# Homework 3

Time Series

**Statistical Data Analysis – 10th of January, 2020**

## Group 5

Antoñanzas Martínez, Guillermo
Pueblas Núñez, Rodrigo
Martínez Capella, Ignacio
Burrell, David

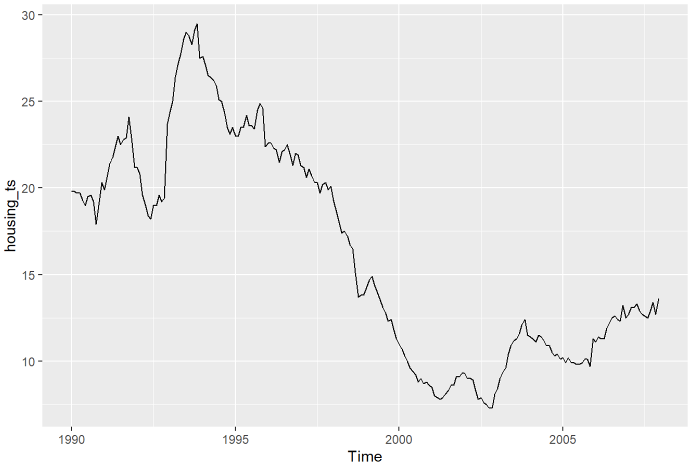
## Brief data description, preparation and manipulation

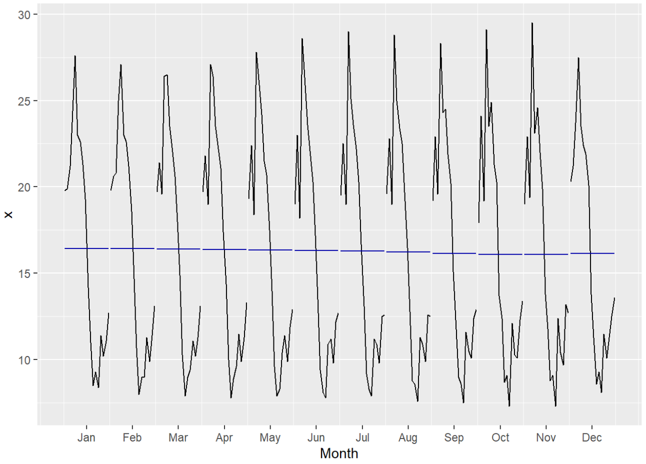
We have been assigned the data set "data_g5.xlsx". In this dataset describes the percentage of subsided housing approvals per month. Observations were taken at the end of each period and correspond to data from January 1990 to December 2007. The source is Banco de España (www.bde.es). No cleaning was required as only two columns, Date and Percentage, were present.

# Research questions, plots and findings

## Question 1

**Plot the series and briefly comment on the characteristics you observe (stationary, trend, seasonality…).**





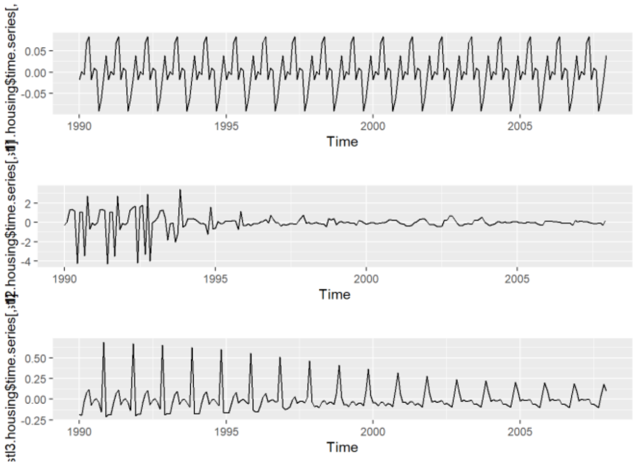
We can address each of the properties separately:

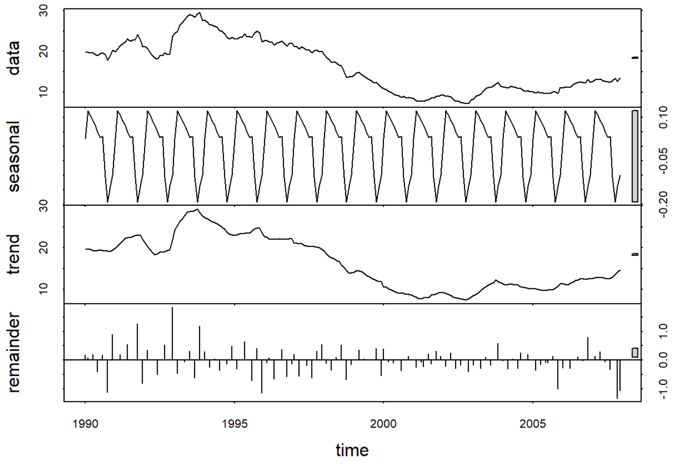
- Trend: There is a trend over time, and we can see different patterns. There is a positive trend between 1991 and 1993, then between 1993 and 2003 a negative trend occurs, followed by a positive slope from 2003 and onwards.
- Seasonality: As there is no regularly repeating pattern no seasonality is present.
- Cyclical components: There is no evidence of any cyclic behaviour in the data.
- Stationary: Due to the presence of a trend this series is non-stationary, even with no seasonal or cyclical components.

## Question 2

**2a) Obtain a plot of the decomposition of the series, using stl(). Use an additive decomposition or a multiplicative one, depending on your data.**

An additive decomposition has been applied since there are no changes in seasonal components. Using the first seasonal window (s.window = periodic), it is seen the seasonal components are small, centred around zero and more or less constant. These results show the seasonal components have little influence this window will be used.

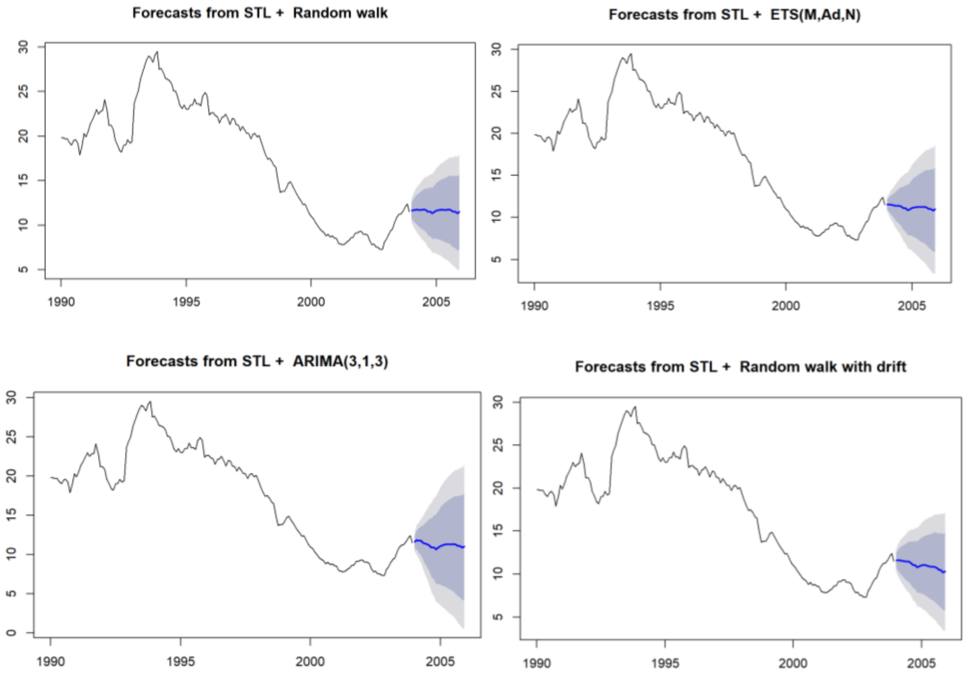




Keeping a periodic window for the seasonal components (assuming they are equal across the years), we can see that using a window with a value of 5 produces the trend that best captures the structure of the data. Thus, it will be chosen.

Observing the decomposition plot produced by the selected windows, we can see that the seasonal components have low constant values that are centred around zero ([-0.192, 0.124]). The structure of the data is captured well by the trend component and the seasonal adjusted series matches the original one (as shown in the attached Rmd file). The remainder values are somewhat large ([-1.34, 1.84]), but have zero mean (-0.0153).

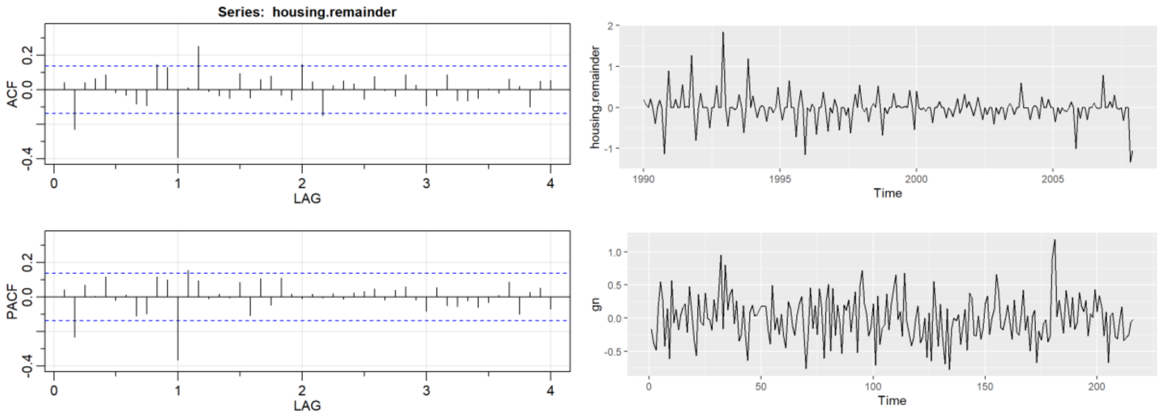
### Use the function forecast() to forecast future values.



Using stl() to forecast, we can see that using the random walk with drift method to predict future values results in the lowest RMSE.

**Does the remainder look like a white noise to you?**

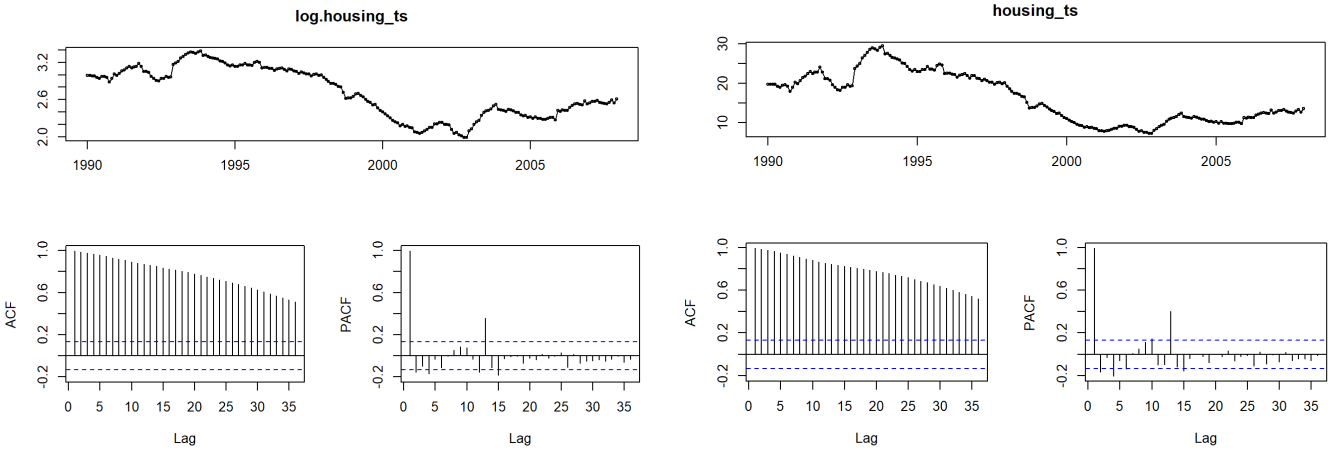
The remainder does not look like gaussian noise, since it does not have a constant variance. In the preceding years, the data has much more variance than in the most recent ones.



## Question 3

**Fit an ARIMA model to your time series. Some steps to follow:**

### 3a) Decide on whether to work with your original variable or with the log transform one.



Applying the log transformation has little effect on the data so the original will be used.

### 3b) Are you going to consider a seasonal component? If the answer is yes, identify s.

As previously stated in question 1 we can see that there is no seasonality. Thus, no season component will be considered, and P, D and Q will be set to 0 for all of our models.

### 3c) Decide on the values of d and D to make your series stationary.

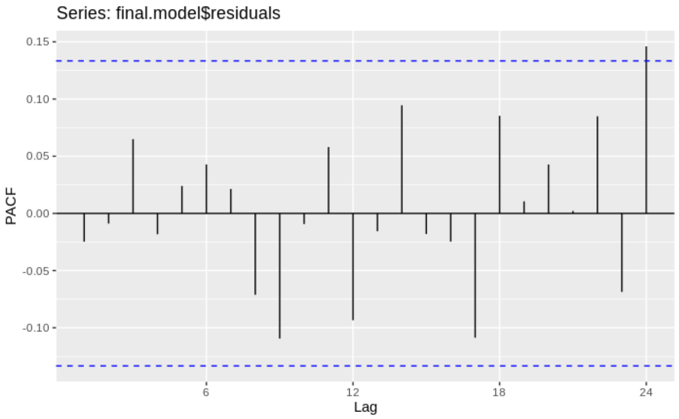
Using the ndiffs() method the estimated number of differences is 1. To confirm this the difference of the series was created using diff() and tests applied to confirm the new data is stationary. Applying the

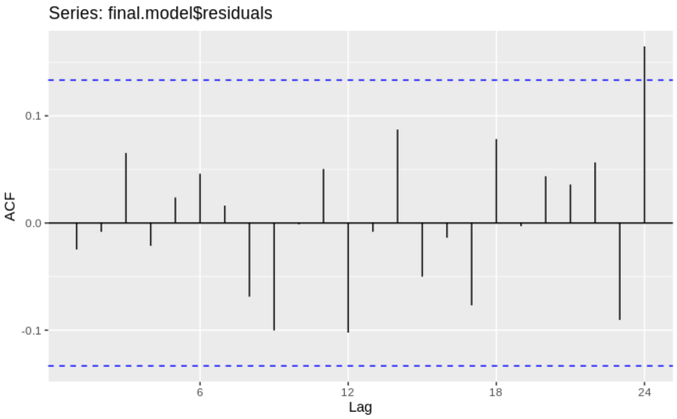
Augmented Dicky-Fuller and KPSS tests return P-values of less than 0.01 indicating the series is indeed stationary. Thus we use d= 1 for the following models.

### 3d) Identify values for p and q for the regular part

Upon investigating a range of different values for p and q it was discovered that the model with the best AICc score of 307.092 was the configuration p = 0, d = 1 and q = 12. Upon inspection of the correlation between coefficients for the model it is seen that the largest correlation is 0.5276921 which is an acceptable value.

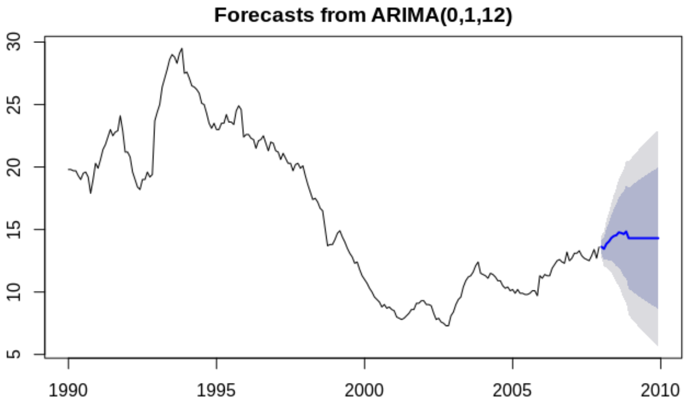
### 3e) Make diagnostic of the residuals for the final model chosen (autocorrelations, zero mean, normality).





Plotting the residuals for the model described above it can be seen that there is no obvious pattern in the series. Calculating the mean of the values gives the acceptable near zero value of -0.0132. Plotting the autocorrelation of the residuals it can be seen that most of the relationships have been captured apart from that at lag 24. Although it passes the threshold it is only slights so it acceptable (attempts were made to remove this using a q = 24 but lead to worse models overall). Though performing well on the other metrics, the residuals are shown to be not normal using a Jarque Bera Test which gives a P-value failing the null hypothesis.

### 3f) Once you have found a suitable model, repeating the fitting model process several times if necessary, use it to make forecasts.

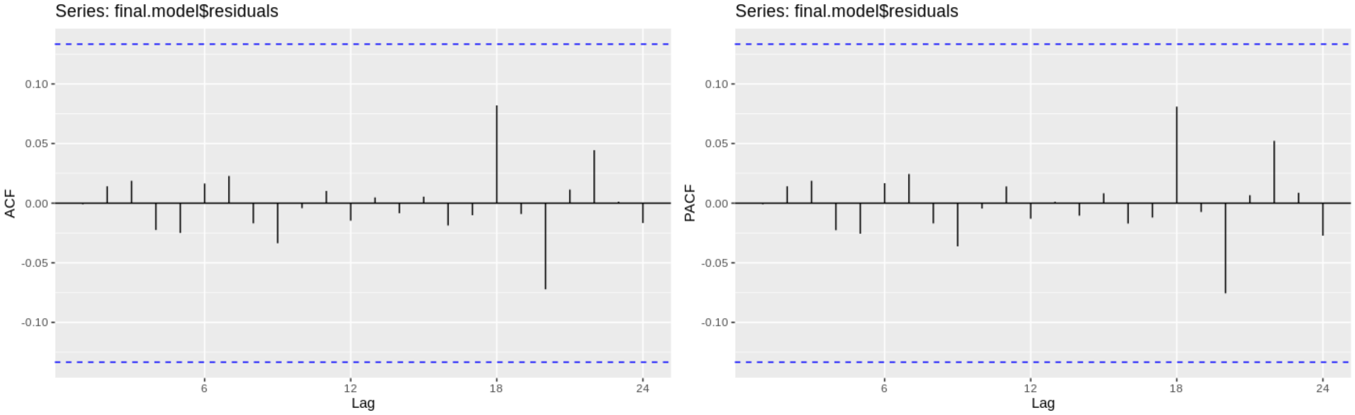


After repeating the fitting process several times, we get to the shown plot. As we see in the image, the dark grey area of the prediction means that the trend will be within those boundaries with an 80% probability. The light grey area represents a confidence interval with a 95% probability.

**3g) Use the function getrmse to compute the test set RMSE of some of the models you have already fitted. Which is the one minimizing it?**

Using the method getrmse on the described model (p = 0, d = 1 and q = 12) an RMSE value of 0.3047258. It is found by increasing p the value is reduced. With p = 1 a RMSE of 0.2977918 is given, and p = 2 a value of 0.2937613 is given. But with a p = 3 the RSME begins to increase, with a value of 0.3091569.

**3h) 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.**



Using the auto.arima(), configured to have a maximum p and q of 24 and d of 2, a model with the variables p = 1, d = -1 and q = 1 is generated. This model as a worse AICc than our chosen model, 397.22, and slightly worse RMSE, 0.3091569. Looking at the correlation between the coefficients it can be seen that there is a value of -0.8810521 between the ar and ma values, which is not desired. Although with this model performing ACF and PACF upon the residuals shows that at all lag intervals the values are within the boundaries.