Exercise 10

Philipp Drebes

04.05.2023

Exercise 10.1

In this exercise we again have a look at the sunspot data (Series 4). We figured out that AR(10) is a suitable model to describe the log transformed data.

```
library(fpp, quietly = T)
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

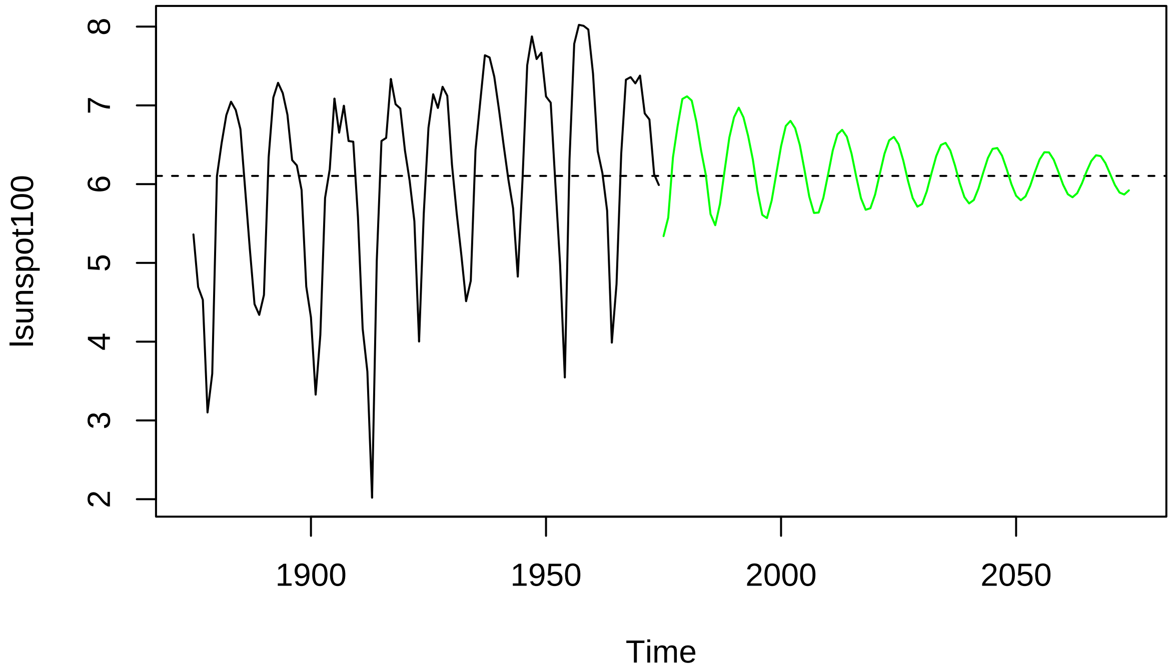
```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
lsunspot100 <- window(log(sunspotarea), start = 1875, end = 1974)
fit.ar10 <- arima(lsunspot100, order = c(10, 0, 0))
```

a) For the AR(10) model, predict the next 100 observations of the log-transformed time series and plot them together with the log-transformed time series. Also add a line for the estimated global mean to the plot. What do you observe?

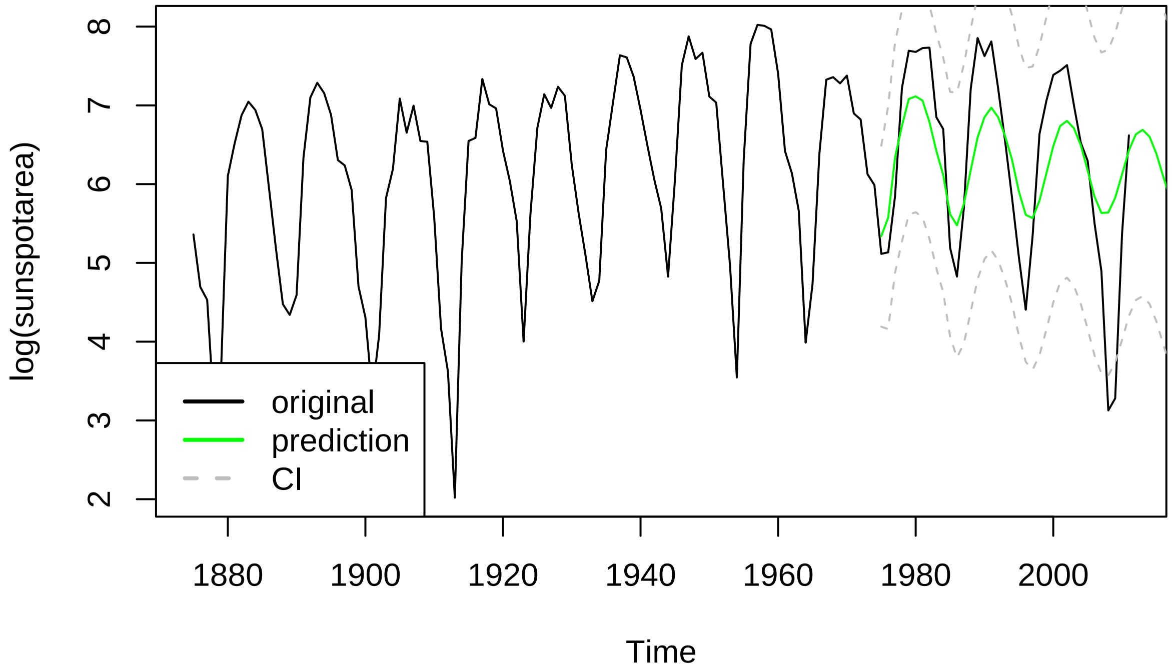
```
pred <- predict(fit.ar10, n.ahead = 100)
plot(lsunspot100, xlim = c(1875, 2074))
abline(h = fit.ar10$coef["intercept"], lty = 2)
lines(pred$pred, col = "green")
```



We observe a regression to the mean for the predicted time series.

b) Perform an out-of-sample evaluation, i.e. compare your prediction with the last 37 observed values of the time series. Plot the full log-transformed time series (1875 - 2011) and add your prediction (1975 - 2011) as well as prediction intervals to the plot and comment on the plot. Also compute the mean squared forecasting error of your prediction.

```
plot(log(sunspotarea))
lines(pred$pred, col = "green")
lines(pred$pred + 1.96 * pred$se, col = 'gray', lty = 2)
lines(pred$pred - 1.96 * pred$se, col = 'gray', lty = 2)
legend(x = 'bottomleft',
       legend = c('original', 'prediction', 'CI'),
       col = c('black', 'green', 'gray'),
       lty = c(1, 1, 2) , lwd = 2, bg = 'white')
```



Again, we see the regression to the mean for our predicted values and how those differ from the actual data. It could be argued that the first one or two oscillations around 1980 - 1990 can be used to predict values. However, after 1990 the model cannot be used to predict the actual values as the difference to the actual recorded data is to large. At around the year 2008, we see that the confidence intervals of our model do not contain the recorded values, whereas before 2008, all values lie inside the confidence interval. One could argue that this could be an outlier, but considering the overall poor performance of this model after 1990, I tend to argue in favor of a model error.

Mean squared forecasting error (MSE)

```
actual <- window(log(sunspotarea), start = 1974, end = 2011)
mean((actual - pred$pred)^2)
```

```
## [1] 0.7557876
```

This value alone does not contain a lot of information. However, we could use it to compare this model to another one.

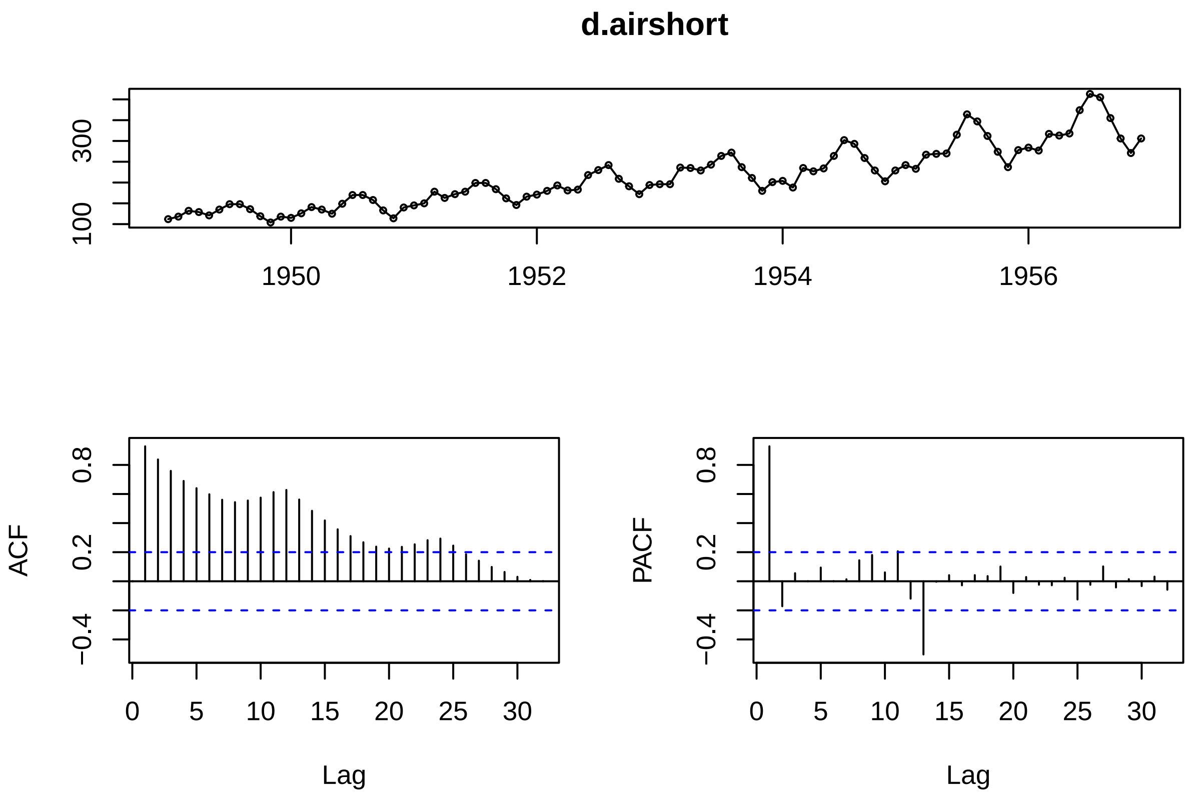
Exercise 10.2

We want to compare methods for forecasting the airplane data: the forecasting using a SARIMA model and the forecasting using an STL-decomposition. For being able to compare the prediction methods with the real data, first read in the `airplaine` data and use only the observations from 1949 to 1956.

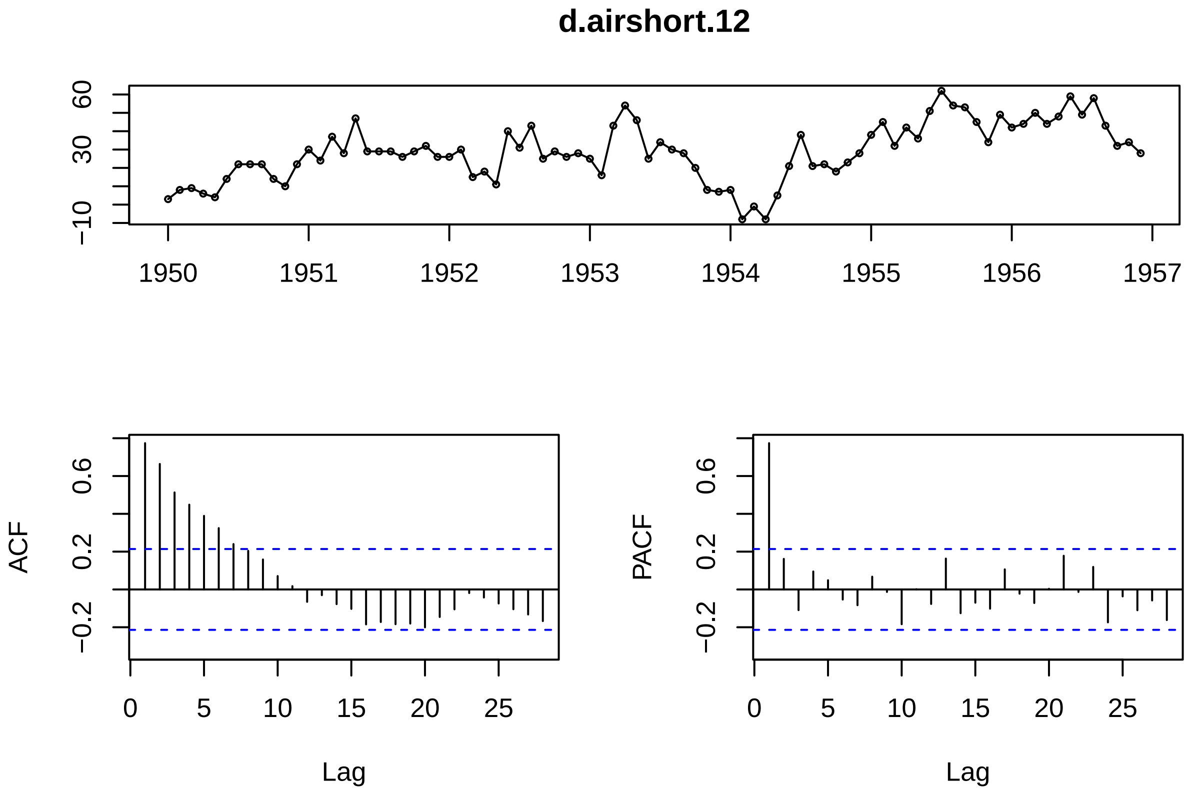
```
d.air <- AirPassengers
d.airshort <- window(d.air, end = c(1956, 12))
```

a) Fit an ARIMA/SARIMA Model for the shorter dataset `d.airshort`. Use transformations if suitable. Compute a prediction for the years 1957-1960 and plot it along with the prediction interval and the actual observations for this period.

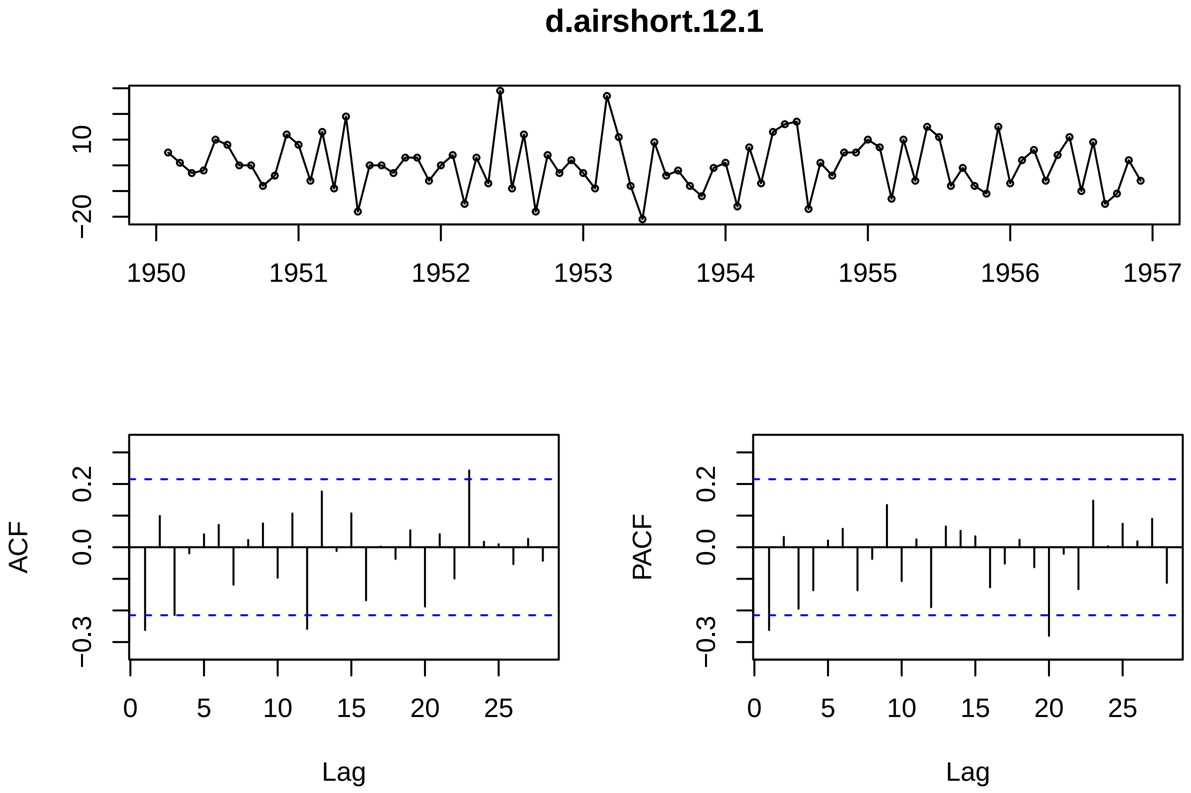
```
tsdisplay(d.airshort)
```



```
# try to get rid of the seasonal effect
# monthly data -> differencing at lag 12
d.airshort.12 <- diff(d.airshort, lag = 12)
tsdisplay(d.airshort.12)
```



```
# there seems to be a trend left
d.airshort.12.1 <- diff(d.airshort.12, lag = 1)
tsdisplay(d.airshort.12.1)
```



We use a SARIMA$(p, d, q)(P, D, Q)^s$-model, where $s = 12$, $D = 1$ and $d = 1$.

P and Q can be determined by looking at the lags which are multiples of 12. Here we choose the following possibilities: $(P, Q) = (0, 1)$ (cutoff in ACF at lag 1, PACF no significant values).

For deciding on the orders p and q, we consider the properties of the ACF and PACF for small lags. Here we use $(p, q) = (1, 1)$.

```
fit <- arima(d.airshort, order = c(1, 1, 1), seasonal = c(0, 1, 1))
fit
```

```
##
## Call:
## arima(x = d.airshort, order = c(1, 1, 1), seasonal = c(0, 1, 1))
##
## Coefficients:
##           ar1     ma1     sma1
##       -0.5902  0.4009  -0.1979
## s.e.   0.3526  0.4023   0.1017
##
## sigma^2 estimated as 90.52:  log likelihood = -305.02,  aic = 618.05
```

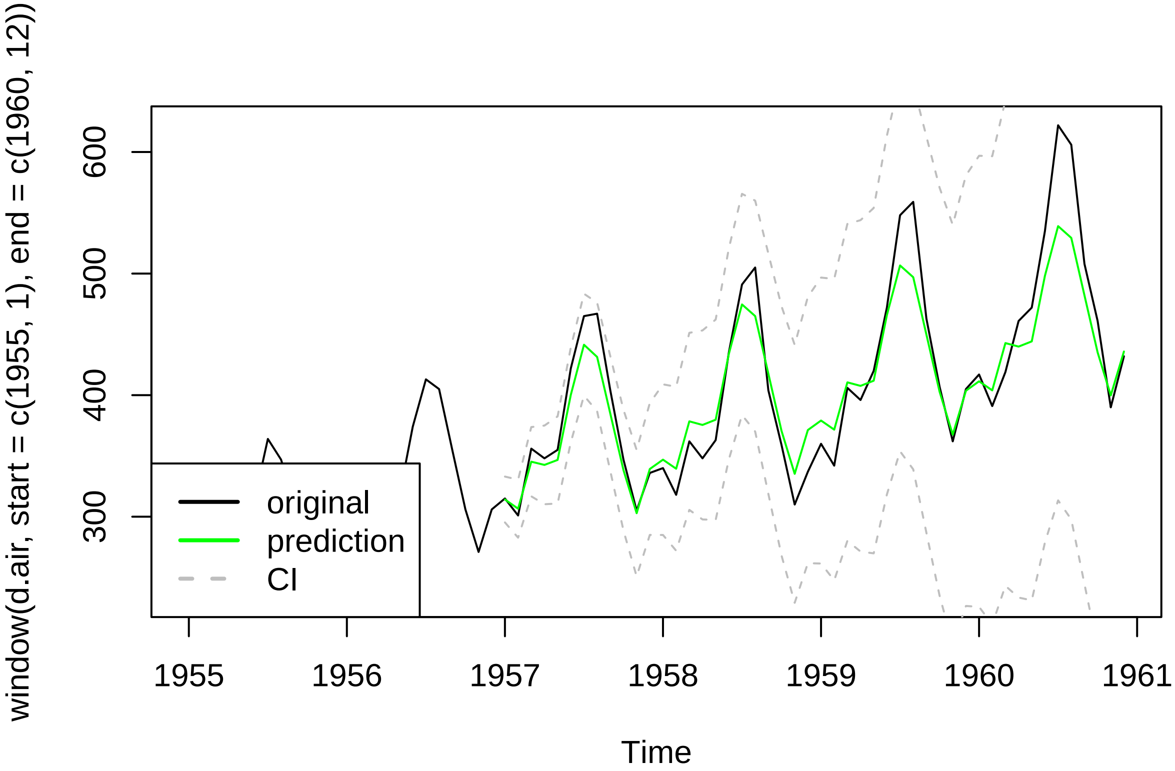
```
fit.auto <- auto.arima(d.airshort, ic = "aic")
fit.auto
```

```
## Series: d.airshort
## ARIMA(1,1,0)(1,1,0)[12]
##
## Coefficients:
```

```
##           ar1     sar1
##       -0.2250  -0.2274
## s.e.   0.1076   0.1081
##
## sigma^2 = 92.5:  log likelihood = -304.98
## AIC=615.97   AICc=616.27   BIC=623.22
```

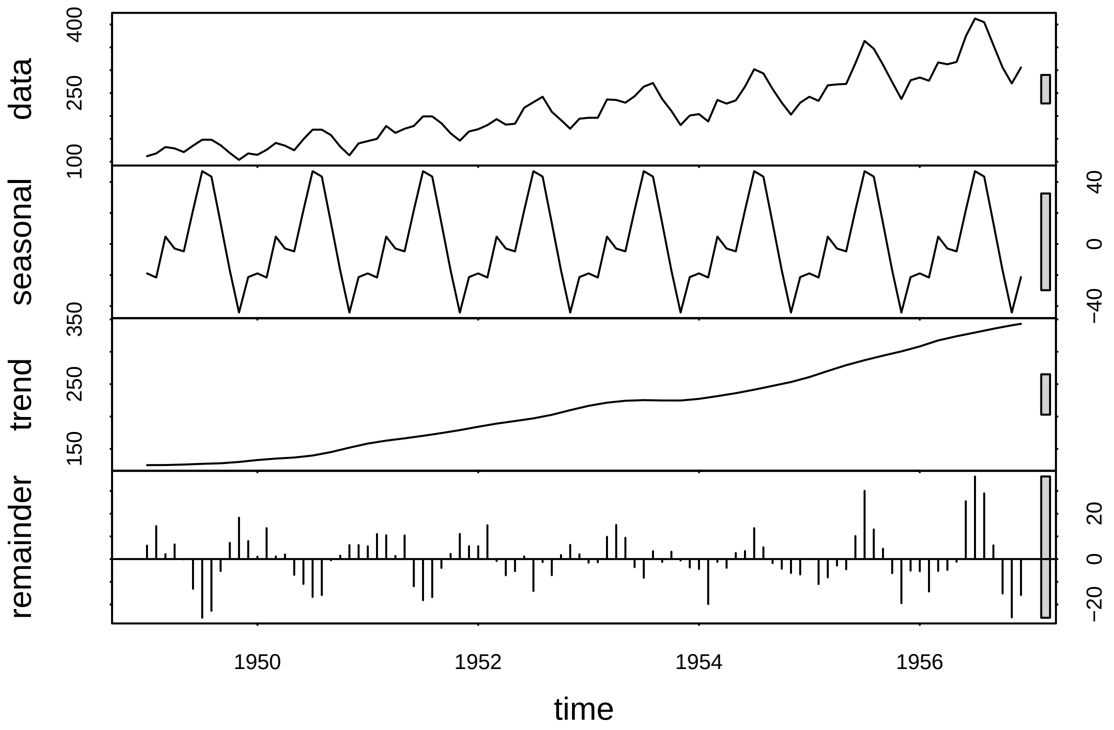
auto.arima() has produced a model with a slightly better AIC. Going forward we will use the model found with the use of auto.arima().

```
air.pred <- predict(fit.auto, n.ahead = 4 * 12)

plot(window(d.air, start = c(1955, 1), end = c(1960, 12)))
lines(air.pred$pred, col = "green")
lines(air.pred$pred + 1.96 * air.pred$se, col = 'gray', lty = 2)
lines(air.pred$pred - 1.96 * air.pred$se, col = 'gray', lty = 2)
legend(x = 'bottomleft',
       legend = c('original', 'prediction', 'CI'),
       col = c('black', 'green', 'gray'),
       lty = c(1, 1, 2) , lwd = 2, bg = 'white')
```



b) Now do prediction for the years 1957-1960 as seen in the lecture with linear extraploation of the trend estimate, continuation of the seasonal effect and and ARMA(p,q) forecast for the stationary remainder. Plot the predicted timeseries (this time without prediction interval, why?) and compare them to the actual observations.

```
decomp <- stl(d.airshort, s.window = "periodic")
plot(decomp)
```

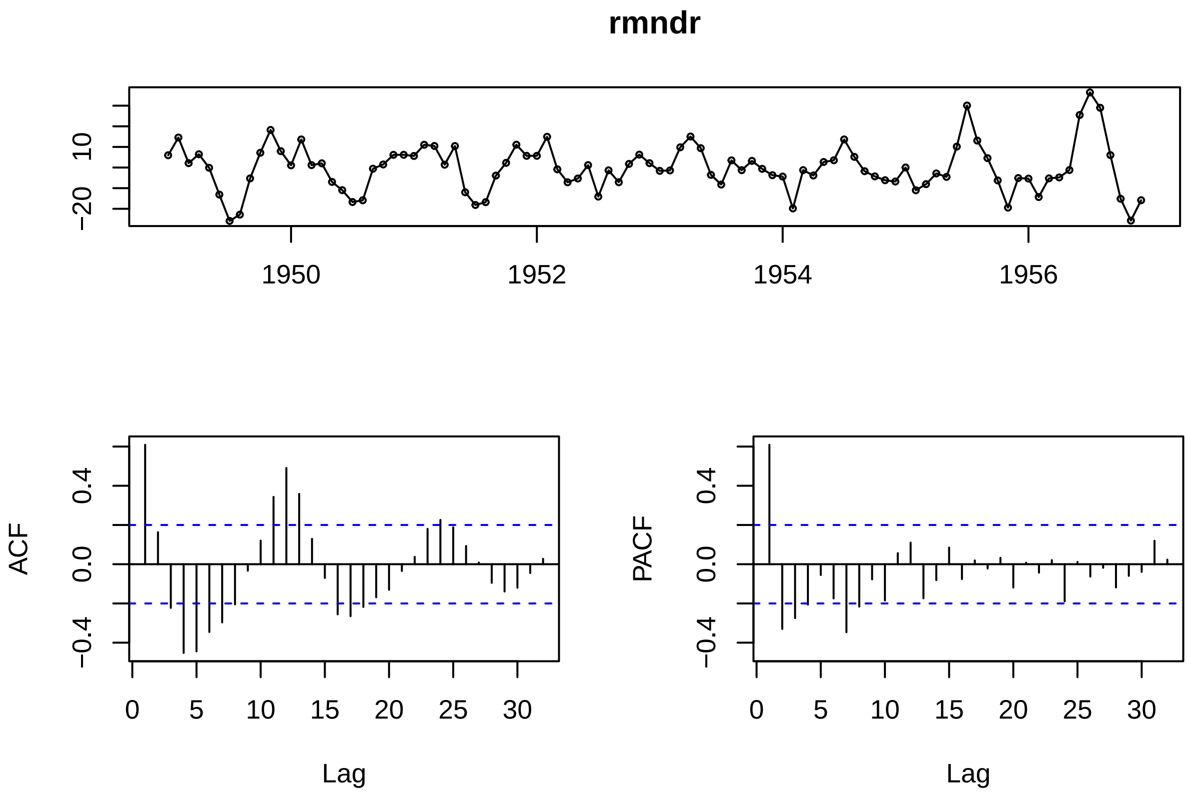


```
# Trend forecast by linear extrapolation
trend <- decomp$time.series[, 2]
xx <- time(trend)
yy <- trend
fit.regr <- lm(yy~xx)
summary(fit.regr)
```

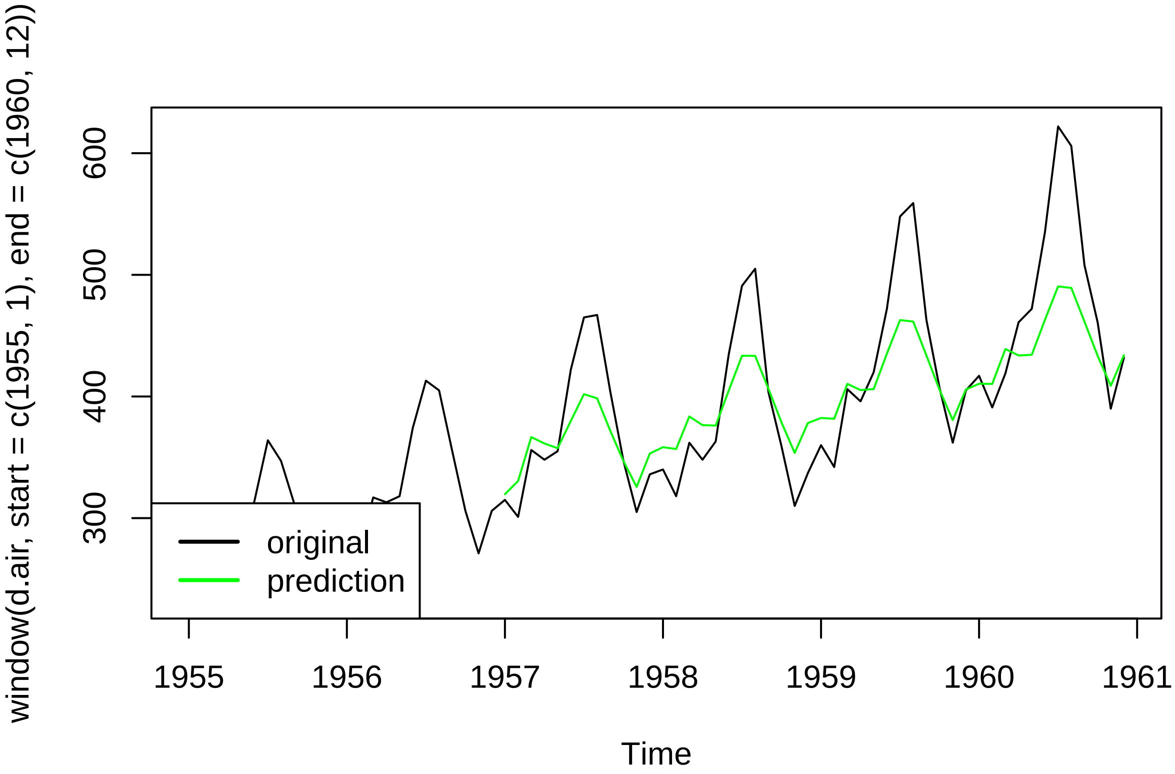
```
##
## Call:
## lm(formula = yy ~ xx)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.349  -4.005  -1.830   4.949  21.991
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.451e+04  8.470e+02  -64.36   <2e-16 ***
## xx           2.802e+01  4.337e-01   64.61   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.813 on 94 degrees of freedom
## Multiple R-squared:  0.978,  Adjusted R-squared:  0.9777
## F-statistic:  4174 on 1 and 94 DF,  p-value: < 2.2e-16
```

```
t.fore <- sort(trend, decreasing = TRUE)[1] + ((1:48)/12) * coef(fit.regr)[2]
```

```
# Seasonal forecast using last values
season <- decomp$time.series[,1]
l2y <- window(season,start=c(1952, 1), end=c(1956,12))
s.fore <- ts(l2y, start=c(1957,1), end=c(1960,12), freq=12)

# Fitting the remainder
rmndr <- decomp$time.series[, 3]
tsdisplay(rmndr)
```



```
fit.rmndr <- arima(rmndr, order=c(4,0,0), include.mean=F)
r.fore <- predict(fit.rmndr, n.ahead=48)$pred

# Adding the 3 components
fore <- t.fore + s.fore + r.fore

plot(window(d.air, start = c(1955, 1), end = c(1960, 12)))
lines(fore, col = "green")
legend(x = 'bottomleft',
       legend = c('original', 'prediction'),
       col = c('black', 'green'),
       lty = c(1, 1) , lwd = 2, bg = 'white')
```



c) Compare the different forecasts. Which of the methods seems to work best for the airplane data and why?

```
d.air.fore <- window(d.air, start = c(1957, 1), end = c(1960, 12))

# SARIMA
mean((d.air.fore - air.pred$pred)^2)
```

```
## [1] 700.7479
```

```
# STL
mean((d.air.fore - fore)^2)
```

```
## [1] 1959.268
```

What we can see visually is also apparent when we compare the MSE of the two methods. The SARIMA model has a much smaller MSE value compared to the STL method. Therefore, in this case, we would prefer this model.