Student Project - Daily Temperature

César Raúl Urteaga Reyesvera

Table of Contents

Introduction

Global warming, the increase of Earth's average surface temperature due to effect of greenhouse gases, is an issue of great importance around the world. Many countries are concerned about this problem, and allocate several resources in order to understand this phenomenon. In particular, some of them measure the temperature through weather stations along their territory. With the aim to comprehend this event and employing the collected data, a possible approach to model the variability of temperature is by the use of time series models.

The purpose of this project is to model the ***daily*** **high temperatures** reported by a weather station in Corpus Christi, TX, through the use of time series analysis and the data provided by the NEAS.

In order to construct our model, the approach that we will use is the model building strategy proposed by Box and Jenkins (1976):

1. **specification**
2. **fitting**, and
3. **diagnostics**

For the first step, we will look at the time plot of the series, compute some statistics from the data, and consider some classes of time series models while trying to attempt to adhere to the principle of parsimony (i.e., the model with the fewest number of parameters that will adequately represent the variation of the data). Secondly, in fitting a model, we will try to find the best possible estimates of the unknown parameters through the observed data. Finally, we will assess the quality of the model that we already have specified and estimated.

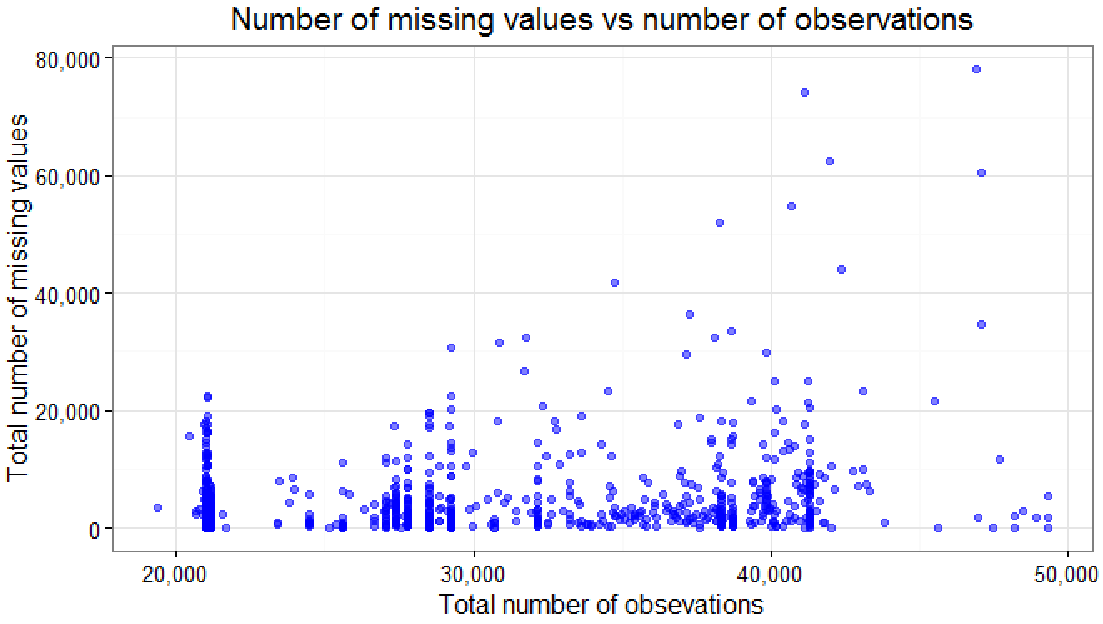
Data

Description

The data was obtained from the NEAS's web site; I download the 14 zip files of daily temperature. There were a total of 1,062 files in CSV format, where each file corresponds to a weather station in the U.S. The reported variables were date of observation, high and low temperatures (in degrees Farenheit), and rainfall precipitation (in hundreds of an inch).

Because each station has a different number of observations, the quality of the data varies among them; specifically, the number of missing values was different for each one.

In order to take advantage of the amount of data, I decided to calculate the total number of missing values (i.e., considering the date and high and low temperatures) and the number of observations. The graph below show this for each weather station.



According to this plot, we can see an increasing pattern of missing values as the number of observations increases. However, there are some stations with few missing values and many observations. Because I am not interested in a particular station, I determined to take the one with many years of data and few missing records. So, I decided to explore these traits among all the files.

The files with less than 20 total missing values were the following:

```
                      file     n date ht lt tm min.y max.y n.y
882 station412015_data.csv 21185    0  0  0  0  1948  2005  58
59  station043257_data.csv 21003    0  1  0  1  1948  2005  58
998 station457938_data.csv 45624    0  0  1  1  1881  2005 125
909 station417945_data.csv 21672    0  1  1  2  1946  2005  60
395 station215435_data.csv 42003    0  1  6  7  1891  2005 115
474 station244055_data.csv 41272    0  1  7  8  1893  2005 113
542 station266779_data.csv 25143    0  2  9 11  1937  2005  69
888 station412797_data.csv 21185    0  6  6 12  1948  2005  58
623 station307167_data.csv 29220    0  6  8 14  1926  2005  80
627 station308383_data.csv 30681    0  5 10 15  1922  2005  84
257 station131635_data.csv 21094    0  9  7 16  1948  2005  58
536 station262573_data.csv 28490    0  7  9 16  1928  2005  78
```

where **date**, **ht**, and **lt** corresponds to the total number of missing values for the record date, low and high temperature variables, respectively. On the other hand, **tm** represents the total number of missing values (i.e., the total number of missing values among the date, ht, and lt variables), **min.y** and **max.y** are the minimum and maximum years of observation, respectively, and **n.y** is the total number of years.

From the above table, I decided to take the station number 412015, which corresponds to the coastal city of Corpus Christi in Texas. Since it does not have any missing value, I did not imputate any datum.

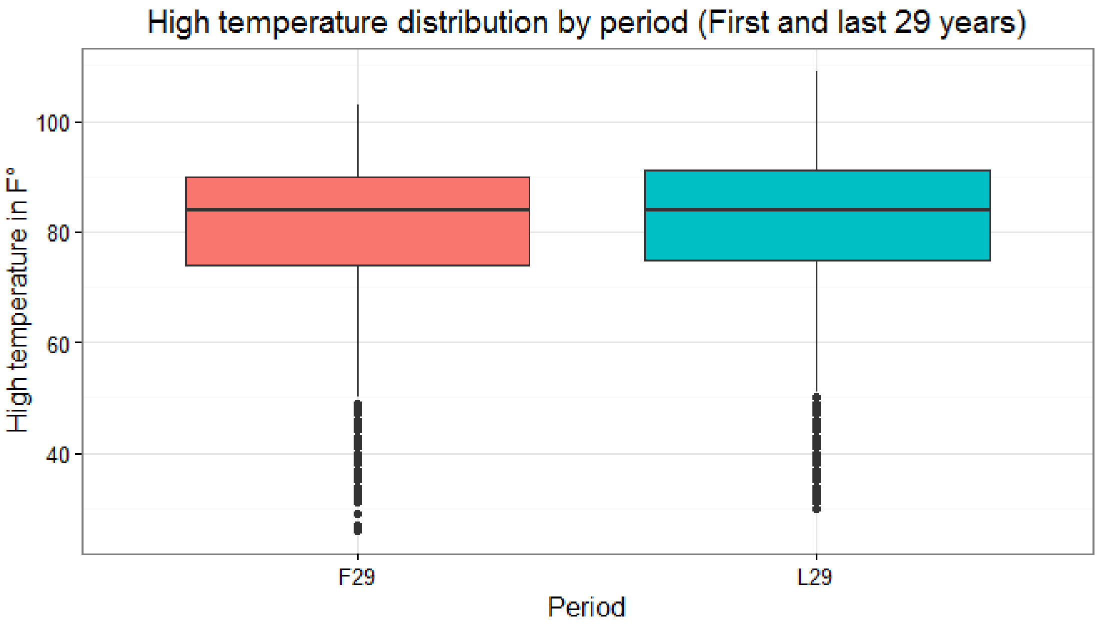
Exploratory Data Analysis

Because I wanted to have a better understanding of the high temperatures in Corpus Christi, I calculated some summary statistics during the 58 years of data (21,185 observations):

```
      date                  ht       
 Min.   :1948-01-01   Min.   : 26.0  
 1st Qu.:1962-07-02   1st Qu.: 75.0  
 Median :1976-12-31   Median : 84.0  
 Mean   :1976-12-31   Mean   : 81.3  
 3rd Qu.:1991-07-02   3rd Qu.: 91.0  
 Max.   :2005-12-31   Max.   :109.0  
```

As we can see from the above table, since the median is greater than the mean, the distribution of the high temperature is left skewed. On the other hand, the standard deviation during this period was 12.06.

I decided to analyse the information by periods of 29 years (1948-1976[F29] and 1977-2005[L29]). From the plot below, we can observe that both periods have the same median (84) with a lower variability for the first 29 years (the standard deviation was 11.948, while 12.17 for the second period).



However, since there were some outliers, the standard deviation was not a good measure of variability, I calculated the interquantile range for the first and last periods, which both had the same value (16). So, it seems reasonable to assume that both periods have the same variability and the same mean.

Analysis

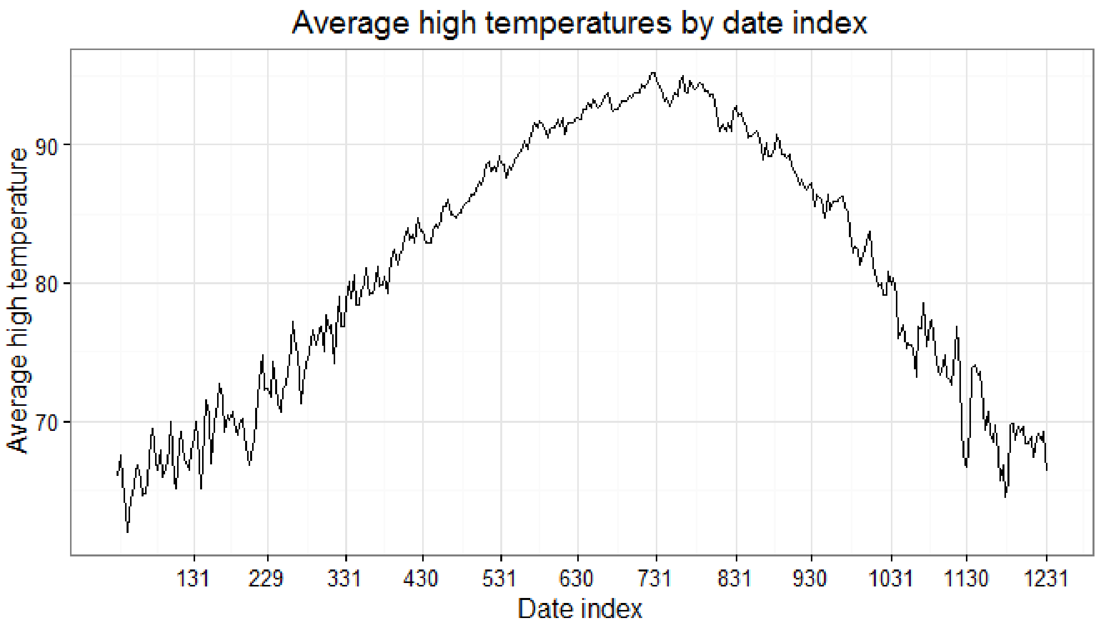
Model

We will assume that high temperatures can be modeled by

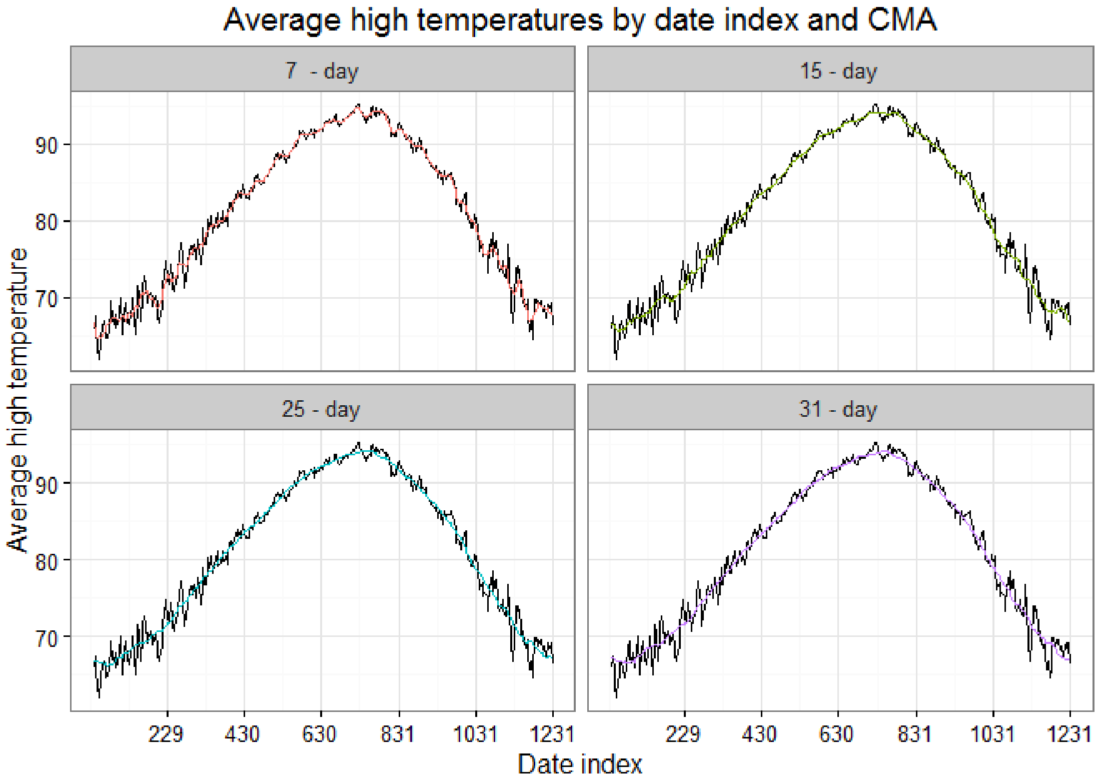
$$Y_t = \mu_t + X_t,$$

where $\mu_t = \beta_i$ is a deterministic seasonal trend with i as the date index defined as $Month \cdot 100 + Day$. Hence, the β_i's give the expected average high temperatures for each day (expressed as an index) of the year. On the other hand, X_t correspond to the stochastic trend of the time series which we will fit through the use of an ARIMA model.

In order to fit the seasonal trend, de-seasonalize the data, and test the robustness of the model, I calculated the average high temperature for each date index using the data for the first period.



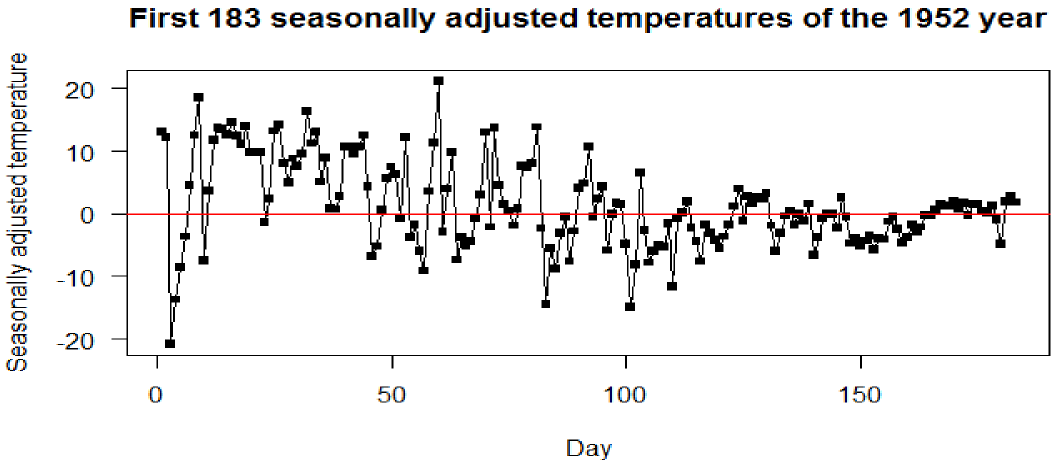
We can see, from the above plot, that the volatiliy of daily temperatures overwhelms the seasonal pattern; the 62.179% of the first differences were positive, while the 37.821% were negative. In order to remove this from the deterministic trend, I used centered moving averages to smooth the sequence of averages. I tried 7, 15, 25, and 31 day centered moving averages, and I found that the 25 centered moving averages work fine.



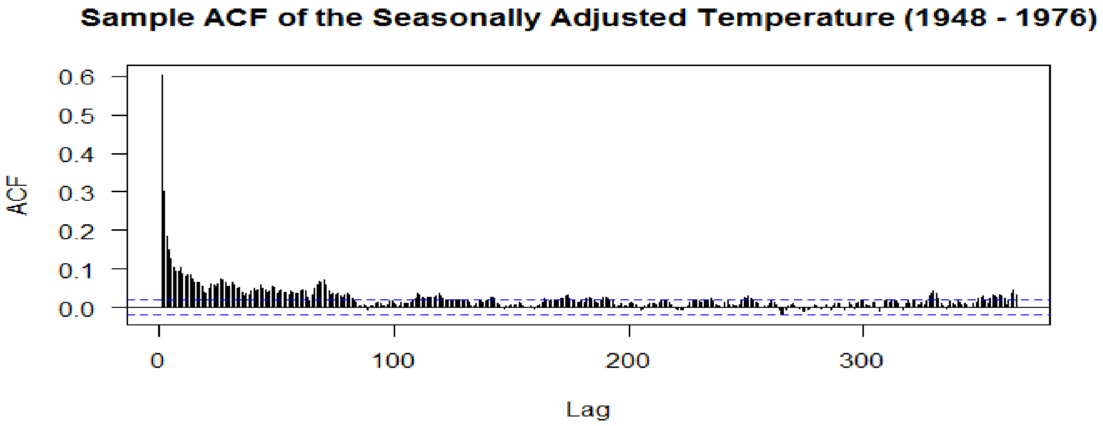
With the 25 day centered moving average, I calculated the seasonally adjusted daily temperatures under the additive model (i.e., the reported temperature minus the smoothed average). Since the mean of the seasonally adjusted temperatures were practically zero (3.04510^{-16}), I did not have to normalize them; their range were (-42.419, 24.457).

Before I fitted the ARIMA model, I checked for two patterns:

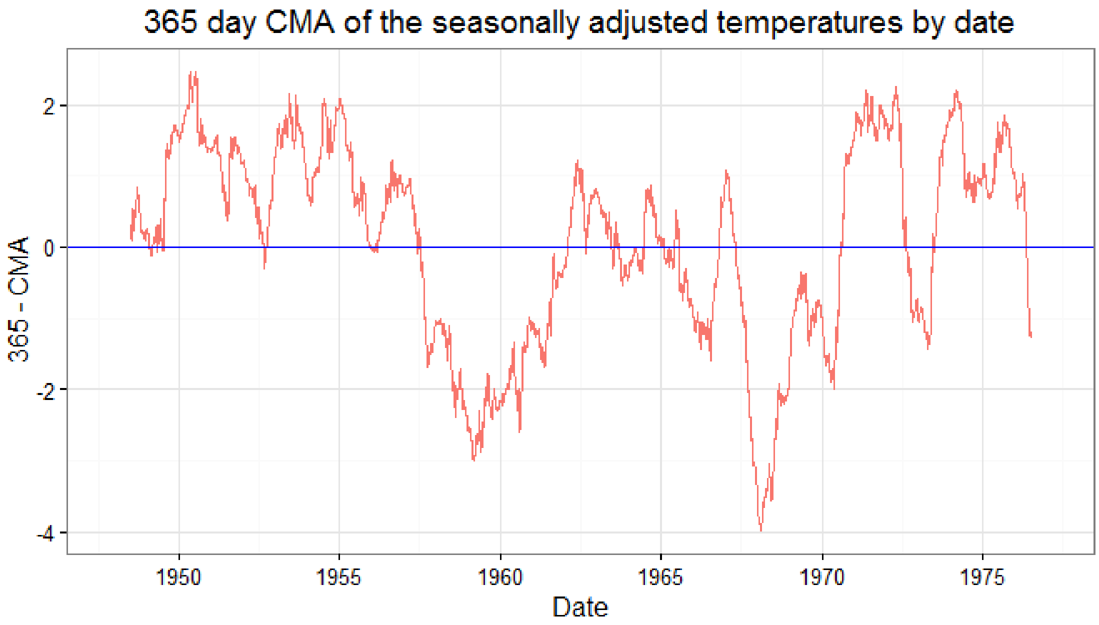
- **White noise**: I plotted the first 183 days of 1952 (leap-year), and I could notice that it was not white noise since positive and negative seasonally adjusted temperatures came in streams.
- **Random walk**: Using the same plot, I also noticed that the mean reversion was strong. So, it was not a random walk.


First 183 seasonally adjusted temperatures of the 1952 year

I calculated the sample autocorrelations and ploted the correlogram. The graph below suggests that the sample autocorrelation functions decays exponentially; in particular, r_1 has a value of 0.603 and r_2 (0.303) is almost r_1^2 (0.364). Even though some sample autocorrelations are not zero, it seems reasonable to assume stationarity.


Sample ACF of the Seasonally Adjusted Temperature (1948 - 1976)

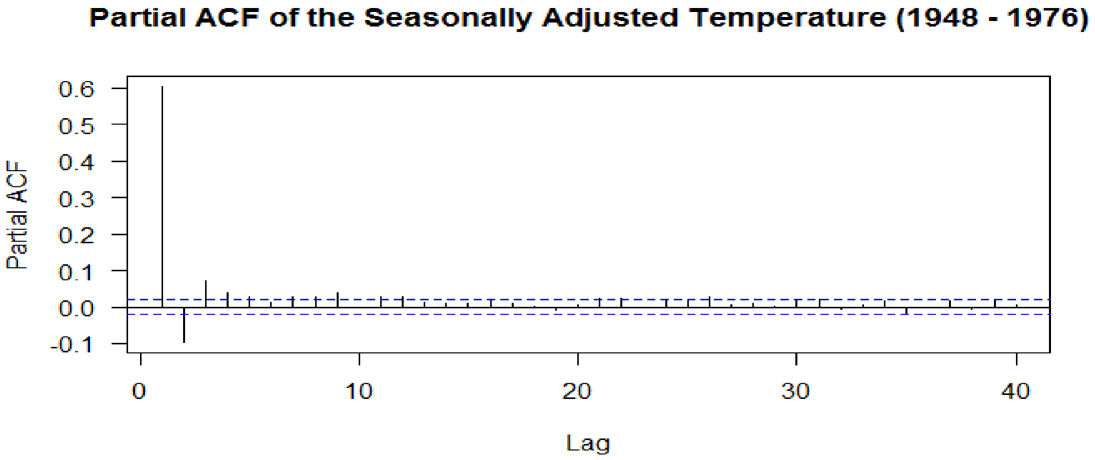
In addition, I decided to plot the 365 centered moving average of the seasonally adjusted temperature of the 29 years of data to check for any long trend. As it can see, from the graph below, it is difficult to know if there is a long trend or is just variability around the mean (blue line). Hence, I decided to take the time series as they are (i.e., without differentiating or detrending any long pattern).



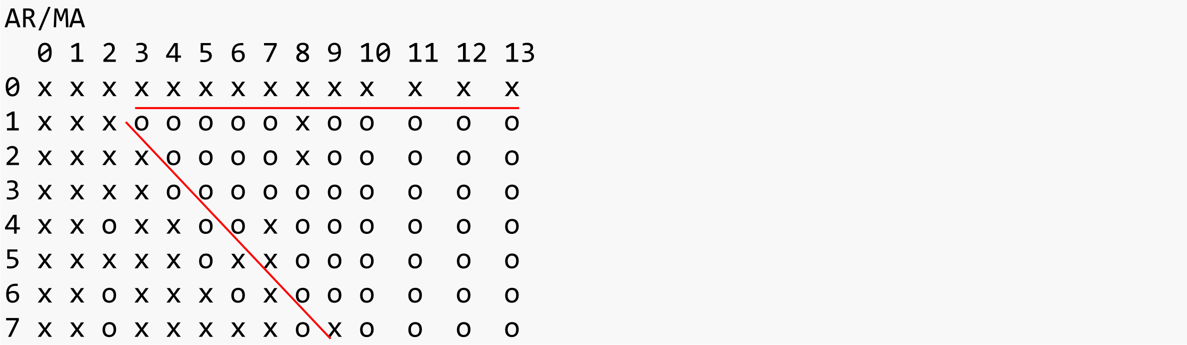
Specification

As shown in the correlogram, I determined that a reasonable parameter d for the ARIMA process would be **0**.

In addition, the exponential decay of the sample autocorrelation function suggests an autoregressive model. I considered an AR(1) and AR(2) models since the partial autocorrelation function suggests them.



I also considered an ARIMA(1,0,3) from the sample extended autocorrelation function:



```
AR/MA
  0 1 2 3 4 5 6 7 8 9 10 11 12 13
0 x x x x x x x x x x x  x  x  x 
1 x x x o o o o o x o o  o  o  o 
2 x x x x o o o o x o o  o  o  o 
3 x x x x o o o o o o o  o  o  o 
4 x x o x x o o x o o o  o  o  o 
5 x x x x x o x x o o o  o  o  o 
6 x x o x x x o x o o o  o  o  o 
7 x x o x x x x x o x o  o  o  o 
```

Fitting

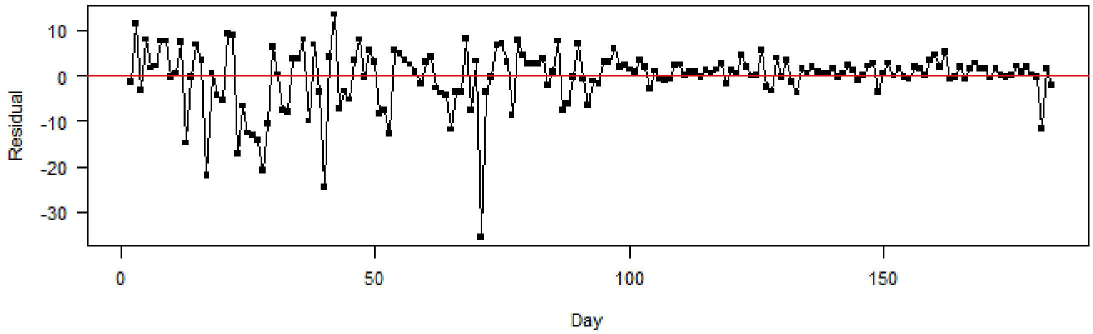
I fitted the proposed models using least squares estimation for the AR(1) and AR(2), while I used the method of maximum likelihood for the ARIMA(1,0,3) model.

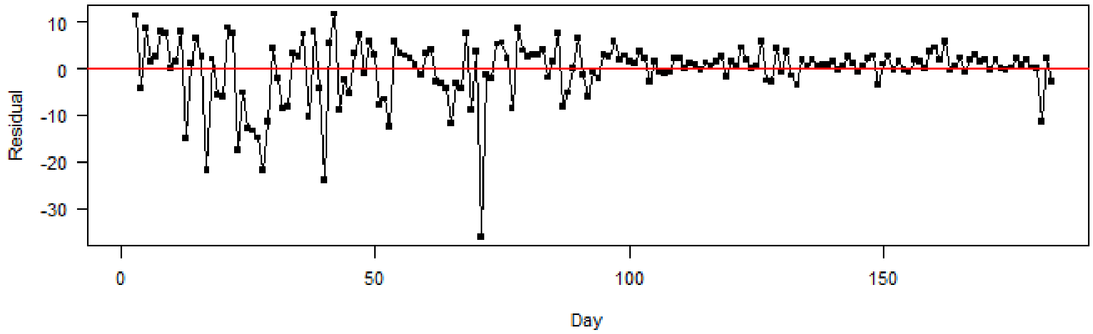
Model	ϕ_1	ϕ_2	θ_1	θ_2	θ_3
AR(1)	0.603				
AR(2)	0.661	-0.096			
ARIMA(1,0,3)	0.911		-0.255	-0.318	-0.135

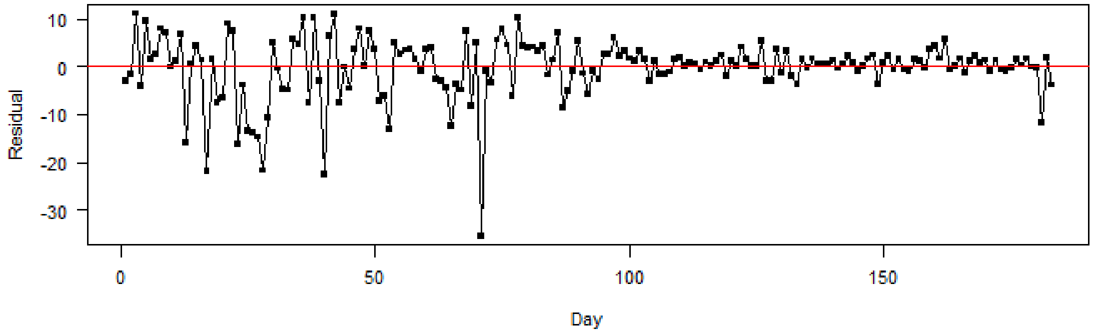
Diagnostics

Plots of the Residuals

For each model, I plotted the first 183 residuals, and I found a rectangular scatter around a zero horizontal level. However, there was an increased variation in the beginning of the period, whereas a reduced variation at the end - not an ideal plot of residuals.

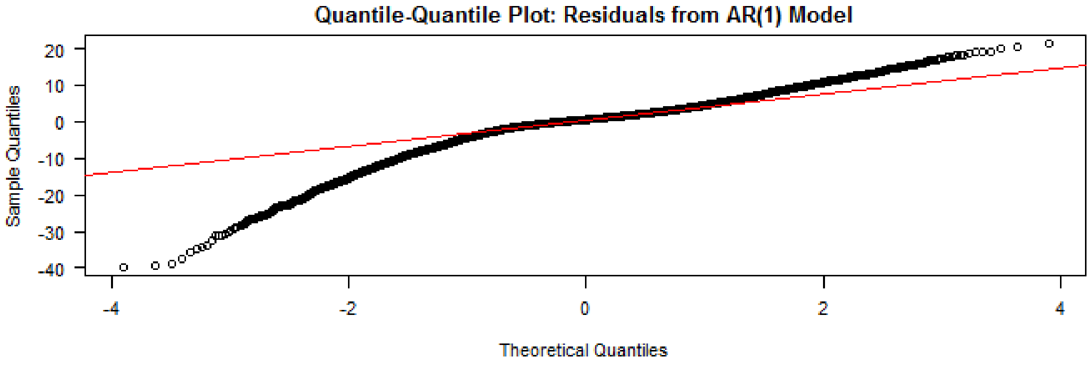

First 183 residuals of the AR(1)

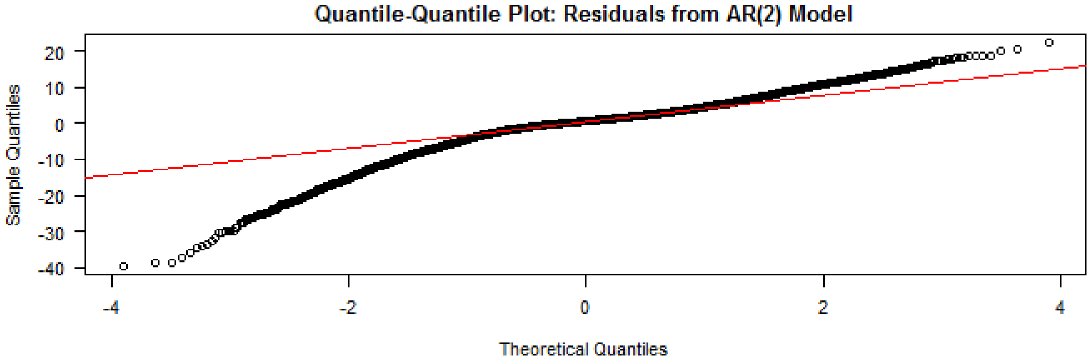

First 183 residuals of the AR(2)

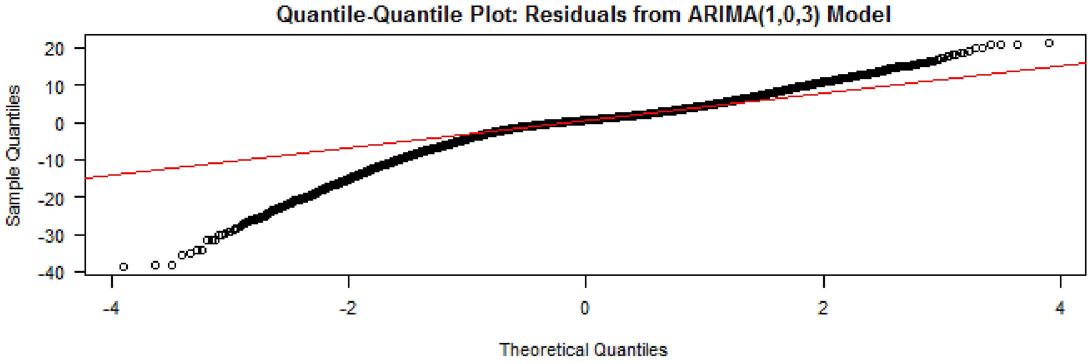

First 183 residuals of the ARIMA(1,0,3)

Normality of the Residuals

Using all the residuals from each model, the quantile-quantile plots do not seem to follow a normal distribution at the tails. The empirical distributions of the residuals are left skewed.







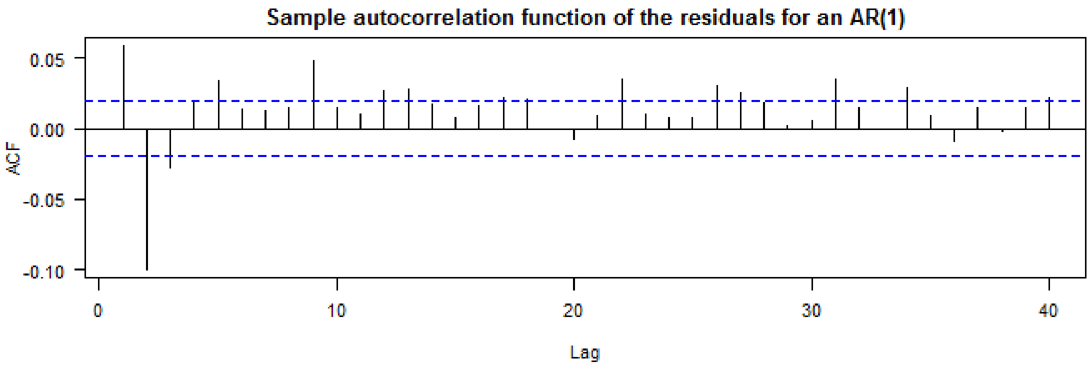
I also tested the normality of the residuals, using the first 5,000 values, through the Shapiro-Wilk test. Since all of them resulted statistically significant (a lower p-value), we rejected the null hypothesis that the residuals of each model were normally distributed.

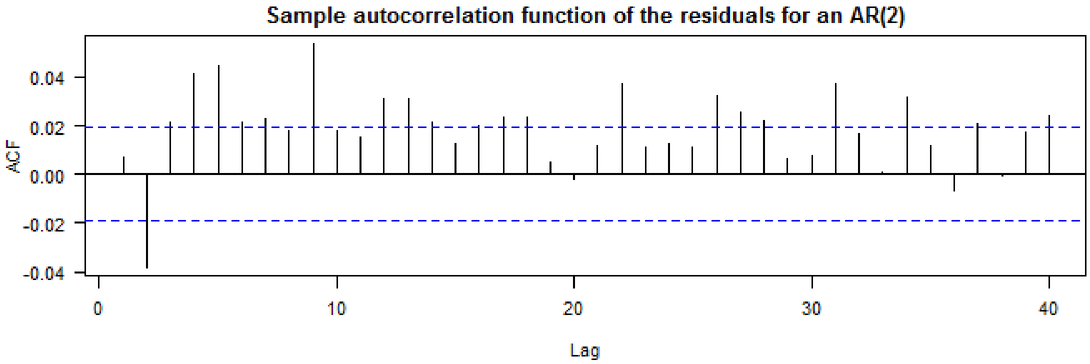
```
	Shapiro-Wilk normality test

data:  m.AR_1$resid[1:5000]
W = 0.9, p-value <2e-16
```

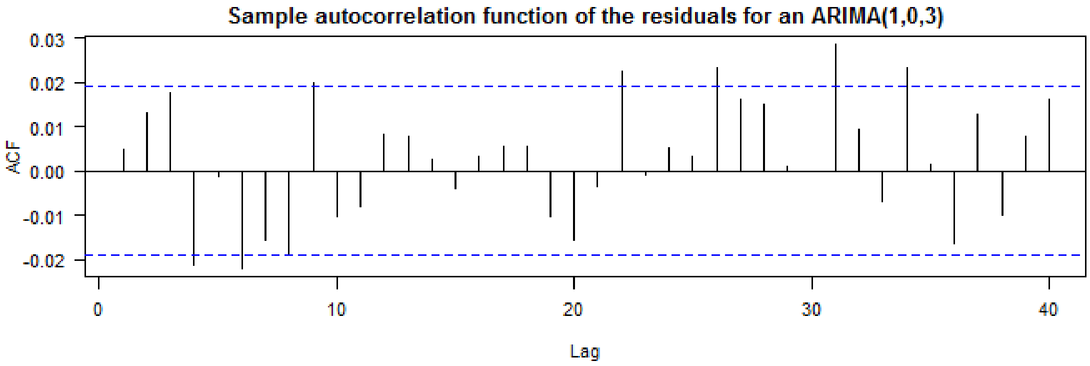
```
    Shapiro-Wilk normality test

data:  m.AR_2$resid[1:5000]
W = 0.9, p-value <2e-16


    Shapiro-Wilk normality test

data:  m.ARIMA_1_0_3$residuals[1:5000]
W = 0.9, p-value <2e-16
```

Autocorrelation of the residuals

Looking the sample autocorrelation function of the residuals for each model, the ARIMA(1,0,3) showed less evidence of autocorrelation among them.


Sample autocorrelation function of the residuals for an AR(1)


Sample autocorrelation function of the residuals for an AR(2)


Sample autocorrelation function of the residuals for an ARIMA(1,0,3)

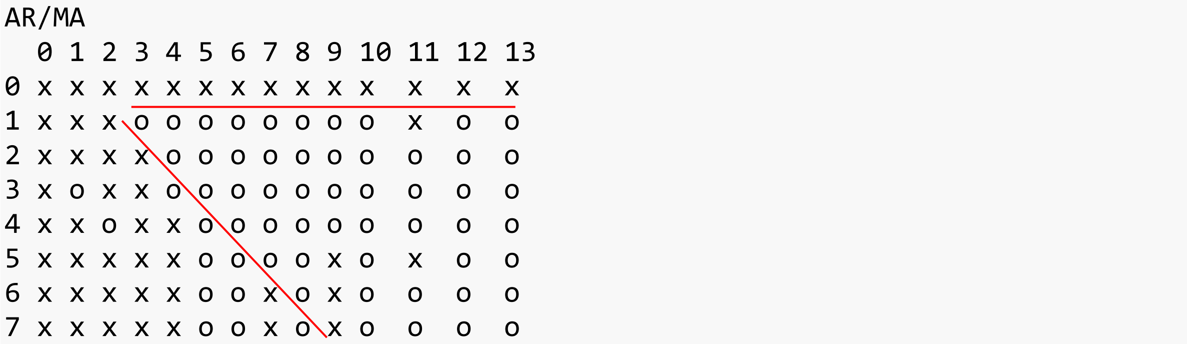
The following table has the Box-Pierce, with their p-values, and the Durbin-Watson statistics for the three models:

Model	Box-Pierce (K = 500)	Box-Pierce (K = 1,000)	Durbin-Watson
AR(1)	831.796(0)	1292.427(8.29610^{-10})	1.883
AR(2)	804.032(1.1110^{-16})	1281.268(2.63710^{-9})	1.985
ARIMA(1,0,3)	537.823(0.095)	985.266(0.59)	1.989

As we can see, the results for the ARIMA(1,0,3) model showed that we do not have evidence to reject the null hypothesis that the error terms are uncorrelated. In contrast, the statistics for the AR(1) and AR(2) were statistically significant. Hence, there are evidence of correlation among the residuals of these models.

Robustness

To test the robustness of the model, I calculated the sample extended autocorrelation function for the last period (1977-2005) using the same procedure described above.



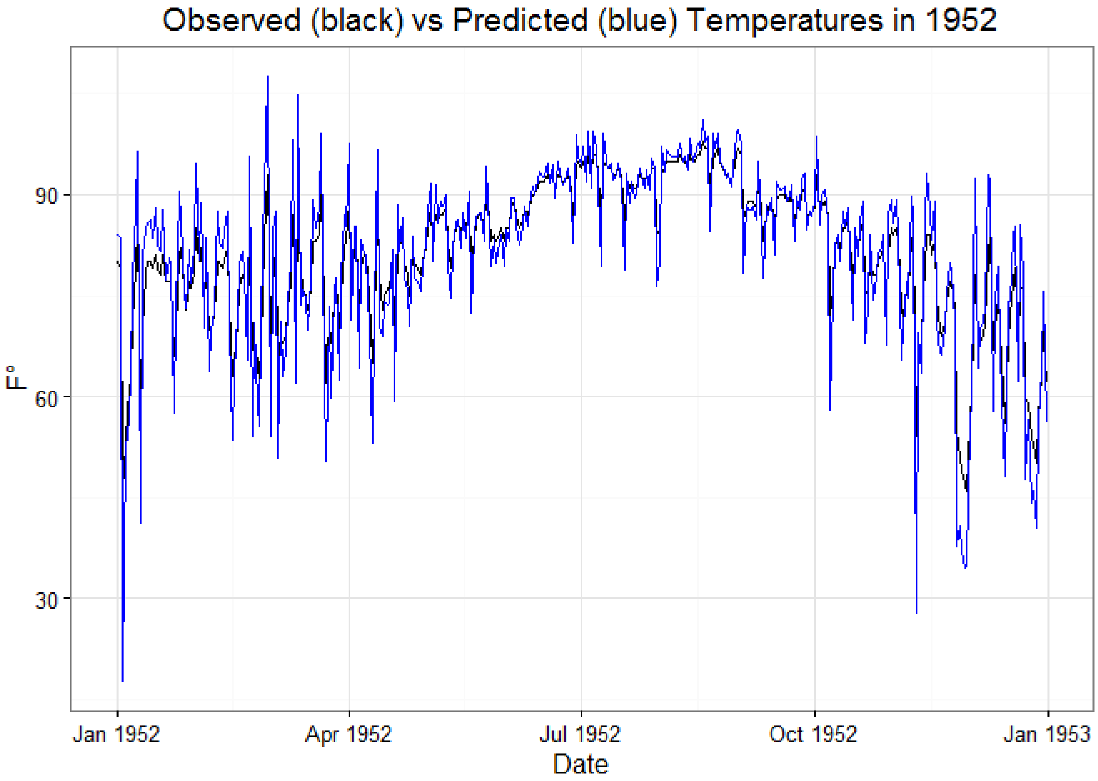
```
AR/MA
  0 1 2 3 4 5 6 7 8 9 10 11 12 13
0 x x x x x x x x x x x  x  x  x
1 x x x o o o o o o o o  x  o  o
2 x x x x o o o o o o o  o  o  o
3 x o x x o o o o o o o  o  o  o
4 x x o x x o o o o o o  o  o  o
5 x x x x x o o o o x o  x  o  o
6 x x x x x o o x o x o  o  o  o
7 x x x x x o o x o x o  o  o  o
```

The results suggest the **same model** obtained with the first 29 years: an ARIMA(1,0,3).

Conclusion

Even though the AR(1) and AR(2) models are simpler (under the principle of parsimony) than the ARIMA(1,0,3) model, I concluded that the best fit for the seasonally adjusted temperatures, based on the previous analysis, was the latter. However, as we saw in the diagnostic section, this model does not explain all the variability of the phenomenon.

The following chart compares the observed daily high temperatures, during the leap-year 1952, with the predicted values under the ARIMA(1,0,3) model.



The model seems to explain well the variability at the middle of the year. However, at the beginning and at the end of the year, the model had a poor fit. It did not capture the high variation of daily temperatures at the borders.