W271-2 – Spring 2016 – **HW 8**

Juanjo Carin, Kevin Davis, Ashley Levato, Minghu Song

April 6, 2016

Build an univariate linear time series model (i.e AR, MA, and ARMA models) using the series in `hw08_series.csv`.

- **Use all the techniques that have been taught so far to build the model, including date examination, data visualization, etc.**
- **All the steps to support your final model need to be shown clearly.**
- **Show that the assumptions underlying the model are valid.**
- **Which model seems most reasonable in terms of satisfying the model underlying Assumptions?**
- **Evaluate the model performance (both in- and out-of-sample)**
- **Pick your "best" models and conduct a 12-step ahead forecast. Discuss your results. Discuss the choice of your metrics to measure "best".**

First we load the series:

```
hw08 <- read.csv('hw08_series.csv', header = TRUE)
str(hw08)
```

```
## 'data.frame':    372 obs. of  2 variables:
##  $ X: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ x: num  40.6 41.1 40.5 40.1 40.4 41.2 39.3 41.6 42.3 43.2 ...
```

```
all(hw08$X == 1:dim(hw08)[1]) # check if 1st column is just an incremental index
```

```
## [1] TRUE
```

```
hw08 <- hw08[, -1]
```

The file has two columns but the first one is just an incremental index so we discard it. The second column (that is stored in a numeric vector called `hw08`) contains 372 observations. 372 is a multiple of 12 ($372/12 = 31$) so **we'll assume that the series contains monthly observations from 31 years (*for labelling purposes only, sometimes we'll also assume that the period goes from 1980 to 2010*)**.

Let's explore the main descriptive statistics of the series, as well as its histogram and time-series plot:

```
# See the definition of the function in ## @knitr Libraries-Functions-Constants
desc_stat(hw08, 'Time series', 'Descriptive statistics of the time series.')
```

Table 1: Descriptive statistics of the time series.

	Time series
Mean	84.83
St. Dev	31.95
1st Quartile	57.38
Median	76.45
3rd Quartile	111.53
Min	36.00
Max	152.60

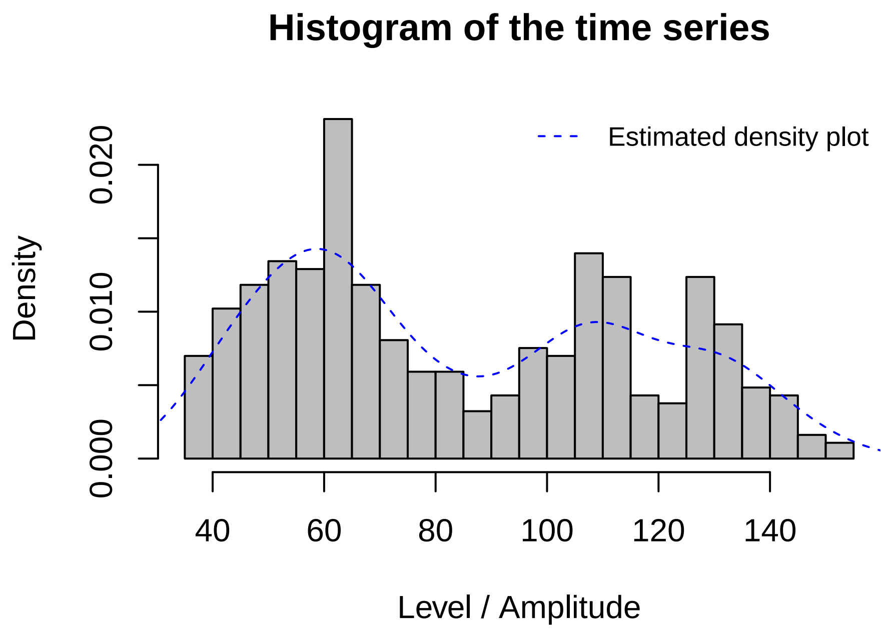


Figure 1: Histogram of the data.

The histogram shows that the distribution of the data is multimodal, and hence far from normal. But as usual, it tells us nothing about the dynamics of the time series: only what values were more or less frequent, but not when they happened.

To label the time-series plot, we will assume (as mentioned) that the data were collected on a monthly basis and will use 1980 as an arbitrary starting point.

```
hw08.ts <- ts(hw08, start = c(1980,1), frequency = 12)
```

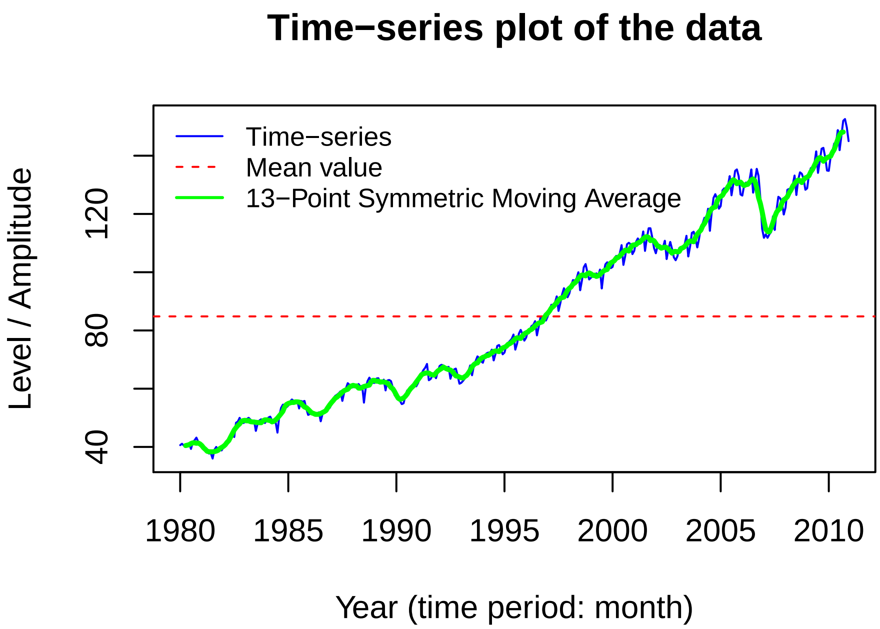


Figure 2: Time-series plot (assuming monthly data, from 1980 until 2010).

Our assumption that the data corresponds to a monthly time series seems reasonable after noticing that there seems to be some seasonality every 12 time periods (see Figure 3 below: the level increases over the first 6 months—especially in February and June—, goes down in July, up from August to October, and down again the last 2 months of the year).

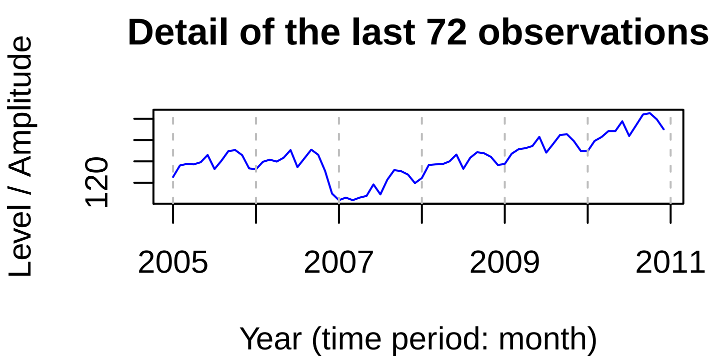


Figure 3: Time-series plot of (the last 72 observations—6 years?—of) the data.

In addition to showing that the time series is **not (mean) stationary** (the mean depends on time, with an increasing trend, and the time series is very **persistent**), Figure 2 in the previous page shows that the time is also **not variance stationary:** the variance is not constant but changes with time (increasing in the last years, especially the last 7); see Table 2 and Figure 4 in the next page.

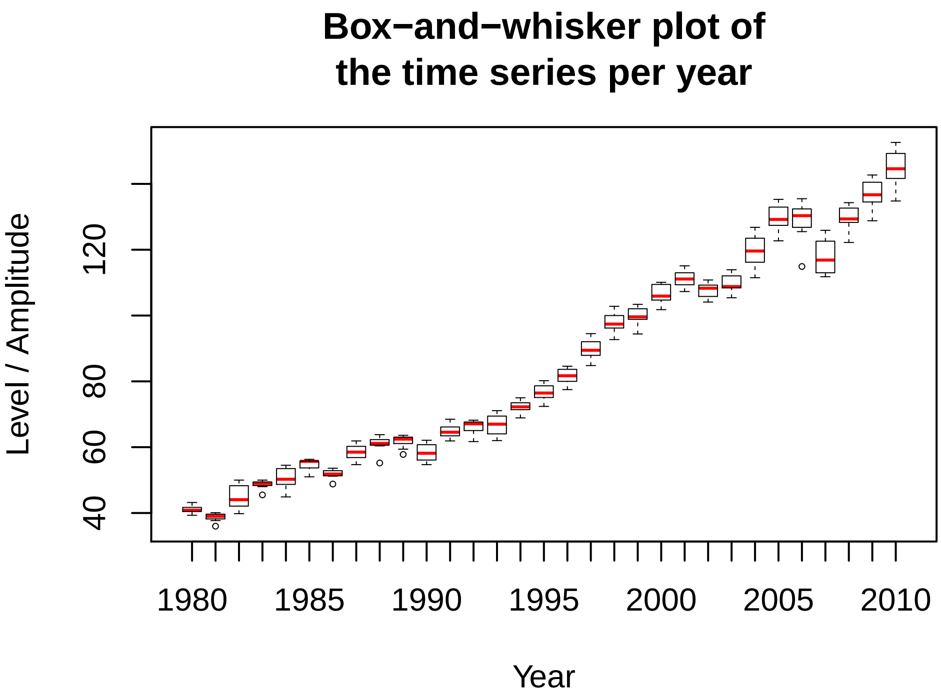


Figure 4: Boxplot of the series, per year (every 12 observations).

Table 2: Variance of the time-series amplitude per year (for the first 30 out of 31).

Year	Mean	Variance	Year	Mean	Variance	Year	Mean	Variance
1980	41.05	1.12	1990	58.29	6.61	2000	106.34	7.84
1981	38.79	1.36	1991	64.82	3.66	2001	111.15	7.23
1982	44.91	13.06	1992	66.25	4.21	2002	107.78	5.03
1983	48.70	1.37	1993	66.70	9.79	2003	109.58	7.79
1984	50.57	8.63	1994	72.24	3.31	2004	119.73	22.65
1985	54.84	2.59	1995	76.51	5.49	2005	129.77	13.84
1986	51.87	1.73	1996	81.67	5.59	2006	129.29	30.61
1987	58.49	4.94	1997	89.83	8.18	2007	117.78	29.39
1988	61.06	4.42	1998	97.77	9.03	2008	129.81	12.05
1989	61.86	3.17	1999	100.00	6.37	2009	137.07	16.68

Both results indicate that the data does not seem to be a realization of a stationary process, so this time-series may not meet the assumptions of an AR-MA process, and thus **an ARMA model may not be a good fit for our data** (maybe it is—as it happened with the USNZ series we analyzed in class—, but it will certainly not be good for forecasting). At the very least, we should transform the data to stabilize the variance, take first differences of the data until they're stationary, and so forth.

To continue the Exploratory Data Analysis, let's decompose the time series to check the growing (though not exactly linear) trend and seasonality:

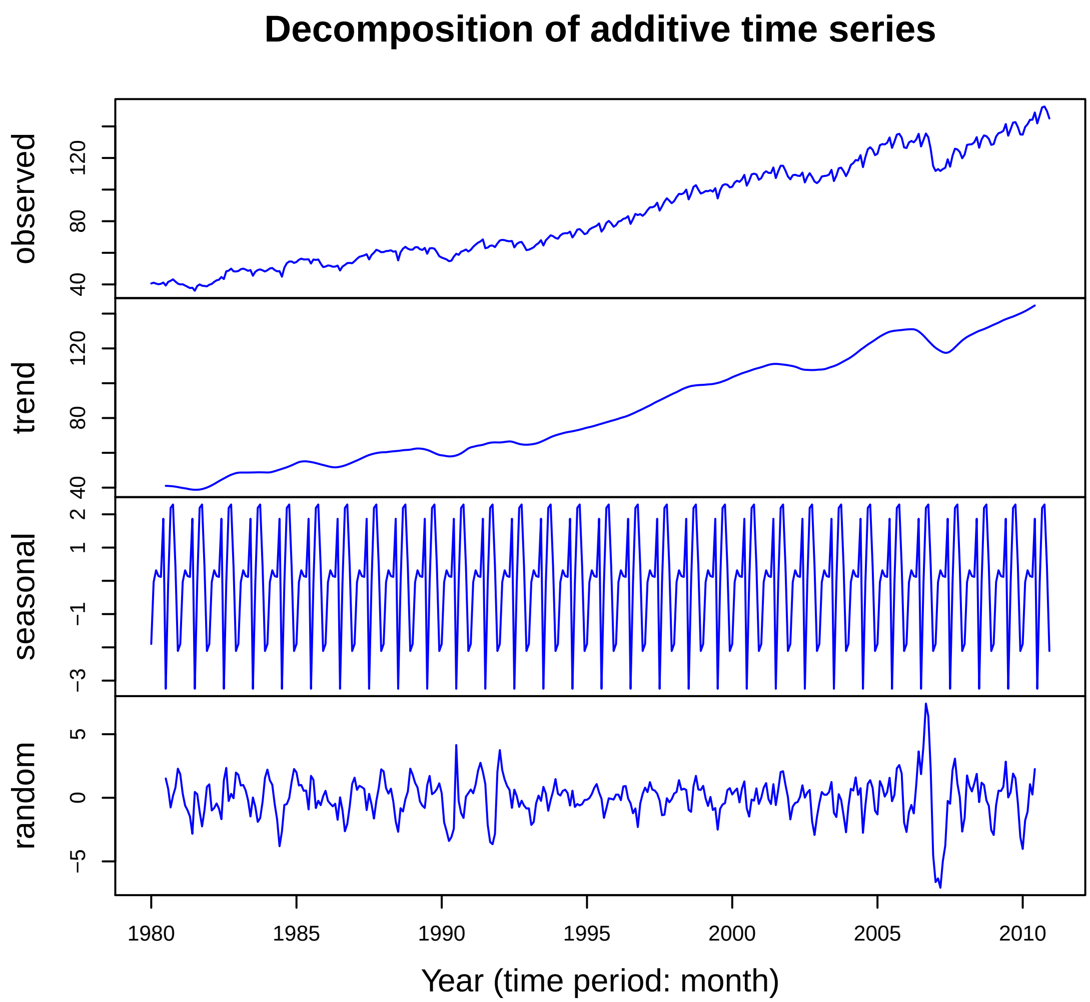


Figure 5: (Additive) decomposition of the time series.

The correlogram (where 2 years—or 24 1-month time displacements—are plotted) also shows how persistent the series is, looking very much like that of a random walk with drift. That is also an indication that an MA model may not be a good fit for this time series. The PACF drops off very sharply after the 1st lag, though the PACF of the 12th lag is also significant (probably due to the seasonal component).

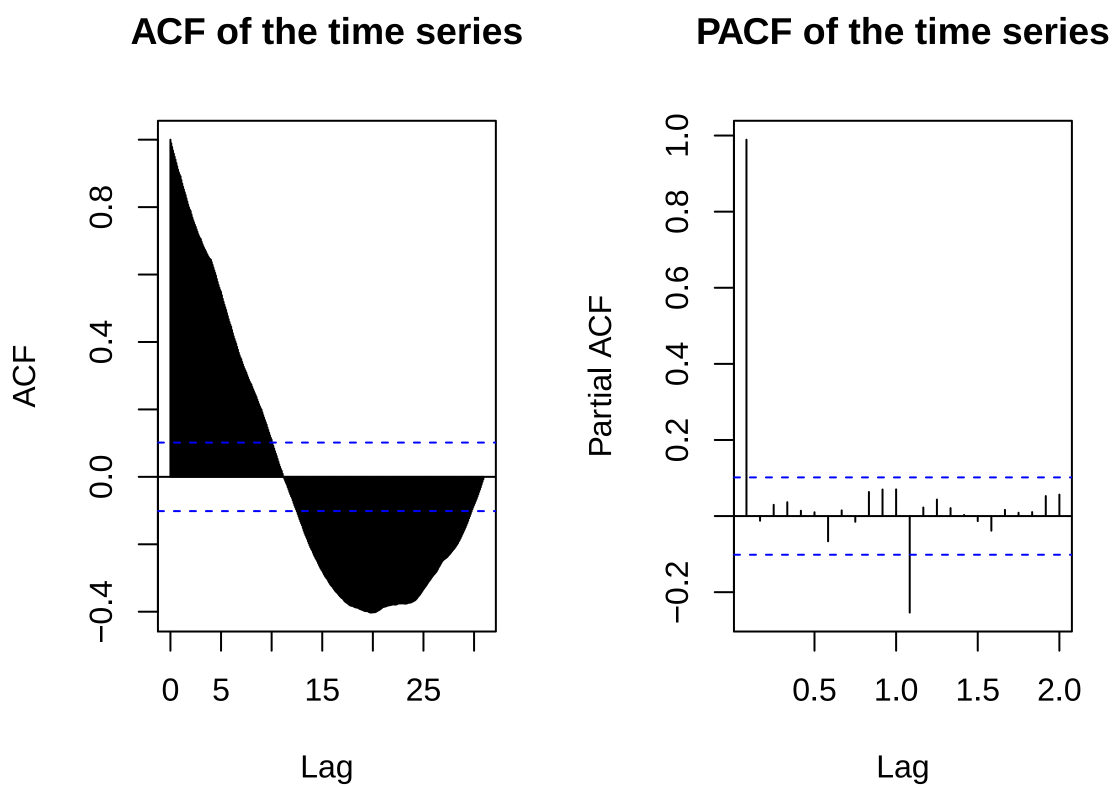


Figure 6: Autocorrelation and partial autocorrelation graphs

We start by trying different orders of AR models. To choose the 'best' AR model, we will try AR models up to degree 8 and use the model with the lowest AIC as the criterion for selecting the prefered model.

```
best.AR.order <- best.MA.order <- best.ARMA.order <- c(0, 0, 0)
best.AR.aic <- best.MA.aic <- best.ARMA.aic <- Inf
for (i in 0:8) {
  # Try the corresponding order, skip if estimating the model yields an error
  # due to non-stationarity
  # Not using try() would cause the loop to break
  try(fit.AR.aic <- AIC(arima(hw08.ts, order = c(i, 0, 0))), silent = TRUE)
  if (fit.AR.aic < best.AR.aic) {
    best.AR.order <- c(i, 0, 0)
    best.ar <- arima(hw08.ts, order = best.AR.order)
    best.AR.aic <- fit.AR.aic
  }
}
best.AR.aic
```

```
## [1] 1786.825
```

```
best.AR.order
```

```
## [1] 3 0 0
```

```
best.ar
```

```
##
## Call:
## arima(x = hw08.ts, order = best.AR.order)
##
## Coefficients:
##          ar1      ar2     ar3  intercept
##       0.9061  -0.0994  0.1922    91.5517
## s.e.  0.0511   0.0692  0.0511    45.2799
##
## sigma^2 estimated as 6.839:  log likelihood = -888.41,  aic = 1786.82
```

Table 3: Coefficients, SEs, and 95% CIs of the estimated AR(3) model

	Coefficient	SE	95% CI lower	95% CI upper
ar1	0.9061	0.0511	0.8038	1.0083
ar2	-0.0994	0.0692	-0.2379	0.0390
ar3	0.1922	0.0511	0.0900	0.2945
intercept	91.5517	45.2799	0.9918	182.1115

The best AR model (with up to 8 coefficients) is the AR(3) one. Note that the 2nd coefficient is not significant.

If we examine the residuals of the AR(3) model we observe that, though their distribution looks normal, they follow a trend and the variance seems to increase over time. Their ACF and PACF do not look like the ones of white noise (especially—but not only—because the ACF and PACF at lag 12).

```r
summary(best.ar$resid)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -11.7300  -0.7425   0.6078   0.3506   1.9630   7.5280
```

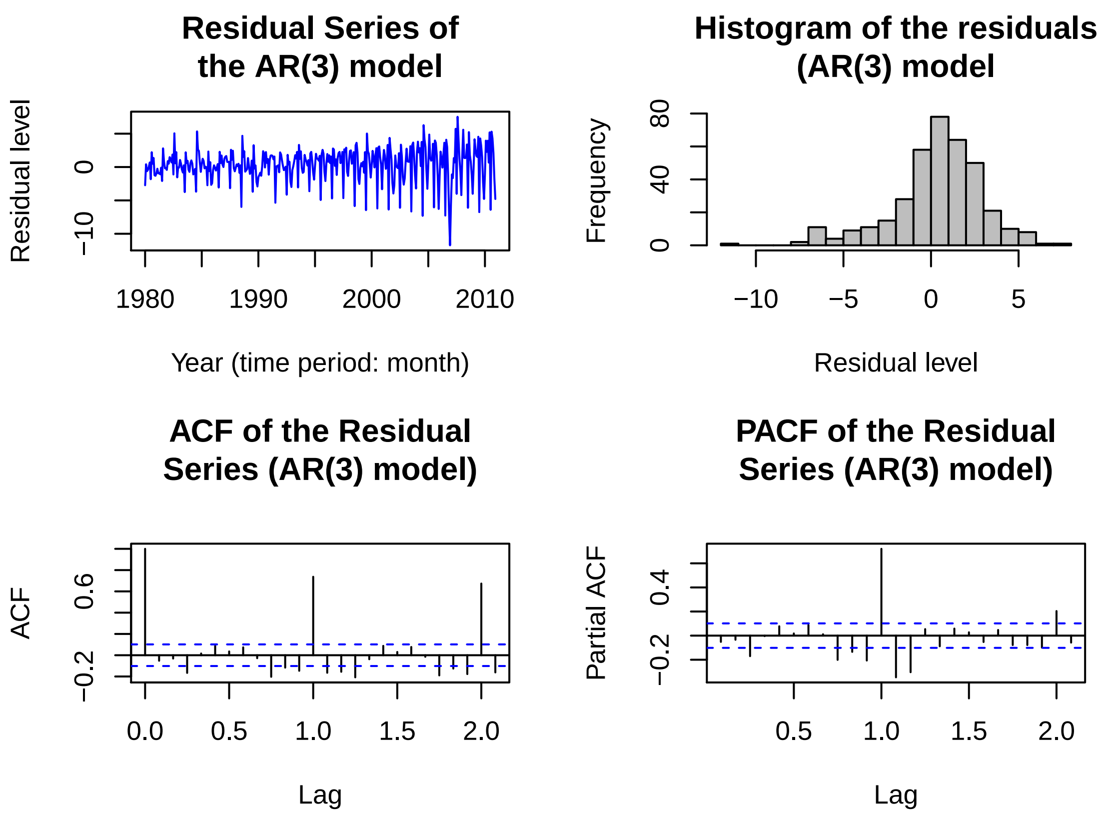


Figure 7: AR(3) model diagnostic based on the residuals

Nonetheless, we cannot reject the hypothesis of independence of the residual series:

```r
Box.test(best.ar$resid, type = "Ljung-Box")
```

```
##
##  Box-Ljung test
##
## data:  best.ar$resid
## X-squared = 1.0155, df = 1, p-value = 0.3136
```

The plot of the AR(3) fitted model versus the original series shows us that the model is well-suited to capturing the in-sample dynamics of this particular series.

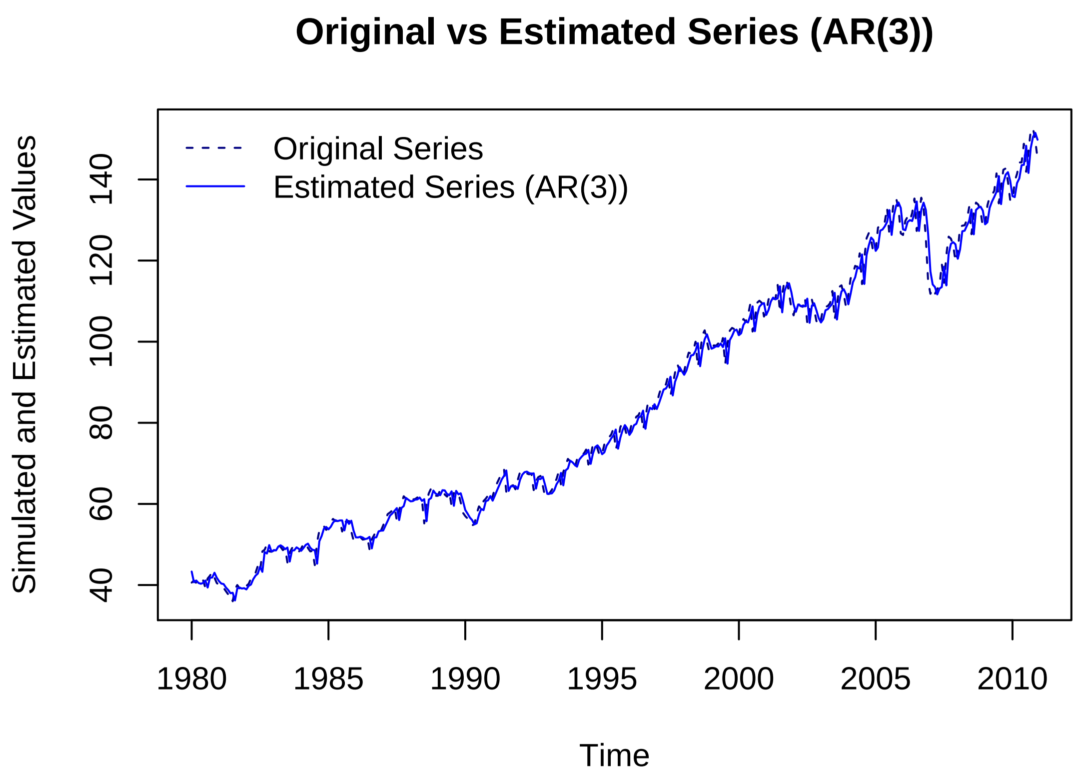


Figure 8: In-Sample Fit of AR(3) Model and Original Series

We follow by trying different orders of MA models, again up to the 8th order and choosing the best model based on its AIC value:

```
for (j in 1:8) {
  # No need to include j = 0 (MA(0) = AR(0) <=> xt = wt)
  # Try the corresponding order, skip if estimating the model yields an error
  # due to non-stationarity
  # Not using try() would cause the loop to break
  try(fit.MA.aic <- AIC(arima(hw08.ts, order = c(0, 0, j))), silent = TRUE)
  if (fit.MA.aic < best.MA.aic) {
    best.MA.order <- c(0, 0, j)
    best.ma <- arima(hw08.ts, order = best.MA.order)
    best.MA.aic <- fit.MA.aic
  }
}
best.MA.aic
```

```
## [1] 2020.487
```

```r
best.MA.order
```

```
## [1] 0 0 8
```

```r
best.ma
```

```
##
## Call:
## arima(x = hw08.ts, order = best.MA.order)
##
## Coefficients:
##          ma1     ma2     ma3     ma4     ma5     ma6     ma7     ma8
##       1.5239  1.6738  2.0420  2.3837  2.0652  1.6387  1.4398  0.8561
## s.e.  0.0338  0.0539  0.0507  0.0565  0.0632  0.0663  0.0496  0.0315
##       intercept
##         84.9092
## s.e.     2.6071
##
## sigma^2 estimated as 12.07:  log likelihood = -1000.24,  aic = 2020.49
```

Table 4: Coefficients, SEs, and 95% CIs of the estimated MA(4) model

	Coefficient	SE	95% CI lower	95% CI upper
ma1	1.5239	0.0338	1.4563	1.5915
ma2	1.6738	0.0539	1.5661	1.7816
ma3	2.0420	0.0507	1.9405	2.1434
ma4	2.3837	0.0565	2.2707	2.4966
ma5	2.0652	0.0632	1.9389	2.1915
ma6	1.6387	0.0663	1.5061	1.7714
ma7	1.4398	0.0496	1.3406	1.5390
ma8	0.8561	0.0315	0.7932	0.9190
intercept	84.9092	2.6071	79.6950	90.1235

The best MA model (with up to 8 coefficients) is the MA(8) one. All its coefficients are significantly different from zero. Note that the AIC value (even for the best model among the eight considered) is larger than the one of the AR(3) model, indicating that the latter is a better choice.

The residuals of this MA(8) model do not look normal like those of a white noise at all: the time plot shows a clear growing trend, the histogram is right-skewed, and many of the auto-correlations (apart from $k = 0$) and partial auto-correlations are significantly different from zero.

```r
summary(best.ma$resid)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -8.7350 -2.6370 -0.5714  0.0464  2.3400  9.2190
```

As expected, the results of a Llung-Box test is that we can reject the hypothesis of independence of the residual series:

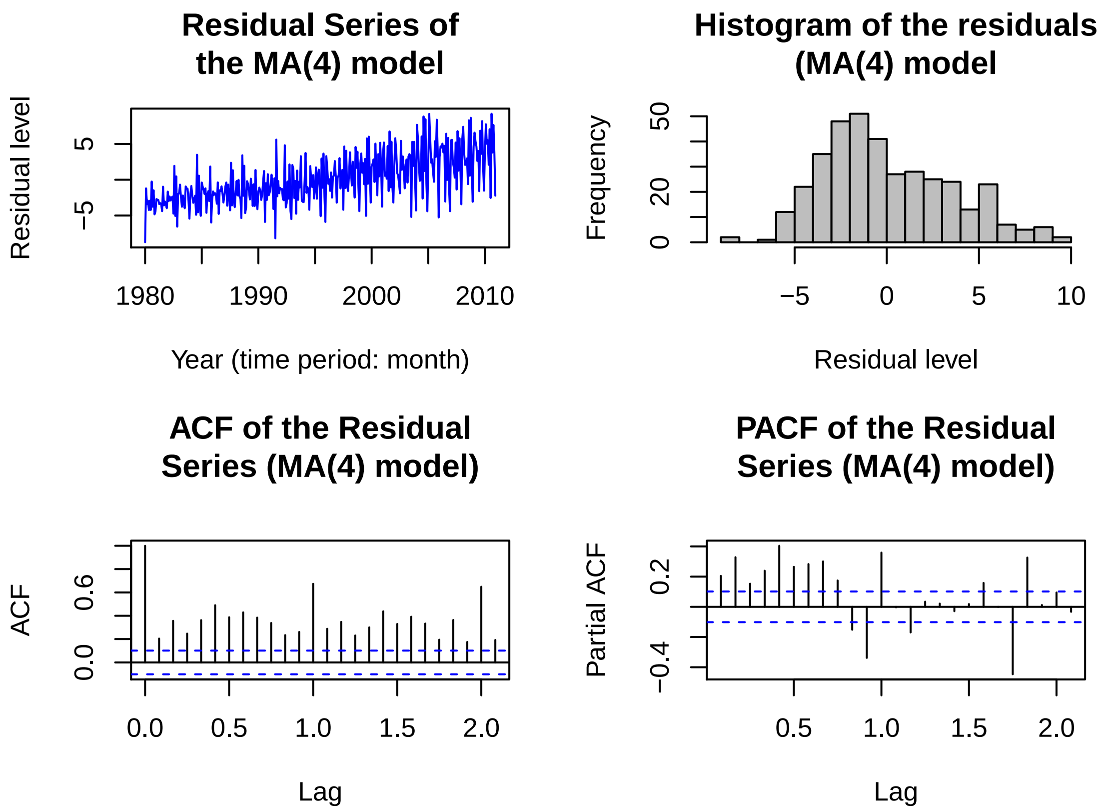


Figure 9: AR(3) model diagnostic based on the residuals

```
Box.test(best.ma$resid, type = "Ljung-Box")
```

```
##
##  Box-Ljung test
##
## data:  best.ma$resid
## X-squared = 15.702, df = 1, p-value = 7.415e-05
```

Similar to the AR(3) model, the plot of the MA(8) fitted model versus the original series shows us that the model is capable of capturing the dynamics of this time-series, although the fit does not look as close as the MA(3) model.

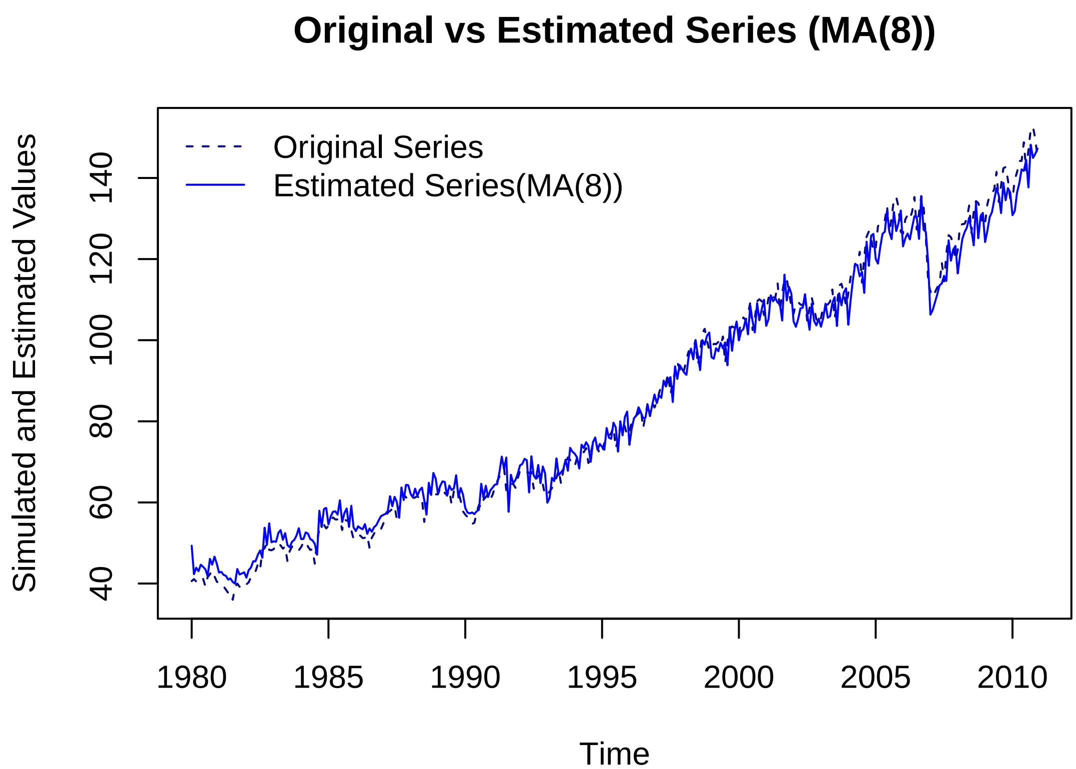


Figure 10: In-Sample Fit of MA(8) Model and Original Series

Finally, we try different ARMA models: we exclude the AR and MA models (i.e., those with $p = 0$ or $q = 0$, already considered), and (because of the nature of ARMA models) only consider orders up to (4,4).......

```
for (i in 1:4)
  for (j in 1:4) {
    # No need to include i = 0 and j = 0 (AR and MA models, already checked)
    # due to non-stationarity
    # Try the corresponding order, skip if estimating the model yields an error
    # Not using try() would cause the loop to break
    try(fit.ARMA.aic <- AIC(arima(hw08.ts, order = c(i, 0, j))), silent = TRUE)
    if (fit.ARMA.aic < best.ARMA.aic) {
      best.ARMA.order <- c(i, 0, j)
      best.arma <- arima(hw08.ts, order = best.ARMA.order)
      best.ARMA.aic <- fit.ARMA.aic
    }
  }
best.ARMA.aic
```

```
## [1] 1728.833
```

```
best.ARMA.order
```

```
## [1] 3 0 4
```

```
best.arma
```

```
##
## Call:
## arima(x = hw08.ts, order = best.ARMA.order)
##
## Coefficients:
##           ar1     ar2     ar3     ma1     ma2      ma3      ma4  intercept
##       -0.3826  0.6441  0.7368  1.4294  0.6315  -0.5088  -0.3335    87.7826
## s.e.   0.0537  0.0335  0.0401  0.0709  0.0860   0.0762   0.0591    48.6021
##
## sigma^2 estimated as 5.608:  log likelihood = -855.42,  aic = 1728.83
```

Table 5: Coefficients, SEs, and 95% CIs of the estimatedARMA(2,1) model

	Coefficient	SE	95% CI lower	95% CI upper
ar1	-0.3826	0.0537	-0.4901	-0.2752
ar2	0.6441	0.0335	0.5771	0.7111
ar3	0.7368	0.0401	0.6566	0.8170
ma1	1.4294	0.0709	1.2877	1.5712
ma2	0.6315	0.0860	0.4594	0.8036
ma3	-0.5088	0.0762	-0.6613	-0.3564
ma4	-0.3335	0.0591	-0.4518	-0.2153
intercept	87.7826	48.6021	-9.4216	184.9869

The best ARMA model (with up to 4 coefficients) is the ARMA(3,4) one. All its coefficients (not considering the mean) are significantly different from zero.

The residuals of this ARMA(3,4) model like similar to the residuals from the AR(3) model and do not look like those of a white noise process: the time plot shows a clear growing trend and many of the auto-correlations (apart from $k = 0$) and partial auto-correlations are significantly different from zero.

The AIC value (1,728.8) for the ARMA(3,4) model is slightly smaller than that of the MA(3) model (which was 1,786.8), so if we considered all of the models together and only used the AIC as a selection criterion, we would select the ARMA(3,4) model. It's worth noting again that the assumption of a stationary process has not been met, and thus there is good reason to believe that the ARMA model will not generalize well out-of-sample or make accurate predictions. However, without transforming the dataset, the ARMA(3,4) model would be selected. .

```
summary(best.arma$resid)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -10.3500  -0.6981   0.4237   0.3928   1.7810   6.4350
```

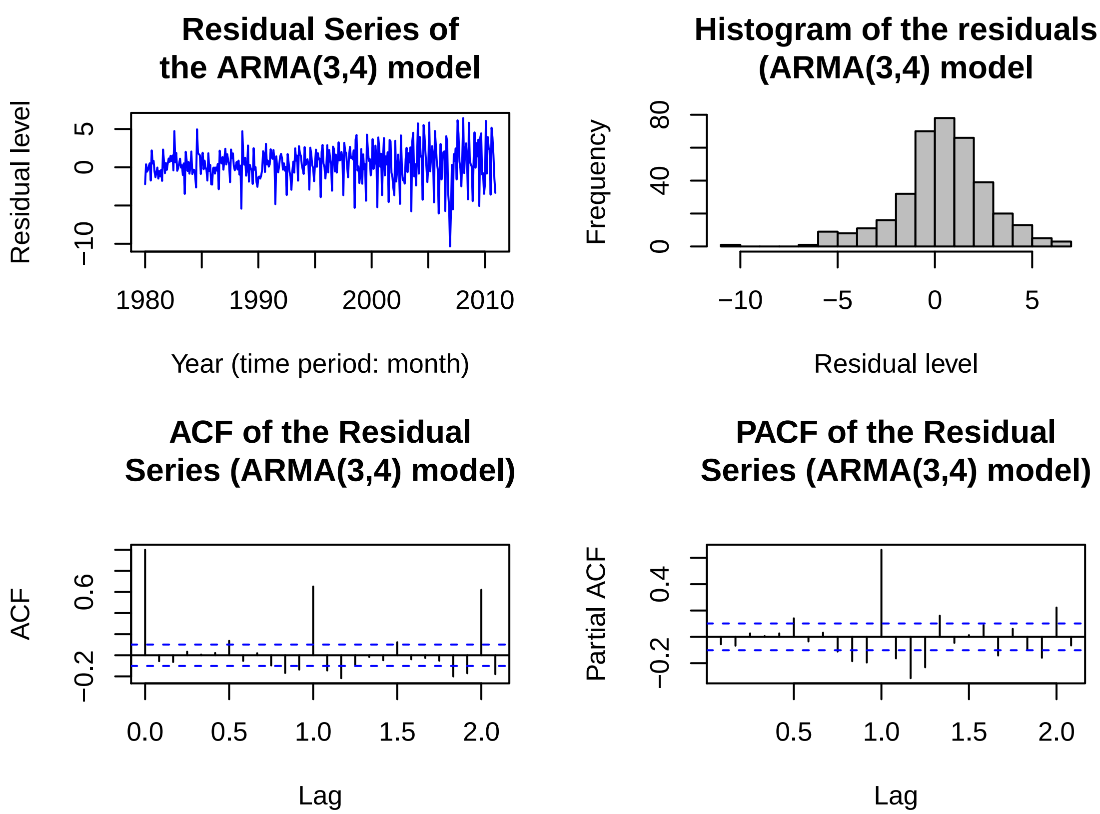


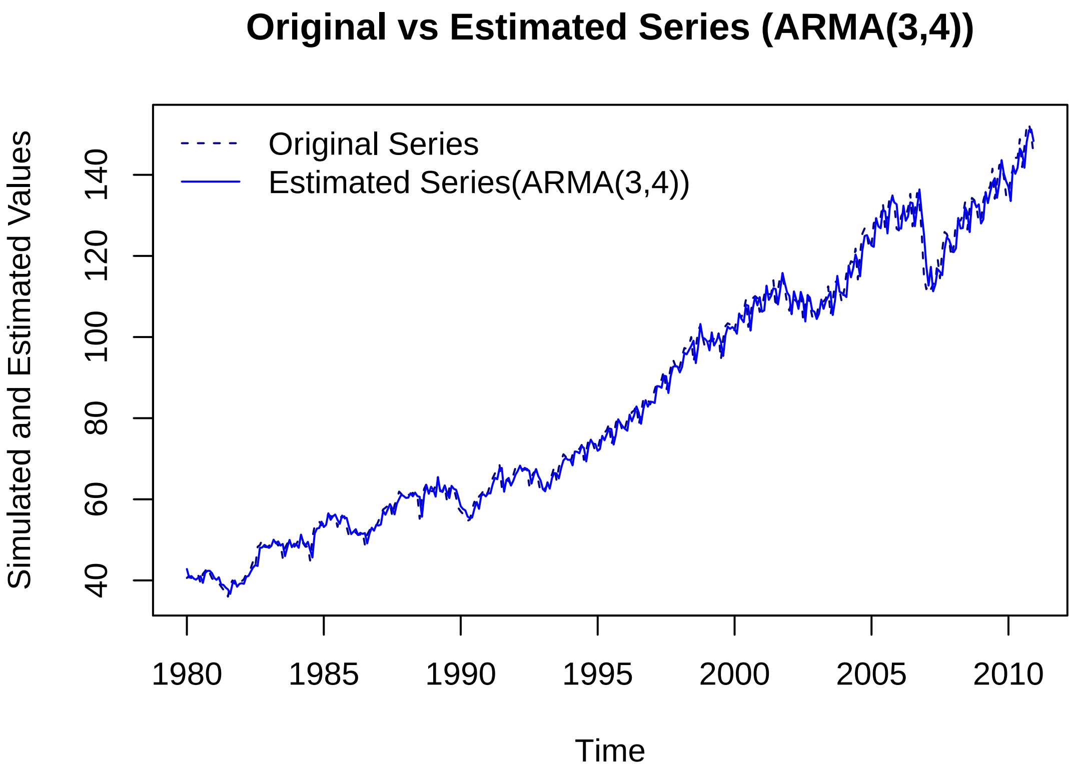
Figure 11: AR(3) model diagnostic based on the residuals

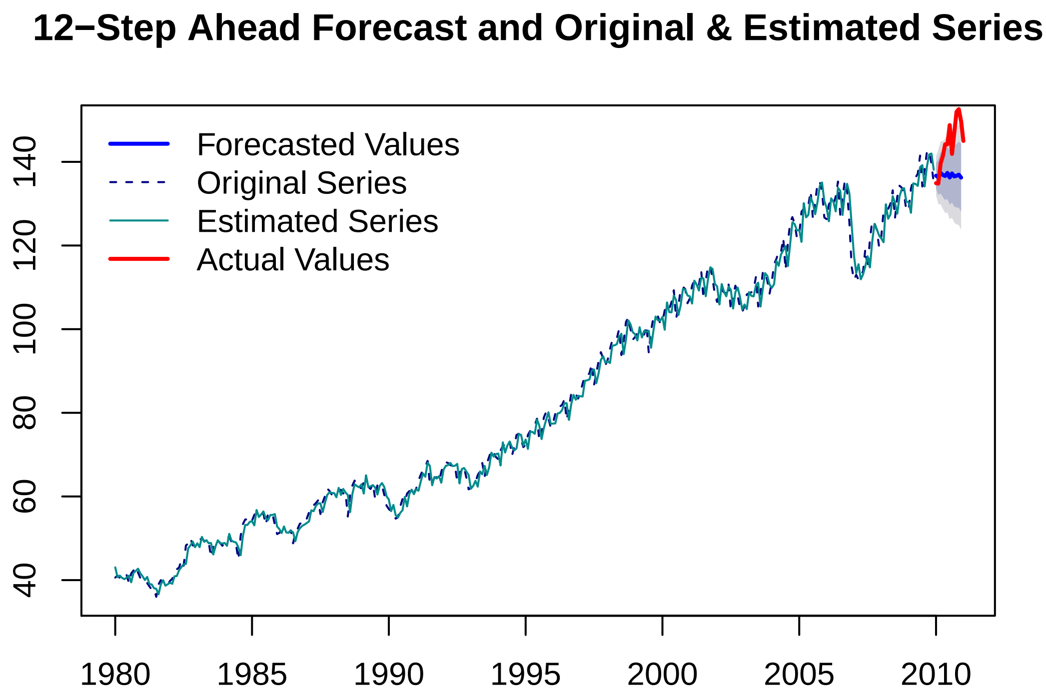
However, the result of a Llung-Box test is that we cannot reject the hypothesis of independence of the residual series:

```r
Box.test(best.arma$resid, type = "Ljung-Box")
```

```
##
##  Box-Ljung test
##
## data:  best.arma$resid
## X-squared = 1.1881, df = 1, p-value = 0.2757
```

The fitted values versus original series plot for the ARMA(3,4) model shows us that the model performs well at matching the within-sample dynamics of the time-series.





One property we are interested in when selecting univariate time-series models in out-of-sample forecasting. For the best fitting model, we will estimate a new model witholding one year of the observations (the last year because of the correlations between incremental observations). We can than use this model to forecast the next observations and compare the forecasted values to the realized data in the series.

Looking at the forecasted versus actual values we can see how the model is unable to capture the dynamics of this non-stationary process when forecasting out-of-sample. The forecasted values tend to be near the mean of the previous few values, while the actual values steadily increase with a few values falling outside of the 95% confidence interval of the forecast. This out-of-sample forecast illustrates why an ARMA type model, while capable of capturing the in-sample dynamics of a random walk with drift process, is not well-suited to forecasting non-stationary processes.
