Relation between Google queries about households, households transactions and land prices in Spain

Alejandro Encalado Masia

21/12/2018

Abstract

Since the arrival of internet in our lives, a lot of things have changed over the time. Nowadays we are able to buy at the supermarket without leaving our room, we don't need newspapers anymore, we have the lasts news of all arround the world in just one click. The fact is that internet has changed the world and in consequence the interaction between people and the world. This is why we have decided in this assignment to have an overview about how searching queries in Google can have a direct relation with the real world of economics. Here is the discussion about the relation between land prices in Spain from 2007 to 2016, the total number of households transactions in the same period and the Google queries of households topics, such as "Flats" or "Houses" for example. We will take a look to the correlations between this diferent data sets, and we will perform a simple prediction to the google queries over time. With all of this knowledge we will discuss how households queries have impacted to land prices.

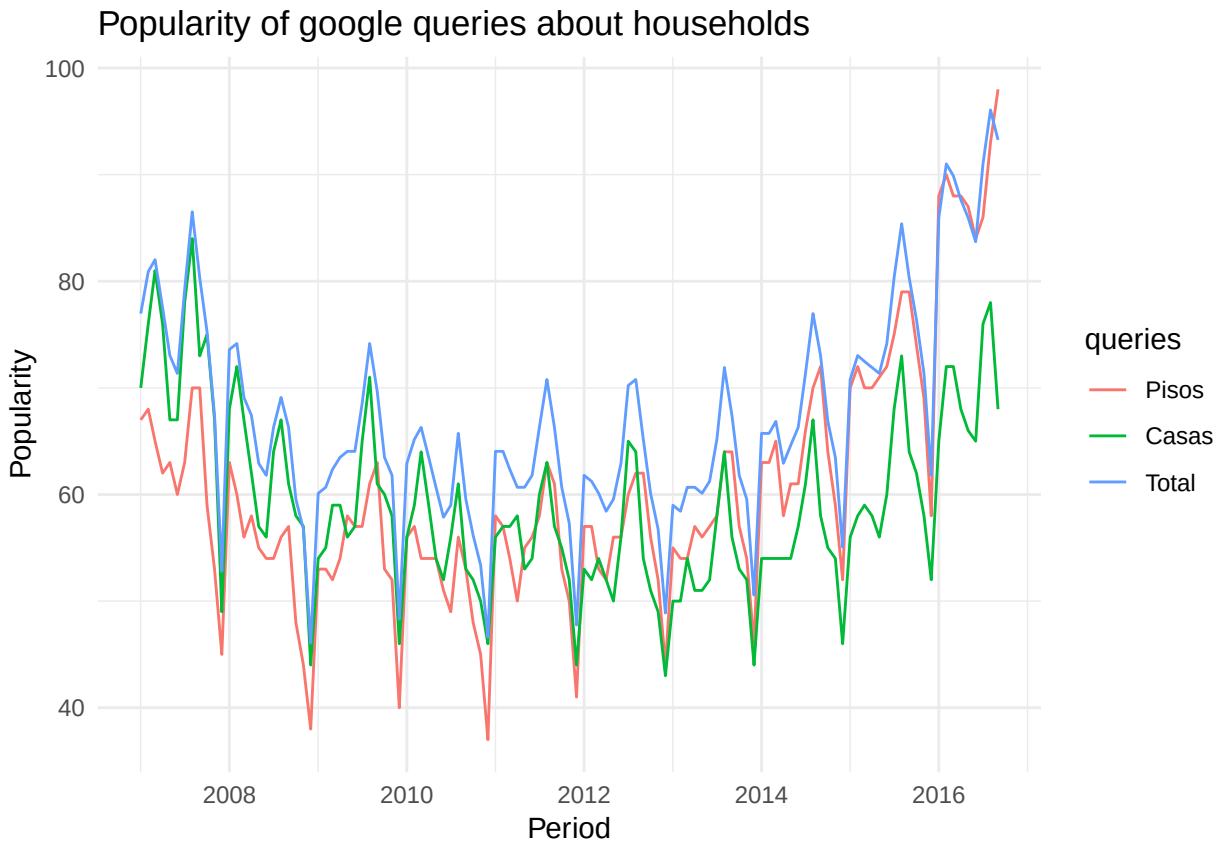
Introduction

In the assignment we pretend to analyse the behaviour of three time series that could be related between them. The main idea is to study the relation between land prices in Spain from January 2007 to September 2016 with the total amount of household transactions in Spain at the same period. And also the relation between land prices mentioned before with the Google households queries. Our hypothesis is that these two data sets are highly correlated due to the law of supply and demand. Moreover, the interesting point of this analysis will be the study of the relation between the total amount of households transactions in Spain and the "Popularity" of the queries on Google containing the words "Pisos" and/or "Casas". This popularity represents, with a number from 1 to 100, how people are interested in this kind of topic in a certain period. Moreover, we think that people tends to search more information on internet about a certain topic when they are more interested in, in this case, if someone is interested in buying a house or a flat, he or she will increase his or her searches on google and once the house or the flat is bought, they stop searching more information on internet. So this could handle into a correlation between this two lasts data sets, if this is true, the trend of the google queries can become a predictor or an indicator for the house prices, and it could be interesting in a economical point of view. The first sample we have used is a time series dataset coming from INE (Instituto Nacional de Estadistica), which contains information about the number of households sells from 2007 to 2016 and information about the price of every quarter from 2007 to 2016 as we mentioned. Then, the last one comes from Google Trends webpage, and it is a time dataset sequence that contains information about the "popularity" of a certain query along the time in Spain. In our case, we have choose those queries which contains the word "Pisos" and "Casas".

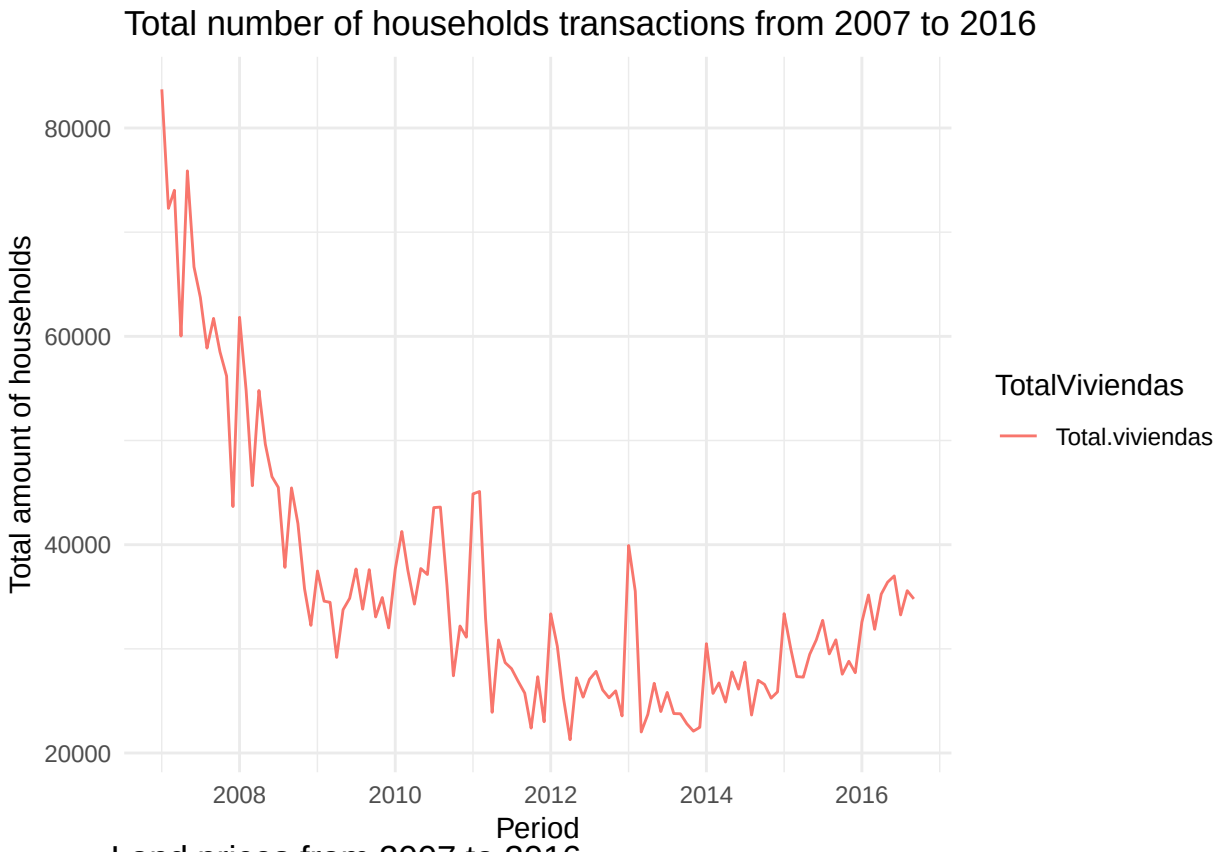
Code used

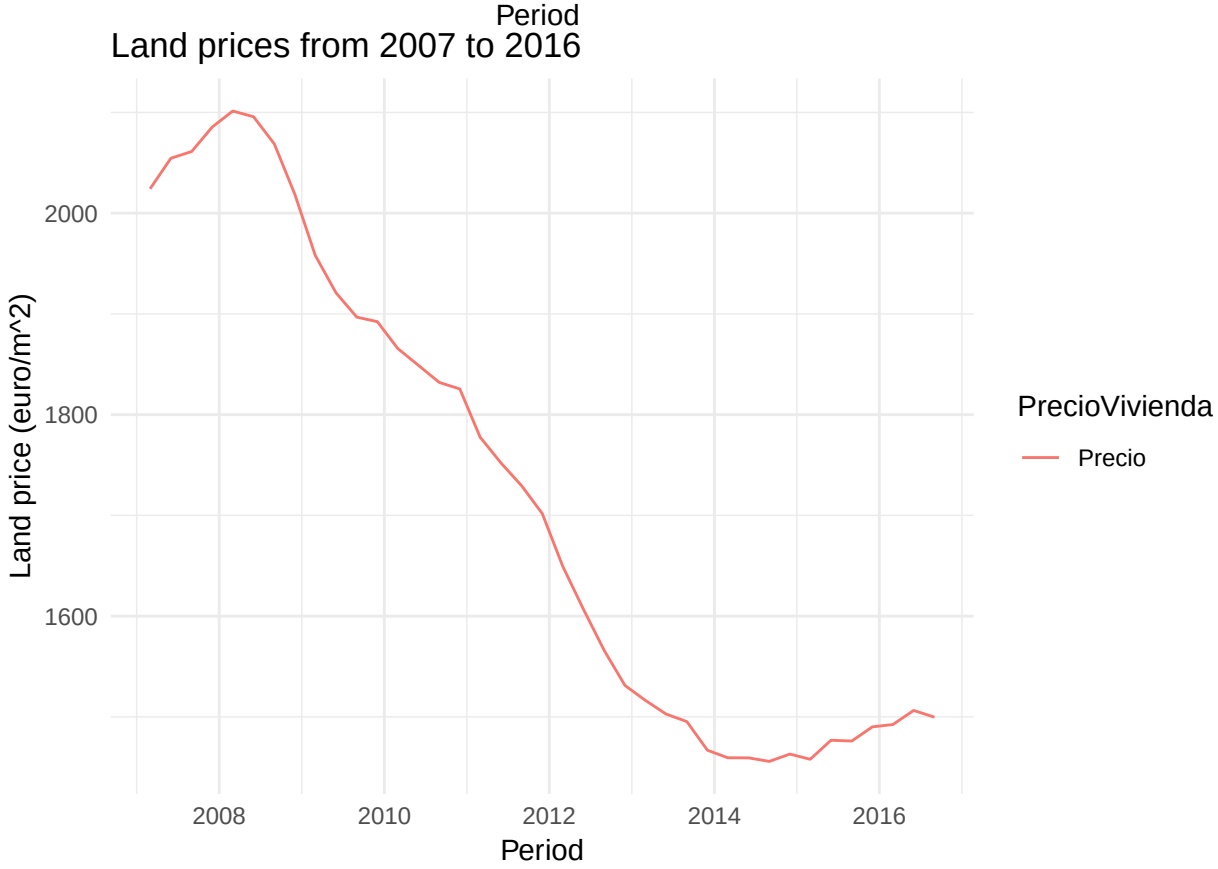
```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```



The figure above we can distinguish between three subsets. The first one and the second one are given by the Google Trends data set. The third one is a combination of both, which is the sum of them divided by the maximum value. Since we are assuming that people who search for households, uses queries containing the word "flat" and the word "house" with the same proportion. So, this transformation will be useful in order to synthesize the popularity of the last two time series, which in fact, contains the same kind of information. As we can see, all the series presents a certain pattern of peaks every six months approximately, moreover, a certain "bouncing" trend is also present in the figure. From now on, we will work only with the total popularity of both queries.



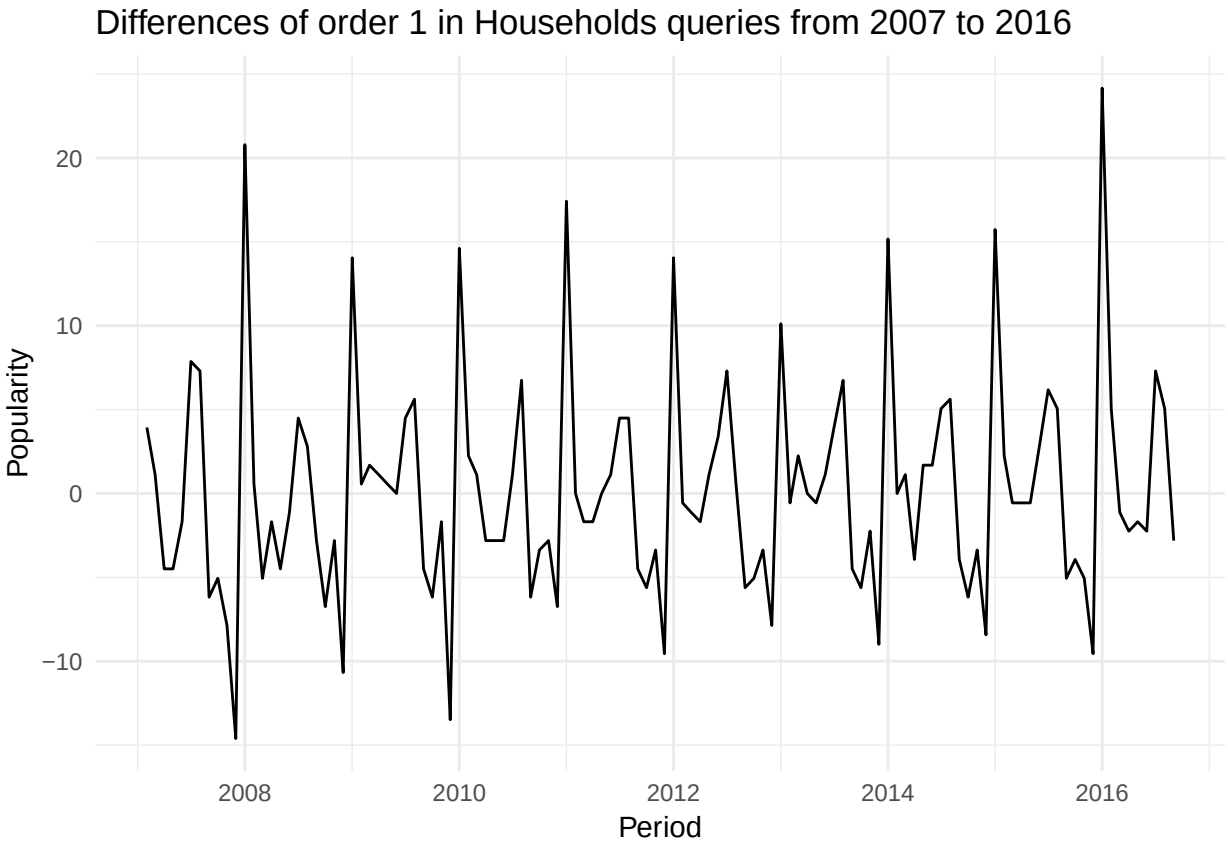


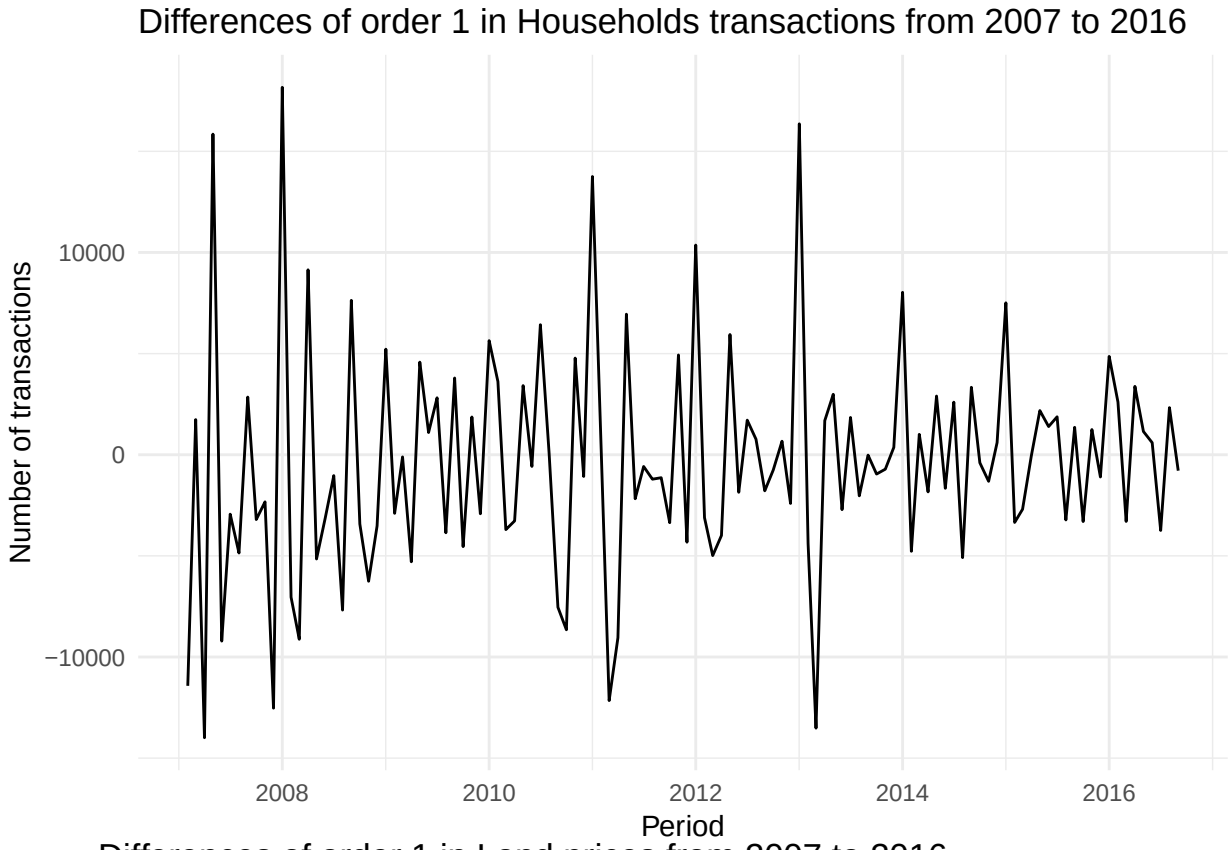
The first of the lasts two figures represents the total number of transactions of householdings from 2007 to

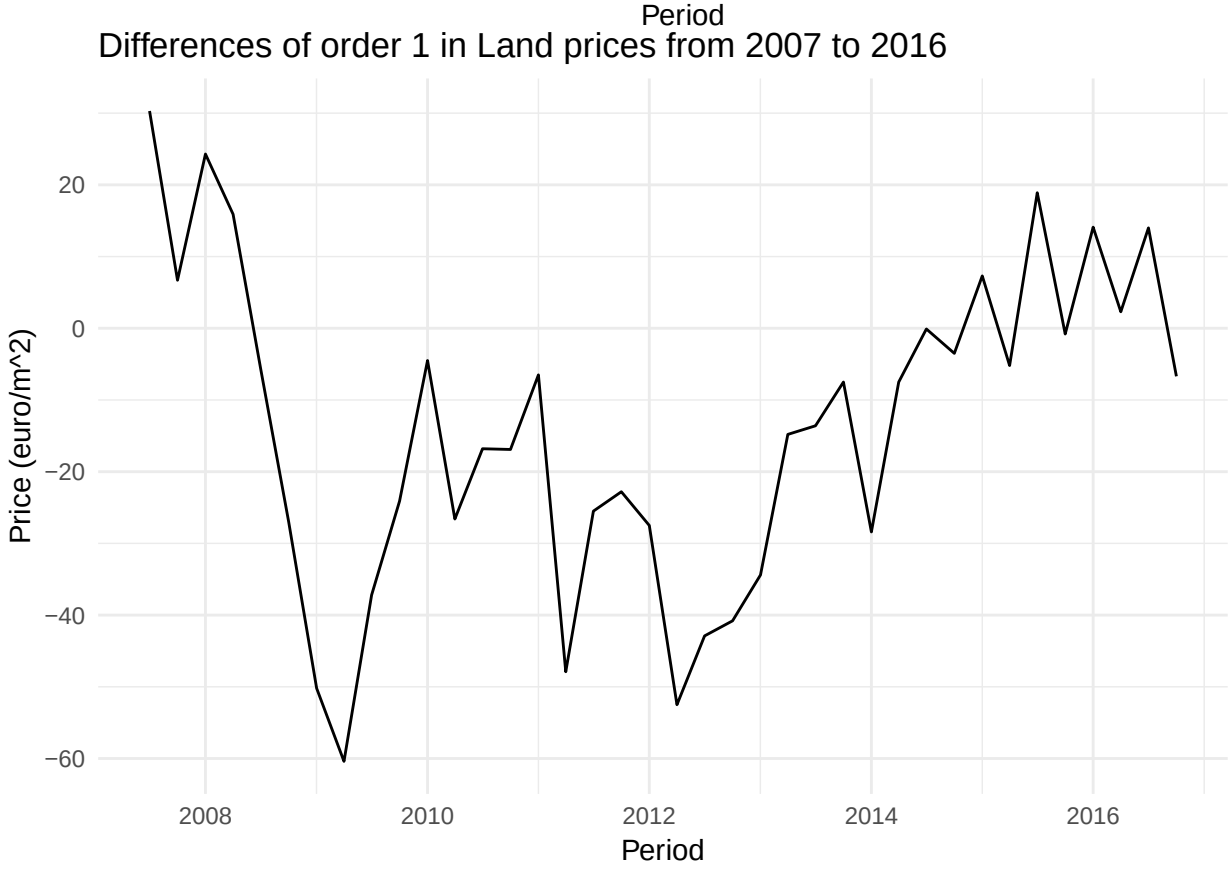
2016. The other one represents the land prices in the same interval but in a frequency of quarters. Both of them shows a decreasing number of transactions from 2009 to 2013, and then a slightly increasing from 2013 to 2014 until 2016 with an exception of a discrepancy between two time series from 2007 to 2009 approximately. It is interesting how both figures are quite similar, and seems to hide a correlation between them. The main analysis in this task will be figure out if there exists any correlation between the three time series and if so, make a prediction of one of them.

Time serie analysis

From now on we are going to take a closer look between the total popularity of total queries, the total amount of transactions and the land prices. To perform the following analysis it will be interesting to observe the plain time series and also the differences between $y_t - y_{t-1}$ in order to look for stationarity.







As we can see in the figures above, for the total queries and the number of transactions, the variance is not

constant, but the mean seems to be almost constant along the time. However for the land prices it is not, so it won't be useful to make comparisons over it.

Correlation analysis

First of all, we will analyse the correlation between land prices and total number of transactions without differenciating the time series and then, a correlation between the differentiated and not differentiated total number of transactions and the queries. We have used Pearson correlation first in order to have a general indicator in how both series are correlated.

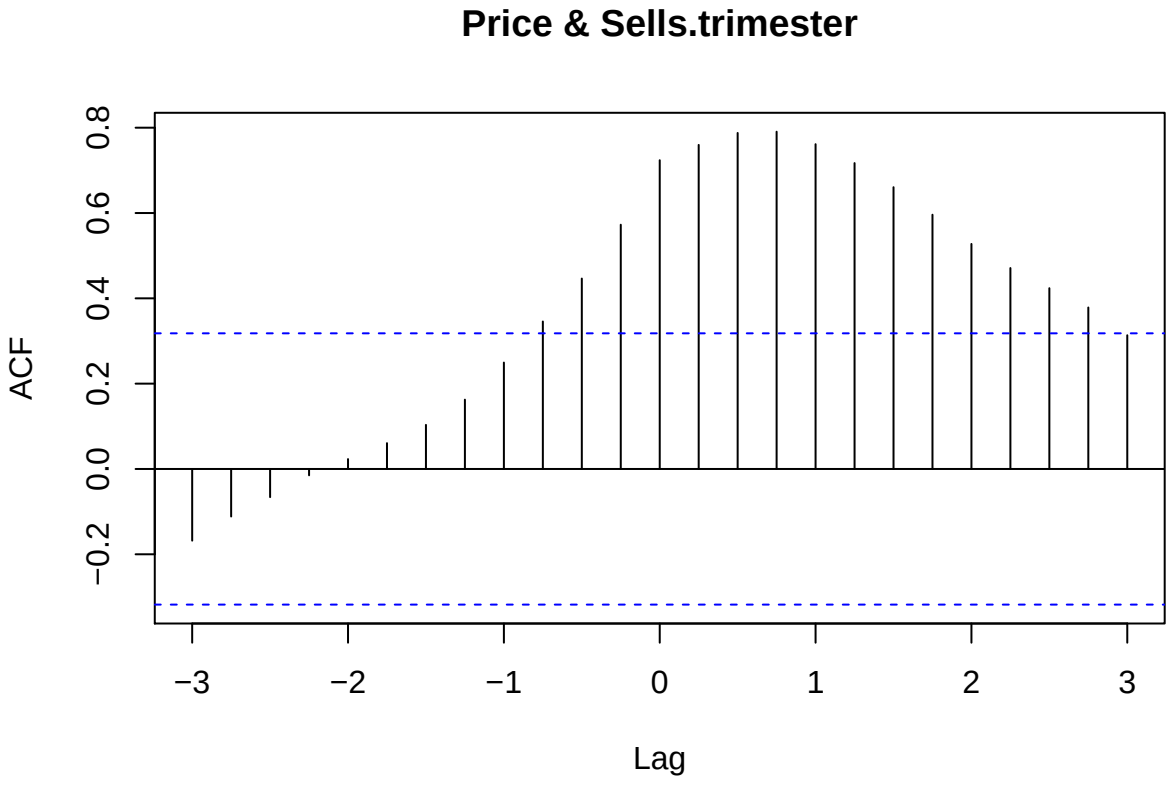
Correlation	C intervals	P value
0.74	(0.56,0.85)	5.18 e-8

Table 1: Correlation betwen Popularity of land prices and total number of transactions

Looking at the table above, the correlation is significant due to the P value, which is small enough compared to the p = 0.05. Also looking to the confidence interval we can confirm our hypothesis that the land prices and the number of households transactions are in fact correlated. We can go further and find which of the lag points are more correlated, in order to find it, we have been used the cross correlation function. This CCF has the following expression :

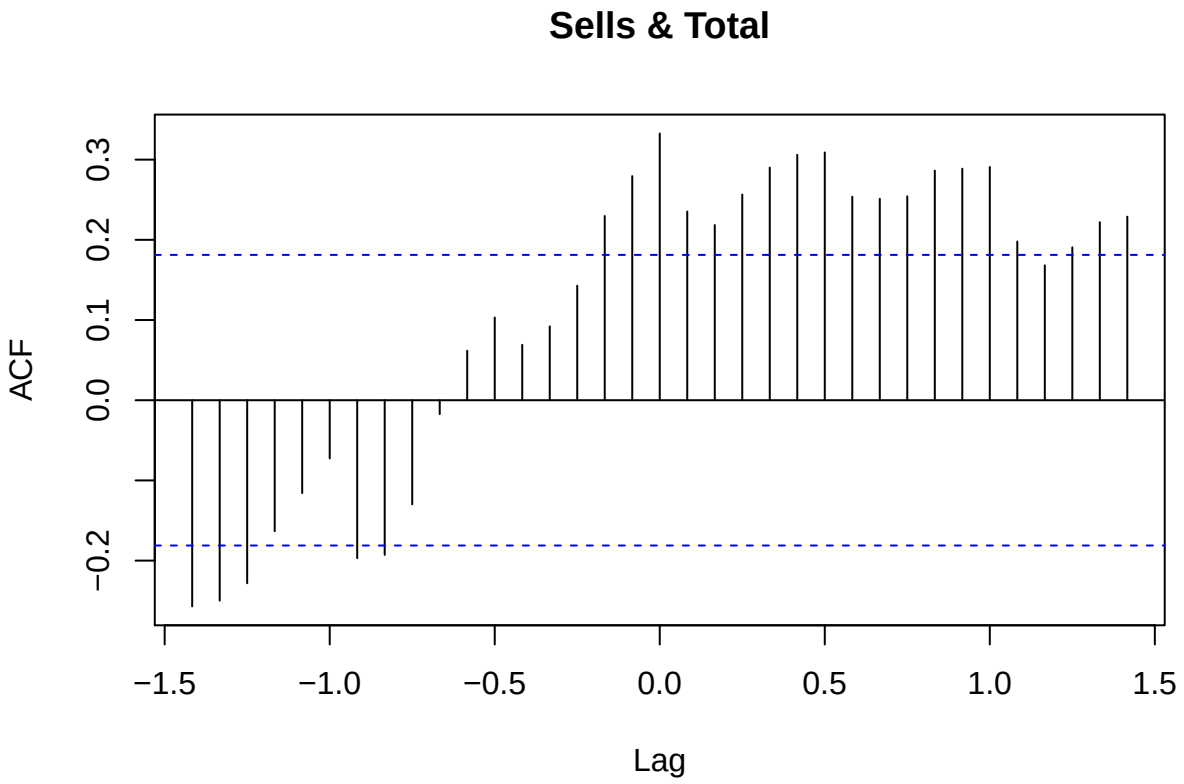
$$\rho_{XY}(\tau) = \frac{< X_t >< Y_{t+\tau} >}{\sigma_X \sigma_Y}$$

The sigma $\rho_{XY}(\tau)$ represents the the correlation between sample Y and sample X in the lag time τ. With this function, we can obtain information about which of the time serie drives the other one.



The cross correlation function shows us how the Land prices from lag $k = 0$ to $k = 3$ are strongly correlated with the $Num.\ Transactions_t$ so, with the information above we can point that the total number of transactions determine how te prices will change in the future.

Using the same criteria, we have analysed the same for the popularity of queries over the Number of transactions. The results are given in the following table:

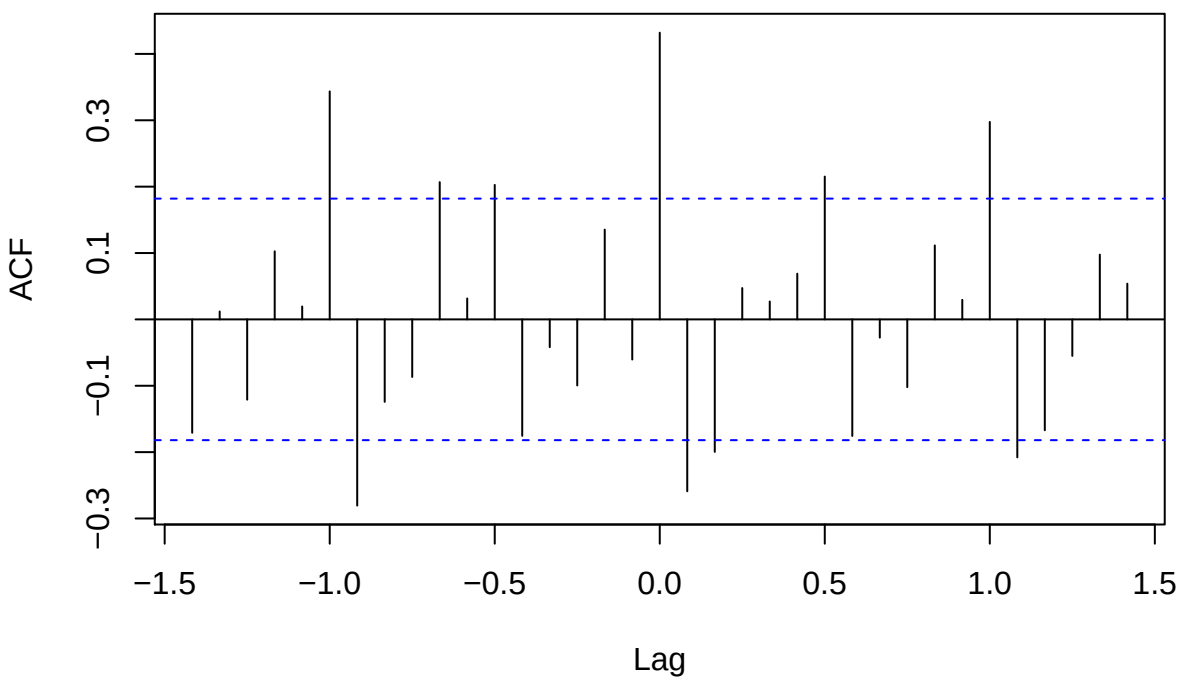


Correlation	C intervals	P value
0.33	(0.16,0.48)	2.48 e-5

Table 2: Correlation betwen Popularity of households queries and total number of transactions

Once again, looking to the statistics of p value and 95% confidence intervals, both time series are significantly correlated, and the cross correlation function shows us how the number of transactions are driven by the popularity of the households queries on Google. Which means that supports hypothesis that the Google queries drives the total number of transactions and then, it could be used as an indicator for the land prices.

Now it is the turn for differentiated time series analysis, the idea behind of using this differentiated instead of the simple ones is to find an stacionary time series. that could be used later for a time series prediction.



Correlation	C intervals	P value
0.43	(0.27,0.56)	1.28 e-6

Table 3: Correlation between Popularity of households queries and total number of transactions

Comparing with the lasts results, we obtained a p value much smaller than the one obtained by the correlation of non differentiated households queries, wich means that applying this transformations, the correlation is statisticaly more significant than the old one. Moreover CCF shows us three lag points which highlight over the other ones. That is in lag $k = 0$ and 1 and -1

Performing a permutation test, we can obtain the p values for each of those lags.

Lag K	Correlation	P value
0	0.43	0
1 (1 year)	0.29	1.6 e-5
-1(-1 year)	0.34	1.7 e-5

Table 4: Correlation betwen differentiated Popularity of households queries and total number of transactions for 1000000 permutations

As we can observe, using 100000 of permutations, the p value for the lag $lag = 0$ is 0, so we can say that both time series are correlated in the same time. As well, we detect a correlation of 0.29 for lag $lag = 1$ with a p value $p = 0.00016$ with is smaller than 0.05 but not so much. The same for $lag = -1$, which the correlation is greater but the p value as well.

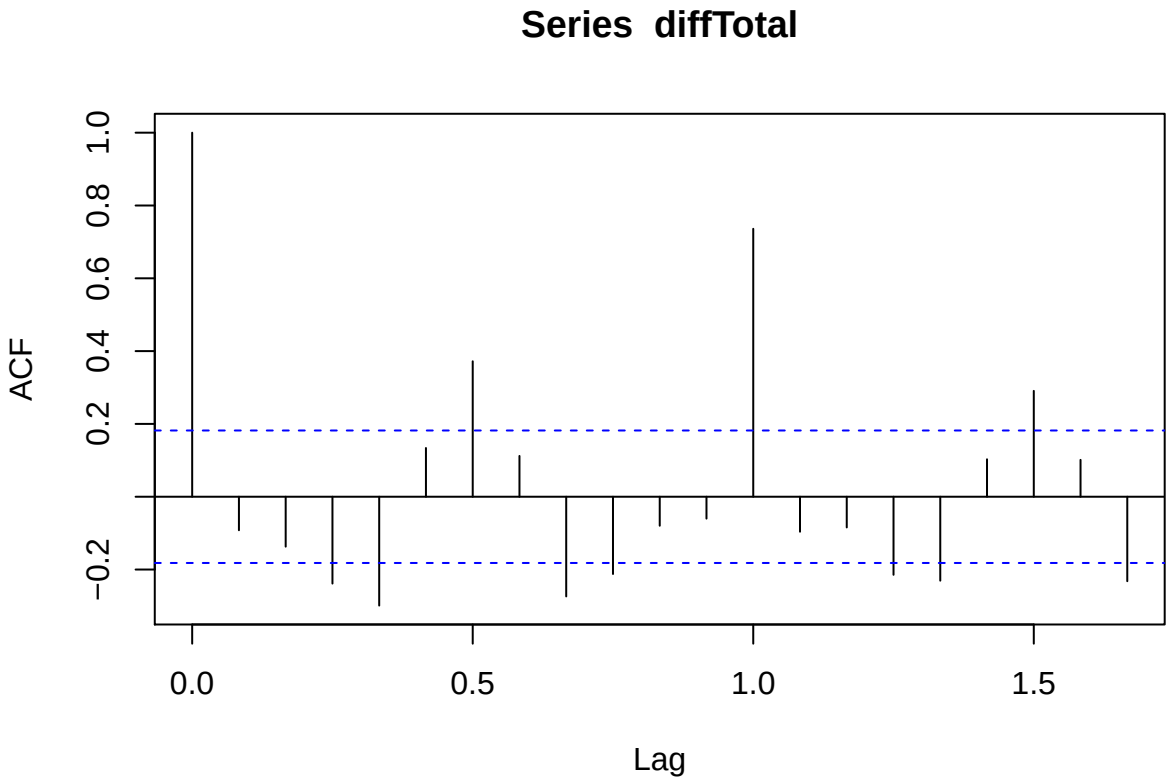
Time Series Prediction using Auto Regressive model

Finally, we have perform a prediction for the popularity of the households queries, we will perform an auto regressive model over the diferentiated time series for the popularity of households queries. Due to stationarity condition, we can simplfy the AR model to one with the constant equal to the sample mean. We have choosen an AR model instead of an ARMA or ARIMA, because of its simplicity and because it fits to the conditions of our sample. As we say, $AR(P)$ models is the acronym for Auto Regressive models of order P. The main idea of this models is to fit a time series into the form of the following expression under the stationarity condition:

$$X_t - \mu = \sum_{i=1}^{p} \phi_i \times X_{t-i} + Z_t$$

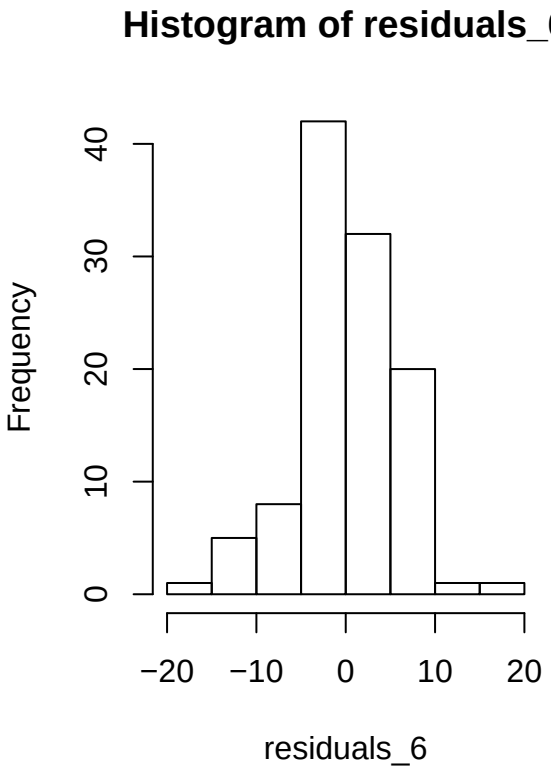
This expression means that a time series could be modeled using some ϕ_i parameters that will deppend on the order of the model and will also deppend on the last X_{t-i} plus a certain noise Z_t.

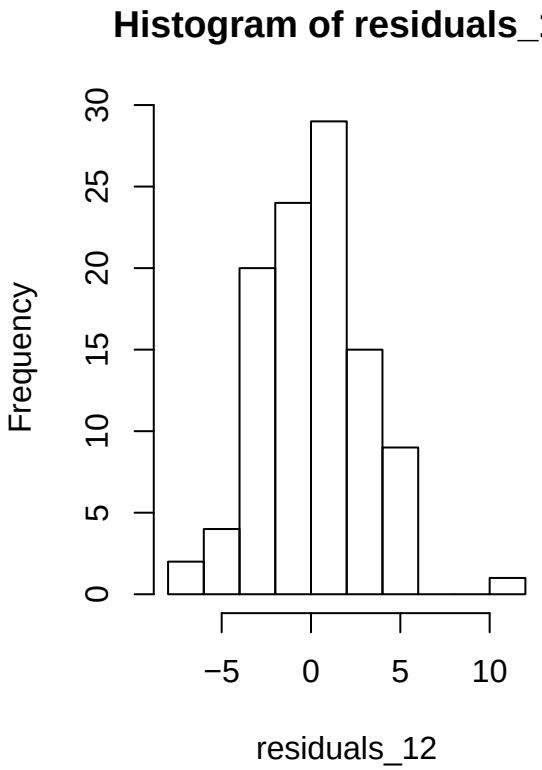
Before to apply the model we have to decide which order fits best for our prediction, to do that, a simple check to the Auto Correlation Function will give us the hints.

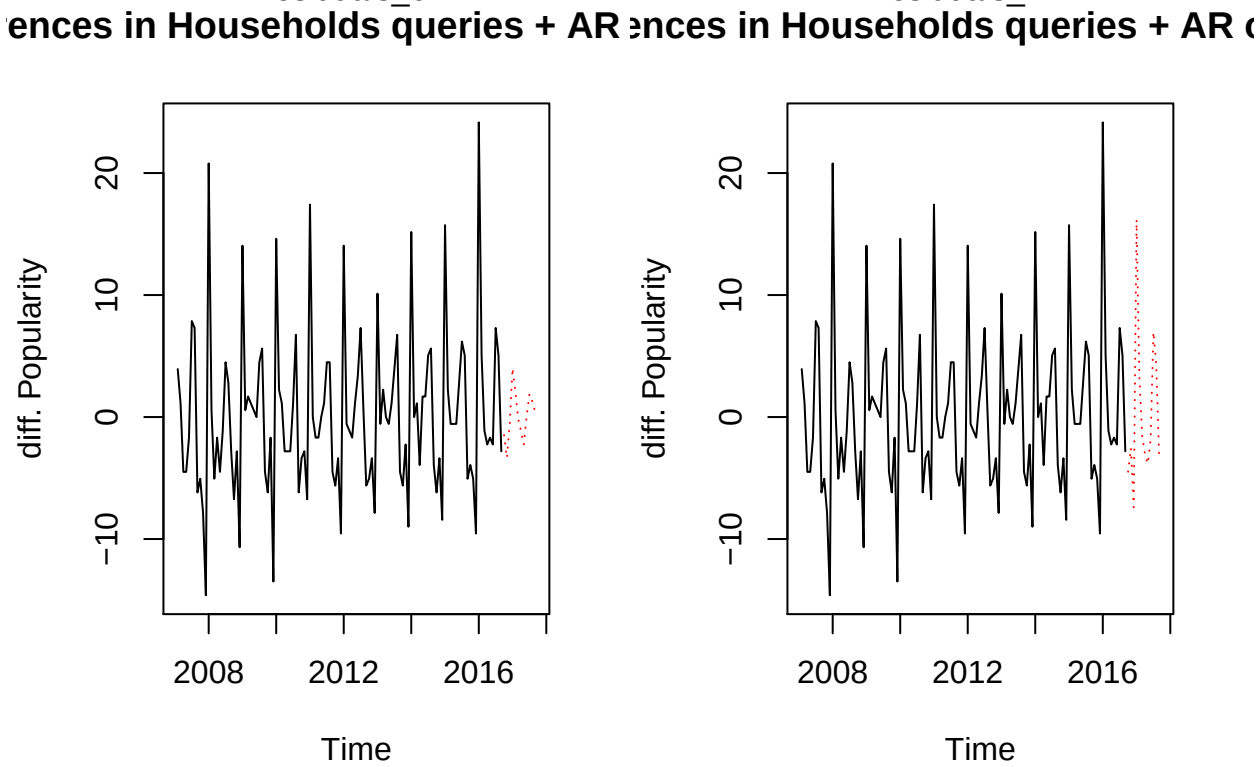


The figure shows a significant correlation in lag 0.5 and lag 1, which represents 6 month an 1 year, but does not show any correlation in between. However, we will apply the model for $P = 12$ and $P = 6$ because of the high correlation in the lags $k = 0.5$ and $k = 1$ just for academical purposes.

Applying the model for $P = 6$ we observe the following:

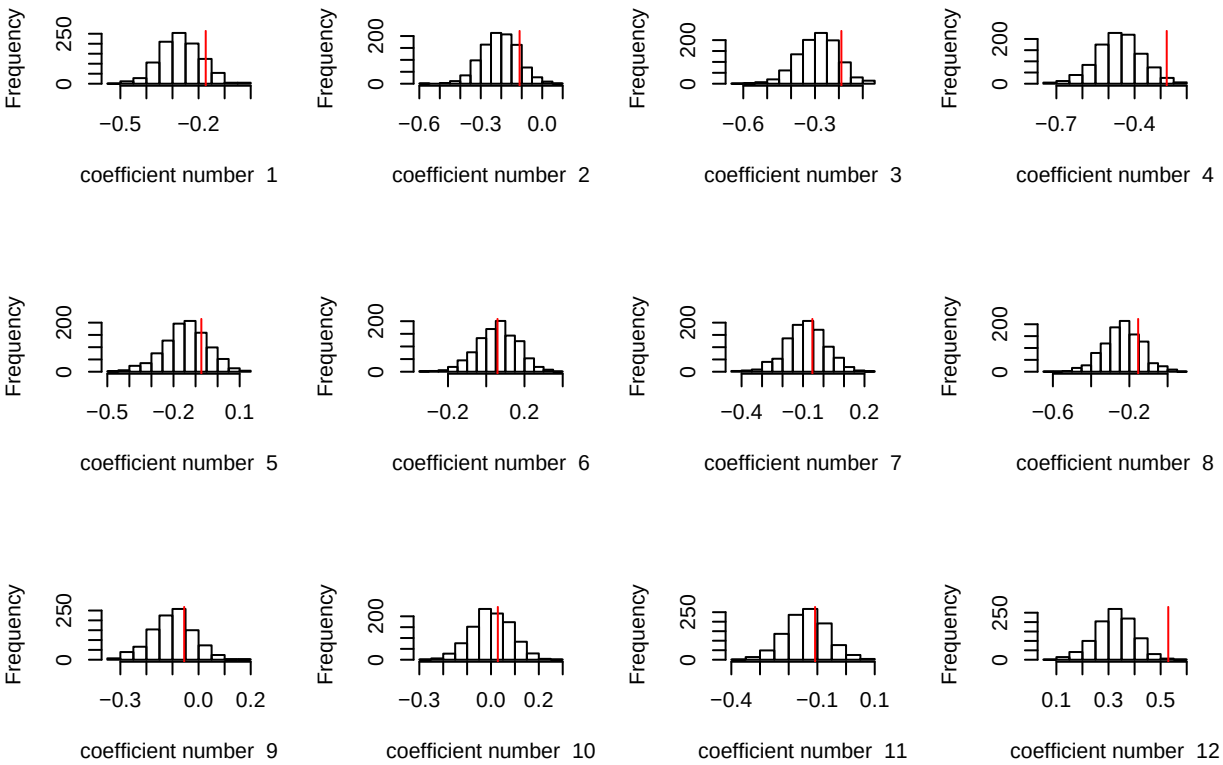






Here we can see how the prediction (in red) for order 12 seems to be in someway "better" than the prediction of AR of order 6. That is because it seems to follow the same pattern as the other points in the time serie.

Notice that the residuals seems to follow a normal distribution, so a bootrapping method of the residuals can be used to obtain the confidence intervals for the rest of the parameters.



The last figure shows us the frequency of the bootstraped coefficients, all of them follows a normal distribution, so we can interpret it confidence interval, the red line shows the value for the fitted coeficient for an AR of order 12. Almost all of the fitted values falls inside the confidence intervals, but the coefficient number 4 and number 12. Also, almost all the confidence intervals have the value 0 inside, so, with this information and with the information given by the auto correlation function, we cannot predict properly the behaivour of the Household queries using an AR model of order 12.

Conclusions

Once analysed the information, we can conclude that the land prices and the total number of households transactions are significantly correlated, and the total number of transactions drives the land prices. As we said in the introductory part, this fact is reasonable due to the law of supply and demand, we knew that this relation exists. But moreover, we have detected a significant correlation of 0.332 between households queries and number of transactions. Whatsmore, the analysis of cross correlation function between the queries and the total transactions, lead us to think that it is probable that the queries drives the total amount of transactions. And also, performing an analysis of the differentiated time series, we have seen that they are significantly correlated for the lags corresponding to 0 months, 1 year and ???1 year. It means that an increase of the queries from one month to the next leads to an increase of a transactions from same month to the next and from the same month to the next but in a year later. As well, applying the Auto regressive model, we conclude using the information given by the Auto correlation function of household queries on Google, and the confidence intervals obtained for a model of order 12, that the prediction done in this assignment is not significant. However, using some other models like ARMA or ARIMA, the prediction results could improve significantly. To sum up, this assignment has just take a simple look into this time series, and the relation between them. We think that this relations could be used in a future research in order to find more significant correlations and better predictions because of the fact that, at least for the period analysed in this assignment, this three datasets are related together and the knowledge of Google trend on households can lead to a better understanding of land prices.

References:

Bibliograpgy

- Autoregressive model. (2018). Retrieved from https://en.wikipedia.org/wiki/Autoregressivemodel
- Cross-correlation. (2018). Retrieved from https://en.wikipedia.org/wiki/Cross-correlation
- Puig,Pere. (2018 Data Visualization and Modelling , Non parametric bootstrap[56-65]

Data sets

- Instituto Nacional de Estadistica. (Spanish Statistical Office). (2018). Retrieved from https://www.ine.es/
- Google Trends. (2018). Retrieved from https://trends.google.com/trends/?geo=US

```
knitr::opts_chunk$set(echo = TRUE)
library(zoo)
library(ggplot2)
library(reshape2)
library(xtable)
library(tinytex)
library(knitr)
library(tseries)
library(ggfortify)
options(xtable.floating = FALSE)
options(xtable.timestamp = "")

Households.queries <- read.csv("GTrendsVivienda.csv", header = T, sep=";")

Households.sells <- read.csv("VivendesVenudesINE.csv", header = T, sep = ";")
Households.sells$Periodo <- gsub("M","-",Households.sells$Periodo)

Households.price <- read.csv("PrecioViviendaINE.csv", header = T, sep = ";")
Households.price$Periodo <- gsub("Q1","-03",Households.price$Periodo)
Households.price$Periodo <- gsub("Q2","-06",Households.price$Periodo)
Households.price$Periodo <- gsub("Q3","-09",Households.price$Periodo)
Households.price$Periodo <- gsub("Q4","-12",Households.price$Periodo)
Households.price$Periodo <- as.Date(as.yearmon(Households.price$Periodo))
Households.price <- subset(Households.price, Periodo >= "2007-01-01")

Households.sells$Periodo <- as.Date(as.yearmon(Households.sells$Periodo))
Households.queries$Periodo <- as.Date(as.yearmon(Households.queries$Periodo))
Households.queries$Total <- (Households.queries$Pisos+Households.queries$Casas)*100/max(Households.queri
Households.queries <- subset(Households.queries, Periodo >= "2007-01-01" & Periodo <= "2016-09-01")
Households.sells <- subset(Households.sells, Periodo >= "2007-01-01" & Periodo <= "2016-09-01")

Houses <-data.frame(Households.queries$Periodo,Households.queries$Casas)
Flats <-data.frame(Households.queries$Periodo,Households.queries$Pisos)
Total <-data.frame(Households.queries$Periodo,Households.queries$Total)

Sells <-data.frame(Households.sells$Periodo,Households.sells$Total.viviendas)
```

```r
Price <- data.frame(Households.price$Periodo,Households.price$Precio)

Houses <- as.ts(read.zoo(Houses, FUN = as.yearmon))
Flats <- as.ts(read.zoo(Flats, FUN = as.yearmon))
Total <- as.ts(read.zoo(Total, FUN = as.yearmon))
Sells <- as.ts(read.zoo(Sells, FUN = as.yearmon))
Price <- as.ts(read.zoo(Price, FUN = as.yearmon))
Sells.trimester <- aggregate(Sells, nfrequency=4)

Total.semester <- aggregate(Total, nfrequency=2)
Total.anual <- aggregate(Total, nfrequency=1)
```

```r
df <- melt(Households.queries ,  id.vars = 'Periodo', variable.name = 'queries')
ggplot(df, aes(Periodo,value)) + geom_line(aes(colour = queries)) + ggtitle("Popularity of google queri
```

```r
df2 <- melt(Households.sells ,  id.vars = 'Periodo', variable.name = 'TotalViviendas')
df3 <- melt(Households.price,id.vars = 'Periodo', variable.name = 'PrecioVivienda' )
par(mfrow=c(2,2))
ggplot(df2, aes(Periodo,value)) + geom_line(aes(colour = TotalViviendas)) + ggtitle("Total number of hou
```

```r
ggplot(df3, aes(Periodo,value)) + geom_line(aes(colour = PrecioVivienda)) + ggtitle("Land prices from 20
#par(mfrow=c(1,1))

#ts.plot(Price)
#ts.plot(Sells)
#par(mfrow=c(1,1))
#ts.plot(Price/max(Price), Sells.trimester/max(Sells.trimester), gpars = list(col = c("black", "red")))
```

```r
#c_Houses = Houses[1]
#c_Flats = Flats[1]
c_Total = Total[1]
c_Sells = Sells[1]
c_Price = Price[1]

#diffHouses = diff(Houses)
#diffFlats = diff(Flats)
diffTotal = diff(Total)
diffSells = diff(Sells)
diffPrice = diff(Price)
```

```r
autoplot(diffTotal)  +ggtitle("Differences of order 1 in Households queries from 2007 to 2016") + labs(y
autoplot(diffSells)  +ggtitle("Differences of order 1 in Households transactions from 2007 to 2016") + l
autoplot(diffPrice)  +ggtitle("Differences of order 1 in Land prices from 2007 to 2016") + labs(y= "Pric
```

```r
library(kableExtra)
correlation_Sells_Price <- cor.test(Sells.trimester,Price, method = "pearson")


#table_1 <- data.frame("Correlation" = correlation_Sells_Price$estimate, "P value" =correlation_Sells_P

#print(xtable(table_1))

correlation_Sells_Price <- ccf(Price,Sells.trimester)

correlation_Sells_queries <- cor.test(Sells,Total)
#correlation_Sells_queries
correlation_Sells_Price <- ccf(Sells,Total)


correlation_Sells_queries <- cor.test(diffSells,diffTotal)
#correlation_Sells_queries
correlation_Sells_Price <- ccf(diffSells,diffTotal)



pvalueforccf <- function(x,y,k,nperm){


  stperm <- numeric(nperm)
  streal <- ccf(x,y, plot = F)[k]$acf
  n=length(y)
  p = 0
  for (i in 1:nperm){

    yperm <- sample(y,n)
    yperm <- ts(yperm, start = as.yearmon("2007-01"), frequency = 12)
    stperm[i] <- abs(ccf(x,yperm, plot = F)[k]$acf)

    if (stperm[i] >= streal){p = p + 1
    p}
  }


  pvalue = p/nperm
  return(list(ccf_k = as.numeric(streal), p_value = as.numeric(pvalue)))
}

#ccf_0 <- pvalueforccf(diffSells,diffTotal,0,1000000)
#ccf_1 <- pvalueforccf(diffSells,diffTotal,1,1000000)
#ccf_menys_1 <- pvalueforccf(diffSells,diffTotal,-1,1000000)



acf(diffTotal)
fit_6<-ar(diffTotal, aic = FALSE, order.max = 6 ,method=c("yule-walker"))
fit_12<-ar(diffTotal, aic = FALSE, order.max = 12 ,method=c("yule-walker"))
```

```r
coeff_6 <- fit_6$ar
coeff_12 <- fit_12$ar
residuals_6 <- fit_6$resid
residuals_12 <- fit_12$resid

par(mfrow=c(1,2))

hist(residuals_6)
hist(residuals_12)

pred_6 <- predict(fit_6,n.ahead=12)
pred_12 <- predict(fit_12,n.ahead=12)

ts.plot(diffTotal, pred_6$pred, lty = c(1,3), col=c(1,2), main="Differences in Households queries + AR c
ts.plot(diffTotal, pred_12$pred, lty = c(1,3), col=c(1,2),main="Differences in Households queries + AR c


  multiplicatoria <- function(x,t,phi){

    res <- 0

    for (i in 1:length(phi)) {



      res <- res + x[t-i + 1]*phi[i]


    }

    return(res)

  }

  Sumatoria <- function(phi){

    res <- 0

    for (i in 1:length(phi)) {

      res <- res + phi[i]

    }
    return(res)

  }

  BootstrapARmodel <- function(x,phi,nb,res){
    mu <- mean(x)
    mub<-numeric(nb)
    xb<-numeric(length(x))
    phib <- list()
```

```
  for (i in 1:length(phi)){phib[[i]] <- numeric(nb)}

  for (j in 1:nb){

    for (i in 1:length(phi)){xb[i] <- x[i]}

    for (t in length(phi):length(x)){

      r<-sample(res[length(phi)+1:length(x)],1,replace=T)
      while(is.na(r)){r<-sample(res[length(phi)+1:length(x)],1,replace=T)}

      xb[t+1] <- multiplicatoria(xb,t,phi) + mu*(1-Sumatoria(phi)) + r
      if (is.na(xb[t+1])) {
        print(multiplicatoria(xb,t,phi))
        print(mu*(1-Sumatoria(phi)))
        print(r)
        return(-12)

      }
    }


    fitb<-ar(xb, aic = FALSE, order.max = length(phi),method=c("yule-walker"))
    mu[j]<-mean(xb)

    for (k in 1:length(phi)){phib[[k]][j]<-fitb$ar[k]}

  }

  return(list(mu = mu,phib = phib))
}


fitb <- BootstrapARmodel(diffTotal,coeff_12,1000,residuals_12)
#Els intervals de confian??a per cada parametre es plotejan tal que asi:
par(mfrow=c(3,4))
for (k in 1:length(coeff_12)){
  hist(fitb$phib[[k]], main = "", xlab = paste("coefficient number ",k))
  abline(v=coeff_12[k],col="red")}
par(mfrow=c(1,1))
##############################################################################
```