Intelligent Data Analysis - Homework 3

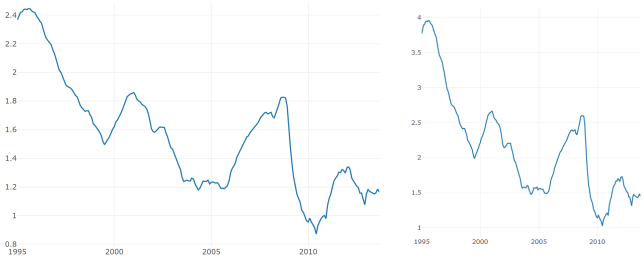
Panagiotis Michalopoulos, Javier de la Rua, Michail Gongolidis,
Ignacio Rodriguez, Daniel Minguez

January 16h, 2019

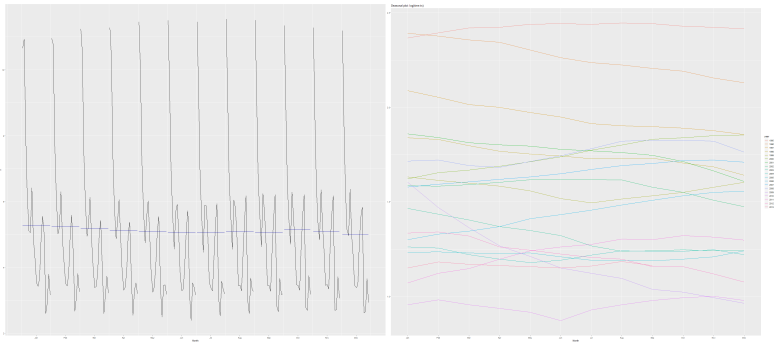
1 Plot the series and comment

As we are group 10, we have used the file with data related to the monthly interest rate on unsubsidised loans for house purchase (percentage) from January 1995 to September 2013. The first thing we did was plot the series:

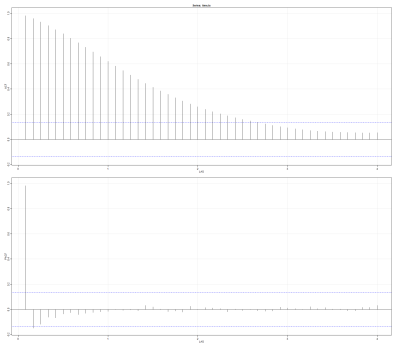
We can appreciate that the time series has an obvious global downward trend without seasonality (so it is not a stationary series). We can appreaciate also upward subtrends in some intervals. After applying the log transformation (left) and the box-cox transformation (right) to the data we obtain the following result:



It does not make so much change, neither in trend nor in variability, so we decided to work with the original time series. We can appreciate also that the time series is not stationary and that it does not have a seasonal trend by checking the seasonal differences and the month differences:



Finally we also used the autocorrelation and partial autocorrelation plots:

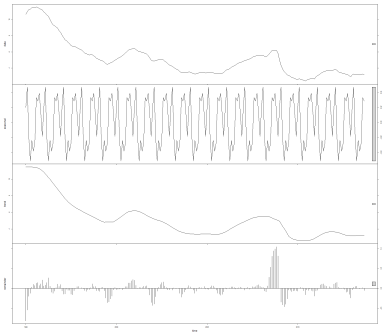


We can see that ACF does not drop quickly to values inside or almost inside the critical values and that r_1 is large and positive. All this are signs of a non-stationary time series. PACF value r_1 is almost 1. This is also a sign of non-stationary process. Therefore the data should be differenced to apply a model. We applied also a stationary test, with a p-vlaue lower than 0.05 we reject the null hypothesis of stationarity:

```
> adf.test(time.ts)

    Augmented Dickey-Fuller Test

data:  time.ts
Dickey-Fuller = -3.4445, Lag order = 6, p-value = 0.04895
alternative hypothesis: stationary
```

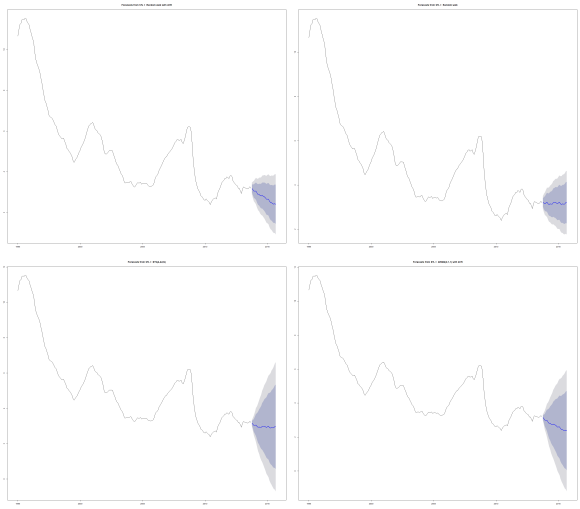
2 Obtain a plot of the decomposition of the series

Since the log transformation does not change the variance of our time series and our series does not have seasonality, we use an additive decomposition approach with *stl* with the parameters *s.window="periodic", t.window=15, robust=TRUE* to the log data. We can appreciate the downtrend and the differences of the remainder across time.:



2.1 Use the function forecast() to forecast future values.

We have used the forecast function with Random walk with drift, random walk, ETS and ARIMA:



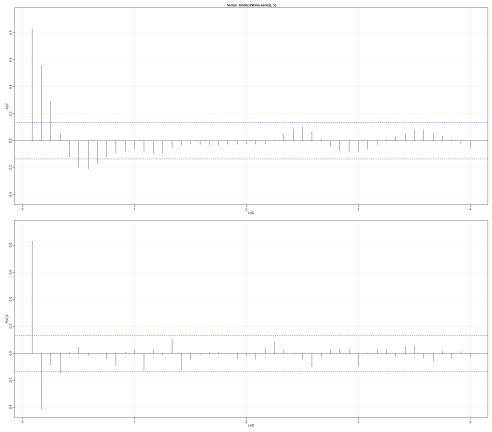
We can appreciate that the confidence intervals of ETS and ARIMA are bigger that the one created with random walk variants.

2.2 Multiplicative decomposition is useful when the logarithmic transformation makes a difference in the appearance of your series and shows more constant variance. Does the remainder look like a white noise to you?

As we saw in the first part, the log transformation does not change the variance of our data, so we applied the additive decomposition "*Time series = Seasonal + Trend + Random*" to the original series

2.3 White noise is just a group of independent, identically distributed variables, with zero mean and constant variance. Answer to this point just visually or plot the ACF and PACF of the remainder part.

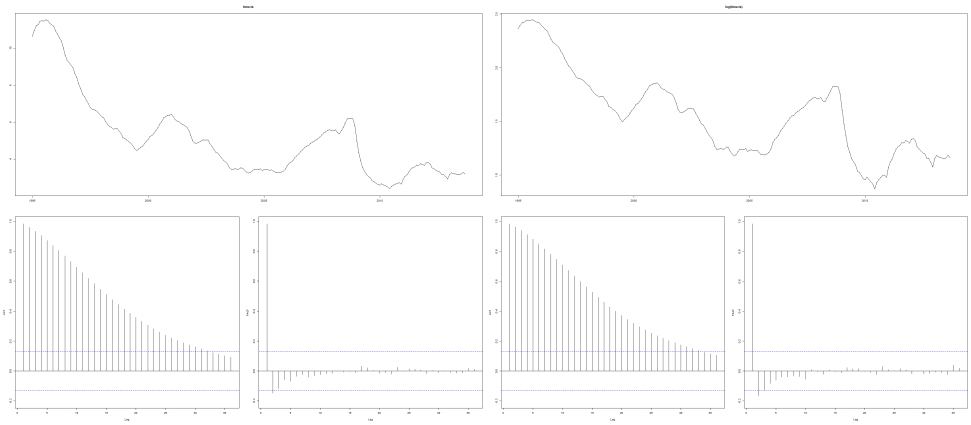
We can see in the remainder plot that there are some parts where the differences does not seem to be random noise because of its variance, using the ACF and PCAF graph we can check as well that the r_1 is close to 0.6 and -0.4 respectively, so we can say that the remainder is not white noise:



3 Fit an ARIMA model to your time series

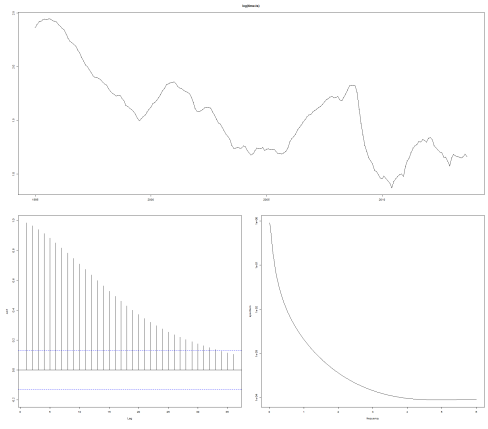
3.1 Decide on whether to work with your original variable or with the log transform one

As we have seen in the previous part, we are going to use the original data. The following is the tsdisplay of the data and log data:



3.2 Are you going to consider a seasonal component?

We are not going to use a seasonal component since, as we checked in the previous part, there is no seasonal component in our series. Using the periodogram we can also intepret this:

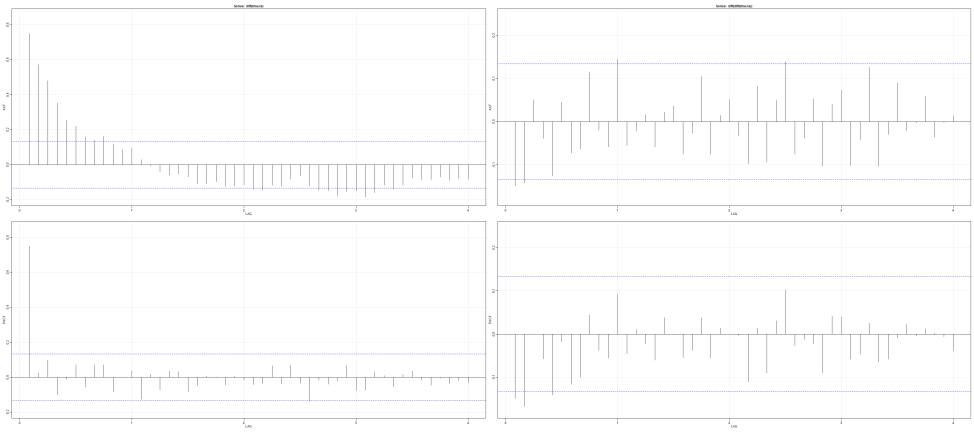


3.3 Decide on the values of d and D to make your series stationary.

In our case, since we do not have a seasonal component, we have $D = 0$, in order to decide about d we took first, second and third differences with diff function, obtaining the following standard deviations:

Difference	Standard deviation
Original log data	2.239662
1st	0.1486072
2nd	0.1030107
3rd	0.1562024

So we have to decide between 1st and 2nd differences by checking ACF and PACF, since both series are stationary by the adf.test (with p-values lower than 0.05, and Dickey-Fuller of -4.3715 and -7.6347 respectively):



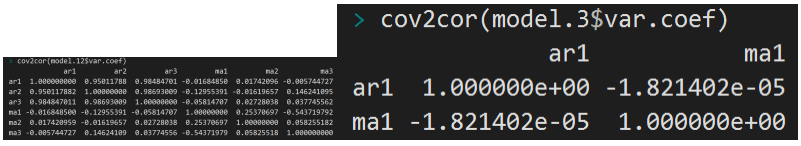
So we have decided to try with second order differences.

3.4 Identify values for p and q for the regular part and P and Q for the seasonal part.

We have fitted different models for different p-q parameters to the train data (data before 2013), we have also calculated the RMSE with the test data (2013 data):

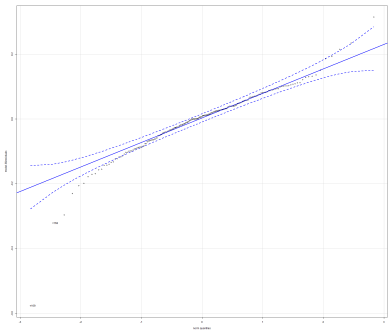
ARIMA Model	AIC	BIC	RMSE
(1,2,0)	-376.4445	-369.7126	0.8832099
(0,2,1)	-377.8681	-371.1361	0.8473354
(1,2,1)	-389.8618	-379.7638	0.51469
(2,2,0)	-378.3003	-368.2024	0.8324347
(0,2,2)	-379.3241	-369.2262	0.8028774
(2,2,1)	-388.1842	-374.7203	0.5174389
(1,2,2)	-388.2293	-374.7654	0.5180392
(2,2,2)	-385.872	-369.0421	0.5150287
(3,2,0)	-376.3377	-362.8738	0.8291334
(3,2,1)	-374.7115	-357.8816	0.8475045
(3,2,2)	-385.057	-364.8612	0.5252064
(3,2,3)	-394.4765	-370.9147	0.4526032
(0,2,3)	-377.8079	-364.344	0.7863561
(1,2,3)	-386.5934	-369.7636	0.5257217
(2,2,3)	-384.8269	-364.631	0.5235576

So, apparently the best model was ARIMA(3,2,3) with the lowest AIC and ARIMA(1,2,1) with the lowes BIC. But if we check the correlation between the coeficients of ARIMA(3,2,3) we obtain high correlation between ma1 and ma3 and ar coefficients. We include also the result with ARIMA(1,2,1), obtaining the following correlation:



```
> cov2cor(model.3$var.coef)
                ar1           ma1
ar1  1.000000e+00 -1.821402e-05
ma1 -1.821402e-05  1.000000e+00
```

Regarding the residuals we checked for normality with test with jarque bera test, but we obtained that the residuals are not normal. We look for outliers:

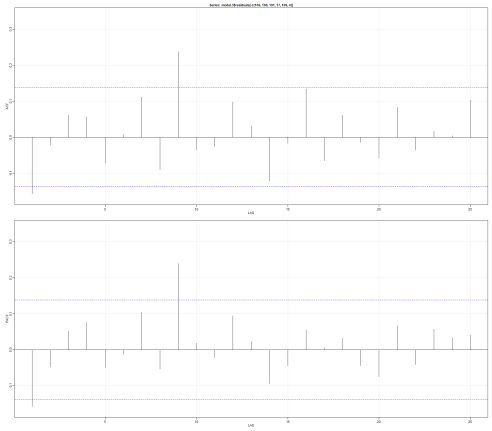


After removing some ouliers (we would need to find an explanation for them), we obtain normal residuals (although the mean test say that mean is not equal to 0):

```
> jarque.bera.test(model.3$residuals[-c(169, 168, 191, 17,189,4)])

        Jarque Bera Test

data:  model.3$residuals[-c(169, 168, 191, 17, 189, 4)]
X-squared = 5.691, df = 2, p-value = 0.0581
```

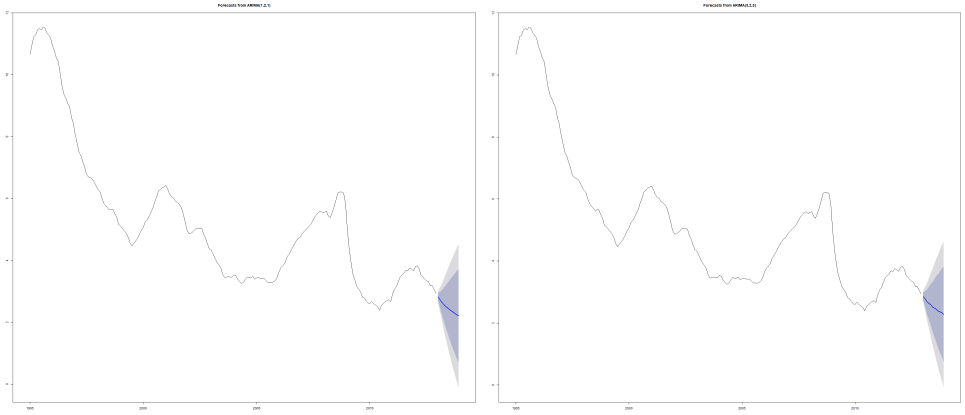
```
> t.test(model.3$residuals[-c(169, 168, 191, 17, 189, 4,187,183)])

        One Sample t-test

data:  model.3$residuals[-c(169, 168, 191, 17, 189, 4, 187, 183)]
t = 0.7909, df = 207, p-value = 0.4299
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.006465853  0.015129010
sample estimates:
  mean of x
0.004331579
```

Also we have plot the autocorrelations



3.5 Once you have found a suitable model, repeating the fitting model process several times if necessary, use it to make forecasts. Plot them.

We have done forecast with ARIMA(3,2,3) and ARIMA(1,2,1):



3.6 Use the function getrmse to compute the test set RMSE of some of the models you have already fitted.

As we can see in the previous section, we obtained that the model ARIMA(3,2,3) is the one that get better RMSE, although the correlation of some of the coeficients is close to 1.

3.7 You can also use the auto.arima() function with some of its parameters fixed, to see if it suggests a better model that the one you have found.

If we execute the auto.arima function we obtain the following result

```
> ggb.auto = auto.arima(train, d=2, D=0, max.p=3, max.q=3, max.P=0, max.Q=0)
  ggb.auto
Series: train
ARIMA(0,2,1)(0,0,1)[12]

Coefficients:
          ma1    sma1
      -0.1807  0.1070
s.e.   0.0775  0.0729

sigma^2 estimated as 0.009818:  log likelihood=192
AIC=-378   AICc=-377.88   BIC=-367.9
```