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's underling assumption?**
- **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 montly 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

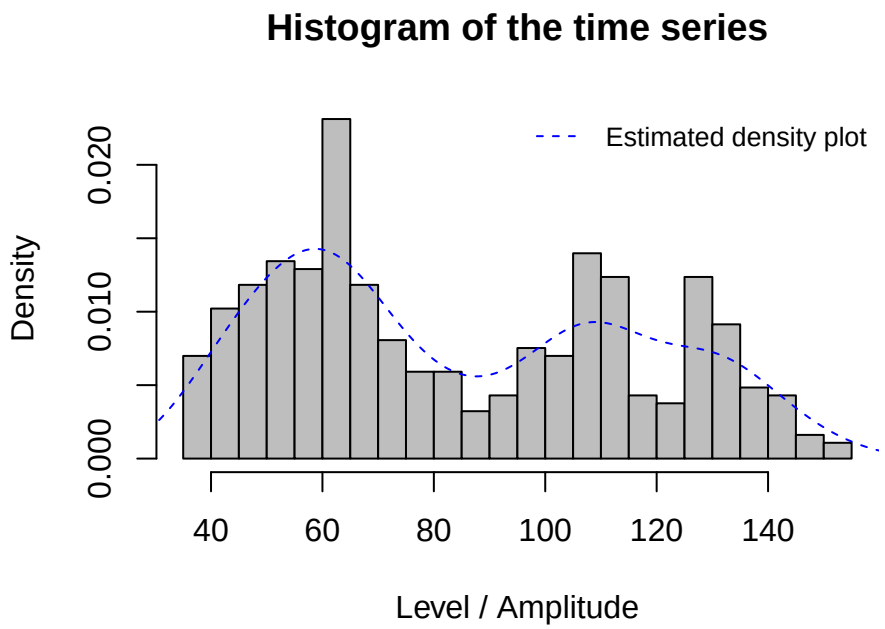


Figure 1: Histogram of the data.

The histogram shows that the distribution of the data is multimodal, and far from normal. But as usual, it tells us nothing about the dynamics of the time series. 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)
```

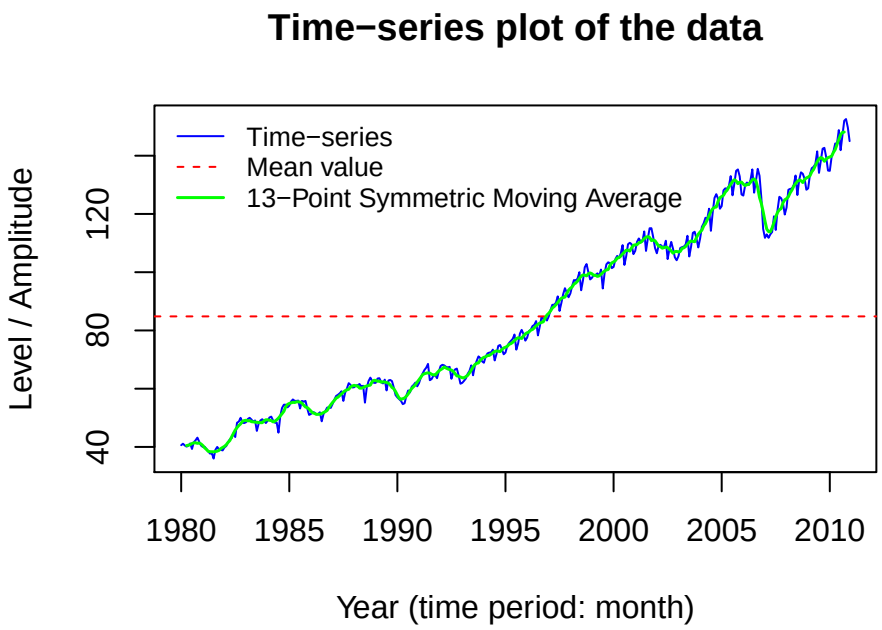


Figure 2: Time-series plot (assuming monthly data, from 1980 until 2010).

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).

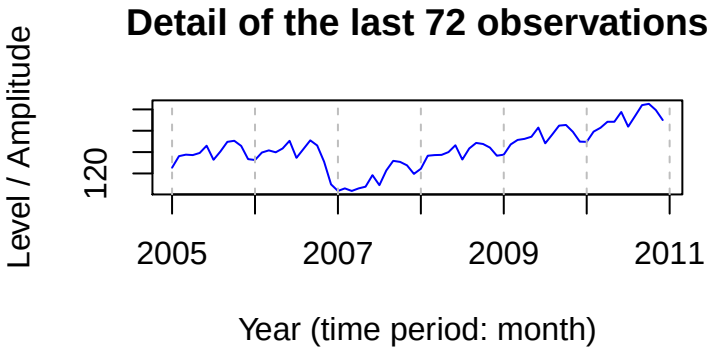


Figure 3: Time-series plot of (the last 72 observations—6 years?—of) the data.

Apart from 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.

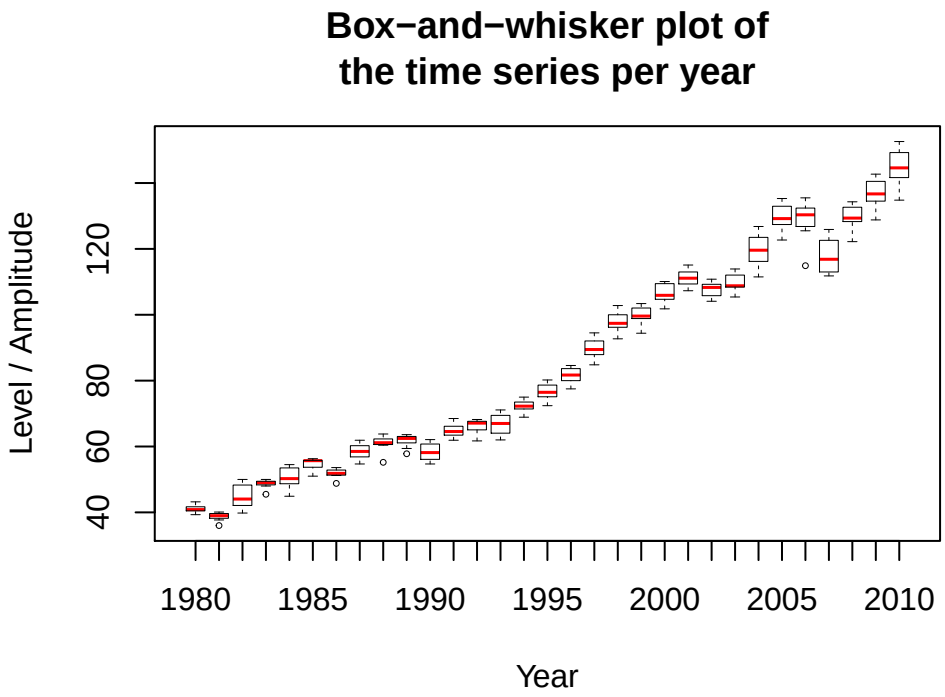


Figure 4: Boxplot of the series, per year (every 12 observations).

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 **an ARMA model may not be a good fit for our data**????

To finish the Exploratory Data Analysis, let's decompose the time series to check the growing trend and seasonality:

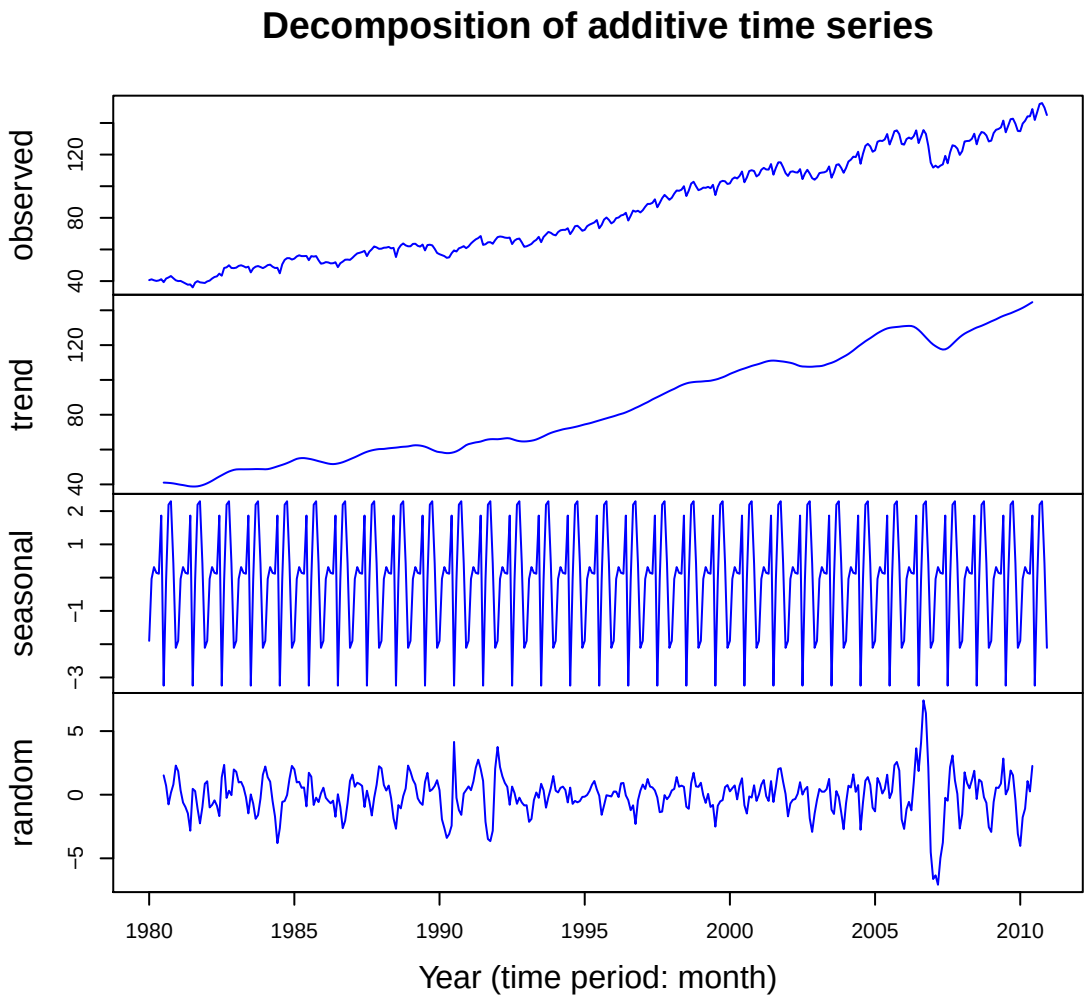


Figure 5: (Additive) decomposition of the time series.

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. The PACF drops off very sharply after the 1st lag.

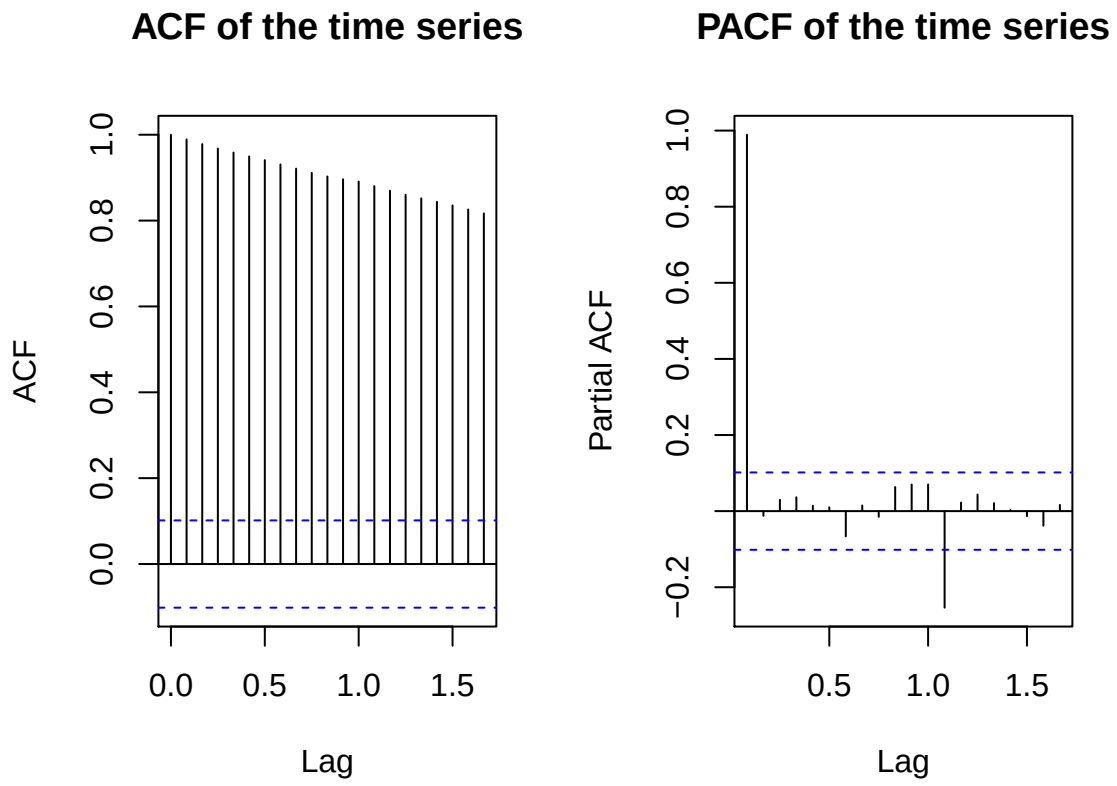
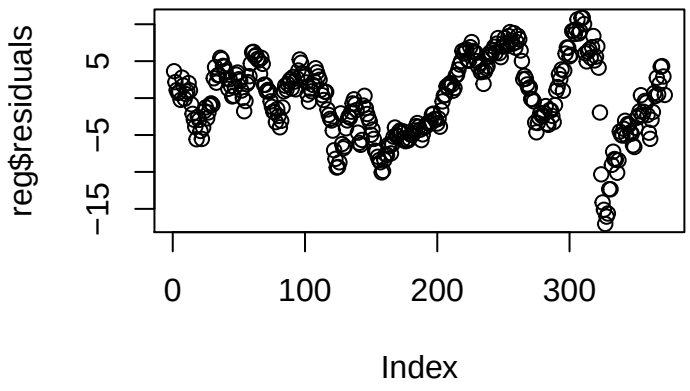
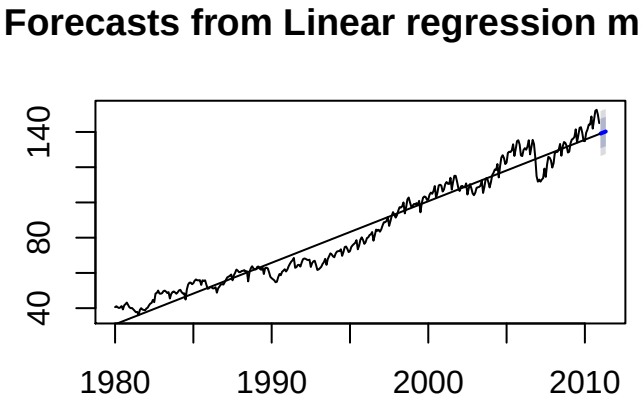


Figure 6: Autocorrelation and partial autocorrelation graphs

```
M <- factor(cycle(hw08.ts))
reg <- lm(hw08.ts ~ time(hw08.ts) + I(time(hw08.ts)^2)+ M)
plot(reg$residuals)
```




```
library(forecast)
m2 <- tslm(hw08.ts~trend)
f <- forecast(m2, h=5,level=c(80,95))
plot(f)
lines(fitted(m2))
```

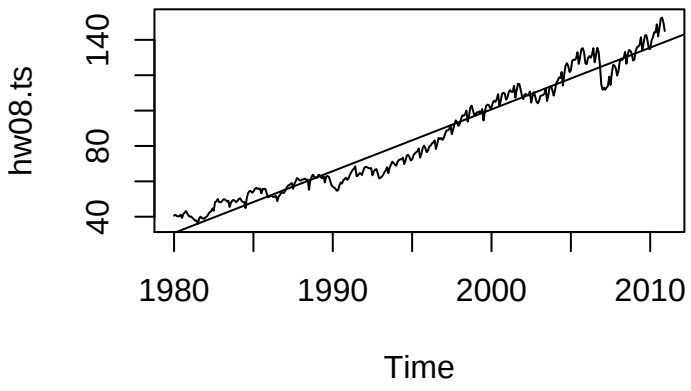



```
summary(m2)
```

```
##
## Call:
## lm(formula = formula, data = "hw08.ts", na.action = na.exclude)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.2504  -5.3285   0.9076   4.8494  14.5360
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 30.55476    0.67154   45.50   <2e-16 ***
## trend        0.29100    0.00312   93.26   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.463 on 370 degrees of freedom
## Multiple R-squared:  0.9592, Adjusted R-squared:  0.9591
## F-statistic:  8697 on 1 and 370 DF,  p-value: < 2.2e-16
```

```
d <- data.frame(x = hw08, time = time(hw08.ts))
m <- lm(x~time, d)
plot(hw08.ts)
abline(m)
```




```
library(car)
linearHypothesis(reg, sapply(c(2:12), function(i) paste0("M", i)))
```

```
## Linear hypothesis test
##
## Hypothesis:
## M2 = 0
## M3 = 0
## M4 = 0
## M5 = 0
## M6 = 0
## M7 = 0
## M8 = 0
## M9 = 0
## M10 = 0
## M11 = 0
## M12 = 0
```

```
##
## Model 1: restricted model
## Model 2: hw08.ts ~ time(hw08.ts) + I(time(hw08.ts)^2) + M
##
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1    369 10674.5
## 2    358  9648.2 11    1026.2 3.4617 0.0001326 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```
