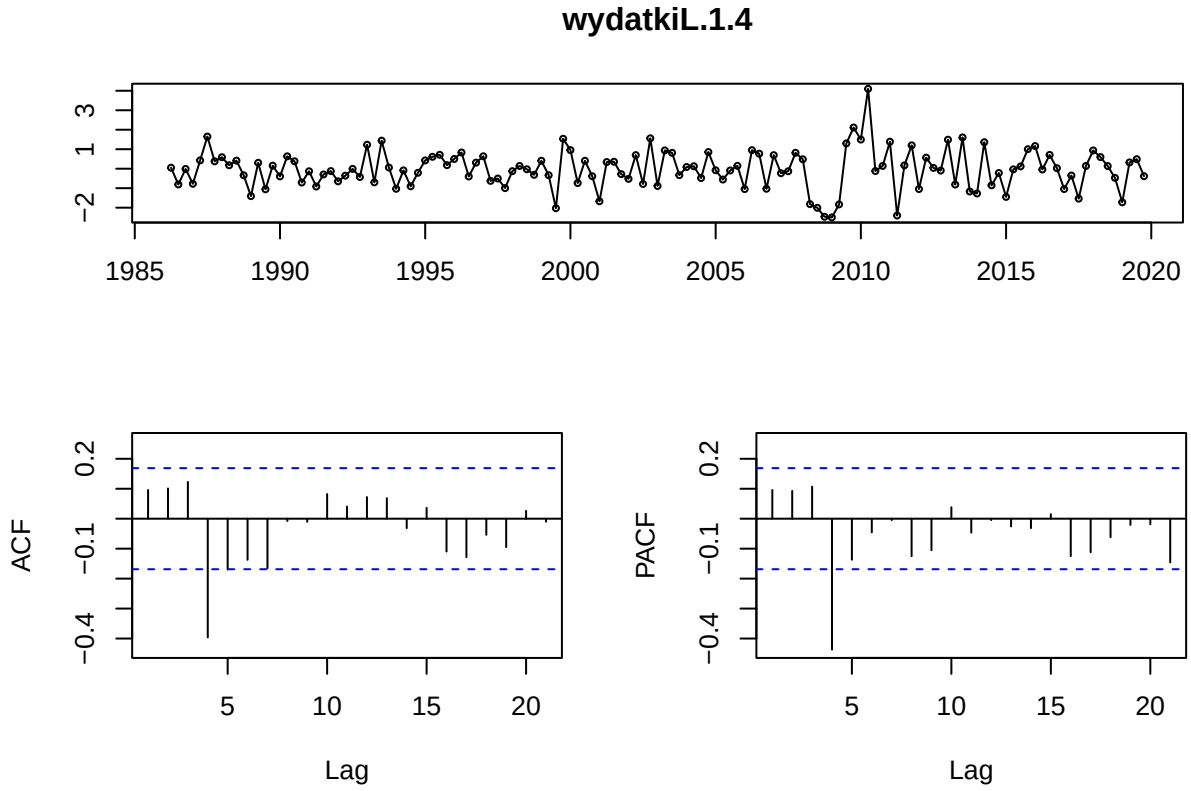

Rysunek 12: Wykres wydatków po transformacji Boxa-Coxa




Rysunek 13: Wykres wydatków na komunikację po transformacji Boxa-Coxa



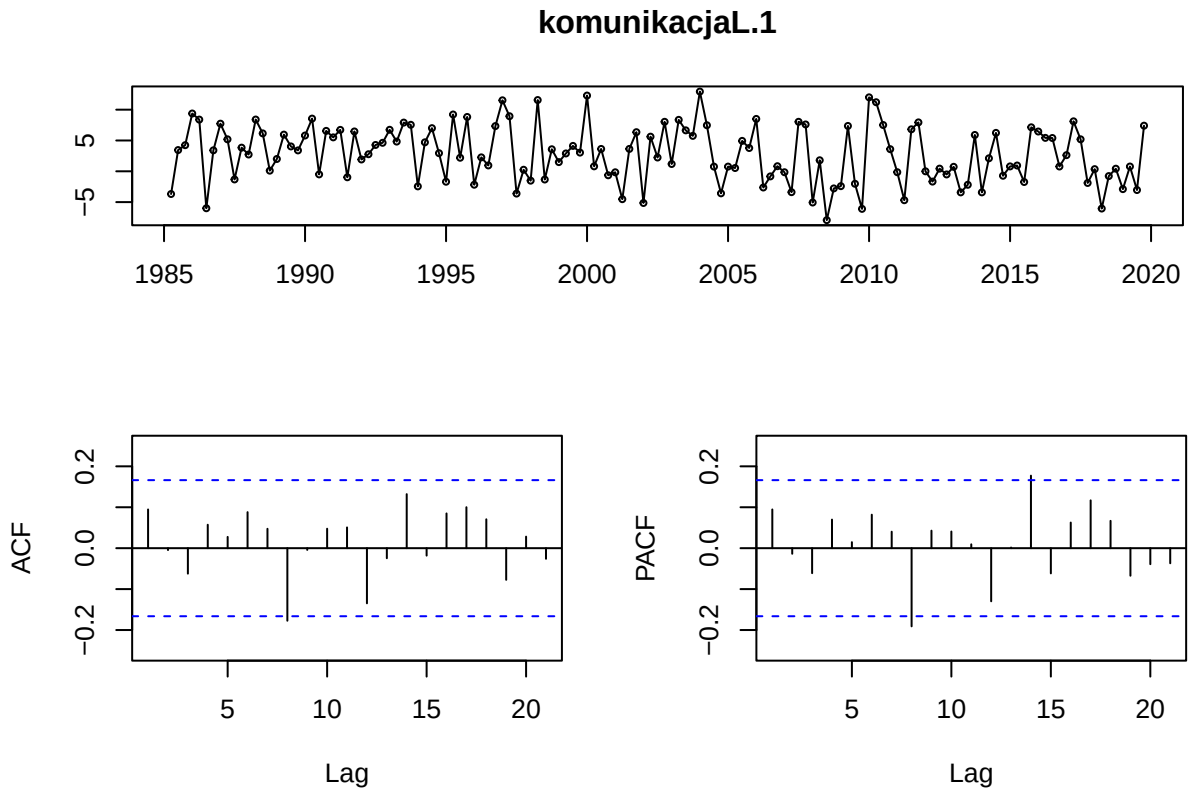

Rysunek 14: Wydatki po transformacji Boxa-Coxa i różnicowaniu z lagiem 1




Rysunek 15: (#fig:wydatki diff1-4)Wydatki po transformacji Boxa-Coxa i różnicowaniu z lagiem 1 i 4

Z powyższych wynika, że współczynniki dla modelu AR powinny wynosić 1 oraz 4 dla szeregu wydatków. Następnie postąpiono analogicznie z szeregiem wydatków na komunikację.

```
komunikacjaL.1 <-diff(komunikacjaL, lag=1)
tsdisplay(komunikacjaL.1)
```




Rysunek 16: Wydatki na komunikację po transformacji Boxa-Coxa i różnicowaniu z lagiem 1

Jak widać po zróżnicowaniu nie ma już potrzeby wyznaczać ręcznie kolejnych współczynników. Jedyną wartością, która nieznacznie odstaje jest dla wartości 8, jednak można ją pominąć.

Wyznaczymy jeszcze współczynniki przez funkcję, która robi to automatycznie.

## 2.4 Wyznaczenie współczynników dla modelu AR

Następnie zostały wyznaczone współczynniki dla modelu AR. W tym celu użyto funkcji `Pacf` na szeregach już odsezonowanych i pozbawionych trendu.

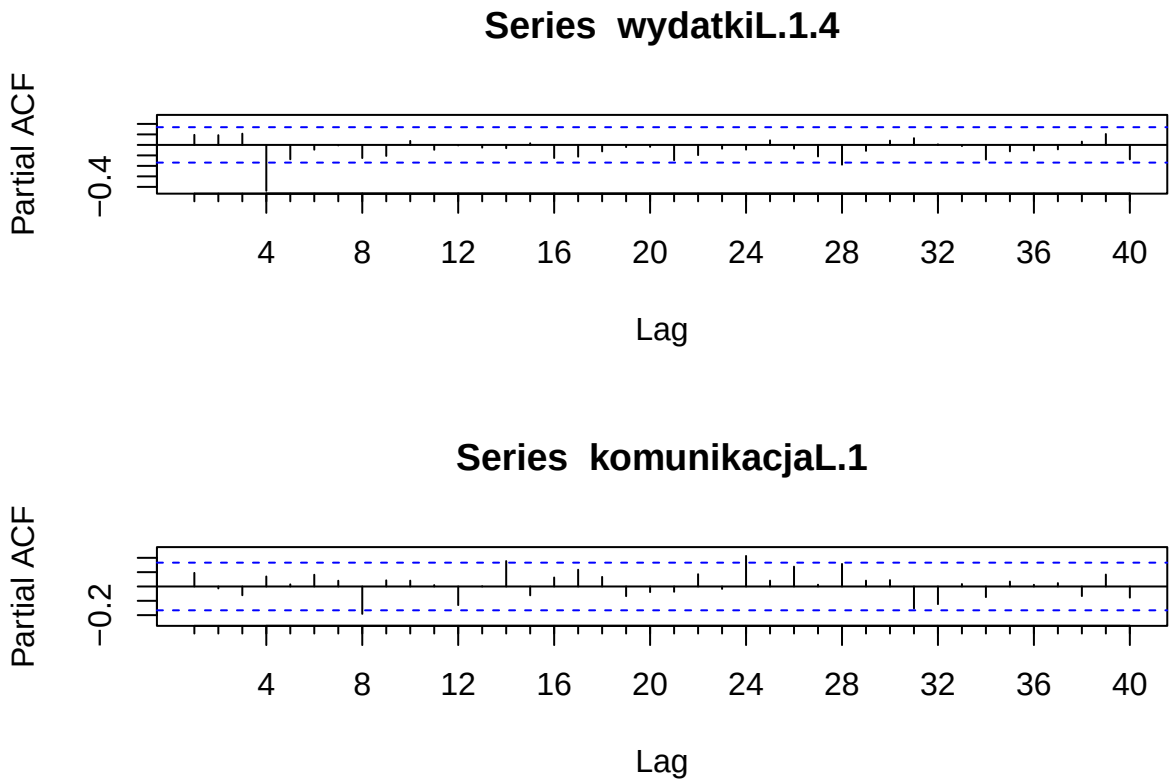
```
par(mfrow=c(2,1))
Pacf(wydatkiL.1.4, lag.max = 40)
Pacf(komunikacjaL.1, lag.max = 40)
```

Z wykresów na rysunku 17 wynika, że dla wydatków należy sprawdzić `AR(4)` i `AR(28)`, natomiast dla komunikacji `AR(8)`, `AR(24)` i rozważyć `AR(14)`.

Sprawdzono także wyniki zaproponowane przez funkcję, która automatycznie dostosowuje współczynniki autoregresji.

```
ar(wydatkiL.1.4, order.max = 24, aic = T, method = "yule-walker")
```

```
##
## Call:
## ar(x = wydatkiL.1.4, aic = T, order.max = 24, method = "yule-walker")
##
## Coefficients:
##       1        2        3        4        5
```




Rysunek 17: Wykresy częściowej autokorelacji

```
##  0.0638   0.1390   0.1570  -0.4199  -0.1372
##
## Order selected 5  sigma^2 estimated as  0.7716
```

```
ar(komunikacjaL.1, order.max = 28, aic = T, method = "yule-walker")
```

```
##
## Call:
## ar(x = komunikacjaL.1, aic = T, order.max = 28, method = "yule-walker")
##
##
## Order selected 0  sigma^2 estimated as  20.5
```

Dla szeregów zostały wyznaczone modele następujących rzędów - 5 i 0. W drugim przypadku znaczy to, że autoregresja nie ma zastosowania.

W celu wyznaczenia wartości kryteriów informacyjnych użyto funkcji `Arima` z rzędem `(p, 0, 0)`. Dla wydatków na komunikację nie sprawdzono dopasowania ze względu na rząd zerowy.

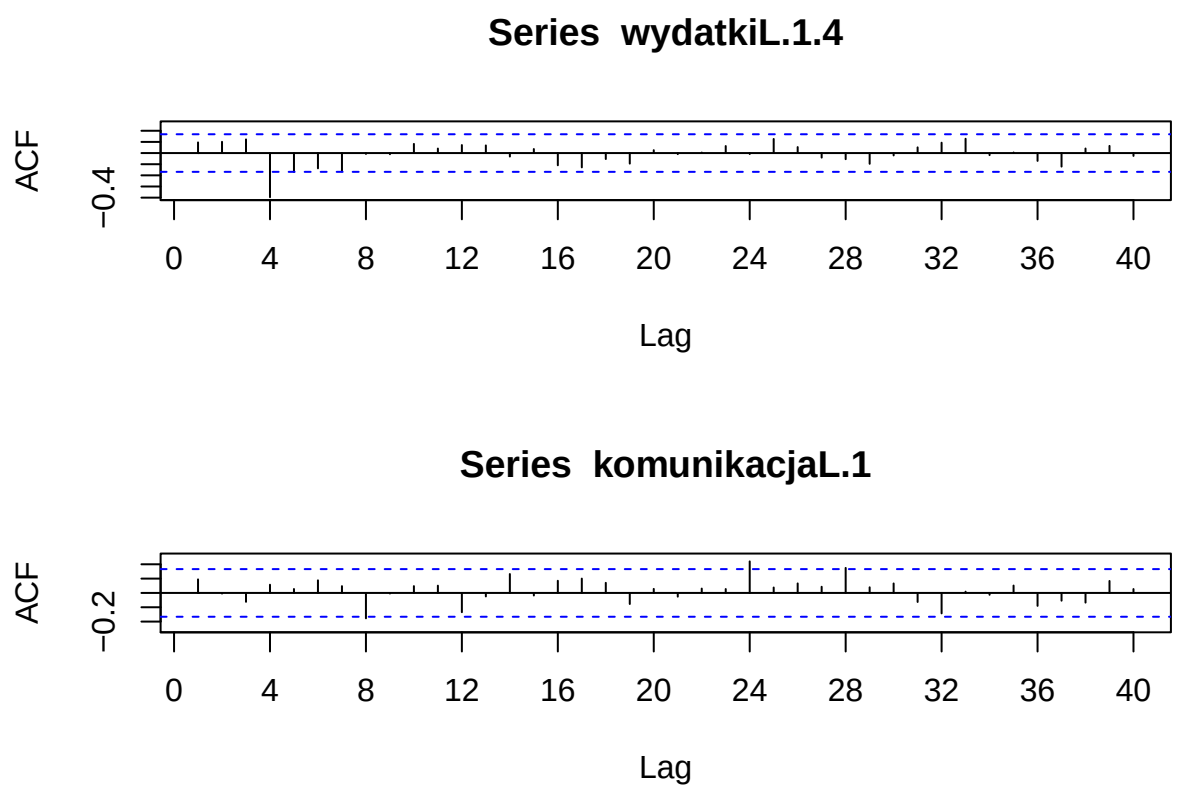
```
Arima(wydatkiL.1.4, order = c(5, 0, 0))
```

```
## Series: wydatkiL.1.4
## ARIMA(5,0,0) with non-zero mean
##
## Coefficients:
##          ar1     ar2     ar3      ar4      ar5     mean
##       0.0593  0.1415  0.1597  -0.4233  -0.1387  -0.0196
## s.e.  0.0850  0.0771  0.0760   0.0768   0.0855   0.0618
##
## sigma^2 estimated as 0.7633:  log likelihood=-170.77
## AIC=355.53   AICc=356.42   BIC=375.87
```

## 2.5 Wyznaczenie współczynników dla modelu MA

W celu ręcznego wyznaczenia współczynników dla modelu MA użyto funkcji `Acf`. Efekty jej działania widoczne są na rysunku @reg(fig:ma).

```
par(mfrow=c(2,1))
Acf(wydatkiL.1.4, lag.max = 40)
Acf(komunikacjaL.1, lag.max = 40)
```




Rysunek 18: Wykresy autokorelacji

Dla szeregu wydatków wyraźnie wybija się szpilka dla `lag=4`, natomiast dla 5 i 6 znajdują się na granicy istotności statystycznej, z tego powodu uwzględniony jedynie współczynnik `MA(4)`.

Dla szeregu dotyczącego wydatków na komunikację widoczne są 3 istotnie odstające szpilki - dla `MA(24)`, `MA(28)` oraz dla `MA(8)`. Po testach okazało się, że `MA(8)` ma najlepsze dopasowanie.

Aby wyznaczyć odpowiednie współczynniki `MA` automatycznie użyto także funkcji ARIMA.

```
Arima(wydatkiL.1.4, order = c(0,0,4))
```

```
## Series: wydatkiL.1.4
## ARIMA(0,0,4) with non-zero mean
##
## Coefficients:
##           ma1     ma2     ma3      ma4     mean
##       -0.0189  0.1231  0.0871  -0.4795  -0.0222
## s.e.   0.1011  0.0999  0.1096   0.1239   0.0539
##
## sigma^2 estimated as 0.7791:  log likelihood=-172.76
## AIC=357.52   AICc=358.17   BIC=374.95
```

```
Arima(komunikacjaL.1, order = c(0,0,8))
```

```
## Series: komunikacjaL.1
## ARIMA(0,0,8) with non-zero mean
```

```
##
## Coefficients:
##          ma1     ma2      ma3      ma4      ma5     ma6     ma7      ma8
##       0.1197  0.0587  -0.0725  -0.0425  -0.0166  0.2924  0.0671  -0.2561
## s.e.  0.0830  0.0862   0.0767   0.0848   0.0902  0.1045  0.0839   0.0798
##         mean
##       2.8298
## s.e.  0.4173
##
## sigma^2 estimated as 19.5:  log likelihood=-399.75
## AIC=819.5   AICc=821.21   BIC=848.84
```

## 2.6 Wyznaczenie optymalnych modeli

```
(wydatki.auto <- auto.arima(wydatkiL.1.4))
```

```
## Series: wydatkiL.1.4
## ARIMA(0,0,0)(0,0,1)[4] with zero mean
##
## Coefficients:
##          sma1
##       -0.5079
## s.e.   0.0849
##
## sigma^2 estimated as 0.7702:  log likelihood=-174.03
## AIC=352.05   AICc=352.14   BIC=357.86
```

```
(komunikacja.auto <- auto.arima(komunikacjaL.1))
```

```
## Series: komunikacjaL.1
## ARIMA(3,1,0)(2,0,1)[4]
##
## Coefficients:
##           ar1      ar2      ar3    sar1     sar2     sma1
##       -0.9135  -0.7616  -0.8165  0.0904  -0.2210  -0.8796
## s.e.   0.0546   0.0854   0.0712  0.1035   0.0926   0.0643
##
## sigma^2 estimated as 19.2:  log likelihood=-398.73
## AIC=811.46   AICc=812.32   BIC=831.95
```

## 2.7 Porównanie modeli

Wszystkie współczynniki zostały sprowadzone do tabeli w celu łatwiejszego porównania

Tablica 1: Porównanie współczynników AIC, AICc i BIC dla szeregu wydatków.

| Model | AIC | AICc | BIC |
|---|---|---|---|
| AR(5) | 355.53 $\pm$ | 356.42 $\pm$ | 375.87 $-$ |
| MA(4) | 357.52 $-$ | 358.17 $-$ | 374.95 $\pm$ |
| ARIMA(0,0,0)(0,0,1) | 352.05 $+$ | 352.14 $+$ | 357.86 $+$ |

Tablica 2: Porównanie współczynników AIC, AICc i BIC dla szeregu wydatków na komunikację.

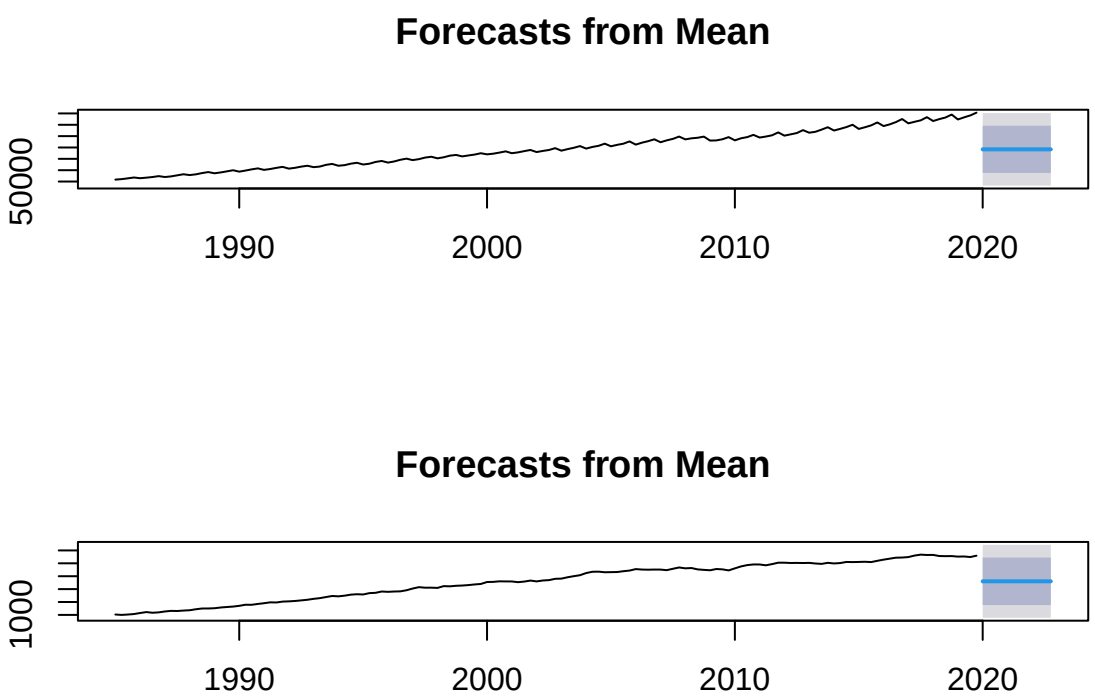
| Model | AIC | AICc | BIC |
|---|---|---|---|
| MA(8) | 819.5 − | 821.21 − | 848.84 − |
| ARIMA(3,1,0)(2,0,1) | 811.46 + | 812.32 + | 831.95 + |

Jak widać w każdym z przypadków najlepszym okazał się automatycznie dobrany model ARIMA.

## 2.8 Prognozowanie naiwne

Zastosowano kilka metod prognozowania. Pierwszą zastosowaną metodą była obliczona na podstawie średniej. Jak widać na rysunku 19 jest ona nieprzystająca do faktycznych danych. Wynika to z faktu, że oba szeregi zawierają trend, natomiast szereg wydatków ogólnych dodatkowo zawiera sezonowość.

```
wydatki.mean.forecast <- meanf(wydatki, h = 12)
komunikacja.mean.forecast <- meanf(komunikacja, h = 12)

par(mfrow=c(2,1))
plot(wydatki.mean.forecast)
plot(komunikacja.mean.forecast)
```

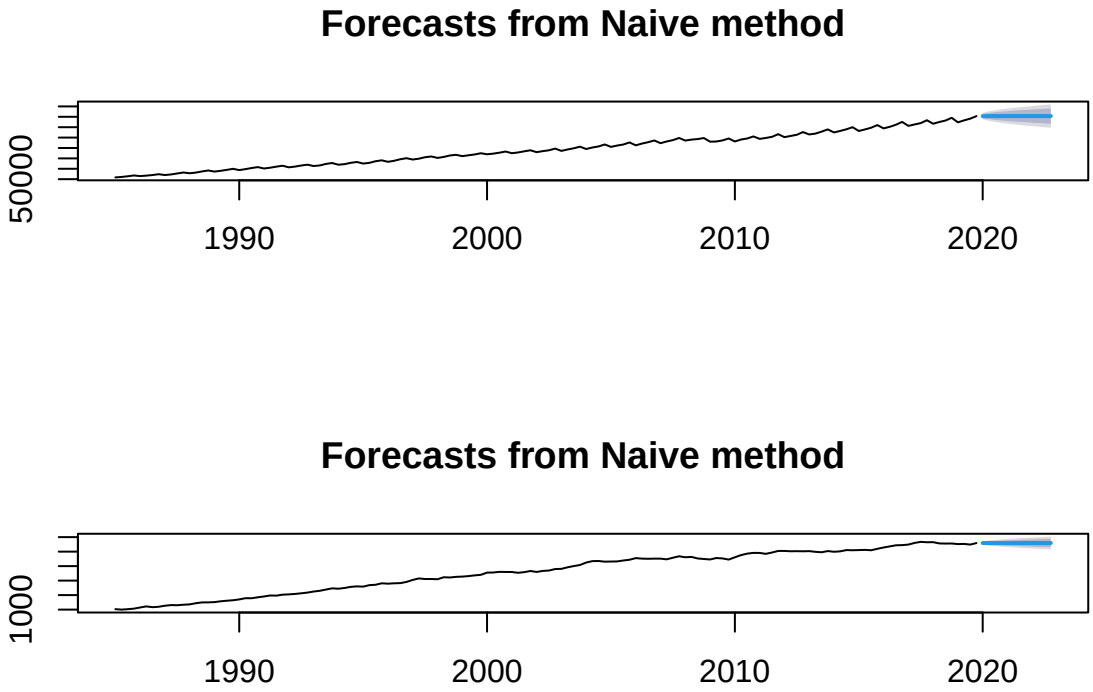



Rysunek 19: Wykresy z przewidywanymi wartościami na podstawie średniej

Jak widać na rysunku 20 metoda naiwna daje znacznie lepsze rezultaty od średniej. Znaczy to, że o wiele lepsze dopasowanie daje poprzedni pomiar niż średnia z całości.

```
wydatki.naive.forecast <- naive(wydatki, h = 12)
komunikacja.naive.forecast <- naive(komunikacja, h = 12)

par(mfrow=c(2,1))
plot(wydatki.naive.forecast)
plot(komunikacja.naive.forecast)
```

Metoda z rysunku 21 jest nieco gorsza od prognozowania na podstawie poprzedniego pomiaru, ponieważ trend ma dużo większe znaczenie w tym przypadku niż sezonowość.




Rysunek 20: Wykresy z przewidywanymi wartościami na podstawie poprzedniego okresu

```
wydatki.snaive.forecast <- snaive(wydatki, h = 12)
komunikacja.snaive.forecast <- snaive(komunikacja, h = 12)

par(mfrow=c(2,1))
plot(wydatki.snaive.forecast)
plot(komunikacja.snaive.forecast)
```

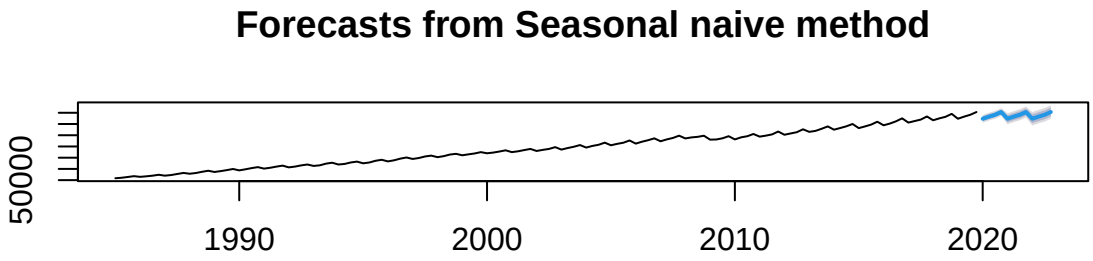
Metoda oparta o dryf wydaje się w tym przypadku najlepszą metodą naiwną.

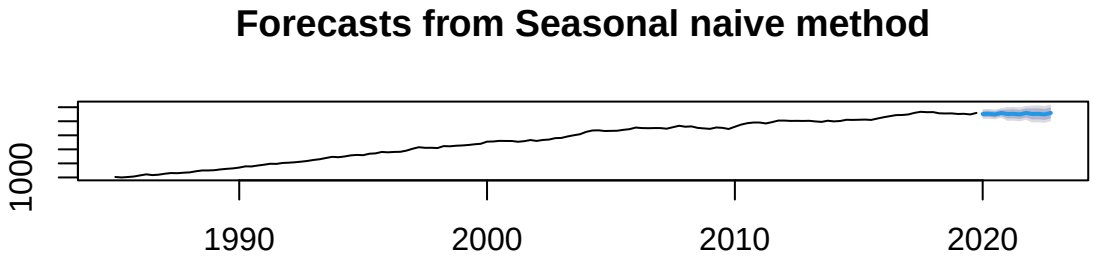
```
wydatki.rwf.forecast <- rwf(wydatki, h = 12, drift = T)
komunikacja.rwf.forecast <- rwf(komunikacja, h = 12, drift = T)

par(mfrow=c(2,1))
plot(wydatki.rwf.forecast)
plot(komunikacja.rwf.forecast)
```

# 3 Wnioski

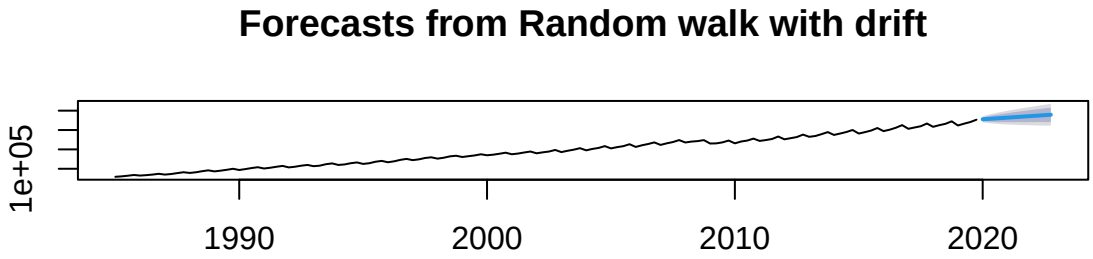
Jak widać na danych, które zostały przeanalizowane, z dość dobrą dokładnością można obliczyć parametry danego szeregu. Pozwala to na przewidzenie zachowania rynku nie tylko jako całości, ale też danego sektora jak np. komunikacja. Dzięki temu możemy lepiej zaplanować strategię biznesową jako firma czy uwzględnić różnice w koszcie życia w danym czasie.

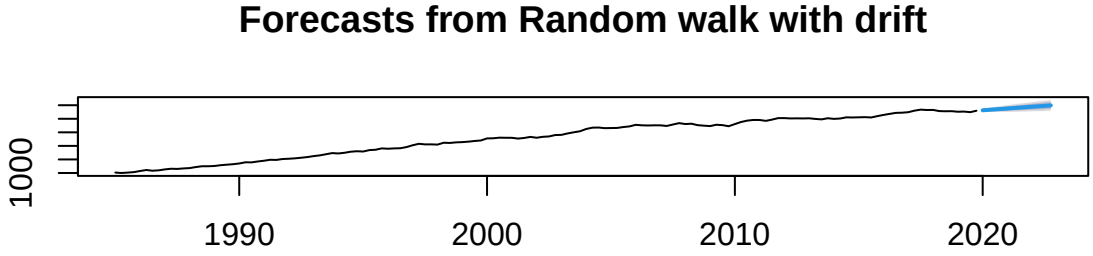
Dzięki takiej analizie nawet "szary Kowalski" może zauważyć, że ceny rosną, więc może się przekonać, że jego oszczędności bez inwestycji za jakiś czas mogą okazać się zupełnie bezwartościowe.




Rysunek 21: Wykresy z przewidywanymi wartościami na podstawie poprzedniego sezonu




Rysunek 22: Wykresy z przewidywanymi wartościami na podstawie dryfu