TIME SERIES & ARIMA MODELING

Anthony Anh Quoc Doan
SGV AI Presentation (Nov 6, 2018)

About Me

Currently a statistician or data scientist with a focus in statistical models.

Have a BS in Computer Science and working part-time on a Master in Applied Statistic. 10+ years of professional work experience as a programmer (public sector for ~2.5 years, private sector 5 years, and 3+ years freelance consultant & ceo) and 2+ years as a part-time statistician doing consultant.

Acknowledgement

- Forecasting: Principles and Practice by *Rob J Hyndman and George Athanasopoulos*
- *All my statistic and math professors at my University (CSULB)*

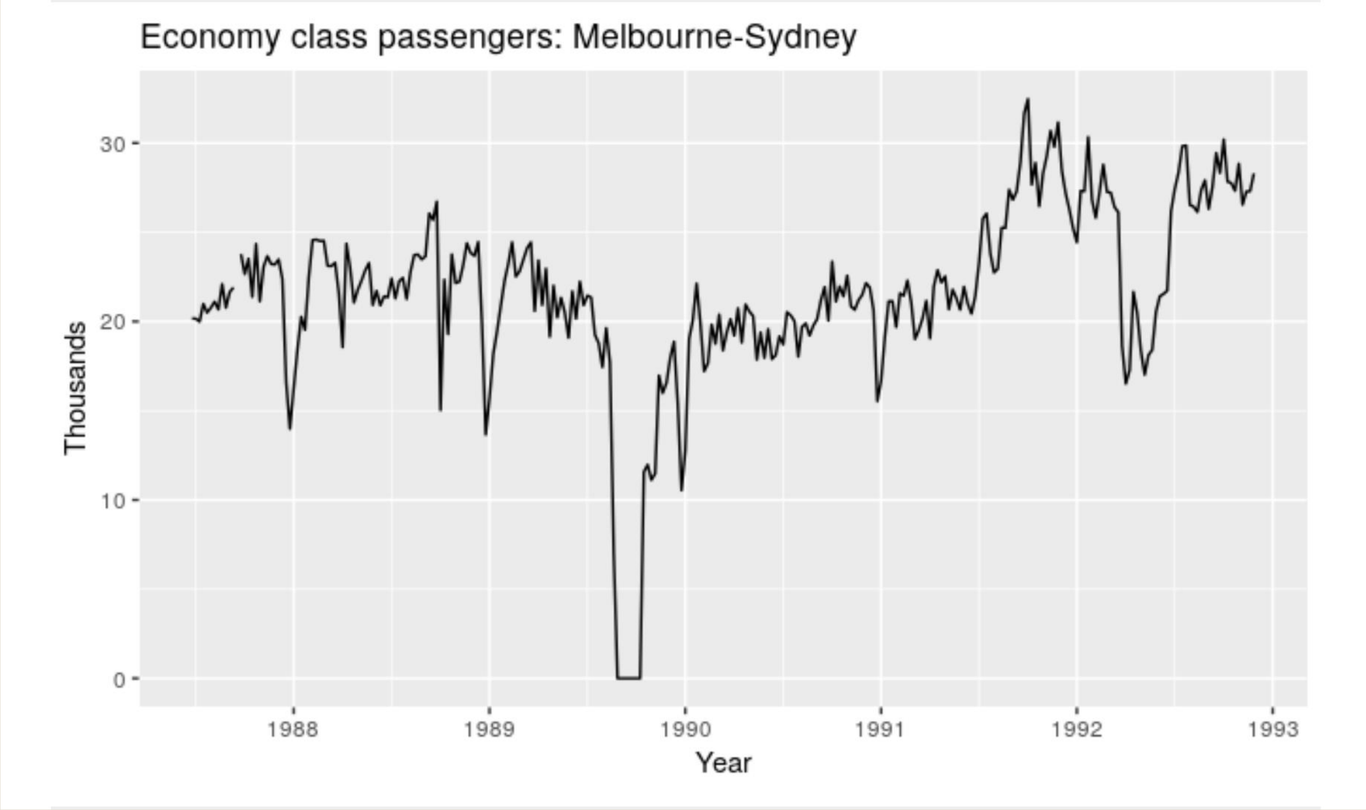
- Almost all of the examples and the pictures and codes are from Hyndman and Athanasopoulos book.

Motivation

- Time series data forecast/predicting models are used in many industry
- Retail industry for inventory
- Uber uses it see how much driver they need in different cities for that week or month or even day
- Stock price prediction

What is Time Series Data?

- Time series data is a sequence of data points that measuring the same thing over time
  - The data is time dependent
  - Every observations/values/responses are affected by past values (there is correlation among them)
  - Order (past, present, future)



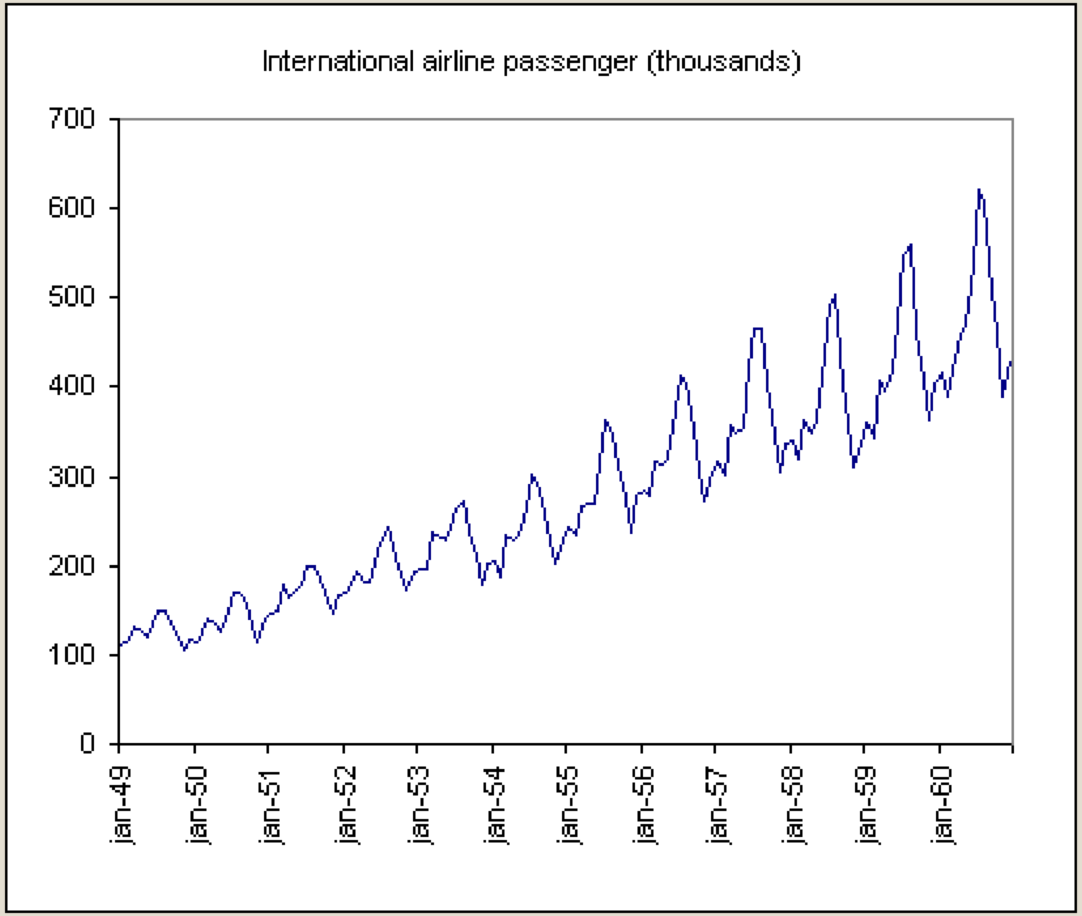
Anatomy of Time Series Data

- Every Time Series Data can be decomposed into 3 components
  - Trend-Cyclical
  - Seasonal
  - Remainder

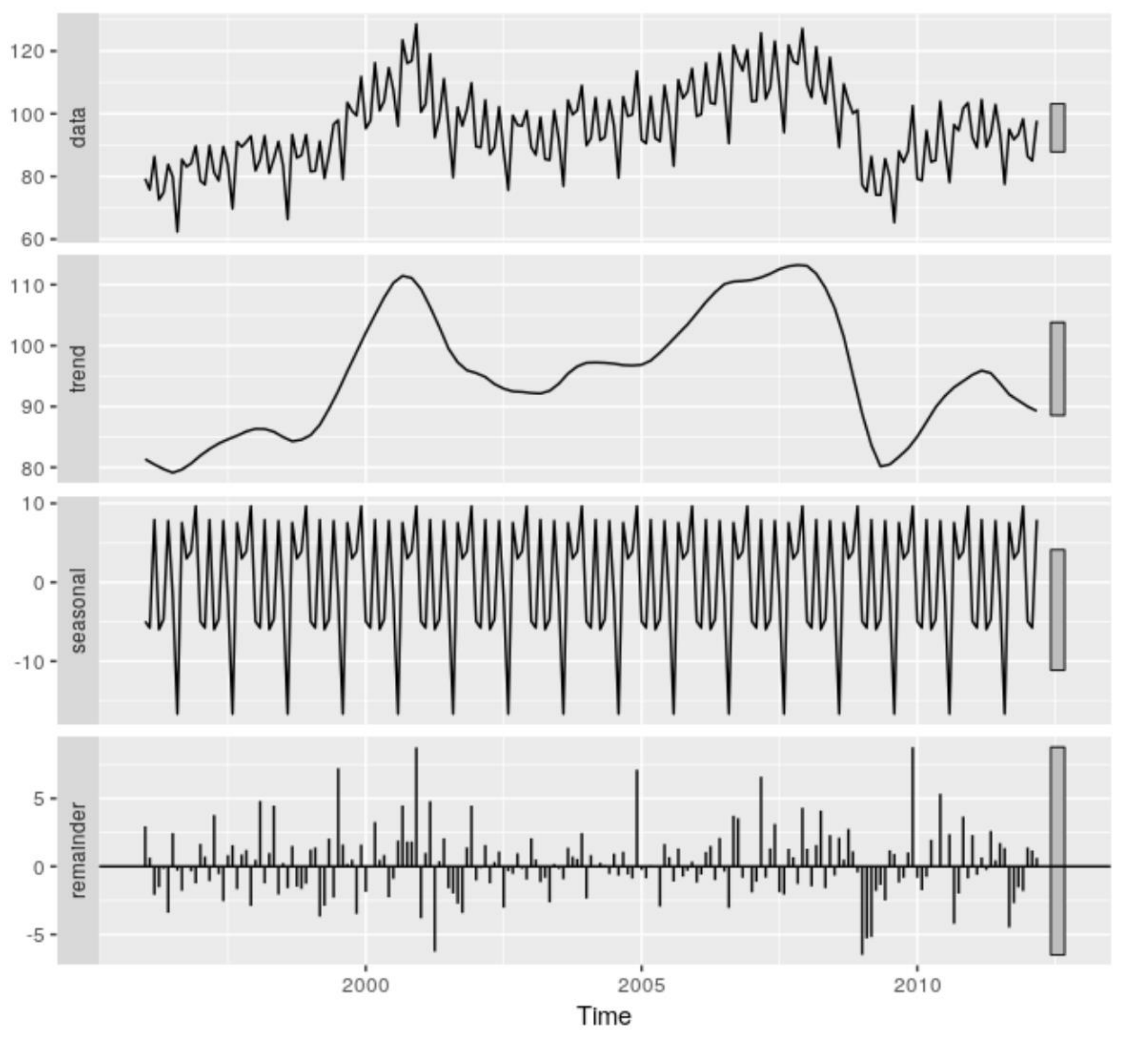
$$y_t = S_t + T_t + R_t,$$

$$y_t = S_t \times T_t \times R_t.$$

Seasonal + Trend



Time Series Decomposition in Action



Why have special models for just Time Series?

- Time series observations are correlated
- Time is ordered past, present, and future. Example: 1999, 2000, 2001, 2002, …, 2018

Example of Linear Regression and why it doesn't work (most of the time)

- Linear regression with non time series data:
  - Y_hat = beta_1*x_1 + … + beta_k*x_k

- Linear regression with time series data:

$$\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 x_{1,t} + \hat{\beta}_2 x_{2,t} + \cdots + \hat{\beta}_k x_{k,t},$$

- It's inefficient it does not account for the correlation (relationship) between the predictors so there are information that it does not utilized
- There are tweaks to linear regression to make it works but tedious
- Can only handle stationary data (ETS can handle nonstationary)

ARIMA Model

- Why? Arima is one of two most used model for time series to forecast
- Dr. Box and Dr. Jenkins created ARIMA in the 1970s they're the giants in Time Series statistic
- Neural Network and non-statistical methods are only in research phase and there aren't any real* improvement in general as seen in Makridakis Competitions (Time series competition)

- * In the M4 competition the best model was a hybrid model between machine learning & statistic model but it's "hand crafted"

ARIMA

- AR - AutoRegressive
- I - Integrated ("to difference" will explain soon)
- MA – Moving Averages

Autoregressive AR(p)

, an autoregressive model of order p can be written as

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t,$$

PACF

- This function plays an important role in data analysis aimed at identifying the extent of the lag in an autoregressive model.
- Shows correlation between time lags.
- We'll see this in later slides in much greater details.

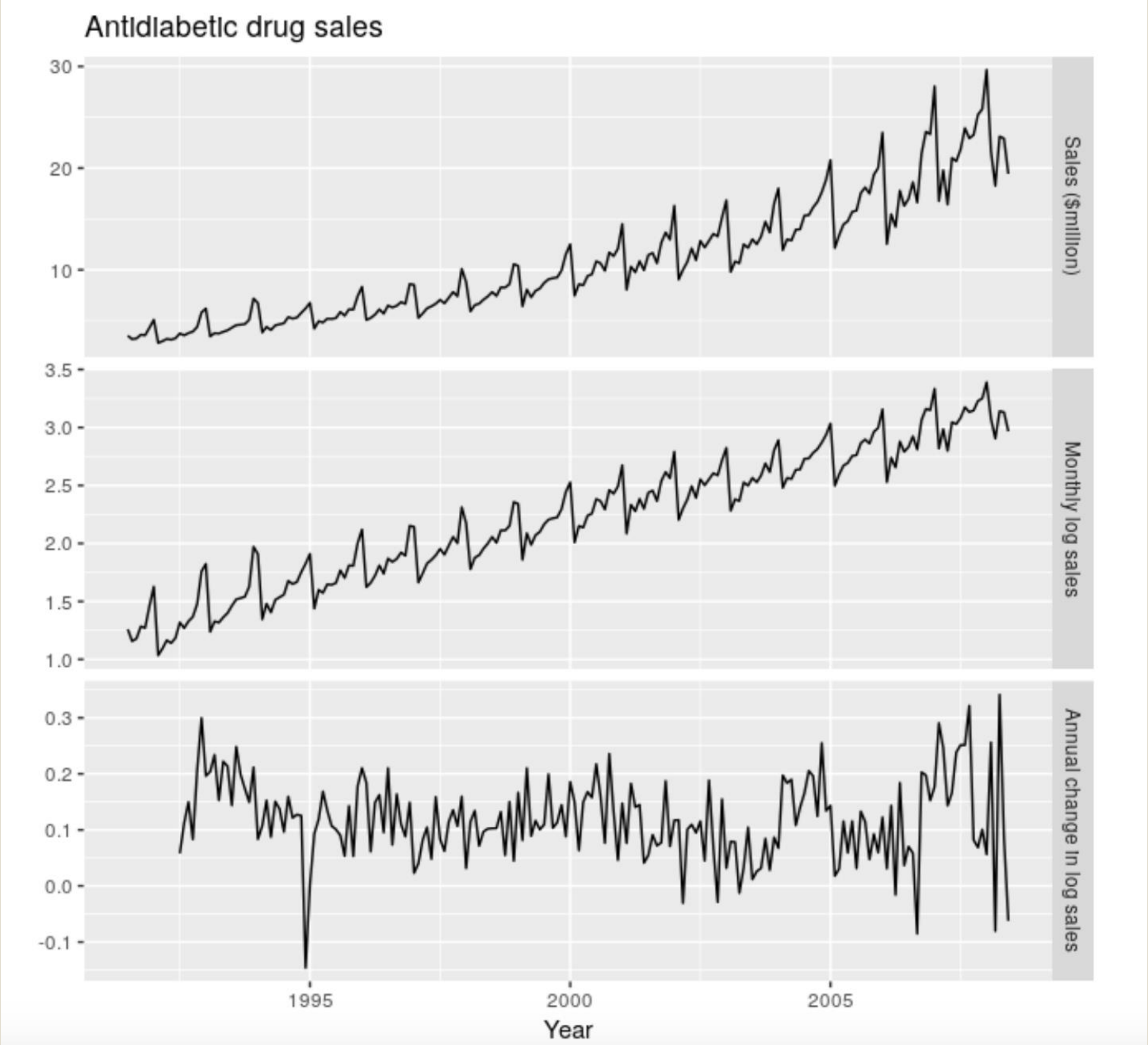
Moving Average MA(q)

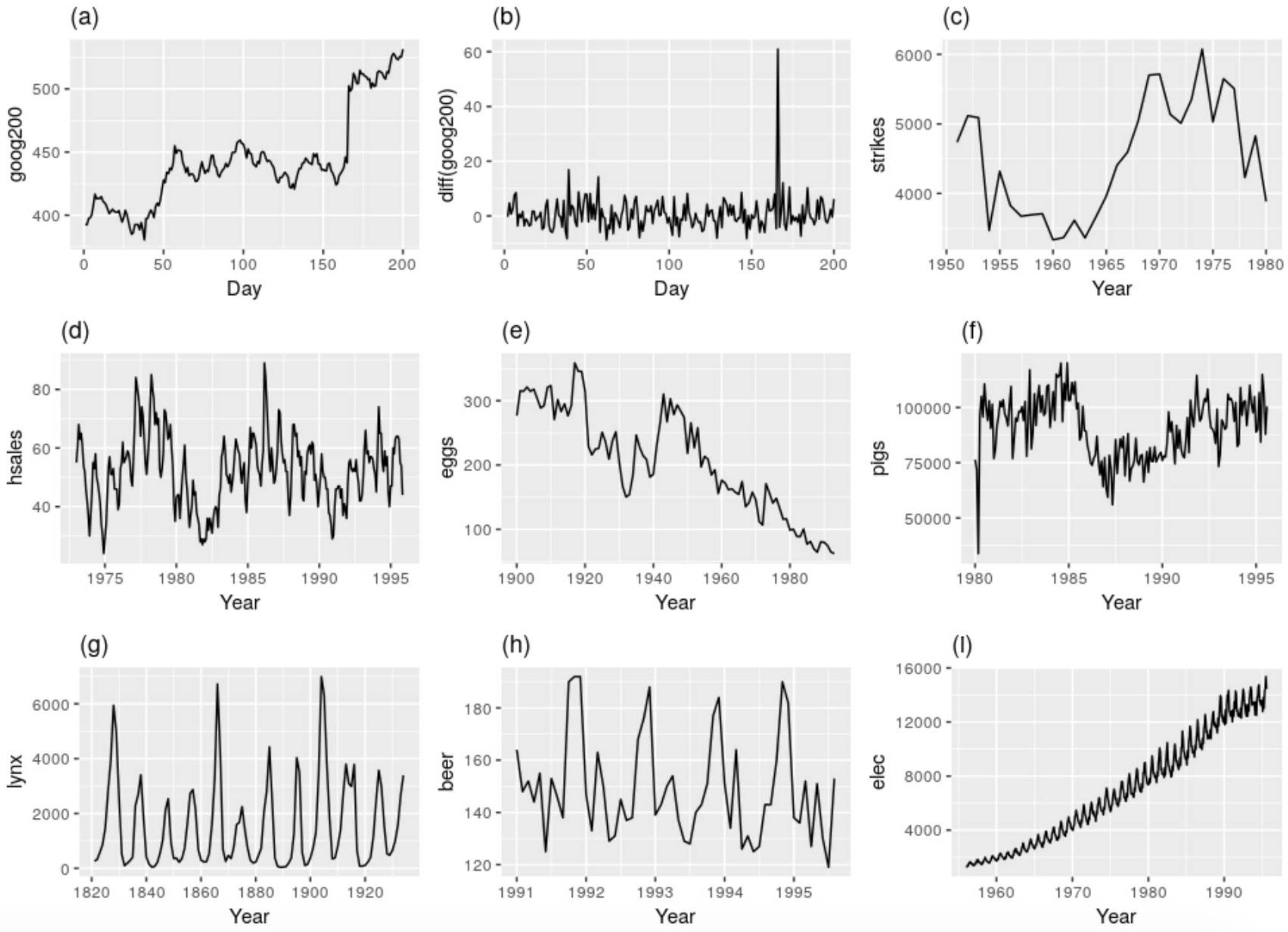
$$y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q},$$

ACF

- This function plays an important role in data analysis aimed at identifying the extent of the lag in Moving Average models.
- Shows correlation between time lags.
- We'll see this in later slides in much greater details.

Stationary Time Series

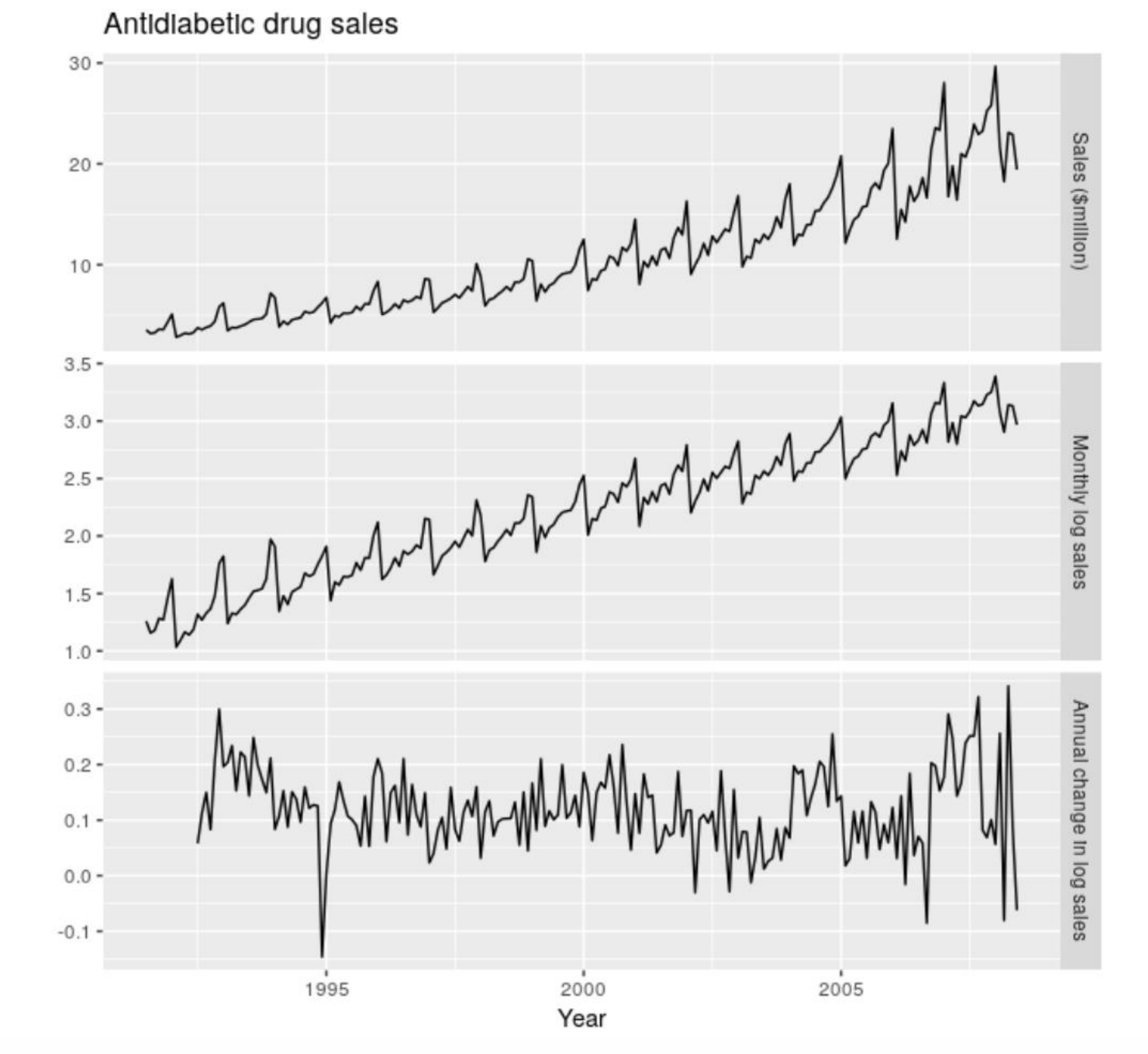



(a)
goog200
Day
(b)
diff(goog200)
Day
(c)
strikes
Year
(d)
hsales
Year
(e)
eggs
Year
(f)
pigs
Year
(g)
lynx
Year
(h)
beer
Year
(i)
elec
Year

Difference

- Makes non-stationary time series into stationary time series by removing trends and seasonality
- If the time series have trend and seasonality. Remove seasonality first to see if it fix it before removing trend on top of seasonlity

```r
cbind("Sales ($million)" = a10,
      "Monthly log sales" = log(a10),
      "Annual change in log sales" = diff(log(a10),12)) %>%
  autoplot(facets=TRUE) +
    xlab("Year") + ylab("") +
    ggtitle("Antidiabetic drug sales")
```



Unit Root Test

- To determine if we need differences

```r
library(urca)
goog %>% ur.kpss() %>% summary()
#>
#> #######################
#> # KPSS Unit Root Test #
#> #######################
#>
#> Test is of type: mu with 7 lags.
#>
#> Value of test-statistic is: 10.72
#>
#> Critical value for a significance level of:
#>                 10pct  5pct 2.5pct  1pct
#> critical values 0.347 0.463  0.574 0.739
```

```r
goog %>% diff() %>% ur.kpss() %>% summary()
#> 
#> #######################
#> # KPSS Unit Root Test #
#> #######################
#> 
#> Test is of type: mu with 7 lags.
#> 
#> Value of test-statistic is: 0.0324
#> 
#> Critical value for a significance level of:
#>                 10pct  5pct 2.5pct  1pct
#> critical values 0.347 0.463  0.574 0.739
```

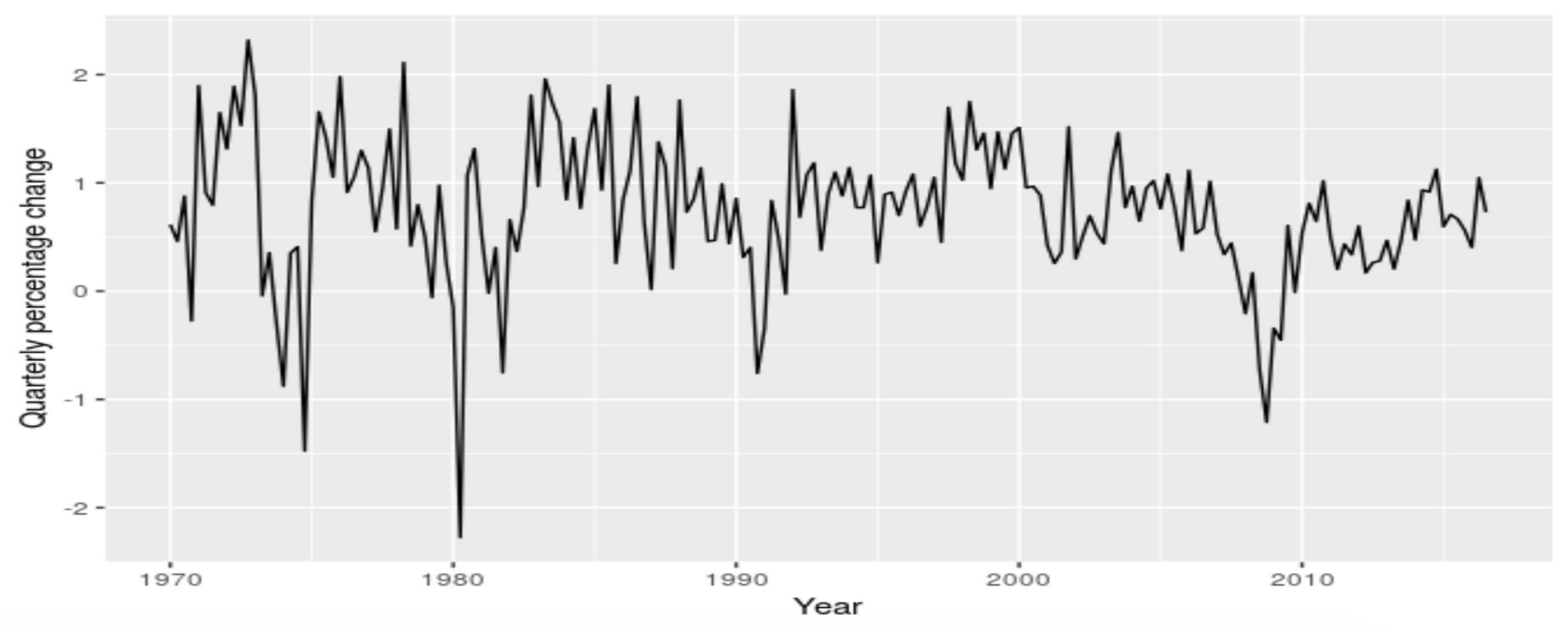
Nonseasonal ARIMA(p,d,q) model

- AR(p)
- I – d
- MA(q)

Example of ARIMA(p,d,q)

US Consumption expenditure

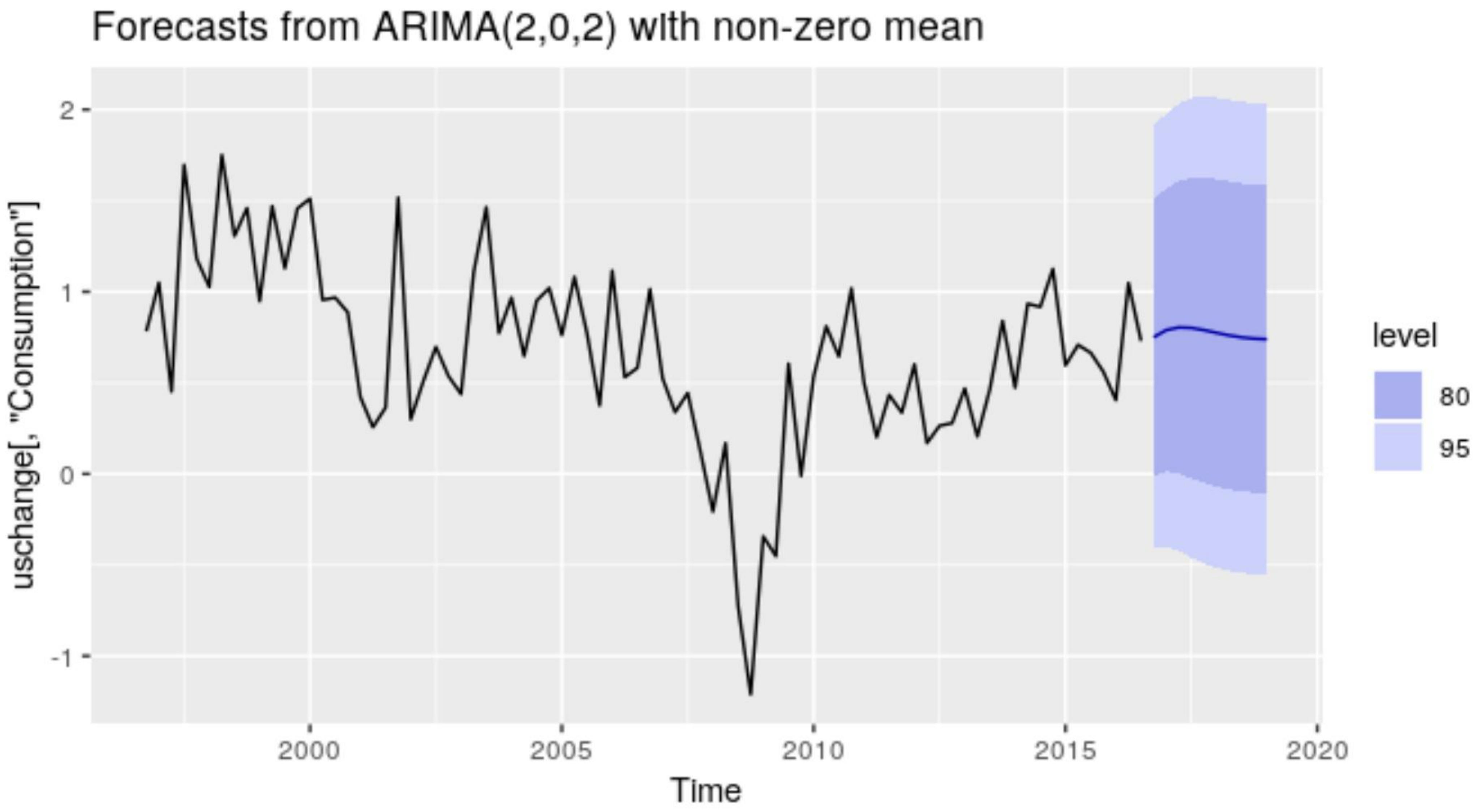
```r
autoplot(uschange[,"Consumption"]) +
  xlab("Year") + ylab("Quarterly percentage change")
```



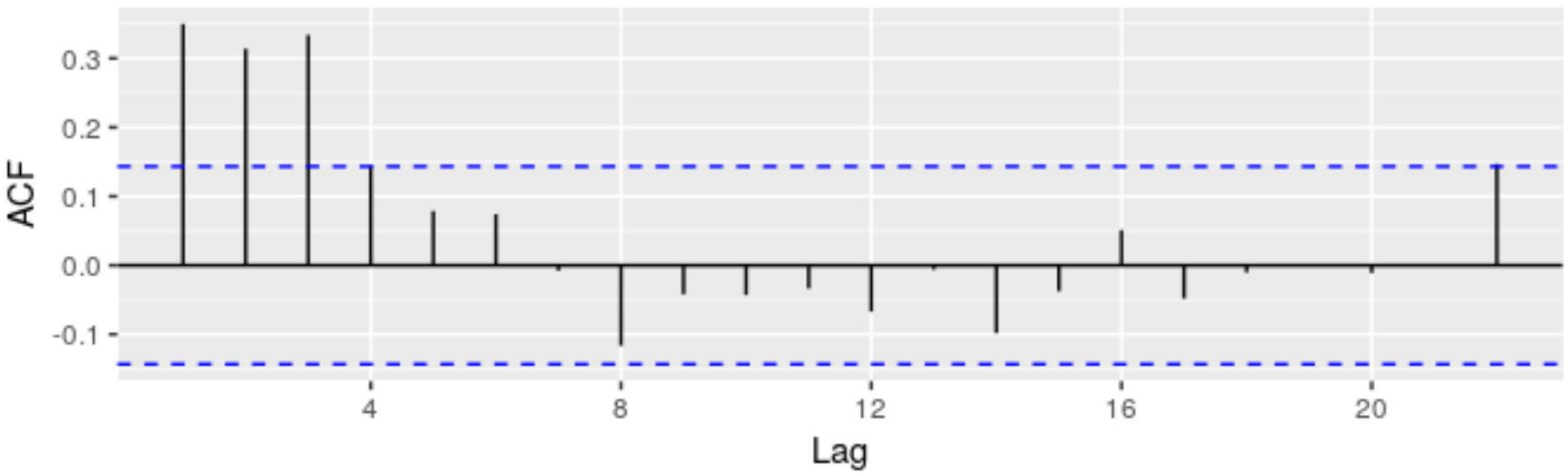
```r
fit <- auto.arima(uschange[,"Consumption"], seasonal=FALSE)
```

```
#> Series: uschange[, "Consumption"]
#> ARIMA(2,0,2) with non-zero mean
#>
#> Coefficients:
#>          ar1     ar2     ma1    ma2   mean
#>        1.391  -0.581  -1.180  0.558  0.746
#> s.e.   0.255   0.208   0.238  0.140  0.084
#>
#> sigma^2 estimated as 0.351:  log likelihood=-165.1
#> AIC=342.3   AICc=342.8   BIC=361.7
```

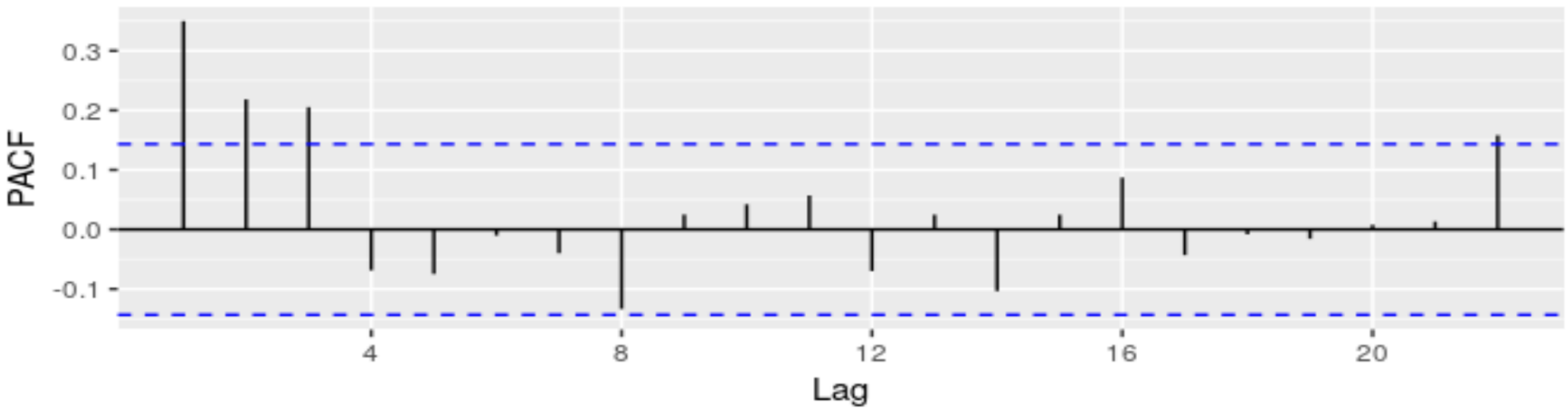
```
fit %>% forecast(h=10) %>% autoplot(include=80)
```



```r
ggAcf(uschange[,"Consumption"],main="")
```



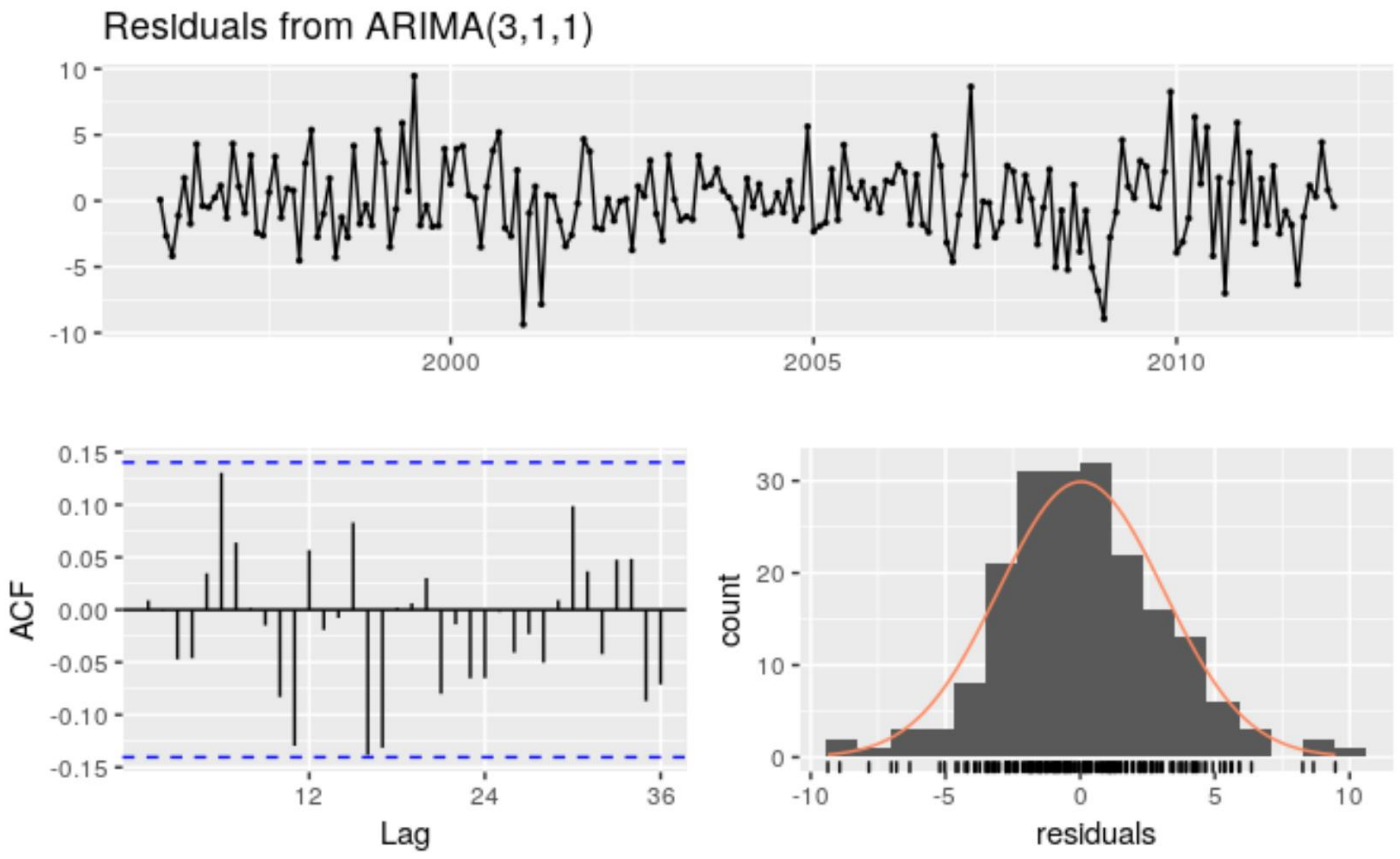
```
ggPacf(uschange[,"Consumption"],main="")
```



```r
(fit2 <- Arima(uschange[,"Consumption"], order=c(3,0,0)))
#> Series: uschange[, "Consumption"]
#> ARIMA(3,0,0) with non-zero mean
#>
#> Coefficients:
#>         ar1    ar2    ar3   mean
#>       0.227  0.160  0.203  0.745
#> s.e.  0.071  0.072  0.071  0.103
#>
#> sigma^2 estimated as 0.349:  log likelihood=-165.2
#> AIC=340.3   AICc=340.7   BIC=356.5
```

```r
(fit3 <- auto.arima(uschange[,"Consumption"], seasonal=FALSE,
  stepwise=FALSE, approximation=FALSE))
#> Series: uschange[, "Consumption"]
#> ARIMA(3,0,0) with non-zero mean
#>
#> Coefficients:
#>          ar1    ar2    ar3   mean
#>        0.227  0.160  0.203  0.745
#> s.e.   0.071  0.072  0.071  0.103
#>
#> sigma^2 estimated as 0.349:  log likelihood=-165.2
#> AIC=340.3   AICc=340.7   BIC=356.5
```

```r
checkresiduals(fit)
```

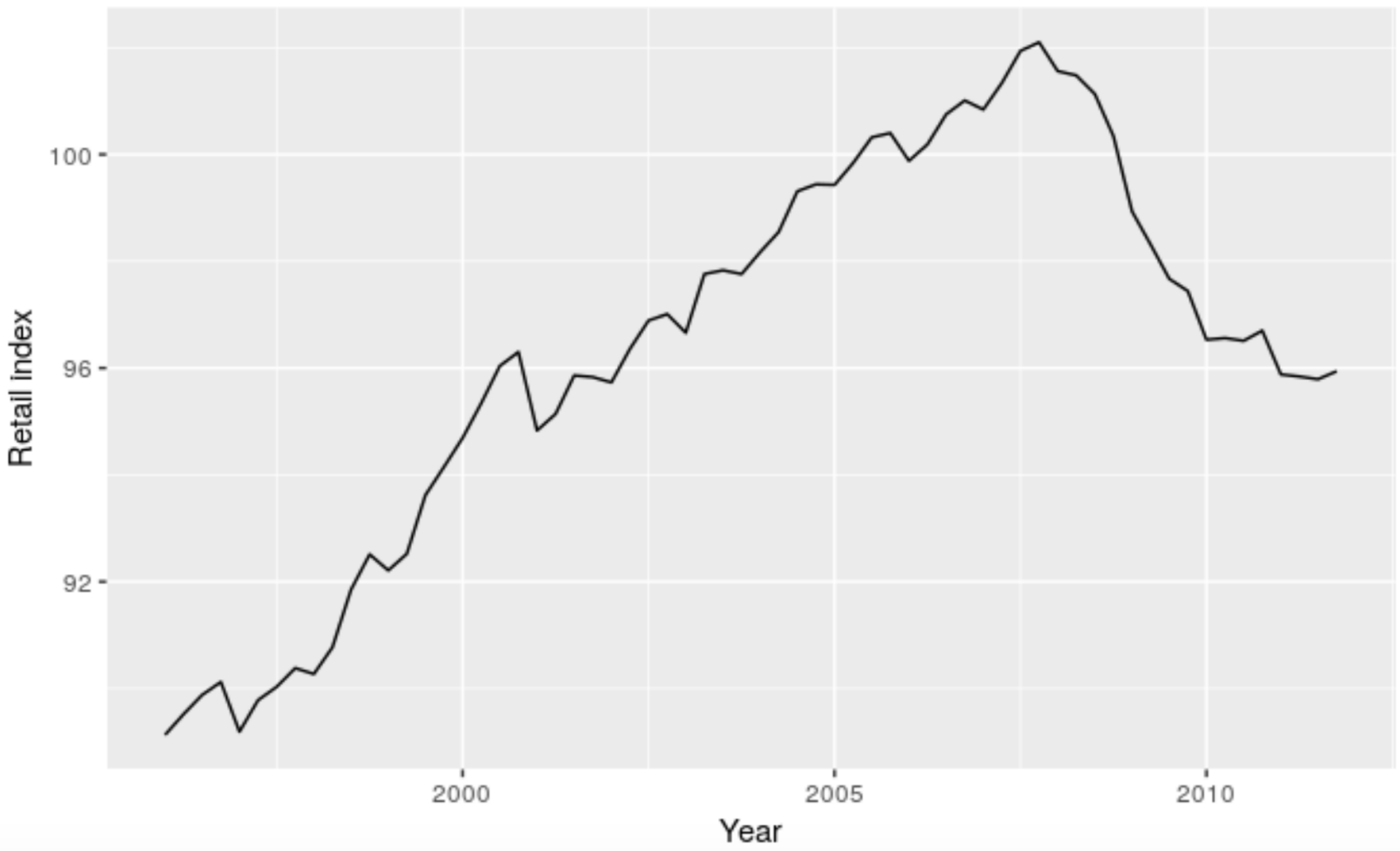


Seasonal ARIMA(p,d,q)(P,D,Q)$_m$

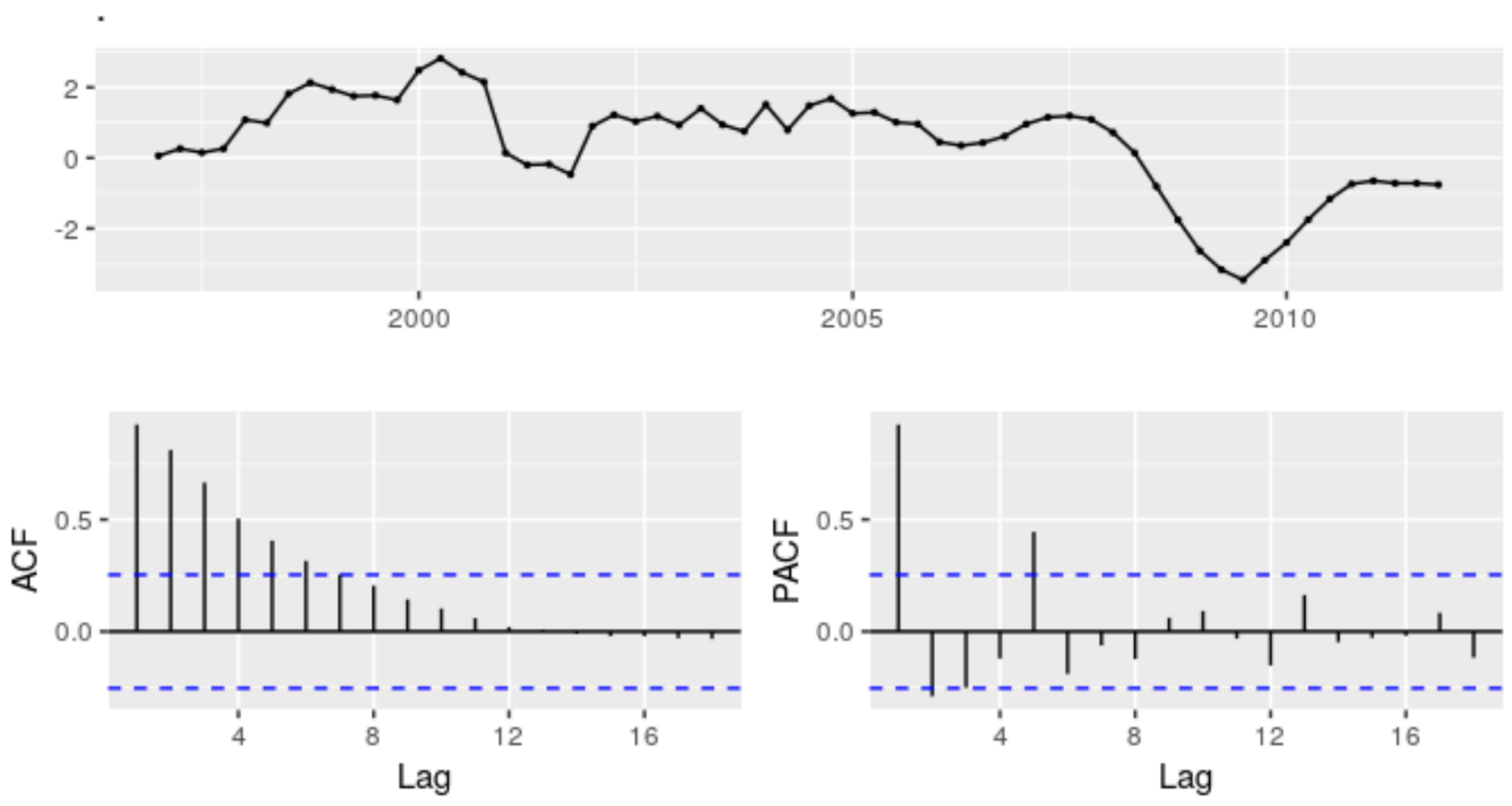
- (p,d,q) - nonseasonal part
- (P,D,Q) - seasonal part
- m is the season (quarter = 4, montly = 12, weekly = 52, etc...)

Example: European quarterly retail trade

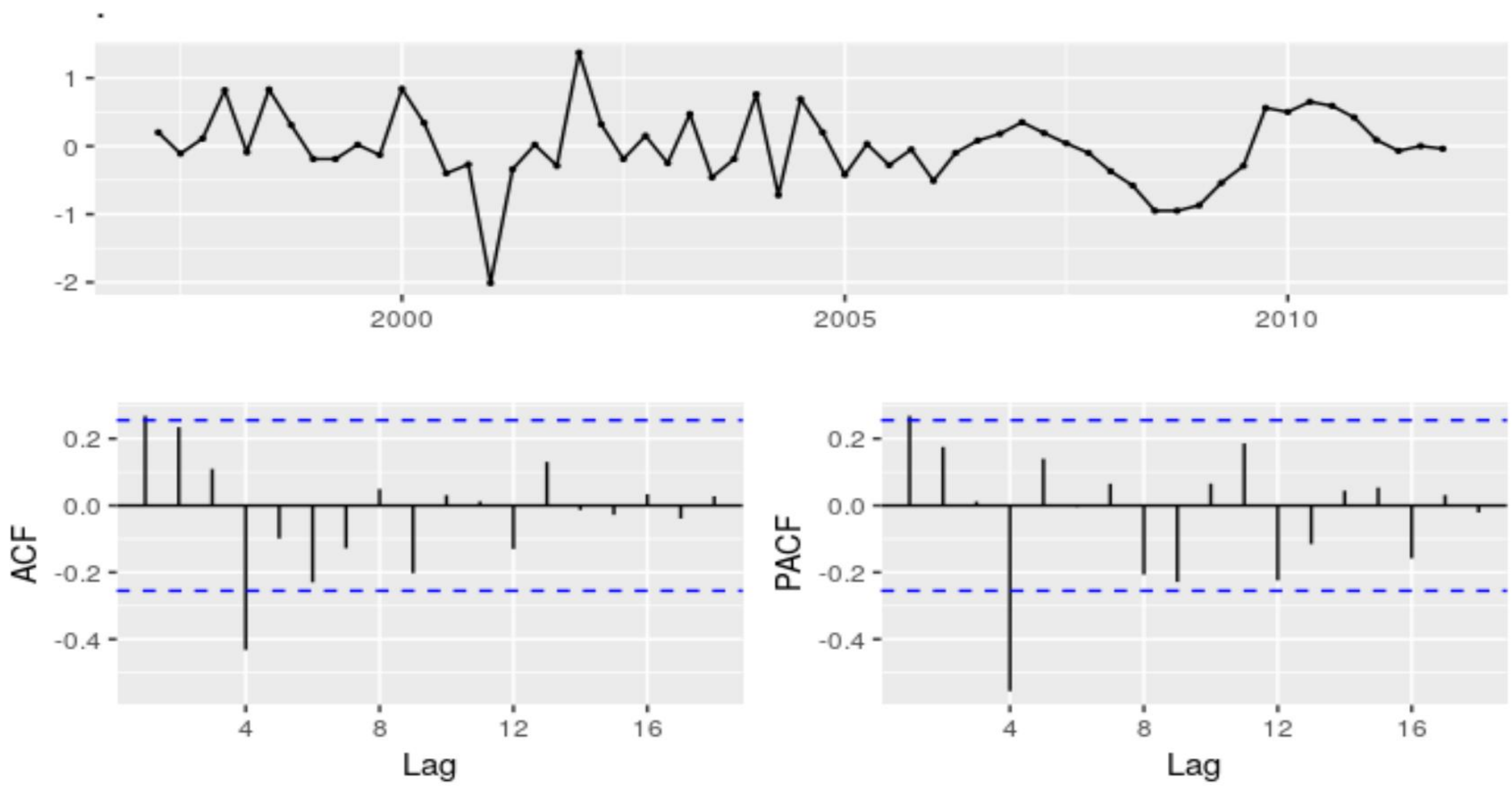
```r
autoplot(euretail) + ylab("Retail index") + xlab("Year")
```



```r
euretail %>% diff(lag=4) %>% ggtsdisplay()
```

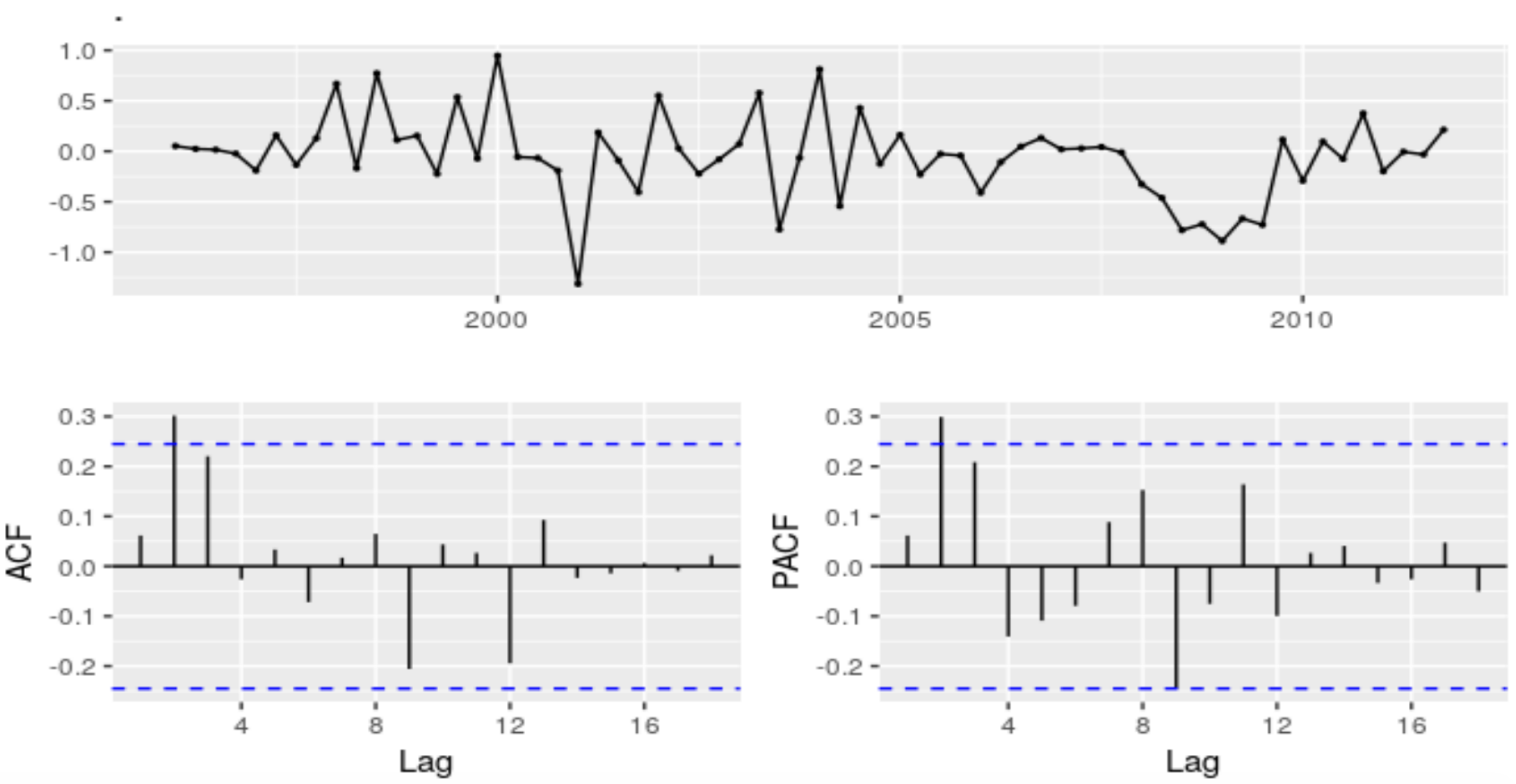


```
euretail %>% diff(lag=4) %>% diff() %>% ggtsdisplay()
```



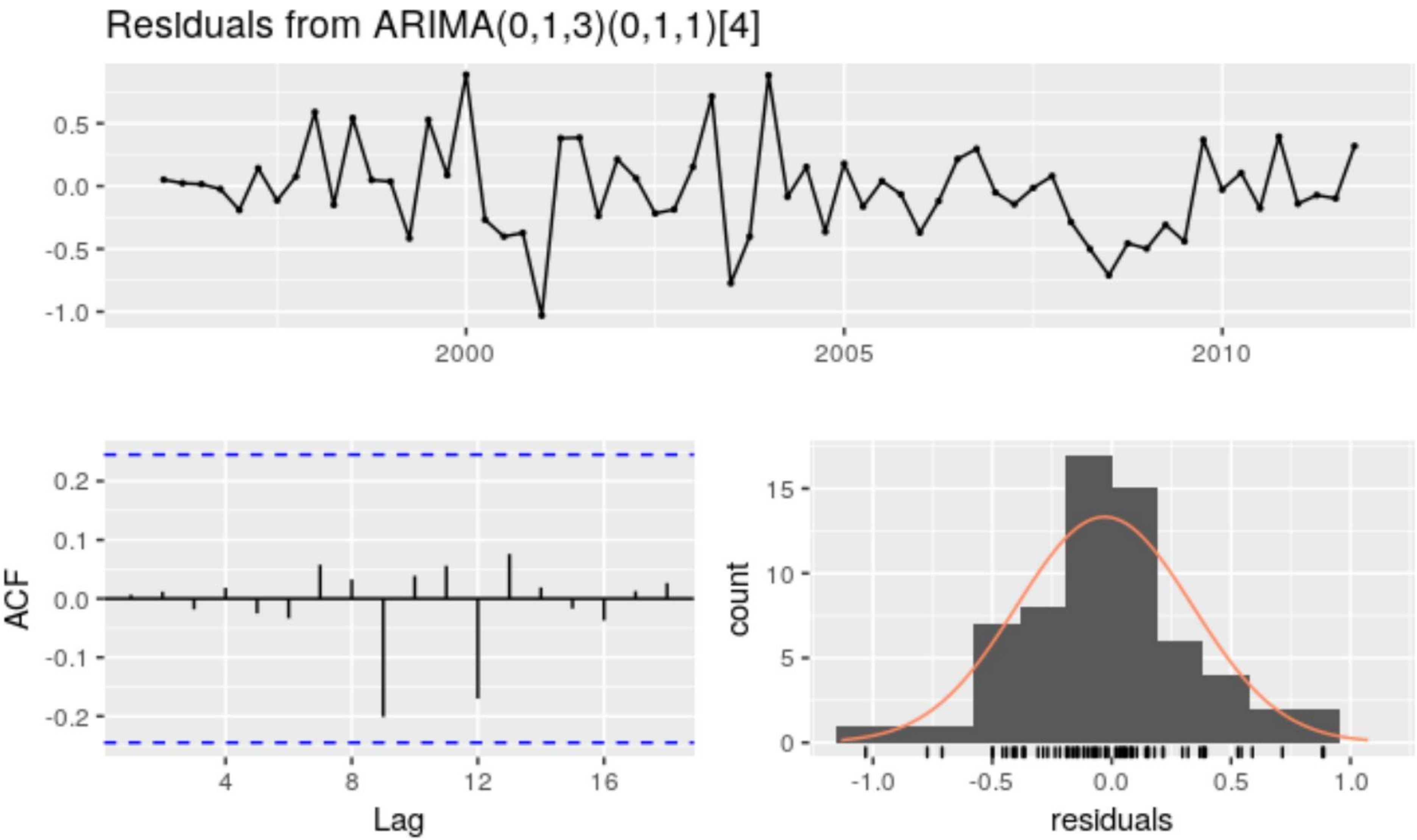
ARIMA(0,1,1)(0,1,1)$_4$ or ARIMA(1,1,0)(1,1,)$_4$

```r
euretail %>%
  Arima(order=c(0,1,1), seasonal=c(0,1,1)) %>%
  residuals() %>% ggtsdisplay()
```



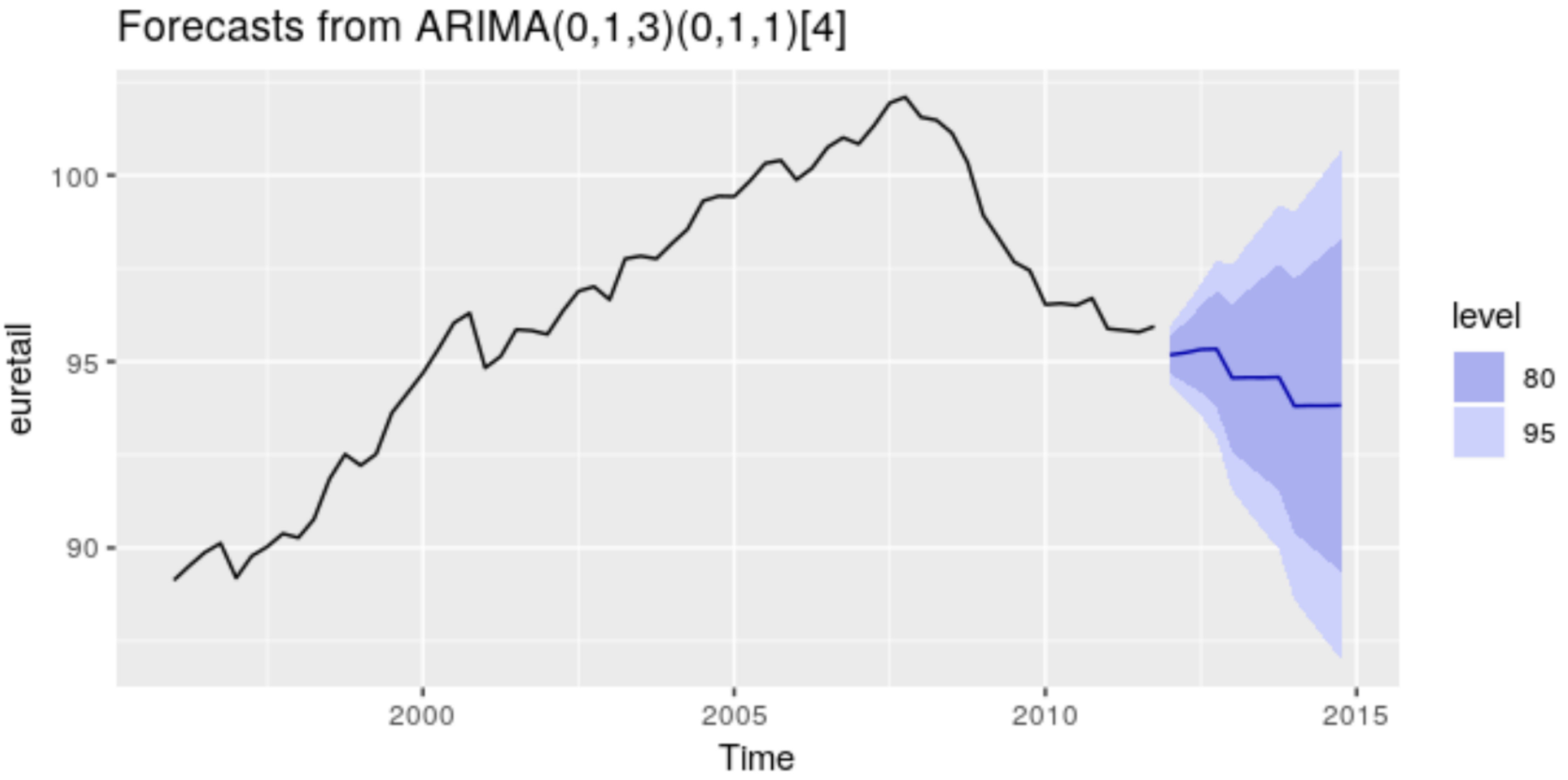
Try ARIMA(0,1,2)(0,1,1)$_4$

```r
fit3 <- Arima(euretail, order=c(0,1,3), seasonal=c(0,1,1))
checkresiduals(fit3)
```



```
#>
#>  Ljung-Box test
#>
#> data:  Residuals from ARIMA(0,1,3)(0,1,1)[4]
#> Q* = 0.51, df = 4, p-value = 1
#>
#> Model df: 4.   Total lags used: 8
```

```r
fit3 %>% forecast(h=12) %>% autoplot()
```



```r
auto.arima(euretail, stepwise=FALSE, approximation=FALSE)
#> Series: euretail
#> ARIMA(0,1,3)(0,1,1)[4]
#>
#> Coefficients:
#>          ma1    ma2    ma3    sma1
#>        0.263  0.369  0.420  -0.664
#> s.e.   0.124  0.126  0.129   0.155
#>
#> sigma^2 estimated as 0.156:  log likelihood=-28.63
#> AIC=67.26   AICc=68.39   BIC=77.65
```