Exercise 2: State Space Models

Iker Caballero

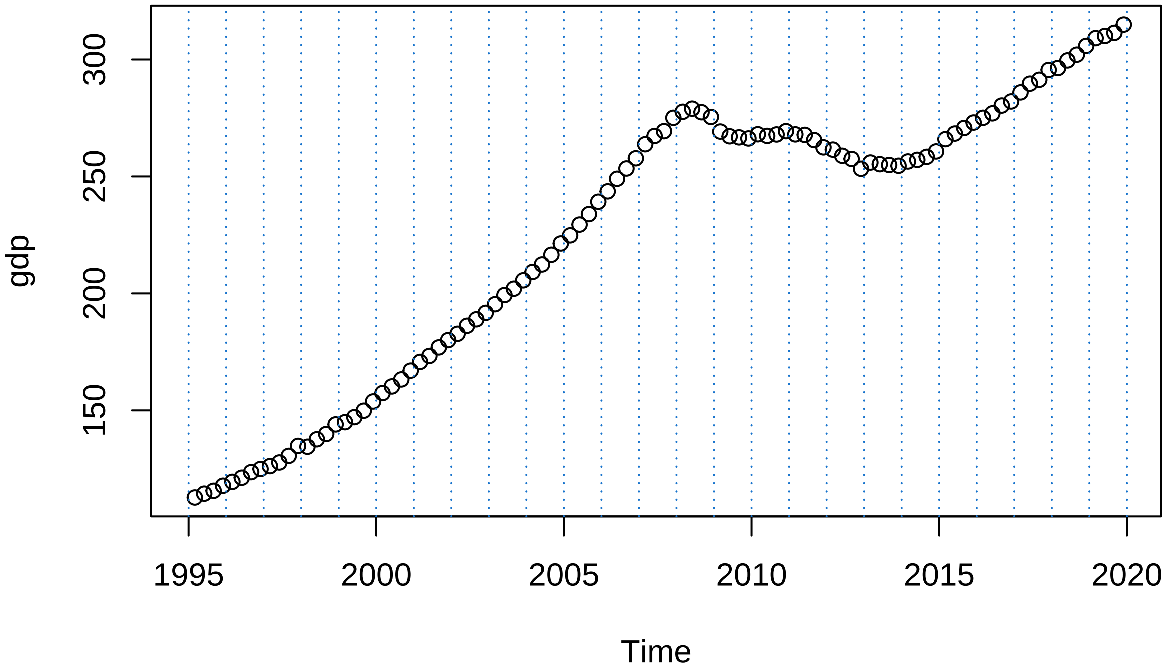
2023-05-28

Quick Note: I highlight that, throughout all the development, I assume $N(0, \sigma^2)$, so that the second parameter in the argument refers to variance. This is the reason why I always use the square when showing the results of the estimation, even though I obtain the squared root. However, this is mathematically equivalent and there is no difference in using one or another but the interpretation at reading.

Phase I: Imputation of missing data. Monthly GDP

Data

```
gdp=ts(c(t(cbind(matrix(NA,length(gdpOri),2),gdpOri))),freq=12,start=c(1995,1),
       end=c(2019,12))
plot(gdp,type="p")
abline(v=1995:2023,lty=3,col=4)
```



Specify the state-space model

State:

The state in this setting is the monthly GDP of Spain (at current prices).

Observations:

The observations in this setting is the quarterly GDP of Spain (at current prices).

Transition Equation for the state:

$$x_t = \phi x_{t-1} + w_t \qquad w_t \sim N(0, Q)$$

where x_t denotes the monthly GDP (state) at time t, ϕ is the autorregresive parameter, and w_t is an iid Gaussian random variable with zero mean and Q the variance. In this case, we assume that the monthly GDP behaves like an AR(1) process, and we want to estimate the parameter ϕ and Q in order to obtain the process. Moreover, we assume that this state process starts with an initial value (an initial value of the monthly GDP), which is distributed as follows:

$$x_0 \sim N(\mu_0, \Sigma_0) \quad where \quad \mu_0 = 112.76 \quad and \quad \Sigma_0 = 100$$

Observation Equation:

$$\begin{pmatrix} y_t^{(1)} \\ 0 \end{pmatrix} = \begin{bmatrix} A_t^{(1)} \\ 0 \end{bmatrix} x_t + \begin{pmatrix} v_t^{(1)} \\ 0 \end{pmatrix} \implies y_{(t)} = A_{(t)} x_t + v_{(t)}$$

$$where \quad v_{(t)} \sim N_2(0, R_{(t)}) \quad and \quad R_{(t)} = \begin{bmatrix} R_{11t} & 0 \\ 0 & I_{22t} \end{bmatrix}$$

This is a general expression, but in this case, $A_t^{(1)} = 1$, $R_{11t} = \frac{1}{1000000}$ and we are dealing with a one dimensional state (not a vector), so we can write the equations as

$$\begin{pmatrix} y_t^{(1)} \\ 0 \end{pmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} x_t + \begin{pmatrix} v_t^{(1)} \\ 0 \end{pmatrix} \implies y_{(t)} = A_{(t)} x_t + v_{(t)}$$

$$where \quad v_{(t)} \sim N_2(0, R_{(t)}) \quad and \quad R_{(t)} = \begin{bmatrix} \frac{1}{1000000} & 0 \\ 0 & 1 \end{bmatrix}$$

For obtaining the observation equation for the quarterly GDP, we partition the vectors and the matrices in two different components: a component for the data observed ($y_t^{(1)}$) and another for the unobserved data or missing data $y_t^{(2)} = 0$, which is fixed to zero in order to maintain the dimensions of the equations (as the state equation and the observation equation would have different dimensions if just accounting for the data observed).

Model Parameters:

In this case, the model parameters are (ϕ, Q). ??? TODOS O SOLO LOS QUE HAY QUE ESTIMAR

Which method has been used to obtain the maximum likelihood estimators?

In this case, the method used to obtain the maximum likelihood estimators is the BFGS algorithm, which is a quasi-newton method that is based on a iterative algorithm for solving nonlinear optimization problems. It determines the descent direction by preconditioning the gradient curvature and approximates better and better the Hessian matrix for the loss function using the secant method.

Estimated Model:

In this case, the estimated model would be the following:

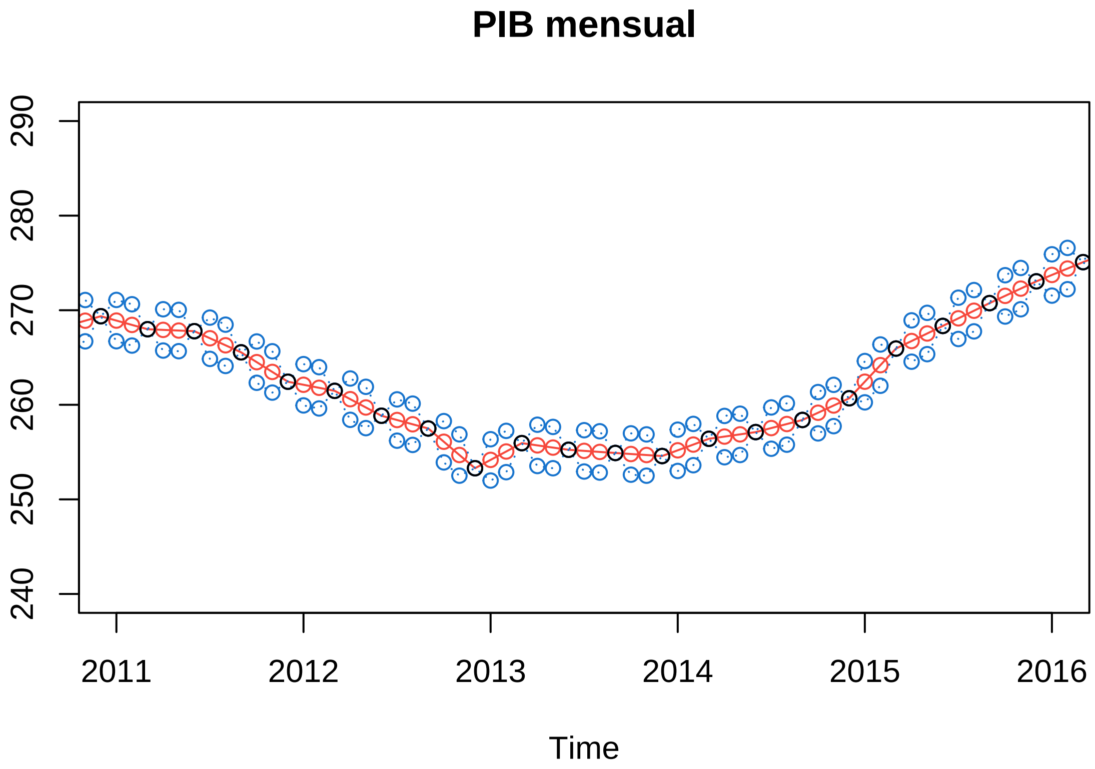
$$x_t = 1.002651 x_{t-1} + w_t \qquad w_t \sim N(0, 1.867361)$$

$$\begin{pmatrix} y_t^{(1)} \\ 0 \end{pmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} x_t + \begin{pmatrix} v_t^{(1)} \\ 0 \end{pmatrix} \implies y_{(t)} = A_{(t)} x_t + v_{(t)}$$

$$where \quad v_{(t)} \sim N_2(0, R_{(t)}) \quad and \quad R_{(t)} = \begin{bmatrix} \frac{1}{1000000} & 0 \\ 0 & 1 \end{bmatrix}$$

Final goal and interpretation/validation of the result

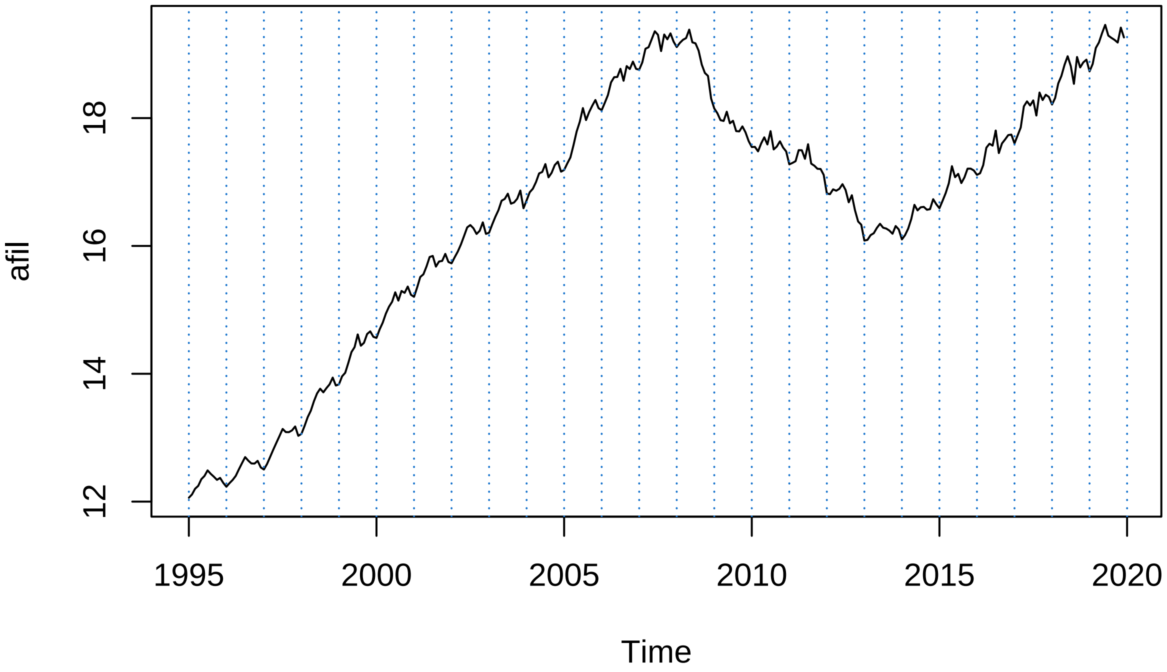
The final goal of this procedure is to obtain monthly values of the Spanish GDP, which is released quarterly. Hence, we want to find the "missing" values for the months which are not observed in the quarterly time series using a state-space model. The following graph shows the estimated values for the missing months (in red) and their confidence interval (in blue). In this case, one can see that the monthly GDP seems to be very "smooth", in the sense that there is no noticeable deviation from the behaviour exhibited by the original time series.



Phase II: Structural Time Series: Deseasonalisation of the affiliates Time series

Data

```
afil=ts(read.table("afiliados.csv",header=F)[,1]/1000000,start=c(1995,1),end=c(2019,12),freq=12)
plot(afil)
abline(v=1995:2023,lty=3,col=4)
```



Specify the state-space model

State:

The state in this setting is the set of monthly local trend, irregular and seasonal components of this time series. More specifically, there is the local trend component at T_t, the irregular component β_t at time t and the seasonal component S_t and its lags $S_{t-1}, S_{t-2}, ..., S_{t-10}$. The functional forms of each component would be the following:

$$T_t = T_{t-1} + \beta_{t-1} + w_t^T$$
$$\beta_t = \beta_{t-1} + w_t^\beta$$

$$S_t + S_{t-1} + S_{t-2} + S_{t-3} + S_{t-4} + S_{t-5} + S_{t-6} + S_{t-7} + S_{t-8} + S_{t-9} + S_{t-10} + S_{t-11} = w_t^S$$

Observations:

The observations in this setting are the monthly number of workers affiliated to the social security in Spain eliminating the seasonal patterns in the time series.

Transition Equation for the state:

$$x_t = \Phi x_{t-1} + w_t$$

$$where \quad x_t = \begin{pmatrix} T_t \\ \beta_t \\ S_t \\ S_{t-1} \\ ... \\ S_{t-10} \end{pmatrix}, \quad x_{t-1} = \begin{pmatrix} T_{t-1} \\ \beta_{t-1} \\ S_{t-1} \\ S_{t-2} \\ ... \\ S_{t-11} \end{pmatrix}, \quad \Phi = \begin{bmatrix} 1 & 1 & 0 & 0 & ... & 0 & 0 \\ 0 & 1 & 0 & 0 & ... & 0 & 0 \\ 0 & 0 & -1 & -1 & ... & -1 & -1 \\ 0 & 0 & 0 & 1 & ... & 0 & 0 \\ & & & & ... & & \\ 0 & 0 & 0 & 0 & ... & 1 & 0 \end{bmatrix}$$

$$and \quad w_t = \begin{pmatrix} w_t^T \\ w_t^\beta \\ w_t^S \\ 0 \\ ... \\ 0 \end{pmatrix} \sim N(0, Q) \quad with \quad Q = \begin{bmatrix} q_1^T & 0 & 0 & 0 & ... & 0 & 0 \\ 0 & q_1^\beta & 0 & 0 & ... & 0 & 0 \\ 0 & 0 & q_1^S & 0 & ... & 0 & 0 \\ 0 & 0 & 0 & 0 & ... & 0 & 0 \\ & & & & ... & & \\ 0 & 0 & 0 & 0 & ... & 0 & 0 \end{bmatrix}$$

where x_t denotes the monthly time series components (state) at time t (even though the lags are obviously not indexed at time t), Φ is the parameter matrix of the state equation, and w_t^T, w_t^β and w_t^S are iid Gaussian random variable with zero mean and q_1^T, q_1^β and q_1^S variances, respectively. Moreover, we assume that this state process starts with an initial value (an initial value for the different components), which is distributed as follows:

$$x_0 \sim N(\mu_0, \Sigma_0) \quad where \quad \mu_0 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ ... \\ 0 \end{pmatrix} \quad and \quad \Sigma_0 = \begin{bmatrix} 10 & 0 & 0 & 0 & ... & 0 & 0 \\ 0 & 10 & 0 & 0 & ... & 0 & 0 \\ 0 & 0 & 10 & 0 & ... & 0 & 0 \\ 0 & 0 & 0 & 10 & ... & 0 & 0 \\ & & & & ... & & \\ 0 & 0 & 0 & 0 & ... & 0 & 10 \end{bmatrix}$$

Observation Equation:

$$y_t = \begin{pmatrix} 1 & 0 & 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} T_t \\ \beta_t \\ S_t \\ S_{t-1} \\ ... \\ S_{t-10} \end{pmatrix} + v_t \quad where \quad v_t \sim N(0, R)$$

In this case, the observed components are both the local trend component and the seasonality at moment t, as the observation is the complete series and we do not have no information about β_t or the lags of S_t.

Model Parameters:

In this case, the model parameters that have to be estimated are $(q_1^T, q_1^\beta, q_1^S, R)$, which are the variances of the random terms for both equations.

Which method has been used to obtain the maximum likelihood estimators?

In this case, the method used to obtain the maximum likelihood estimators is the BFGS algorithm, which is a quasi-newton method that is based on a iterative algorithm for solving nonlinear optimization problems. It determines the descent direction by preconditioning the gradient curvature and approximates better and better the Hessian matrix for the loss function using the secant method.

Estimated Model:

In this case, the estimated model would be the following:

$$x_t = \Phi x_{t-1} + w_t$$

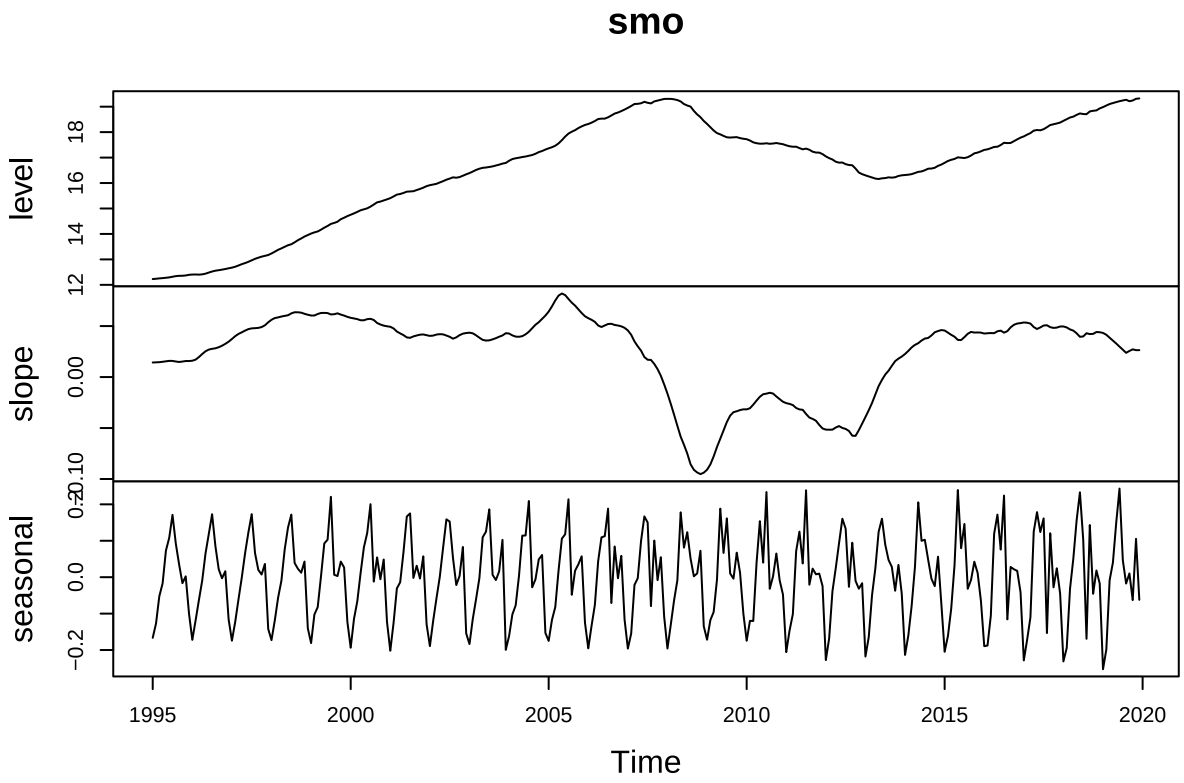
$$where \quad x_t = \begin{pmatrix} T_t \\ \beta_t \\ S_t \\ S_{t-1} \\ ... \\ S_{t-10} \end{pmatrix}, \quad x_{t-1} = \begin{pmatrix} T_{t-1} \\ \beta_{t-1} \\ S_{t-1} \\ S_{t-2} \\ ... \\ S_{t-11} \end{pmatrix}, \quad \Phi = \begin{bmatrix} 1 & 1 & 0 & 0 & ... & 0 & 0 \\ 0 & 1 & 0 & 0 & ... & 0 & 0 \\ 0 & 0 & -1 & -1 & ... & -1 & -1 \\ 0 & 0 & 0 & 1 & ... & 0 & 0 \\ & ... & ... & ... & ... & ... & ... \\ 0 & 0 & 0 & 0 & ... & 1 & 0 \end{bmatrix}$$

$$and \quad w_t = \begin{pmatrix} w_t^T \\ w_t^\beta \\ w_t^S \\ 0 \\ ... \\ 0 \end{pmatrix} \sim N(0, Q) \quad with \quad Q = \begin{bmatrix} 0.03990327 & 0 & 0 & 0 & ... & 0 \\ 0 & 0.009785646 & 0 & 0 & ... & 0 \\ 0 & 0 & 0.03055968 & 0 & ... & 0 \\ 0 & 0 & 0 & 0 & ... & 0 \\ ... & ... & ... & ... & ... & ... \\ 0 & 0 & 0 & 0 & ... & 0 \end{bmatrix}$$

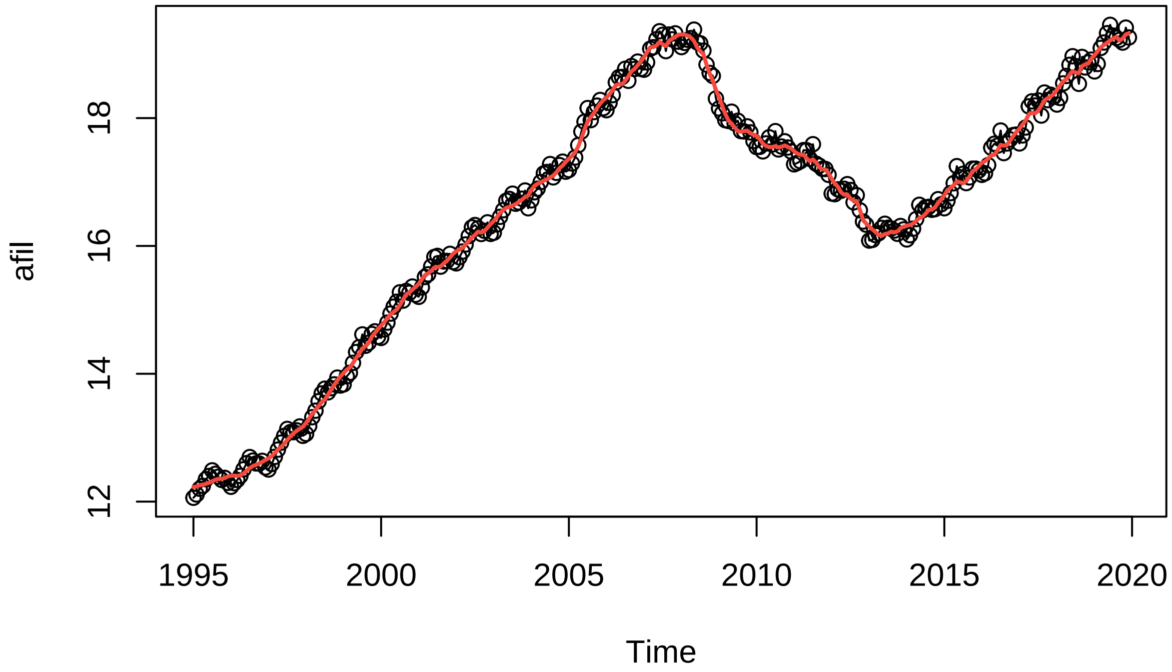
$$y_t = \begin{pmatrix} 1 & 0 & 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} T_t \\ \beta_t \\ S_t \\ S_{t-1} \\ ... \\ S_{t-10} \end{pmatrix} \quad as \quad v_t \sim N(0, 0) = 0$$

Final goal and interpretation/validation of the result

The final goal of this procedure is to estimate both the local trend, the irregular and the seasonal component in order to obtain a time series without the seasonal pattern. The procedure yields the estimation of the following components, which are represented graphically in the next diagram:



Hence, once we have the estimation for each component, we can substract the seasonal component for the original series and obtain the following series without seasonality, shown in red:

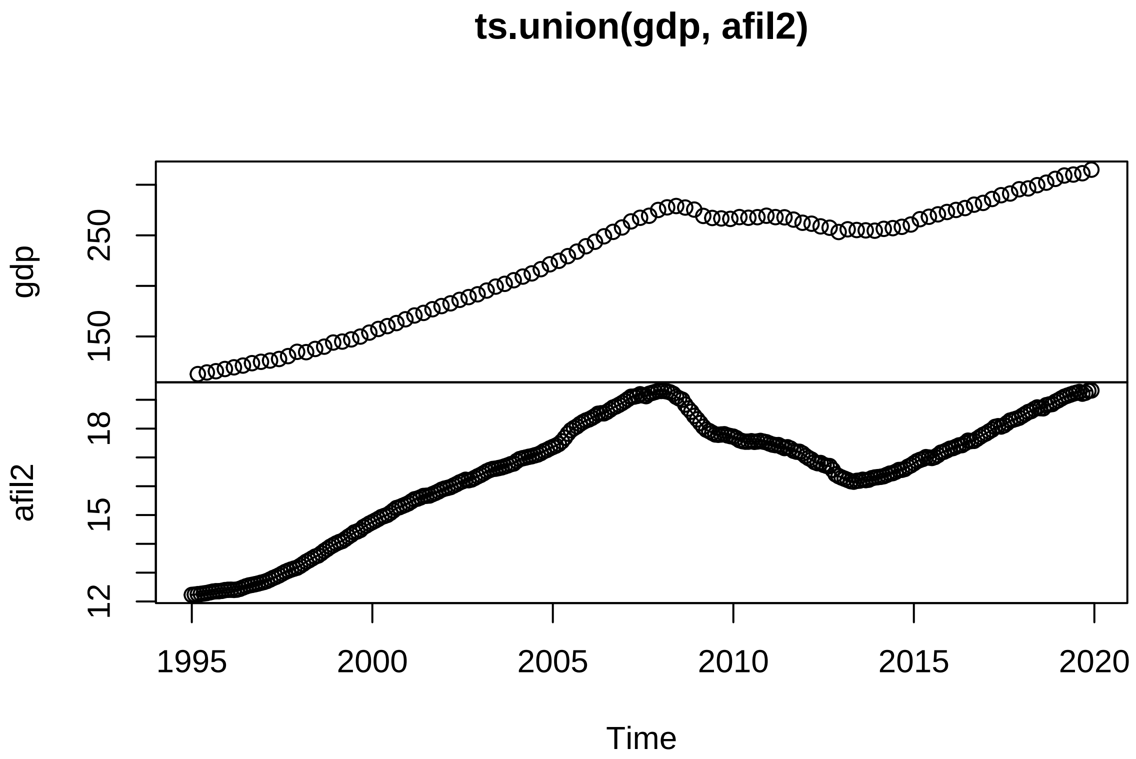


As we can see, the effect of this elimination is that there is no small variations inside the span of a year, which account for months where employment is increased (summer) and reduced (after summer), and hence is smoother than the original time series.

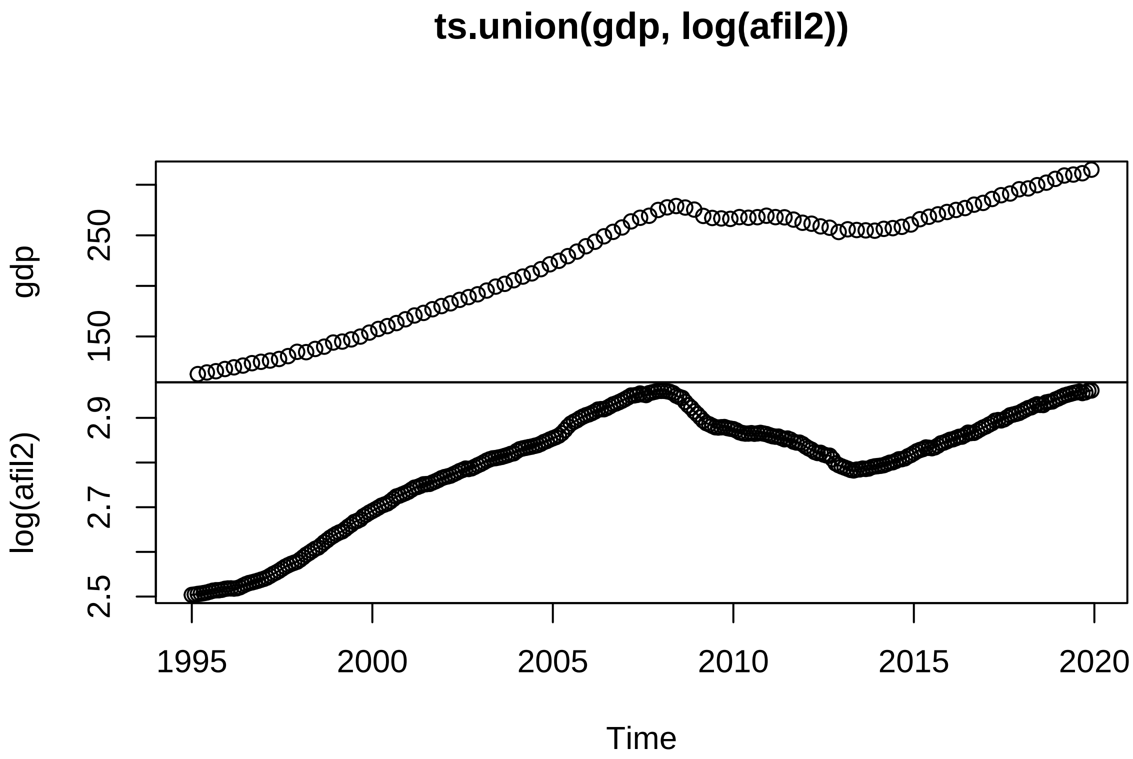
Phase III: Multivariate Time Series. Monthly GDP by using the affiliates time series

Data

```
plot(ts.union(gdp,afil2),type="o")
```



```
plot(ts.union(gdp,log(afil2)),type="o")
```



```r
serie=cbind(gdp,log(afil2))
dimnames(serie)[[2]]=c("GDP","lnAfil")
```

Specify the state-space model

State:

The state in this setting is the monthly GDP of Spain (at current prices). However, we note that this is not the same case as Part I, because the transition equation will try to estimate the monthly GDP of Spain taking into account only the quarterly data and taking into account both the quarterly GDP and the monthly number of affiliates (for comparison).

Observations:

The observations in this setting are both the quarterly GDP of Spain (at current prices) and the monthly number of affiliates to the social security in Spain.

Transition Equation for the state:

$$x_t = \Phi x_{t-1} + w_t \qquad w_t \sim N_2(0, Q)$$

$$where \quad x_t = \begin{pmatrix} x_t^{(1)} \\ x_t^{(2)} \end{pmatrix}, \quad x_{t-1} = \begin{pmatrix} x_{t-1}^{(1)} \\ x_{t-1}^{(2)} \end{pmatrix}, \quad \Phi = \begin{bmatrix} 1 & \phi_{12} \\ \phi_{21} & 2 \end{bmatrix}$$

$$and \quad w_t = \begin{pmatrix} w_t^{(1)} \\ w_t^{(2)} \end{pmatrix} \quad with \quad Q = \begin{bmatrix} 2 & q_{12} \\ q_{21} & 1 \end{bmatrix}$$

where $x_t^{(1)}$ denotes the monthly GDP (state) at time t that will be estimated using only the quarterly GDP, $x_t^{(2)}$ denotes the monthly GDP (state) at time t that will be estimated using both the quarterly GDP and the monthly affiliates, and w_t is an iid Gaussian random vector with zero mean and Q covariance matrix (interpreted analagously as the x_t vector).

Given that we do not have $\phi_{21}, \phi_{12}, q_{21}$ nor q_{12}, we fix these values to zero and then we will obtain estimates for these.

Moreover, we assume that this state process starts with an initial value (an initial value of the monthly GDP), which is distributed as follows:

$$x_0 \sim N(\mu_0, \Sigma_0) \quad where \quad \mu_0 = \begin{pmatrix} 112.76000 \\ 12.25371 \end{pmatrix} \quad and \quad \Sigma_0 = \begin{bmatrix} 100 & 0 \\ 0 & 100 \end{bmatrix}$$

Observation Equation:

$$\begin{pmatrix} y_t^{(1)} \\ y_t^{(2)} \end{pmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} x_t + \begin{pmatrix} v_t^{(1)} \\ v_t^{(2)} \end{pmatrix}$$

$$where \quad v_t \sim N_2(0, R) \quad and \quad R = \begin{bmatrix} 1e-06 & 0 \\ 0 & 1e-06 \end{bmatrix}$$

For obtaining the observation equation for the quarterly GDP and the monthly affiliates, we partition the vectors in two different components: a component for the data of quarterly GDP ($y_t^{(1)}$) and another for the data of monthly affiliates $y_t^{(2)}$.

Model Parameters:

In this case, the model parameters to be estimated are $(\phi_{21}, \phi_{12}, q_{21}, q_{12})$.

Which method has been used to obtain the maximum likelihood estimators?

In this case, the method used to obtain the maximum likelihood estimators is the EM algorithm, which is based on a step of expectation (creates a function for expectation of log-likelihood for current estimates) and another of maximization (computes parameters based on maximizing the expected log-likelihood). There are explicit equations developed for applying this algorithm in different context, but all are based on the same idea of obtaining maximum likelihood estimates when the model depends of unobserved latent variables.

Estimated Model:

$$x_t = \Phi x_{t-1} + w_t \qquad w_t \sim N_2(0, Q)$$

$$where \quad x_t = \begin{pmatrix} x_t^{(1)} \\ x_t^{(2)} \end{pmatrix}, \quad x_{t-1} = \begin{pmatrix} x_{t-1}^{(1)} \\ x_{t-1}^{(2)} \end{pmatrix}, \quad \Phi = \begin{bmatrix} 0.9915730730 & 0.1568424 \\ -0.0005000956 & 1.0082471 \end{bmatrix}$$

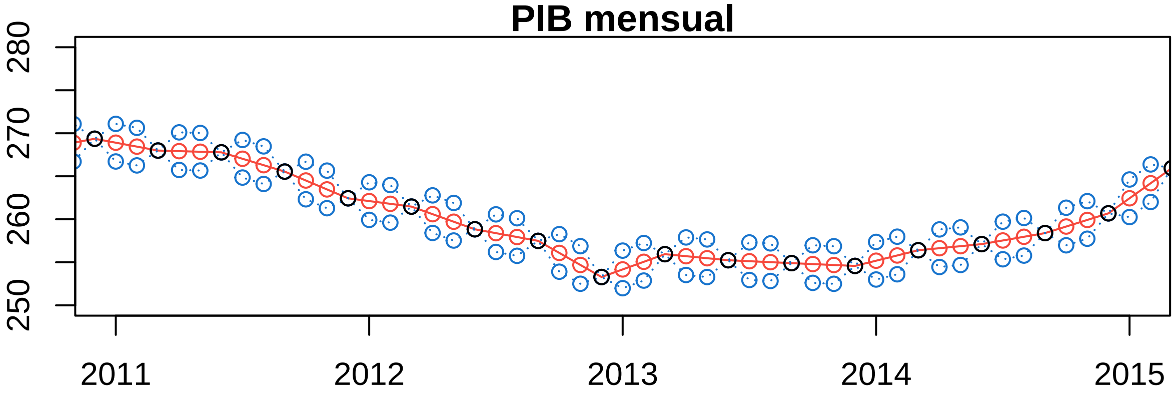
$$and \quad w_t = \begin{pmatrix} w_t^{(1)} \\ w_t^{(2)} \end{pmatrix} \quad with \quad Q = \begin{bmatrix} 0.9413148 & 0.026373498 \\ 0.0263735 & 0.002238424 \end{bmatrix}$$

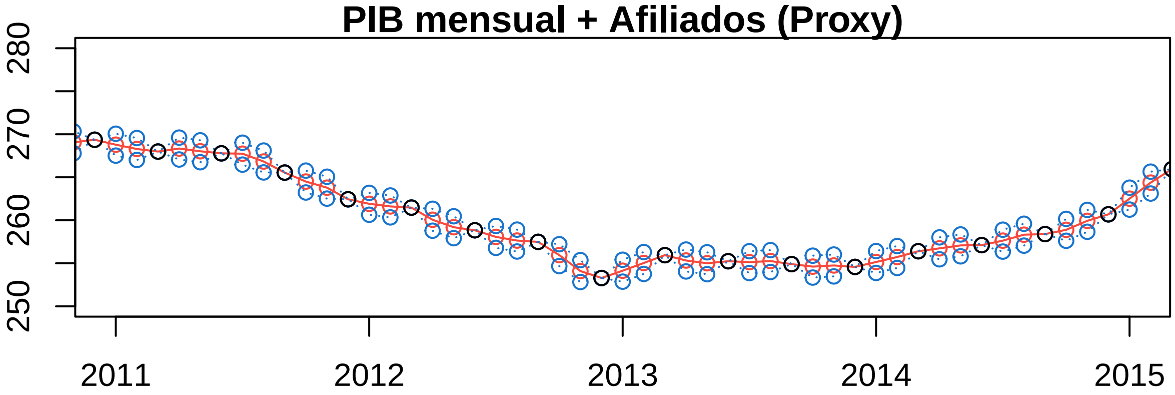
$$\begin{pmatrix} y_t^{(1)} \\ y_t^{(2)} \end{pmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} x_t + \begin{pmatrix} v_t^{(1)} \\ v_t^{(2)} \end{pmatrix}$$

$$where \quad v_t \sim N_2(0, R) \quad and \quad R = \begin{bmatrix} 1e-06 & 0 \\ 0 & 1e-06 \end{bmatrix}$$

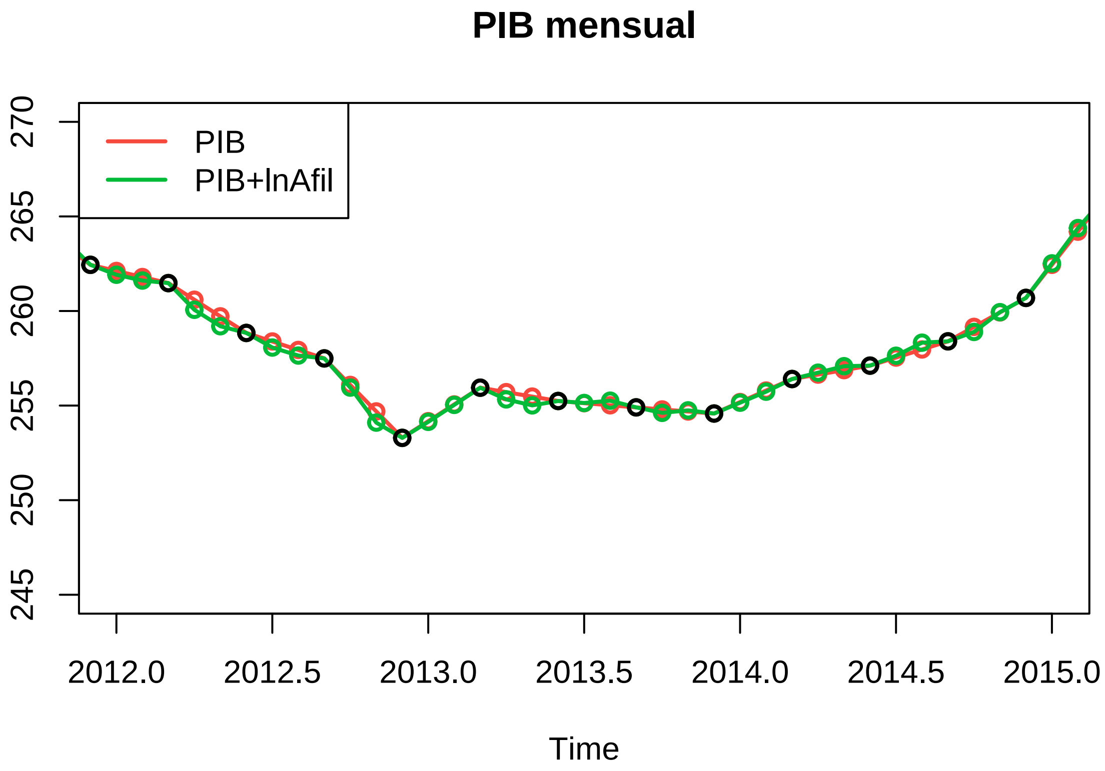
Final goal and interpretation/validation of the result

The final goal of this model is to compare the obtention of the monthly GDP data for Spain using just the quarterly data with the obtained time series by using the quarterly data and the monthly data of affiliates. Both of this approaches are represented in the following graphs:





We can see how the confidence intervals from the first are wider than that of the second obtained time series, showing there is less variability. By looking at the following graph, we can see how there are subtle differences between each estimated serie, even though there are pretty similar.


PIB mensual
PIB
PIB+lnAfil
270
265
260
255
250
245
2012.0
2012.5
2013.0
2013.5
2014.0
2014.5
2015.0
Time