

Grau en Estadística

Títol: Anàlisi, detecció i estudi de bombolles múltiples en sèries temporals de freqüència de cerques a Google

Autora: Sabina Criballés Vilamitjana

Directora: Montserrat Guillén Estany

Departament: Econometria, Estadística i Economia Aplicada

Convocatòria: Juny 2020



En primer lloc, a la Montse Guillén, per la seva ajuda, consells i tot el temps dedicat a la realització d'aquest treball. En segon lloc a la meva família, per a la seva paciència i a les meves companyes de carrera per haver-nos recolzat mútuament durant tot aquest temps.

Resum (català)

La detecció de períodes d'exhuberància i crisi en sèries temporals permet una anàlisi més a fons d'aquestes, i un desenvolupament de millors mètodes de predicció. Dins el comportament d'una sèrie trobem les bombolles: pics espontanis que precedeixen a una crisi. Al 2011 Phillips, Wu i Yu desenvolupen el PWY test per a detecció de bombolles, i més endavant, al 2015, Phillips, Shi i Yu el milloren augmentant la seva potència de detecció en sèries amb múltiples bombolles. En aquest treball s'estudia l'aplicació d'aquest test PSY en sèries temporals de Google Trends. D'aquesta manera, es pretén entendre i detectar el comportament d'un tema digitalment viral, per a poder tractar-lo de forma diferent o per a futures aplicacions i estudis. També es compara l'anàlisi clàssica amb la determinista, i com la presència o no de múltiples bombolles pot afectar-hi. La implementació del test es fa a través del paquet *psymonitor* de R, on es modifiquen els nivells de significació i algunes funcions. Per últim, s'implementa una Shiny de visualització d'unes dades determinades tractades i una Shiny que permet interactuar i escollir la paraula al moment.

Paraules clau

bombolles múltiples, test PWY (2011), test PSY (2015), viral, trending topic, Google Trends, anàlisi sèries temporal

Abstract (english)

The detection of exuberance periods and crisis allows a deeper analysis in time series and an improvement of better forecast methods. In the time series' behaviour, we could find bubbles which are spontaneous peaks that precede a crisis. In 2011 Phillips, Wu and Yu developed a test to detect these bubbles. And in 2015, Phillips, Shi and Yu improved the test, increasing its power to detect multiple bubbles. This thesis investigates the application of this test on time series from Google Trends. The aim is to understand and detect the behaviour of a viral topic so it can be treated in a different way for future implementations and studies. This thesis also examines how the presence of multiple bubbles may affect the classical analysis, in both stochastic and deterministic approaches. The test is implemented through the package *psymonitor* of R, where the significance levels and some functions are modified. The last step is to create two Shiny App's, one to display the analysed data and the other to interact with and choose the word in the same moment.

Key words

multiple bubbles, test PWY (2011), test PSY (2015), viral, trending topic, Google Trends, time series analysis

Classificació AMS

37M10 Time series analysis

ÍNDIX DE CONTINGUTS

| | | |
|-------|---|----|
| I. | INTRODUCCIÓ | 7 |
| II. | METODOLOGIA..... | 12 |
| 2.1 | Test de detecció de bombolles múltiples | 12 |
| 2.1.1 | Obtenció i visualització de la sèrie de Google Trends | 12 |
| 2.2.2 | Aplicació del test de bombolles..... | 14 |
| III. | RESULTATS..... | 23 |
| 3.1 | Sèrie iphone | 23 |
| 3.2 | Sèrie pants | 25 |
| 3.3 | Sèrie lithium | 25 |
| 3.4 | Sèrie obama | 27 |
| 3.5 | Sèrie bitcoin | 28 |
| 3.6 | Sèrie spread | 29 |
| IV. | ANÀLISI DE LES SÈRIES TEMPORALS | 31 |
| 4.1 | Determinista | 31 |
| 4.1.1 | Anàlisi determinista de la sèrie spread | 35 |
| 4.1.2 | Anàlisi determinista de la sèrie pants | 36 |
| 4.1.3 | Anàlisi determinista de la sèrie obama | 37 |
| 4.1.4 | Anàlisi determinista de la sèrie iphone | 38 |
| 4.1.5 | Anàlisi determinista de la sèrie lithium | 39 |
| 4.1.6 | Anàlisi determinista de la sèrie bitcoin | 40 |
| 4.2 | Predicció determinista | 42 |
| 4.2.1 | Sèries tipus III..... | 42 |
| 4.2.2 | Sèries tipus IV | 45 |
| 4.3 | Estocàstic | 47 |

| | | |
|-------|---|----|
| 4.4 | Predicció estocàstica..... | 49 |
| 4.4.1 | Predicció estocàstica de la sèrie spread..... | 49 |
| 4.4.2 | Predicció estocàstica de la sèrie pants..... | 52 |
| 4.5 | Conclusió de l'anàlisi clàssica..... | 55 |
| V. | SHINY | 57 |
| 5.1 | Metodologia..... | 57 |
| 5.2 | Visualització | 57 |
| 5.3 | Interactiva | 59 |
| VI. | CONCLUSIONS | 62 |
| VII. | ÍNDIX DE FIGURES..... | 64 |
| VIII. | ÍNDIX DE TAULES..... | 66 |
| IX. | REFERÈNCIES..... | 67 |
| 8.1 | Articles | 67 |
| 8.2 | Articles electrònics..... | 67 |
| 8.3 | Apunts..... | 68 |
| 8.4 | Llibres..... | 68 |
| 8.5 | Pàgines web | 68 |
| IX. | ANNEX..... | 69 |
| 9.1 | Script Shiny "Visualització" | 69 |
| 9.2 | Script Shiny "Interactiu" | 71 |

I. INTRODUCCIÓ

Avui en dia tenim l'abast informació de qualsevol tipus a un sol *clic* de distància. Ja sigui del caire científic o curiositat sensacionalista, a Internet amb facilitat ho trobem tot.

El DIEC defineix la *tendència* com a “inclinació natural, disposició en virtut de la qual hom és atret devers un fi i hi tendeix, un cos és portat a moure's en una certa direcció”.

Amb aquesta definició, es prova que quan un tema comença a ser molt freqüent i conegut en la societat, pot esdevenir-se el que digitalment s'anomena com el famós fenomen *trending topic*. Aquesta designació, traduïda en català com a “tendència o tema del moment”, no és res més que una altra freqüència de cerques digitals relacionades amb aquest tema, en un període determinat.

Un bon sinònim podria ser “viral”. Aquest fenomen es pot produir a diferent escala i sobre qualsevol tema.

Un bon exemple i molt clar pot ser el coronavirus. D'ençà que va començar a finals de gener, tot el món ha anat rebent i cercant informació del virus amb diferent intensitat. A través de l'eina Google Trends (s'explica detalladament més avall), he obtingut els gràfic que es mostren a la Figura 1.1. Es pot veure com ha evolucionat el nombre de cerques de la paraula “coronavirus” a Xina (gràfic de dalt) respecte a l'evolució de cerques a Espanya. Amb la quantitat d'informació que s'ha tingut durant la pandèmia, és fàcil identificar el motiu de cada pic.

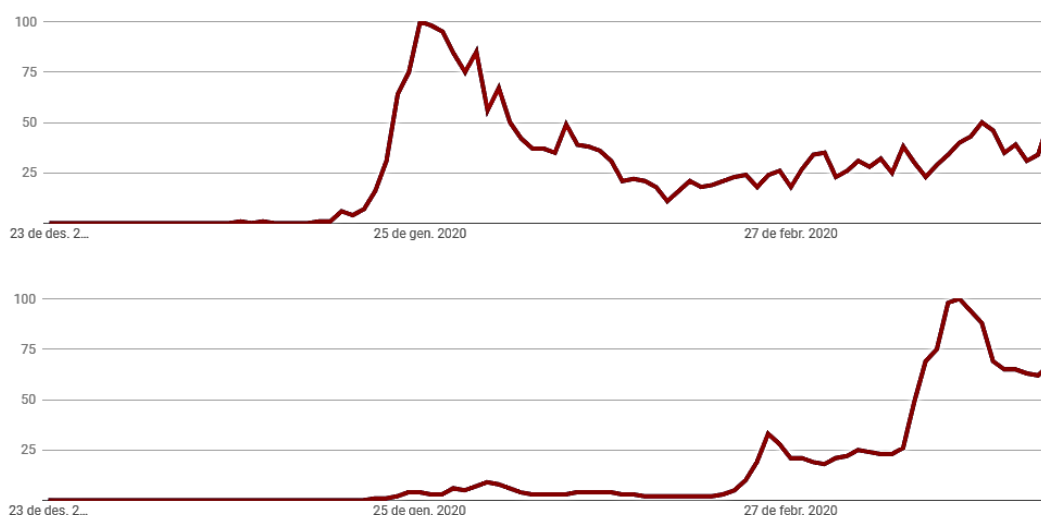


Figura 1.1. Cerques del terme “coronavirus” segons Google a Xina (superior) i a Espanya (inferior).

A finals de gener, concretament el dia 23, va ser quan es va aïllar la regió de Wuhan a Xina. Es pot veure reflectit que és quan s'ha produït el màxim pic a Xina, i on s'intueixen

les primeres cerques del terme a Espanya. A Espanya també es veuen unes petites pertorbacions a finals de gener-principis de febrer que es deuen als primers casos aïllats que es van confirmar al país. A Espanya, el virus va “atacar” i entrar seriosament al 26 de febrer (primer pic), i va tornar a ser molt cercat al voltant de l’11 de març quan l’Organització Mundial de la Salut va declarar la pandèmia.

Tot i aquest clar exemple, aquests esdeveniments també es poden succeir a escala més petita. Per exemple, quan un grup de música o un artista treu un nou àlbum, el més probable és que se’n parli amb més intensitat durant un temps i que consegüentment, se’n produeixi un nombre més elevat de cerques digitals.

Què és una bombolla? Sovint, una crisi financera està precedida per una inflació en el valor de les accions o en l’increment del preu d’alguns actius. Aquesta bombolla que es produeix no va ser excepció en la crisi que va començar al 2007, i és per això que els economistes han acordat una sèrie de mesures per estabilitzar el sistema financer un cop s’estigui produint això. El que ens interessa és poder detectar quan s’estan produint aquesta sèrie d’esdeveniments per a poder detectar quan començar a haver-hi una situació d’exuberància econòmica, o saber quins moments s’han produït per a poder canviar el nostre mètode de l’anàlisi que estem fent ja sigui òbviant aquests moments, o simplement tractar-los diferent.

Phillips, Wu i Yu (2011, PWY test) van proposar un mètode recursiu que pot detectar aquestes situacions en la fase inicial, en la inflació. Aquest mètode utilitza el suprem de l’estadístic ADF (augmented Dickey-Fuller), el qual no és gaire potent si els períodes d’exuberància i col·lapse es repeteixen. És per això que al Phillips, Shu-Ping i Yu (2011, PSY) presenten una millora d’aquest últim test SADF, generalitzant-lo: GSADF. Aquest últim test està basat en la idea de repetir el SADF però ampliant la mostra amb un rang més flexible. En comptes de fixar el punt inicial (primera observació de la mostra), el GSADF amplia la seqüència de mostres canviant el punt d’inici i el punt de finalització. D’aquesta manera, la finestra de la mostra és més flexible i la detecció de bombolles, si és que n’hi ha múltiples, millora.

Phillips, Shi i Yu (2015) escriuen l’article “Test for Multiple Bubbles: historical episodes of exuberance and collapse in the S&P 500”, on apliquen l’estadístic GSADF (generalització del PWY 2011). Com ja he explicat, aquest funciona amb submostres de les dades amb diferents punts d’inicialització a les regressions i té més capacitat per a detectar quan hi ha múltiples bombolles.

$$GSADF(r_0) = \sup_{\substack{r_2 \in [r_0, 1] \\ r_1 \in [0, r_2 - r_0]}} \{ADF_{r_1}^{r_2}\}$$

r_1 i r_2 són els punts d'inici i finalització de la submostra en aquell instant.

r_0 és el punt on inicial d'origen, és a dir, el punt on s'hi associarà l'estadístic obtingut.

Una bombolla no deixa de ser un fenomen que es dona en una sèrie temporal qualsevol i és per això que no només se'n pot parlar lligat a temes econòmics. Aquesta es pot presentar i ser detectada sigui quina sigui la naturalesa de les dades. Per exemple, les dades poden provenir de l'aplicació Google Trends. En aquest cas, la detecció de bombolles passa a basar-se en trobar moments d'exuberància d'una cerca concreta. Parlaríem de *trending topic* o el que podríem dir-ne "bombolles" però de la freqüència de cerques de Google.

Així doncs, la finalitat d'aquest treball és poder detectar aquestes "bombolles" en cerques a Google a partir de l'historial de freqüència de cerques d'una paraula concreta, quan hi ha hagut aquests períodes de crisi al llarg de tota la sèrie temporal i fer una identificació a posterior de quin patró sembla que han seguit aquests períodes en la sèrie en qüestió.

Per a l'obtenció de les dades de Google Trends en aquest treball utilitzaré el paquet d'R *gtrendsR*. Aquest és un paquet que serveix per a exportar i mostrar la informació disponible online de l'aplicació Google Trends. "Trends" és el nombre donat a les cerques d'un període i d'una localització geogràfica concreta que pot mostrar.

La funció *gtrends* d'aquest mateix paquet permet obtenir, en forma de base de dades, el nombre de cerques (a la funció s'anomenen *hits*) d'una paraula, o conjunt de paraules, concreta amb una periodicitat determinada. A més, es pot delimitar el territori i la llargada de la sèrie temporal. Els paràmetres pels quals es pot ajustar la funció i que jo utilitzo són els següents:

keyword: la paraula/conjunt de paraules que volem analitzar.

time: vector on s'especifica el període del temps del que en volem saber les cerques. Es pot especificar el dia concret, valors per defecte que ja dona Google Trends com ara: la última hora, els últims quatre dia, els últims 30 o 90 dies, etc. També hi ha l'opció de "all" que significa des del principi que hi ha disponibilitat de les dades a Google Trends, que és al 2004.

hl: vector on s'especifica el llenguatge del codi. Sempre he utilitzat "es" per referir-me a l'espanyol.

Aquesta funció torna un list que conté un data frame amb el nombre de cerques per cada data concreta.

És rellevant comentar la periodicitat amb la que la funció retorna els data frames. Com ja he comentat, es pot especificar el període a través del paràmetre *time*. Si per exemple s'especifica que es volen les dades de l'hora o les quatre hores anteriors, torna el nombre de cerques de cada minut: 59 o 239 valors, respectivament. Amb l'opció "now 1-d" o "now 7-d" que retorna tot el dia, el format és la informació de cada 8 minuts (180 valors). Si són els últims 7 dies, retorna valors de cada hora (168 en total). Si s'escull de l'últim o dels tres últims mesos anteriors, tenim informació diària. A partir d'un interval superior als nou mesos (270 valors) retorna valors mensuals.

Per a l'aplicació del test de detecció de bombolles múltiples en aquest treball utilitzaré el paquet *psymonitor*. Aquest serveix per aplicar els testos mencionats anteriorment que es van proposar a l'article de Phillips, Shi and Yu al 2015. La finalitat és obtenir unes taules on apareixen els períodes que s'hi ha detectat un comportament explosiu, de bombolla. Més endavant explico com s'apliquen les funcions i els dos paquets d'R mencionats. Les funcions que conté aquest paquet les he de modificar ja que presenten alguna incompatibilitat de format, l'opció d'operar amb computació paral·lela, també afegiré l'opció de modificar el nivell de significació dels valors crítics que escullo i varis errors o modificacions que explico a l'apartat **2.2.2**.

En primer lloc, començo fent l'anàlisi de presència o no bombolles en un total de 7 sèries diferents obtingudes amb Google Trends. Aquestes sèries les escullo tenint en compte que tindran un caràcter diferent entre elles i així poder veure què i què no considera bombolla el test PSY (2015).

Seguidament, presento l'anàlisi clàssica de les mateixes 7 sèries temporals. L'anàlisi és la determinista (amb Excel) i l'estocàstica (R). Aquesta anàlisi la faig per a veure si la presència de bombolles o el caràcter explosiu i irregular en certs períodes, afecta a una anàlisi clàssica i a les seves conseqüents prediccions.

A continuació, l'objectiu és crear dues aplicacions de Shiny d'R on s'hi aplica el test PSY. En una simplement hi mostraré les sèries que hauré analitzat prèviament i de les quals tinc els resultats del test guardats. Serà com una Shiny de visualització amb explicacions del seu comportament. L'altra està pensada per a que sigui interactiva i es pugui fer la consulta de la paraula que es vulgui i s'hi apliqui el test per analitzar-la al moment.

Objectiu global

Aplicació del test de Phillips, Shi i Yun (2015) per a la detecció de bombolles en sèries temporals referents a la freqüència de cerques en Google.

Objectius específics

- Entendre el funcionament del test de Phillips, Shi i Yun (2015).
- Generalització del test PSY (2015) per a poder aplicar-lo a les sèries de Google Trends.
- Anàlisi clàssica (determinista i estocàstica) en un conjunt de dades de Google Trends concretes i comparació entre elles.
- Implementació del test PSY (2015) en un conjunt de dades de Google Trends concretes.
- Creació d'una aplicació Shiny per a visualitzar els resultats obtinguts.
- Creació d'una aplicació Shiny interactiva que permeti la descàrrega immediata d'una sèrie de Google Trends i aplicar el test al mateix moment.

II. METODOLOGIA

2.1 Test de detecció de bombolles múltiples

2.1.1 Obtenció i visualització de la sèrie de Google Trends

En primer lloc es tracta d'obtenir la sèrie temporal. Com s'ha explicat anteriorment, les dades les obtinc a través del paquet `gtrendsR`, el qual permet l'obtenció i el tractament de les dades instantàniament des de Google Trends. A més, el format obtingut és de fàcil tractament ja que sempre torna un data frame amb les següents variables. Per fer la cerca, utilitzo les següents línies de codi:

```
resm <- gtrends(keyword = c("manel"),time = "all",hl="es")

> head(resm[[1]])
  date hits  geo time keyword gprop category
1 2004-01-01 19 world all manel web 0
2 2004-02-01 17 world all manel web 0
3 2004-03-01 23 world all manel web 0
4 2004-04-01 17 world all manel web 0
5 2004-05-01 16 world all manel web 0
6 2004-06-01 23 world all manel web 0
```

Figura 2.1. Part superior del data frame d'exemple.

A la Figura 2.1 es mostra el contingut obtingut. *date* és la variable que conté el moment del temps d'aquell valor. *hits* és el nombre de cerques d'aquell moment de la sèrie temporal. És important saber que aquest valor és una freqüència relativa. És a dir, està sobre una escala de 100. Segons el conjunt total de valors que pren la sèrie es recalculen amb el màxim d'aquests. Això pot ser un problema ja que a l'hora de "concatenar" més d'una sèrie, o voler comparar-ne més d'una, s'ha de fer l'exportació des del Google Trends alhora. D'aquesta manera, els valors obtinguts han siguts tots "reescalats" de la mateixa forma i amb la mateixa base. Seguidament, per a l'anàlisi de la presència de bombolles només m'interessa la data i el valor que pren la sèrie temporal, em quedo amb un data frame amb aquestes dues columnes.

```
dfm <- data.frame(t=resm[["interest_over_time"]]$date,
y=(resm[["interest_over_time"]]$hits))
```

A la columna "t" hi guardo les dates de la sèrie i a "y" els valors que pren la variable resposta de la sèrie.

A continuació, hi ha alguns casos en que hi ha moments del temps en que el valor és tant petit que en comptes d'aparèixer un zero, la funció torna per defecte un nou nivell

anomenat “<1”. Això comporta problemes a l'hora de fer-ne una gràfica. Per tant, faig una modificació i els hi assigno el valor numèric 0 als individus que presenten aquest nivell.

```
levels(df$y)[1] <- "0"
```

Com acabo de dir, a continuació faig un gràfic amb el paquet *ggplot2*. Al tractar-se d'una visualització senzilla per veure la forma de la sèrie al llarg del temps,

```
ggplot(data = dfm, aes(x = t, y = y)) + geom_line(color = "black", size = 0.4)
```

Seguint amb l'exemple de l'script que estic explicant, obtinc el gràfic de la Figura 2.2X. És interessant fer una ullada les dades perquè ens pot donar alguna pista d'on es poden localitzar les bombolles, si la sèrie presenta tendència, estacionalitat, etc.

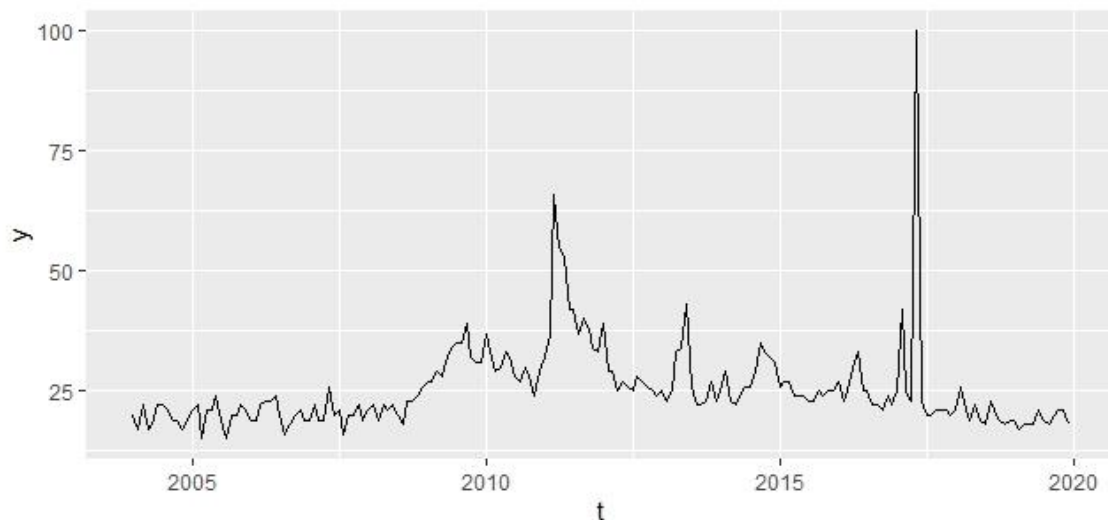


Figura 2.2. Gràfic de la sèrie temporal d'exemple.

En aquest cas, he anat analitzant la paraula “manel”. No sembla de massa interès fins al voltant del 2007-2008 que va començar a donar-se a conèixer un grup de música amb aquest nom a Barcelona. Llavors, es pot intuir que no és fins a finals de l'any 2008 que la sèrie presenta un comportament massa destacable. S'intueix una variabilitat normal i constant, però sense seguir cap patró ni amb tendència a la sèrie. A partir d'allà hi ha un creixement destacable i al 2011 es veu veure un pic força senyalat. Això coincideix amb el llançament del seu segon àlbum, amb el qual encara es van donar a conèixer més, esdevenint *trending topic* a nivell espanyol. La sèrie va presentant pics, decreixements, etc., fins a mitjans del 2016 on hi ha el màxim valor d'aquesta sèrie, i a més, es presenta de forma esporàdica i molt puntual. Això coincideix amb el llançament del seu quart àlbum.

És important fer aquesta anàlisi exploratòria per a poder conèixer i entendre el comportament de les dades: pics destacables, tendència local o general, etc. D'aquesta manera he pogut saber com s'ha comportat al llarg de tota la sèrie i després puc comparar el que primer he intuït a ull nu amb els resultats dels testos estadístics que apliqui.

2.2.2 Aplicació del test de bombolles

Així doncs, el següent pas és passar de l'anàlisi i aplicar el test PSY (2015) per poder veure la presència o no de bombolles, i a més, poder comparar-ho amb les suposicions que s'han fet.

Per a l'aplicació d'aquest test he de determinar una sèrie de paràmetres que s'utilitzen per a modificar com d'estricta es vol el test i amb quins quantils es volen comparar els valors obtinguts.

```
y <- dfm$y
```

S'assigna a *y* el vector amb els valors de la sèrie.

```
obs <- length(y)
```

obs és el nombre de valors de la sèrie, la llargada.

```
swindow0 <- floor(obs * (0.01 + 1.8 / sqrt(obs)))
```

S'obté amb la següent fórmula:

$$swindow0 = \left\lfloor n^{\circ} obs \cdot \frac{0.01 + 1.8}{\sqrt{n^{\circ} obs}} \right\rfloor$$

swindow0 és un número enter que depèn directament del nombre total de dades que conté la sèrie. Es tracta de la longitud mínima que fa la "finestra petita" que utilitza l'algoritme per a poder comprovar si hi ha, o no, un comportament explosiu en el valor analitzat en aquell moment.

```
IC <- 0
```

IC serveix per a seleccionar el nombre de retards. Si pren el valor 0, s'utilitza un retard fixat a 1. El valor 1 indica que utilitza el Criteri d'informació d'Akaike (AIC) i el valor 2 fa servir el Criteri d'Informació Bayesiana (BIC).

```
adflag <- 2
```

adflag és el nombre màxim de retards.

```
yr      <- 2
```

yr serveix per a calcular la finestra gran (*Tb*, a continuació). Es tracta d'un valor que influeix molt en la rigidesa dels valors crítics que es calculen. Com major sigui *yr*, més estricte és el test. Per tant, com menor sigui més facilitat de trobar una bombolla a la sèrie analitzada. En aquest cas, utilitzo un valor de 2.

```
Tb      <- 12*yr + swindow0 - 1
```

El paràmetre *Tb* el calculo a partir dels valors de la finestra petita (*swindow0*) i del valor anterior.

```
nboot   <- 99
```

El número de rèpliques que faig amb bootstrap es guarda al paràmetre *nboot*.

A continuació, una vegada determinats tots els paràmetres, faig els càlculs dels estadístics associats a través de la funció PSY2. Primer se li han d'assignar quatre paràmetres: el vector amb els valors de la sèrie, la longitud mínima de la finestra (*swindow0*), el criteri *IC* i *adflag*.

De manera simple, aquesta funció és un algorisme que està compost de dos bucles. En primer lloc, hi ha r_2 , el qual pren valors des de l'observació núm. *swindow0* (valor mínim) fins al total de llargada de la sèrie (nombre total de dades). r_1 pren valors des de $[1, r_2 - swindow0 + 1]$. És a dir, per a cada valor r_2 , hi ha un conjunt de valors de r_1 . Cada combinació de r_1 i r_2 serveix per a calcular l'estadístic ADF. Aquests dos paràmetres, els quals serveixen d'índexs, els utilitzo per a delimitar un conjunt de valors de la sèrie, els quals en calcularé l'estadístic.

Per entendre-ho de forma més visual, si per exemple hi ha una sèrie amb 7 valors i el valor de la finestra mínima (*swindow0*) val 3, l'esquema dels valors calculats seria el que es mostra la Taula 2.1.

| r_2 | r_1 | $y[r_2 : r_1]$ | | | | |
|-------|-------|----------------|--------|--------|--------|--------|
| 3 | 1 | y[1:3] | | | | |
| 4 | 1:2 | y[1:4] | | y[2:4] | | |
| 5 | 1:3 | y[1:5] | y[2:5] | y[3:5] | | |
| 6 | 1:4 | y[1:6] | y[2:6] | y[3:6] | y[4:6] | |
| 7 | 1:5 | y[1:7] | y[2:7] | y[3:7] | y[4:7] | y[5:7] |

Taula 2.1. Exemple del funcionament de l'algorisme.

De cada conjunt de valors de la sèrie (y), en calculo l'estadístic ADF i em quedo amb el valor màxim de cada fila. És a dir, per a cada valor de r_2 , que és el que realment acaba recorrent sense repetició els valors als quals correspon un estadístic, em quedo amb un únic valor.

Per a $r_2 = 4$, em quedo amb el $\max\{\text{ADF}(y[1:4]), y[2:4]\}$ i així successivament.

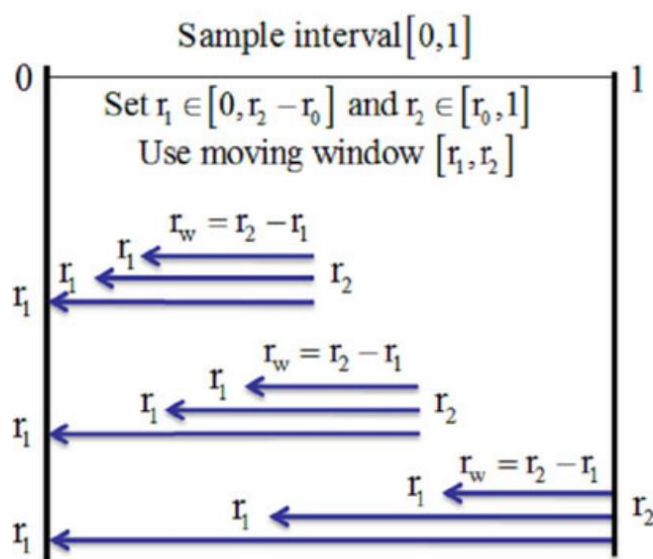


Figura 2.3. Esquema de com funcionen les submostres del test PSY (Phillips, Wu i Yu (2015)).

Com s'aprecia a l'esquema de la Figura 2.3, a cada valor de la sèrie calculo uns quants estadístics. Es tracta d'un mètode recursiu, el test GSADF.

Sempre hi haurà menys estadístics que la llargària original ja que amb aquest mètode es perden unes quantes dades de l'inici del qual no es pot calcular. Això és degut a que per a calcular la presència o no de bombolla en un punt es tenen en compte els valors passats. Llavors, clarament no es pot calcular aquest estadístic de la primera dada, ja que se'n desconeix el passat.

Una vegada ja he calculat els estadístics, m'interessa calcular-ne els punts de tall per determinar a partir de quin valor (valor crític) en aquell punt hi haurà, o no, presència de bombolla. Per això s'utilitza la funció *quantilsPSY*.

quantilsPSY necessita els paràmetres anteriors de l'amplada de la finestra petita, el criteri IC i el paràmetre *adflag*. Quan he hagut d'utilitzar aquesta funció m'he trobat amb el problema que està codificada per a utilitzar-la amb computació paral·lela. La computació paral·lela és quan els càlculs es divideixen en més d'una computadora per així eliminar l'excés de càrrega i millorar la velocitat de càlcul i l'obtenció dels resultats. A l'hora d'aplicar aquesta funció *quantilsPSY*, com que ho executo des del meu ordinador i sense aquesta opció de computació en paral·lel, ho he de modificar. Al tractar-se d'una funció ja definida pel paquet no es pot modificar com a tal. És per això que he de copiar el codi i modificar-lo per a que funcioni de redefinir. La nova funció, definida amb el mateix nom, ja no presenta aquest problema.

Amb aquesta funció també vaig veure que el "nivell de significació" que utilitza la funció per a obtenir els quantils estava marcada al 99%, 95% i 90%. Per a poder fer una anàlisi més profunda he volgut canviar aquests valors i a la mateixa funció *quantilsPSY*, hi he afegit un nou paràmetre el qual anomeno *val.quantils*. En aquest hi assigno un vector amb tots els valors, que pertanyen a l'interval $[0, 1]$, de significació que vull comprovar. Amb aquest canvi, puc rebaixar el nivell d'exigència per a la detecció de bombolles, ja que hi ha sèries temporals on al gràfic s'hi veia un caràcter viral/de bombolla en algun període, no hi detectava res. Després de provar varis canvis amb els valors dels paràmetres que he descrit anteriorment, he vist que modifiquen molt la dificultat amb la que el test pot trobar o no bombolles. Així doncs, canviant els valors d'alguns paràmetres per al valor de les finestres els resultats canviaven significativament, però tot i així els valors crítics donaven valors molt allunyats dels estadístics calculats amb la funció anterior.

Del paquet *psymonitor*, utilitzo quatre funcions: *PSY*, *cvPSYwmboot*, *ADFres* i *locate*. A continuació, mostro les línies de codi que he modificat.

La funció *PSY* no l'he modificat, però l'he reanomenat *PSY2*.

La funció *cvPSYwmboot* és l'encarregada d'aplicar l'algorisme que retorna els valors dels estadístics calculats per cada valor de la sèrie. Aquesta funció és la que té més pes computacional ja que utilitza bootstrap per al càlcul d'aquests i això fa que necessiti temps. Aquesta funció està originalment pensada per a funcionar amb computació paral·lela, però en el meu cas no puc fer-ho. Quan executava això em produïa errors ja que no trobava l'opció, així que ho he modificat.

La funció calcula per defecte els quantils (nivell de significació) 99, 95 i 90%, però hi he afegit un nou paràmetre anomenat *val.quantils* on quan crido a la funció, l'hi assigno a aquest paràmetre un vector amb tots els valors dins l'interval [0 , 1], és a dir, amb els nivells de significació que vull per a l'execució concreta. Això és interessant perquè es pot veure a partir de a quin nivell de significació un període comença a considerar-se amb un caràcter explosiu i a ser considerat "bombolla" segons els paràmetres donats.

El codi següent aquesta funció esmentada *quantilsPSY*, (antiga *cvPSYwmboot*) i es mostra en vermell el codi que he eliminat i en ver el que he afegit/modificat.

```
quantilsPSY <- function (y, swindow0, IC = 0, adflag = 0, Tb, nboot = 199,
                        useParallel = FALSE, nCores, val.quantils)
{
  qe <- as.matrix(val.quantils)
  nboot <- nboot
  result <- ADFres(y, IC, adflag)
  beta <- result[[1]]
  eps <- result[[2]]
  lag <- result[[3]]
  T0 <- length(eps)
  t <- length(y)
  dy <- as.matrix(y[2:t] - y[1:(t - 1)])
  g <- length(beta)
  if (missing(swindow0)) {
    swindow0 <- floor(t * (0.01 + 1.8/sqrt(t)))
  }
  if (missing(Tb)) {
    stop("Missing a value for 'Tb'", call. = FALSE)
  }
  suppressWarnings(RNGversion("3.5.0"))
  set.seed(101)
  rN <- matrix(sample(1:T0, Tb*nboot, replace = TRUE), nrow = Tb, ncol = nboot)
  wn <- matrix(rnorm(1), nrow = Tb, ncol = nboot)
  dyb <- matrix(0, nrow = Tb - 1, ncol = nboot)
  dyb[1:lag, ] <- rep(dy[1:lag], times = nboot)
  for (j in 1:nboot) {
    if (lag == 0) {
      for (i in (lag + 1):(Tb - 1)) {
        dyb[i, j] <- wn[i - lag, j] * eps[rN[i - lag, j]]
      }
    }
    else if (lag > 0) {
      x <- matrix(0, nrow = Tb - 1, ncol = lag)
      for (i in (lag + 1):(Tb - 1)) {
        x <- matrix(0, nrow = Tb - 1, ncol = lag)
        for (k in 1:lag) {
          x[i, (k)] <- dyb[(i - k), j]
        }
        dyb[i, j] <- x[i, ] %*% beta[2:g,1] + wn[i-lag,j] * eps[rN[i-lag,j]]
      }
    }
  }
}
```

```

yb0 <- rep(y[1], times = nboot)
dyb0 <- rbind(yb0, dyb)
yb <- apply(dyb0, 2, cumsum)
if (useParallel == TRUE && missing(nCores)) {
  nCores <- detectCores() - 1
}
else {
  nCores <- 1
}
cl <- makeCluster(nCores)
registerDoParallel(cl)
dim <- Tb - swindow0 + 1
i <- 0
MPSY <- foreach(i = 1:nboot, .inorder = FALSE, .combine = rbind) %dopar%
{
  PSY(yb[, i], swindow0, IC, adflag)
}
stopCluster(cl)
MPSY <- 0
for(i in 1:nboot){
  aux <- PSY2(yb[, i], swindow0, IC, adflag)
  MPSY <- rbind(MPSY, aux)
}
SPSY <- as.matrix(apply(MPSY, 1, max))
Q_SPSY <- as.matrix(quantile(SPSY, qe))
return(Q_SPSY)
}

```

D'altra banda, hi ha la funció *locate*, la qual associa els estadístics que són superiors als nivells de significació i els hi associa la data a la qual pertanyen. En verd hi ha ressaltat les línies de codi que he afegit per a que funcionés.

```

locate2 <- function (index, dates) {
  maxi <- max(index)
  lc <- which.max(index)
  if (maxi == 1) {
    count <- 1
    EP <- matrix(0, nrow = 30, ncol = 2)
    i <- lc + 1
    EP[count, 1] <- dates[lc]
    while (i <= length(index)) {
      if (index[i - 1] == 1 && index[i] == 0) {
        EP[count, 2] <- dates[i - 1]
        i <- i + 1
      }
      else if (index[i - 1] == 0 && index[i] == 1) {
        count <- count + 1
        EP[count, 1] <- dates[i]
        i <- i + 1
      }
      else {
        i <- i + 1
      }
    }
  }
  OT <- EP[1:count, ]

```

```

if(length(OT) == 2){
  OT <- matrix()
  OT <- rbind(EP[1:count,])
}
v <- nrow(OT)
if (OT[v, 2] == 0) {
  OT[v, 2] <- dates[length(dates)]
}
OT <- as.Date(OT, origin = "1970-01-01")
}
else if (maxi == 0) {
  OT <- NULL
  warning("No bubble or crisis periods found", call. = FALSE)
}
return(OT)
}

```

Continuant amb l'exemple donat, he decidit calcular els valors crítics per als quantils 60%, 80%, 90% i 95%. A continuació, a la Figura 2.4 mostra l'evolució dels valors dels estadístics calculats i amb vermell, hi ha afegits els valors crítics:

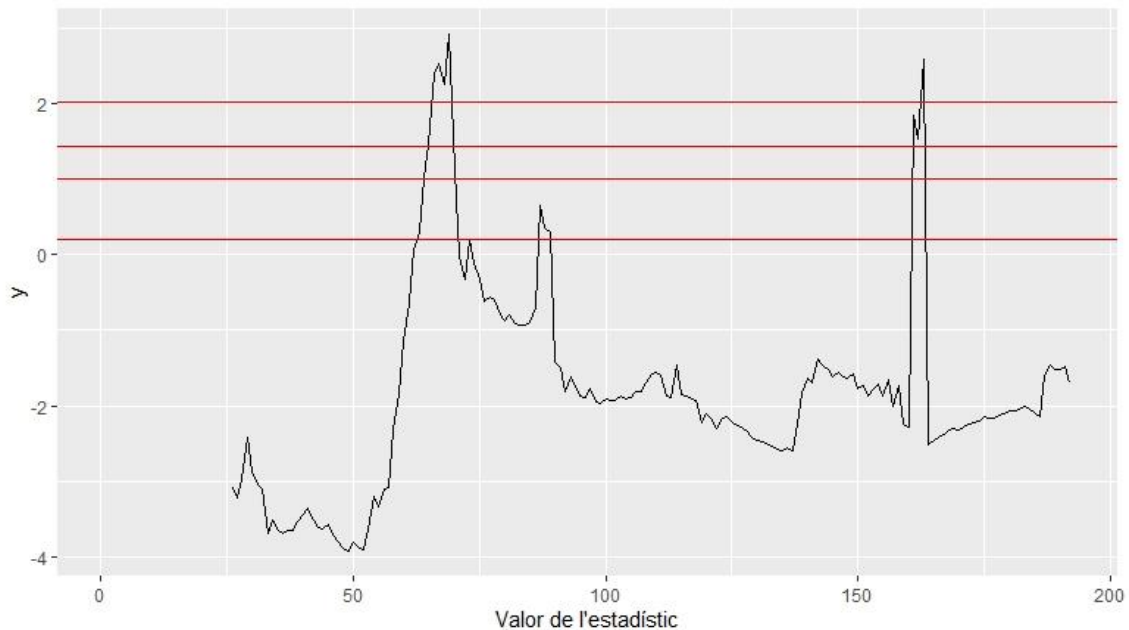


Figura 2.4. Evolució dels estadístics d'exemple amb els punts crítics al 95, 90, 80 i 60%.

Per als tres nivells de confiança més elevats, el 95, el 90 i el 80% les bombolles detectades són les mateixes. Per al nivell del 80% la primera bombolla potser comença algun valor abans, però totes van referenciades a unes dates similars, si no les mateixes. Per al valor del 60% també detecta una bombolla just després del primer període de crisi. Com es pot veure, hi ha un *gap* al principi de la sèrie, com ja he comentat, que correspon als estadístics que no podem calcular. És important remarcar que és fàcil la confusió quan hi ha el gràfic de la sèrie temporal i el dels estadístics seguits fent contrastos de "forma vertical".

Per últim, a la Taula 2.2 es mostren tots els períodes de bombolla trobats amb la data d'inici i de final, classificats segons el nivell escollit.

| Nivell | start | end |
|---------------|--------------|------------|
| 95 % | 2009-06-01 | 2009-09-01 |
| | 2017-07-01 | 2017-07-01 |
| 90 % | 2009-05-01 | 2009-09-01 |
| | 2017-05-01 | 2017-07-01 |
| 80 % | 2009-04-01 | 2009-10-01 |
| | 2017-05-01 | 2017-07-01 |
| 60 % | 2009-03-01 | 2009-10-01 |
| | 2011-03-01 | 2011-05-01 |
| | 2017-05-01 | 2017-07-01 |

Taula 2.2. Evolució dels estadístics de la sèrie d'exemple (manel) amb els punts crítics al 95, 90, 80 i 60%.

Als quatre nivells de significació que he escollit (95, 90, 80 i 60) es detecta una bombolla a l'any 2009. Aquesta comença, segons els nivell, entre juny i març i s'allarga fins al setembre o a l'octubre.

Com ja he comentat al principi quan he presentat el perquè he escollit aquesta sèrie, es podria relacionar aquest pic del 2009 amb el *boom* que van tenir després de llençar al seu primer àlbum i la gira que van fer durant aquell any. A més a més, tots quatre nivells detecten una bombolla de maig a juliol (o només juliol en el cas del 95%) de l'any 2017. És curiós veure com per exemple, observant la Figura 2.2 i amb el que he comentat anteriorment, al 2011 sembla que s'hauria de poder intuir o almenys detectar un període de crisi i en canvi, el Test PSY detecta un període explosiu al 2009 on a primer cop d'ull no sembla massa rellevant.

A la Figura 2.5 es mostren les bombolles detectades.

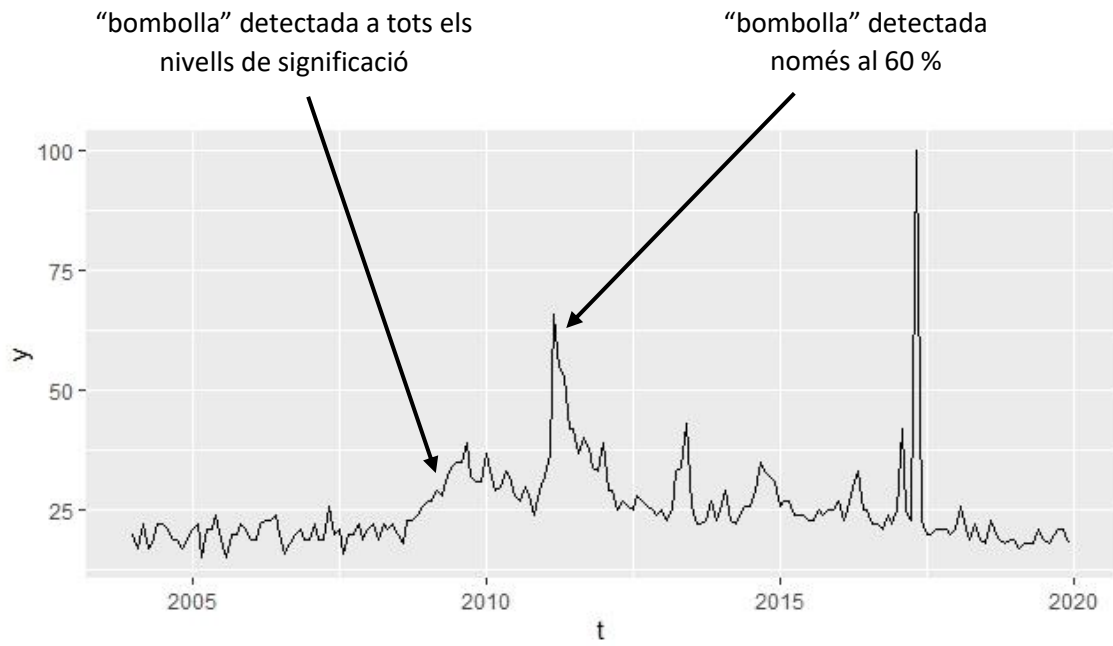


Figura 2.5. Gràfic de la sèrie manel amb períodes comentats.

III. RESULTATS

En aquest apartat, mostro els resultats igual que a la Figura X de la sèrie *manel*, però de la resta de sèries que he analitzat. En cada una especifico quins valors dels paràmetres he utilitzat i què obtingut. A més, he mirat d'estudiar el comportament de cada una i intentat explicar el perquè dels períodes de crisi detectats.

Fins ara el test de PSY de detecció de múltiples bombolles l'he explicat amb la sèrie de la paraula "manel", però aquest procediment també l'he repetit variant alguns paràmetres del test i les dates amb altres paraules. A continuació presento aquests resultats i comento el perquè del comportament de cada del test aplicat a la les sèries.

Les primeres sèries presentades són: *iphone*, *pants*, *lithium* i *obama*; les quals les he obtingut a través de Google Trends. Després, hi ha la sèrie *spread*, la qual és la que el paquet *psymonitor* utilitza com a exemple per a aplicar el test. Per últim, hi ha la sèrie *bitcoin* la qual també l'he extret de Google Trends, però la periodicitat és setmanal. A través de *Yahoo Finance* es pot veure com ha emergit i va incrementar el seu valor, i consegüentment va fer-se viral, al voltant del 2018. És per això, que he determinat un interval més curt de manera que obtinc més valors i la sèrie és més "precisa" durant aquell temps.

3.1 Sèrie iphone

He triat analitzar la sèrie temporal de la paraula *iphone* perquè he pensat que segurament és una paraula que presentaria més d'un pic ja que cada any o cada dos anys Apple llença al mercat un nou iphone i sempre crea tema de debat online i es viralitza aquesr terme. Es tracta d'una sèrie de periodicitat mensual i 144 observacions.

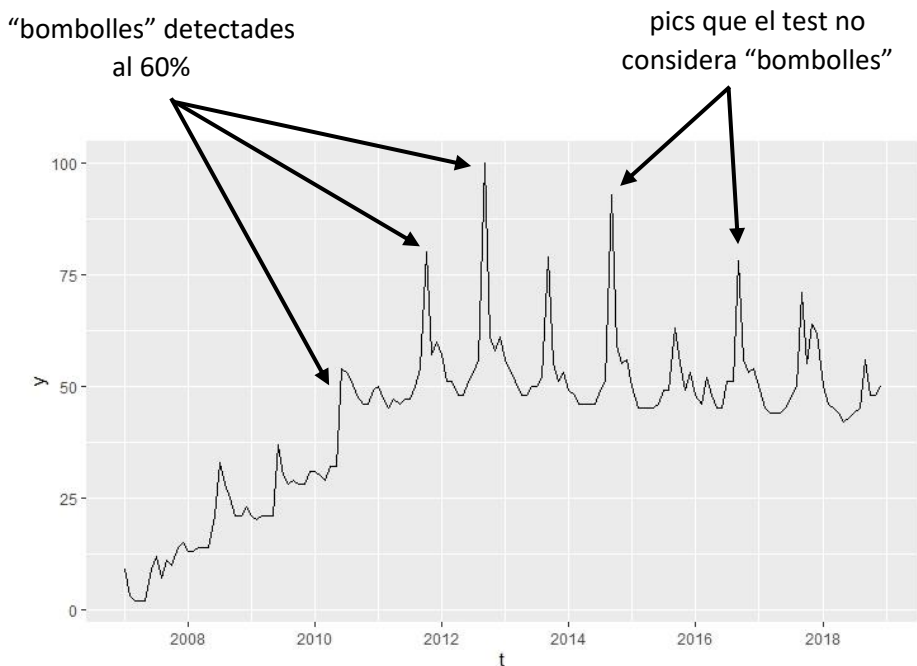


Figura 3.1. Gràfic de la sèrie iPhone.

Com es pot veure a la Figura 3.1, cada un temps determinat la sèrie presenta un augment considerable, que coincideix amb el llançament del nou model d'iPhone al mercat. Tot i així, al nivell 90 i 95 % el test no detecta cap comportament de bombolla.

| Nivell | start | end |
|--------|-------------------------------------|------------|
| 95 % | "No bubble or crisis periods found" | |
| 90 % | "No bubble or crisis periods found" | |
| 80 % | 2010-07-01 | 2010-08-01 |
| | 2011-10-01 | 2011-10-01 |
| 60 % | 2010-06-01 | 2010-08-01 |
| | 2011-10-01 | 2011-10-01 |
| | 2012-09-01 | 2012-09-01 |

Taula 3.1. Períodes detectats pels estadístics de la sèrie iPhone amb els punts crítics al 95, 90, 80 i 60%.

Aquest comportament de tenir un pic aproximadament una vegada a l'any, sembla que comença a repetir-se de forma similar a partir del 2011. El test considera els valors anteriors al punt analitzat per a calcular els estadístics, i és per això que a la que aquest comportament comença a ser repetit, aquests "pics" deixen de ser bombolles i a formar part del comportament normal de la sèrie.

Aquests períodes detectats coincideixen amb el llançament al mercat de l'iPhone 4, 4s i el 5 als anys 2010, 2011 i 2012, respectivament.

3.2 Sèrie pants

La sèrie *pants* l'he escollit per veure com es comporta una paraula que aparentment no sembla que hagi de tenir cap pic destacable o no cap moment "viral", ja que és una paraula comuna. Conté un total de 107 observacions, de periodicitat mensual. A la Figura 3.2 de a continuació es mostra la sèrie.

És curiós veure com sí que sembla haver-hi un cert comportament sobretot amb tendència creixent i estacional. Cada línia vertical blanca, igual que al gràfic anterior, és el canvi d'un any a l'altra. Sembla veure's que a finals d'any, segurament en època prèvia al Nadal sol haver-hi sempre un augment de les cerques.

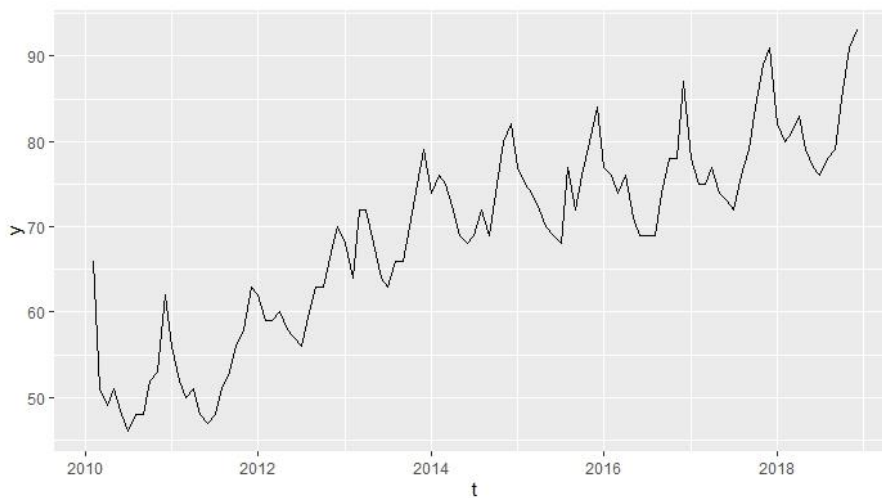


Figura 3.2. Gràfic de la sèrie pants.

Una vegada he aplicat el test a la sèrie, observo amb la Taula 3.2 que en cap moment ha trobat cap bombolla. És a dir, la sèrie sembla no haver estat viral en cap moment.

| Nivell | start | end |
|--------|-------------------------------------|-----|
| 95 % | "No bubble or crisis periods found" | |
| 90 % | "No bubble or crisis periods found" | |
| 80 % | "No bubble or crisis periods found" | |

Taula 3.2. Períodes detectats pels estadístics de la sèrie pants amb els punts crítics al 60, 80, 90 i 95%

3.3 Sèrie lithium

Al voltant de l'any 2018 es va produir un gran augment en la demanda de liti, com a matèria prima, a causa de la gran demanda i increment de cotxes elèctrics. Aquests cotxes necessiten unes bateries potents per a poder funcionar, les quals estan fabricades amb liti. A conseqüència d'això, el preu de mercat del liti va patir una inflació

a l'any 2018 i analitzar si es van fer gaires cerques respecte això pot ser interessant, també per veure quan va començar.

Altres cop, la sèrie és mensual amb un total de 96 observacions.

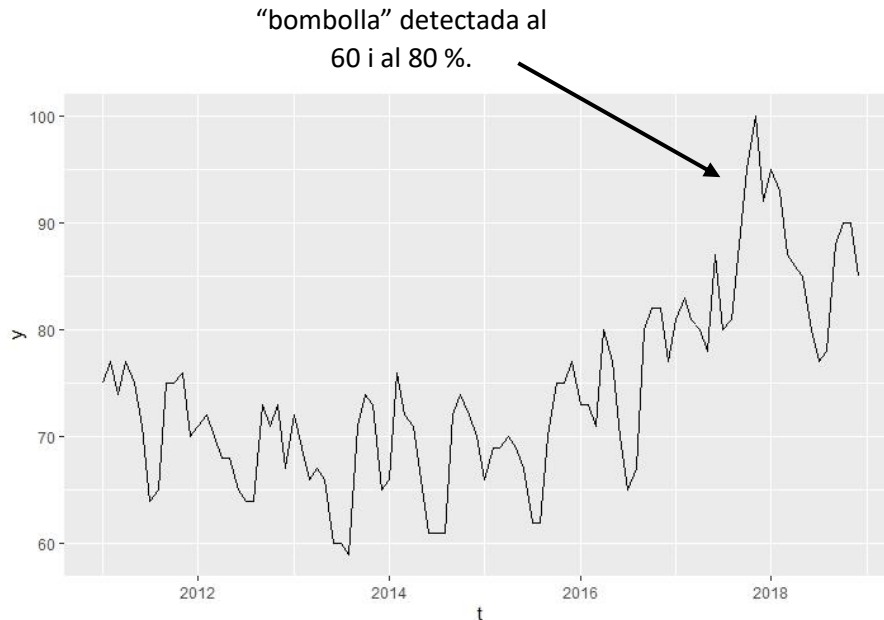


Figura 3.3. Gràfic de la sèrie lithium.

A la Figura 3.3 es mostren els valors de la sèrie des del 2011 fins al 2019. No he trobat necessari afegir els anys anteriors ja que a nivell visual no es podia apreciar res destacable, més aviat un comportament normal i estable. Com es pot apreciar, a partir del 2016 sembla que es comença a produir un augment del nombre de cerques, amb el punt màxim just abans del 2018.

Amb el test PSY, com s'aprecia a la Taula 3.3, només troba una bombolla en aquest punt, just abans del 2018.

| Nivell | start | end |
|--------|-------------------------------------|------------|
| 90 % | "No bubble or crisis periods found" | |
| 80 % | 2017-11-01 | 2017-11-01 |
| 60 % | 2017-11-01 | 2017-11-01 |

Taula 3.3. Períodes detectats pels estadístics de la sèrie lithium amb els punts crítics al 90, 80 i 60%.

Per a gent que inverteix en el mercat de productes com el liti, o simplement per saber com fluctuarà aquesta "tendència", és interessant poder detectar quan comença a fer-se viral.

3.4 Sèrie obama

En Barack Obama, més conegut i buscat digitalment només pel seu cognom “obama”, va presentar-se a les eleccions a la presidència dels EEUU al gener del 2009 amb el partit Demòcrata. Com es pot observar clarament a la Figura X, des que comença el 2009 que es produeix un gran augment del nombre de cerques en comparació amb les que hi havia abans. També es veu com a finals del 2009 encara es produeix un pic més elevat, segurament per accions preses durant la seva legislatura. Al cap de 4 anys, quan tornen a haver-hi eleccions es torna a presentar i torna a guanyar.

Es tracta d'una sèrie mensual altra vegada, amb 144 observacions.

Com es pot veure a la Figura 3.4, coincideix amb el segon pic més elevat, que es produeix cap a finals del 2012. Per últim, es veu un altre període amb un nombre més elevat de cerques pel final del 2016-principis del 2017, que és quan es van tornar a celebrar les eleccions.

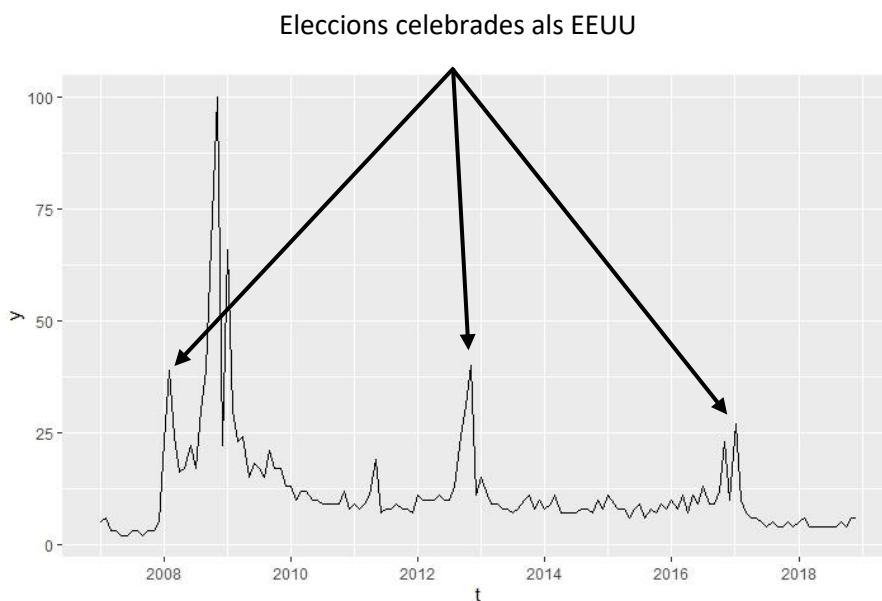


Figura 3.4. Gràfic de la sèrie obama.

El primer pic (al 2009) en comparació amb el segon (2012) i el tercer (2016-2017) és molt més elevat, tot i que veient els resultats de la Taula 3.4 aquesta tercera bombolla és detetada per a tots els nivells analitzats. En canvi, aquest període amb el valor màxim (2009) que semblaria ser quan la bombolla és més clara, només la detecta al 60%.

| Nivell | start | end |
|--------|------------|------------|
| 90 % | 2017-01-01 | 2017-02-01 |
| 80 % | 2016-12-01 | 2017-03-01 |
| 60 % | 2008-11-01 | 2008-11-01 |
| | 2012-11-01 | 2012-11-01 |
| | 2013-02-01 | 2013-02-01 |
| | 2016-11-01 | 2017-03-01 |

Taula 3.4. Períodes detectats pels estadístics de la sèrie obama amb els punts crítics al 90, 80 i 60%.

3.5 Sèrie bitcoin

A través de Google Trends, he especificat un període més curt, que vaig del 2015 al 2018 incloïnt ambdós sencers. Aquesta sèrie presenta periodicitat setmanal amb un total del 209 observacions.

Com ja he explicat abans, la bombolla del *bitcoin* va produir-se a finals del 2017-principis del 2018. A més, tot i que la moneda ja existia abans, no hi ha cap comportament sospitós o si més no, rellevant per a voler analitzar.

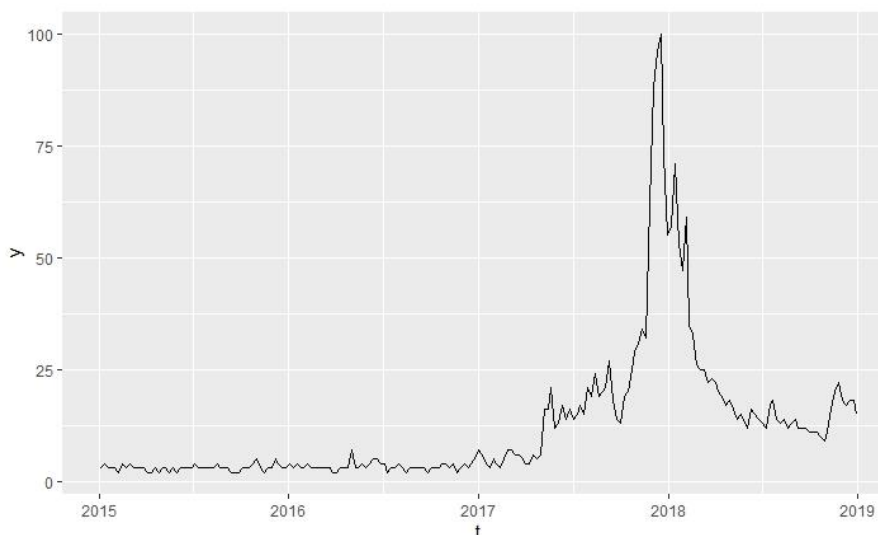


Figura 3.5. Gràfic de la sèrie bitcoin.

Com s'observa a la Figura 3.5 i coincidint amb el comentat, l'augment es produeix mica en mica a partir del 2017 fins que a finals d'aquest mateix any pren els valors més elevats fins que acaba "explotant" i tornant a decreixer.

| Nivell | start | end |
|--------|------------|------------|
| 95 % | 2017-11-26 | 2017-12-17 |
| | 2018-01-07 | 2018-01-14 |
| 90 % | 2017-05-21 | 2017-05-21 |
| | 2017-09-10 | 2017-09-10 |
| | 2017-11-26 | 2018-01-14 |
| 80 % | 2017-05-21 | 2017-05-21 |
| | 2017-06-11 | 2017-06-11 |
| | 2017-08-13 | 2017-08-13 |
| | 2017-09-10 | 2017-09-10 |
| | 2017-11-05 | 2018-01-14 |
| 60 % | 2017-05-21 | 2017-05-21 |
| | 2017-06-11 | 2017-06-11 |
| | 2017-08-13 | 2017-09-10 |
| | 2017-10-29 | 2018-01-14 |

Taula 3.5. Períodes detectats pels estadístics de la sèrie bitcoin amb els punts crítics al 95, 90, 80 i 60%.

A la Taula 3.5 es veu com el mateix “període” en el qual es diria fàcilment que només hi ha una bombolla, el test en detecta més d’una en tots els nivells. Com més baix el nivell de tolerància amb la qual decideixo aplicar el test, més “ample” es fa l’interval del període total de la bombolla.

3.6 Sèrie spread

Per últim, la sèrie *spread* l’he pres del mateix paquet *psymonitor*, on per explicar les funcions definides que he explicat abans i com aplica el test, utilitza aquesta funció com a exemple.

Aquesta sèrie té periodicitat mensual amb un total de 229 observacions.

Aquests valors indiquen com ha fluctuat l’índex de risc que té la gent que compra el deute públic d’un conjunt de països. D’ençà la crisi econòmica que va començar al 2007-2008, el risc va anar augmentant i augmentant, juntament amb l’expansió de la crisi. Aquest crac va provocar que el risc que tenia la gent que comprava deute públic de molts països, entre d’altres Espanya, Itàlia o Grècia, va passar de valors inferiors a 1 a fins a 6.

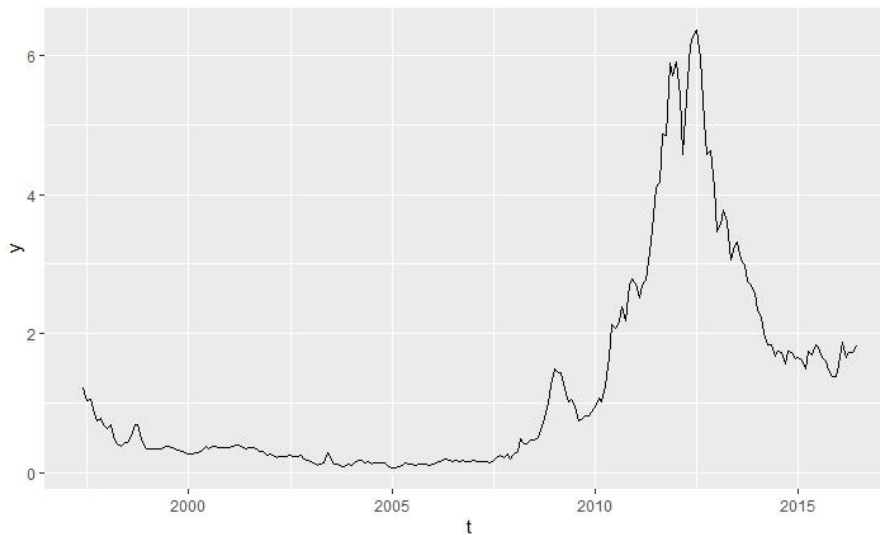


Figura 3.6. Gràfic de la sèrie spread.

Igual que la sèrie anterior, el comportament (només tractant el període que es mostra i l'analitzat) que es veu a la Figura 3.6, s'aprecia com va augmentar des del 2008, on es veu un petit pic, arribant al punt màxim al voltant del 2012.

| Nivell | start | end |
|--------|------------|------------|
| | 2008-03-01 | 2008-03-01 |
| 99 % | 2008-10-01 | 2009-03-01 |
| | 2010-05-01 | 2012-08-01 |
| 95 % | 2008-03-01 | 2008-03-01 |
| | 2008-09-01 | 2009-04-01 |
| | 2010-05-01 | 2012-09-01 |
| 90 % | 2008-03-01 | 2008-03-01 |
| | 2008-09-01 | 2009-04-01 |
| | 2010-05-01 | 2012-09-01 |

Taula 3.6. Períodes detectats pels estadístics de la sèrie spread amb els punts crítics al 99, 95 i 90%.

Tal i com mostra la Taula 3.6, a la sèrie d'exemple utilitzada el test troba alguna bombolla, i amb un nivell de fins al 99% amb més o menys amb coincidència amb els altres valors, la sèrie presenta una bombolla en aquest petit pic. Després una altra molt marcada i més llarga que va des del maig del 2010 fins a l'agost-setembre del 2012.

IV. ANÀLISI DE LES SÈRIES TEMPORALS

Definició. Una sèrie temporal és una seqüència d'observacions sobre una variable ordenats cronològicament. Aquestes poden mesurar des de magnituds econòmiques, meteorològiques (precipitacions, temperatures d'una regió, etc.), o de producció industrial, per exemple. La denotem com $\{Y_1, Y_2, \dots, Y_T\}$, on el subíndex t fa referència al moment del temps al que es troba el valor Y . Aquestes observacions estan preses amb uns intervals de temps regular: mensualment, anualment, setmanalment, etc. I el total d'observacions el denoto amb el paràmetre T . Les sèries temporals són de gran interès per a poder fer prediccions d'un futur i a través d'un nivell de significació també per interval.

Tanmateix, és necessari fer un estudi previ de com ha evolucionat aquesta variable al llarg del temps, per poder decidir quin tipus de sèrie és i poder utilitzar uns models de predicció o uns altres. Hi ha dues corrents que distingeixen les formes de predicció i és la forma en que s'enfoca la predicció i interpretació de la variable.

Per una banda hi ha l'anàlisi determinista. El determinisme, segons el DIEC, és la <<Teoria segons la qual, conegudes exactament unes certes condicions, els fets que se'n deriven poden ésser previstos amb exactitud>>. Malgrat l'exactitud que es considera en el determinisme, en Estadística s'ha de considerar un terme d'error que farà que els esdeveniments posteriors no puguin ser previstos amb un 100% d'exactitud, sinó amb un error.

D'altra banda tenim l'anàlisi estocàstica. Aquesta considera la sèrie temporal com un procés estocàstic i per tant, la finalitat és la modelització d'aquest i veure quin model s'hi ajusta millor minimitzant l'error.

4.1 Determinista

Començo fent una anàlisi clàssica de la sèrie temporal de la cerca de la paraula "manel" i explicar-lo detalladament amb tots els seus passos.

En aquest cas, la sèrie temporal es tracta d'una seqüència de la variable Y que determina la "freqüència relativa de cerques de la paraula". Per aquest motiu, el domini de Y és $[0, 100]$. El valor màxim de 100 el pren quan s'han produït el nombre màxim de cerques en el període de temps determinat. És a través d'aquest valor màxim que es relativitzen els altres.

En l'anàlisi determinista el primer pas és classificar el tipus de sèrie que és segons la seva tendència i la seva estacionalitat. En aquesta branca es considera la sèrie temporal com una funció amb quatre components que expliquen la distribució d'aquesta,

$$Y_t = f(T_t, C_t, S_t, I_t)$$

La tendència d'una sèrie temporal (T_t) és la component que representa l'evolució al llarg del temps de la sèrie. Aquesta pot ser creixent o decreixent. Per a mirar si una sèrie presenta tendència sempre és útil fer un gràfic en primer lloc i analitzar a simple vista com fluctuen els valors al llarg de tot el domini de t . Després, pot afegir-s'hi la "recta de regressió" que ajuda a decidir si en té o no.

L'estacionalitat (S_t) és la component que recull les oscil·lacions de cada període complet i es repeteixen en els períodes successius. Per exemple, si es tracta d'una sèrie mensual, es pot mirar si la sèrie presenta estacionalitat d'any en any comparant l'oscil·lació d'un any sencer amb la de l'any següent.

El cicle (C_t) són el conjunt de moviments oscil·latoris relatius a la tendència que es mantenen en períodes superiors a l'any complet.

La component irregular (I_t) són les variacions que no estan recollides a les components anteriors que no es pot controlar, i que per tant tenen un caràcter residual. Poden ser components que provenen d'efectes aleatoris no explicats (aleatòries) o bé components imprevisibles que poden ser identificades a posteriori.

Per a l'anàlisi determinista que faig només contemplo l'estudi de l'estacionalitat i la tendència. Per tant, després de l'estudi del comportament de la sèrie classifico segons el següent criteri:

- SÈRIE TIPUS I: ni tendència, ni estacionalitat.
- SÈRIE TIPUS II: no tendència, però sí amb estacionalitat.
- SÈRIE TIPUS III: tendència, però no amb estacionalitat.
- SÈRIE TIPUS IV: tendència i estacionalitat.

Per a poder determinar quin tipus de sèrie és cada una de les analitzades en primer lloc faig representació gràfica on també afegeixo la línia de tendència (regressió lineal). A través del gràfic es pot intuir quin tipus de sèrie és, però és necessària l'evidència estadística per a poder classificar-la bé. Per això, realitzo dos contrastos estadístics: contrast de Daniel per avaluar la tendència i el Test de Kruskal-Wallis per a l'anàlisi de l'estacionalitat.

A continuació, explico detalladament l'anàlisi determinista de la sèrie temporal de la paraula "manel". Es tracta d'una sèrie amb periodicitat mensual.

Primer faig una anàlisi exploratòria gràfica i a continuació, avaluo l'estacionalitat i la tendència. Tots els càlculs i els gràfics els faig amb l'Excel.

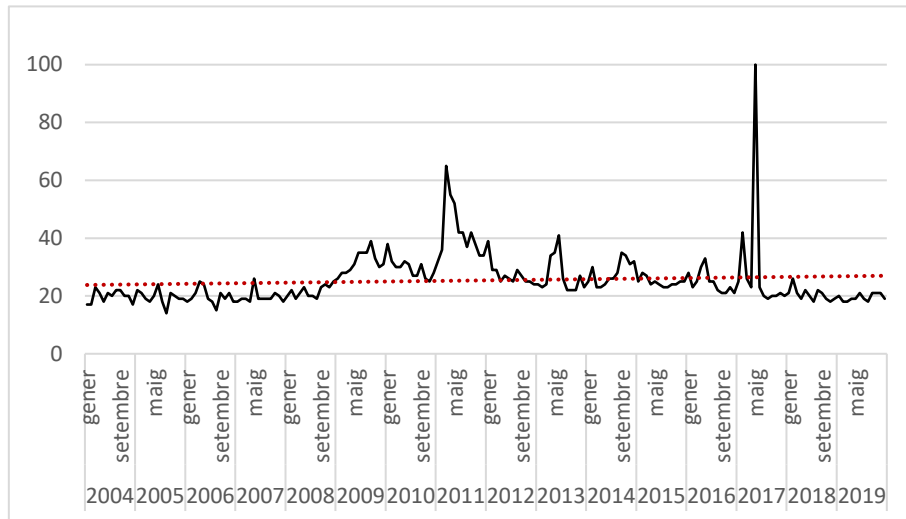


Figura 4.1. Gràfic de la sèrie manel amb Excel amb línia de la tendència.

Com s'aprecia al gràfic de la Figura 4.1 la sèrie sembla que tingui una certa tendència creixent gairebé inapreciable, molt suau, però present. A més, no sembla que hi hagi cap presència d'estacionalitat. Fins al 2008 semblaria seguir una distribució similar: aleatòria i amb gairebé res a destacar. Després al 2009 ja comença a augmentar una mica, i al 2011 es produeix un altre pic destacable. No és fins al 2017 que es produeix un altre pic molt destacable i més breu, però llavors es torna a suavitzar la sèrie.

Per a l'anàlisi estadística, començo amb el contrast de Daniel. Continuant amb l'exemple de la sèrie temporal de la paraula "manel", es plantegen les següents hipòtesis:

$$\begin{cases} H_0: \text{la sèrie no presenta tendència. No té la component de la tendència.} \\ H_1: \text{la sèrie presenta tendència. Té la component de la tendència.} \end{cases}$$

Aquest test es basa en la posició de la variable ordenada cronològicament i la posició que ocupa cada valor ordenada segona la variable "Rang". Rang és la posició que ocupa si la sèrie està ordenada de menor a major valor. A continuació, es calcula la diferència d'aquestes dues variables i es calcula l'estadístic T_s :

$$T_s = 1 - \frac{6 \sum_{t=1}^T d_t^2}{T(T^2 - 1)}$$

d_t és la diferència de la variable Rang i la variable índex t . T és el nombre total d'observacions de la sèrie. A continuació, s'estandarditza l'estadístic Ts de manera que es pugui comparar amb una Normal amb mitjana 0 i variància 1.

$$Z = \sqrt{T-1} * Ts \sim N(0,1)$$

L'estadístic obtingut amb aquesta sèrie pren un valor de $Z = 9.06$. Es compara amb el valor de la distribució de referència que deixa un error del 5%, és a dir, al 95% de confiança (5% de significació). Aquest valor és el 1.96. Com que l'estadístic obtingut $Z > 1.96$, tinc suficient evidència estadística per a rebutjar la hipòtesi nul·la i acceptar l'alternativa que ens diu que la sèrie presenta tendència. A través del gràfic es veu que és una tendència creixent.

Seguint amb l'anàlisi, continuo mirant la presència o no d'estacionalitat amb el Test de Kruskal-Wallis. Plantejo les següents hipòtesis:

H_0 : la sèrie no presenta estacionalitat. No té la component de l'estacionalitat.

H_1 : la sèrie presenta estacionalitat. Té la component de l'estacionalitat.

L'estadístic de Kruskal-Wallis, KW , té en compte la posició de la variable Rang que ocupen els valors de la mateixa posició de cada període. És a dir, com que en la sèrie manel, l'estacionalitat la miro a nivell anual (si es repeteix algun tipus de patró any rere any), es tracta de comparar tots els valors del mes de "Gener" dels diferents anys, amb tots els "Febrer", i així successivament amb tots els mesos o subperíodes dins el període gran o dins l'any.

$$KW = \frac{12}{T(T+1)} * \left[\sum_{i=1}^s \frac{(R_i)^2}{T_i} \right] - 3 * (T+1) \sim \chi_{s-1}^2$$

On T és altra vegada el nombre total d'observacions i s és el nombre de períodes estacionals dins de cada any (període gran). En el meu cas s és 12 ja que cada any té un total de 12 mesos. Si per exemple les dades fossin trimestrals, s seria 4. R_i és la suma de tots els valors de la variable Rang que corresponen al període estacional. T_i és el nombre d'observacions que corresponen a l'estació i -èsima mateix. És a dir, R_i és la suma de tots els valors de la variable El Rang de per exemple, el mes de gener. I si hi ha 15 observacions del mes de gener, T_i és 15. Normalment tots els valors de T_i prenen valors força similars, si no el mateix, d'aquesta manera tinc aproximadament les mateixes dades per a cada estació i -èsima. Per últim, comparo el valor obtingut amb una distribució Chi-Quadrat de $s-1$ graus de llibertat.

En el cas de la sèrie temporal “manel”, com ja he dit abans, és una sèrie mensual, però per fer el Test de Kruskal-Wallis, s’analitza l’estacionalitat d’any rere any. Cada mes té un total de 16 observacions i l’estadístic KW obtingut pren un valor de 6.76, i comparant-lo amb el valor crític d’una distribució Chi-Quadrat d’11 graus de llibertat amb un nivell de significació del 5%, que és 19.675, no puc rebutjar la hipòtesi nul·la. Per tant, no hi ha la suficiència estadística per a acceptar la hipòtesi alternativa, i se suposarà que la sèrie no té estacionalitat. Això coincideix amb la primera valoració de la sèrie a través del gràfic.

Per tant, la sèrie temporal de la paraula “manel” es classifica com a sèrie tipus III.

Per a les sèries de les paraules “spread”, “pants”, “obama”, “iphone”, “bitcoin” i “lithium”, també s’ha realitzat la mateixa anàlisi: gràfica, contrast de Daniel i Test de Kruskal-Wallis. A continuació, mostro els resultats de totes les paraules.

4.1.1 Anàlisi determinista de la sèrie spread

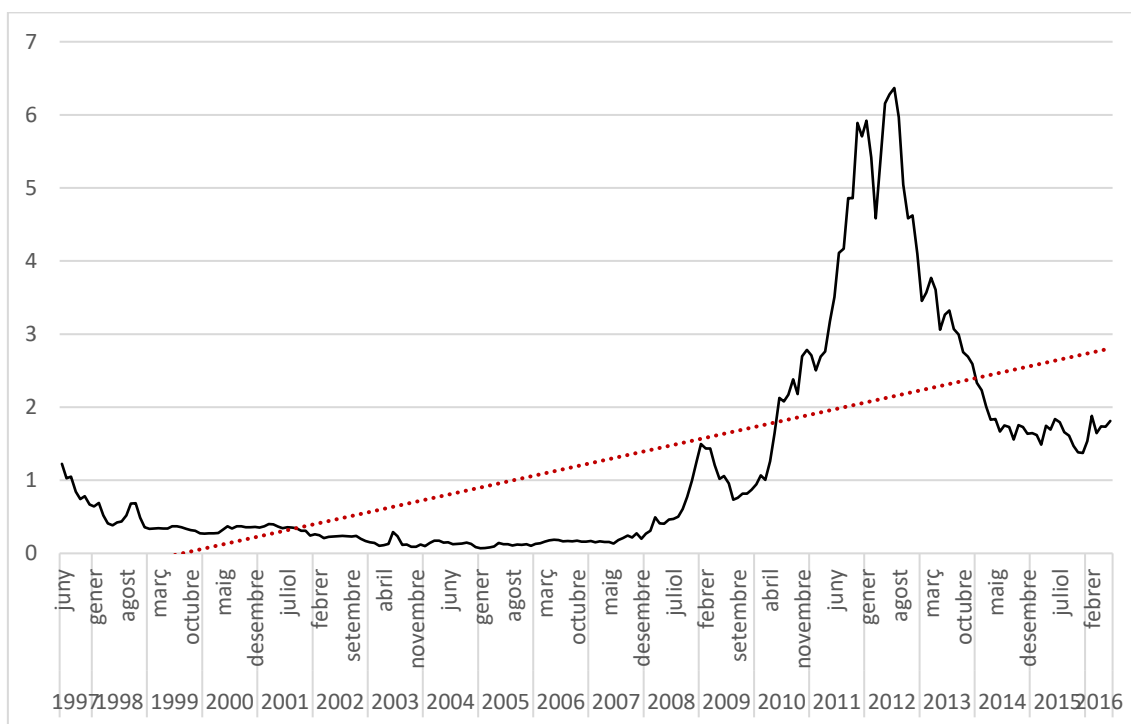


Figura 4.2. Gràfic de la sèrie spread amb Excel amb línia de la tendència.

A través del gràfic de la Figura 4.2 s’observa una clara tendència creixent, i pel que sembla bastant produïda per l’augment de la freqüència que apareix a partir de finals de l’any 2007. Segueix incrementant en valors fins al voltants del 2012, que torna a decreixer. No sembla que la sèrie presenti cap mena d’estacionalitat.

Amb el contrast de Daniel obtinc un estadístic $Z = 9.18$. Comparat amb el 1.96 de la distribució $\text{Normal}(0,1)$, rebutjo la hipòtesi nul·la i s'accepta que la sèrie presenta tendència.

Amb el Test de Kruskal-Wallis obtinc un estadístic $KW = 0.72$, i comparat amb el 19.68 de la Chi-Quadrat d'11 graus de llibertat, no es pot rebutjar la hipòtesi nul·la de no estacionalitat.

La sèrie de la paraula "spread" es classifica com a sèrie tipus III amb tendència i sense estacionalitat. Això coincideix amb el que a primer cop d'ull s'havia intuït amb el gràfic.

4.1.2 Anàlisi determinista de la sèrie pants

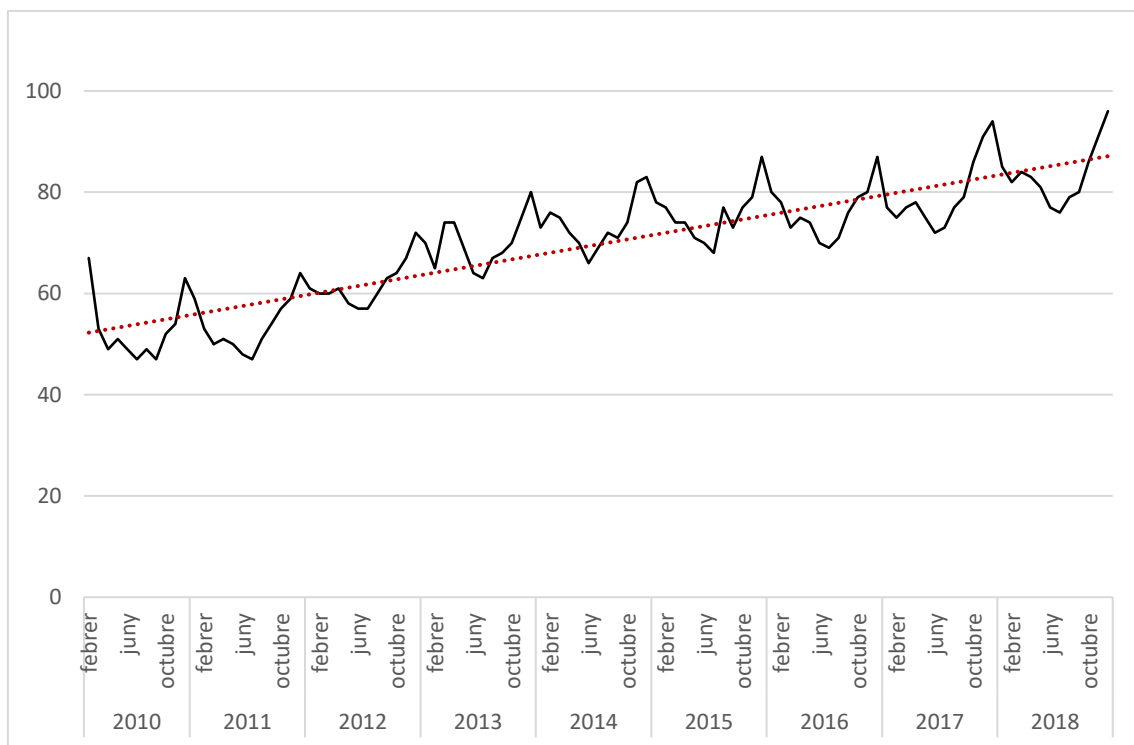


Figura 4.3. Gràfic de la sèrie pants amb Excel amb línia de la tendència.

Aquesta sèrie presenta una clara tendència creixent. A més, també sembla que presenti un comportament estacional, ja que hi ha una mena de patró que es va repetint. Podria tractar-se d'una sèrie tipus IV, tal i com es pot intuir a la Figura 4.3.

El contrast de Daniel dona un valor $Z = 9.02$, que comparat amb el valor crític de 1.96, puc acceptar amb força evidència la presència de tendència a la sèrie.

Amb l'anàlisi de l'estacionalitat amb el test de Kruskal-Wallis, obtinc un valor KW de 17.692, i per poc no tinc la suficiència estadística de rebutjar la hipòtesi nul·la, comparat

amb el valor crític de 19.675. Per tant, la sèrie no presenta estacionalitat amb el nivell de confiança del 95%.

Finalment aquesta sèrie és tipus III, amb tendència i sense estacionalitat. Com ja he comentat i vist al gràfic, sí que sembla que presenti un cert patró estacional, però no està suficientment marcat com per a poder acceptar que la sèrie presenta una component estacional amb el nivell de confiança convencional.

4.1.3 Anàlisi determinista de la sèrie obama

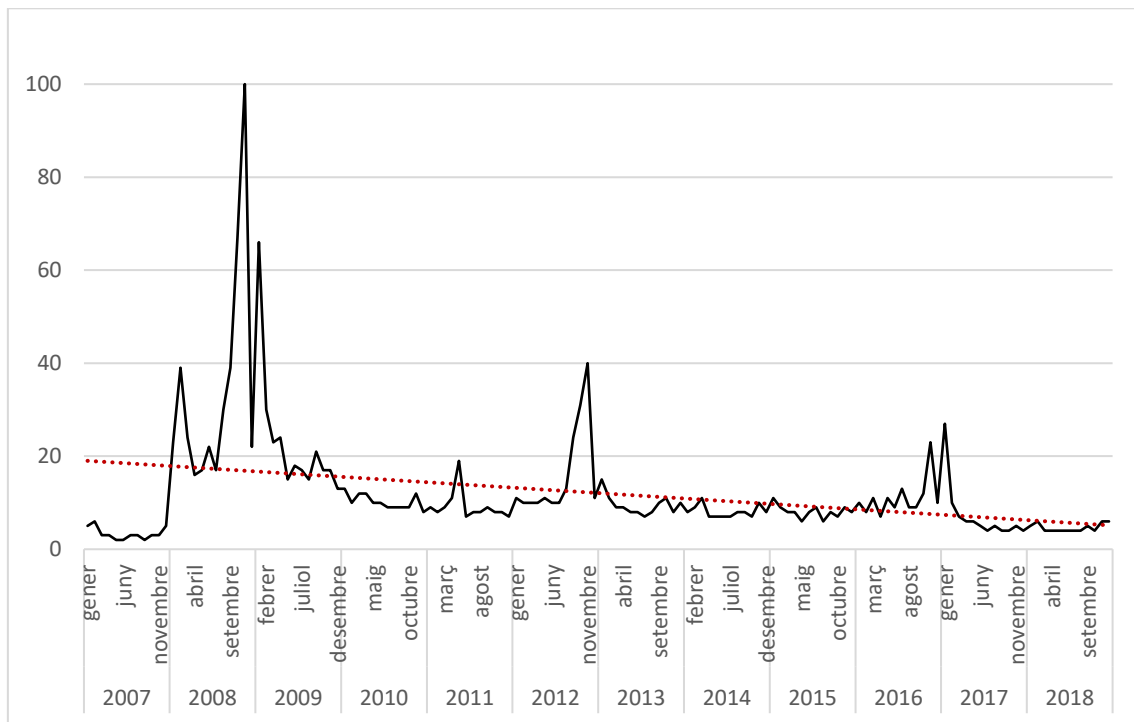


Figura 4.4. Gràfic de la sèrie obama amb Excel amb línia de la tendència

A través de la Figura 4.4, es veu com la recta de regressió d'aquesta sèrie té una tendència decreixent, tot i que bastant suau. Pot ser que presenti aquest pendent negatiu a causa dels pics que hi ha a l'inici, en comparació amb els més petits que hi ha cap al final de la sèrie. No sembla que hi hagi estacionalitat de cap tipus.

Analitzant la tendència amb el contrast de Daniel obtinc un estadístic $Z = 2.52$, que comparat amb el valor crític de la Normal (0,1) de 1.96, hi ha suficiència estadística per a poder acceptar la hipòtesi alternativa de presència de tendència.

Per l'altra banda, amb el Test de Kruskal-Wallis obtinc un estadístic $KW = 6.28$, que comparat amb el 19.68 de la distribució de referència, no puc rebutjar la hipòtesi nul·la.

Coincidint amb la primera impressió, la sèrie és de tipus III, sense estacionalitat i amb una tendència decreixent.

4.1.4 Anàlisi determinista de la sèrie iphone

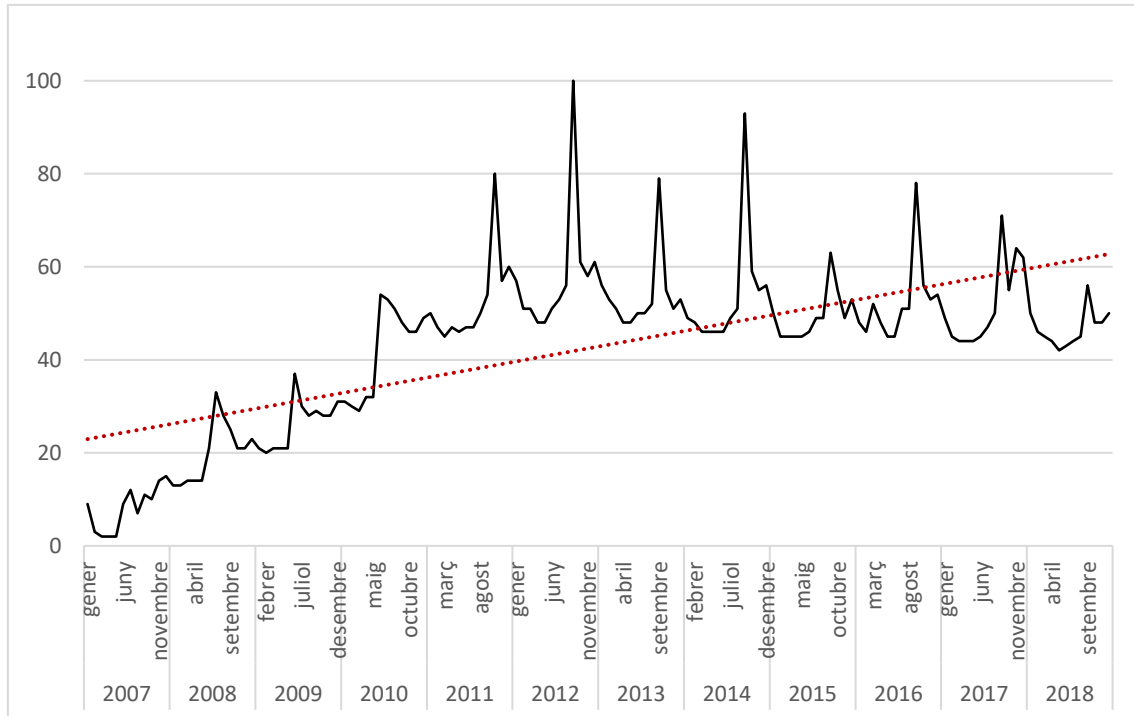


Figura 4.5. Gràfic de la sèrie iphone amb Excel amb la línia de la tendència.

A través de la Figura 4.5 es pot apreciar una tendència creixent. Sembla que al principi el creixement sigui més notable que després, que s'estabilitza. Tot i així, el més destacable sembla una mena d'estacionalitat que provoca uns pics cap als mesos de juliol-setembre. Per tant, es diria que sembla una sèrie tipus IV amb estacionalitat i tendència.

L'estadístic del contrast de Daniel pren un valor de 8.997 que comparat amb el valor crític de la Normal (0,1) de 1.96, es pot rebutjar amb certesa la hipòtesi nul·la i acceptar l'alternativa que diu que la sèrie presenta tendència.

Seguidament, amb l'anàlisi de l'estacionalitat obtinc un estadístic KW de 26.882 que comparant-lo amb el valor crític de 19.675 es pot acceptar la hipòtesi alternativa d'estacionalitat.

Per tant, la sèrie temporal de la paraula 'iphone' té estacionalitat i tendència creixent, i és tipus IV.

4.1.5 Anàlisi determinista de la sèrie lithium

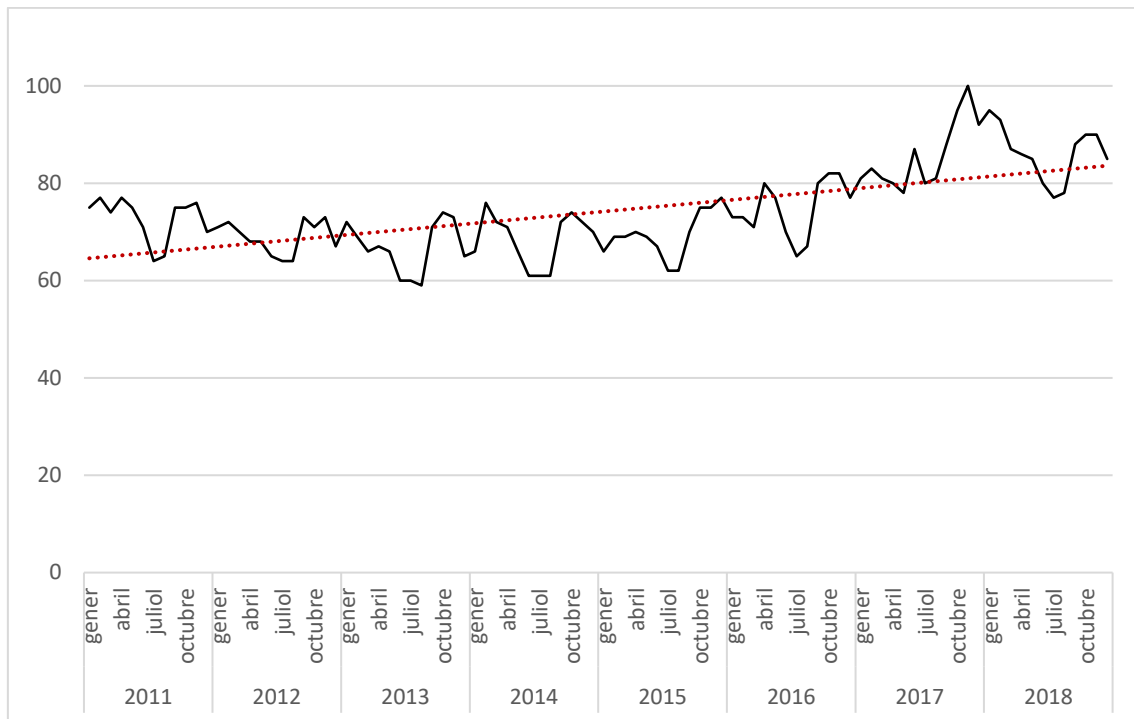


Figura 4.6. Gràfic de la sèrie lithium amb Excel amb la línia de la tendència.

Amb el gràfic de la Figura 4.6 es pot veure que la sèrie té una lleugera tendència creixent, però no sembla presentar cap patró estacional gaire definit.

Amb el contrast de Daniel s'obté un estadístic $Z = 5.83$, i per tant, es pot rebutjar la hipòtesi nul·la, i es diu que la sèrie presenta tendència.

Respecte l'estacionalitat, el valor de l'estadístic de Kruskal-Wallis corresponent a aquesta sèrie és de 22.6, i comparant-lo amb el valor crític de la distribució de referència de la Chi-Quadrat d'11 graus de llibertat, de 19.675, es pot rebutjar la hipòtesi nul·la.

La sèrie presenta tendència i estacionalitat i per tant, es tracta d'una sèrie tipus IV. Amb el gràfic sembla no intuir-se massa l'estacionalitat, però això a vegades és confús degut a l'escala que s'utilitza. Si es canvia, potser és més fàcil intuir-la a cop d'ull:

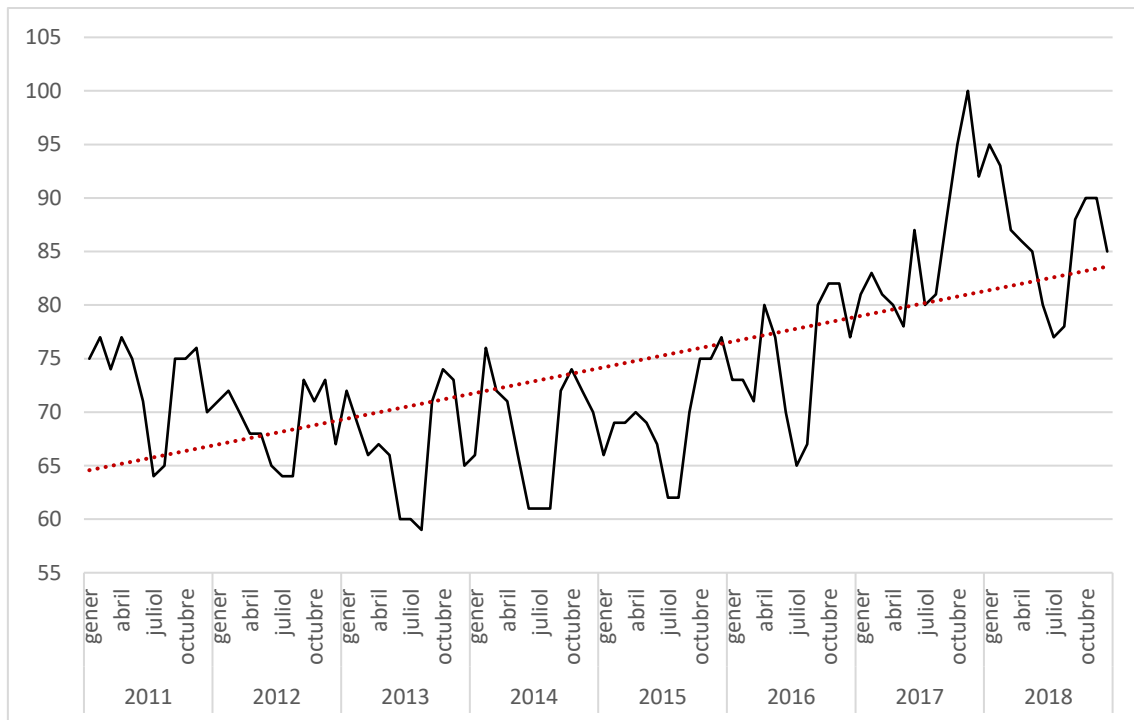


Figura 4.7. Gràfic de la sèrie lithium canviat d'escala amb Excel amb la línia de la tendència.

Un cop canviant l'escala, Figura 4.7, sembla que potser sí que es detecta un comportament estacional al llarg de la sèrie, on sembla prendre valors més baixos al voltant dels mesos d'estiu.

4.1.6 Anàlisi determinista de la sèrie bitcoin

Com ja he explicat en l'apartat on presento els resultats del test de cada sèrie, la sèrie bitcoin és la que es diferencia una mica de la resta. En comptes de ser extreta de Google Trends, és de Yahoo Finance. Es tracta d'una sèrie amb periodicitat setmanal que va des de l'any 2015 fins al 2018 sencer.

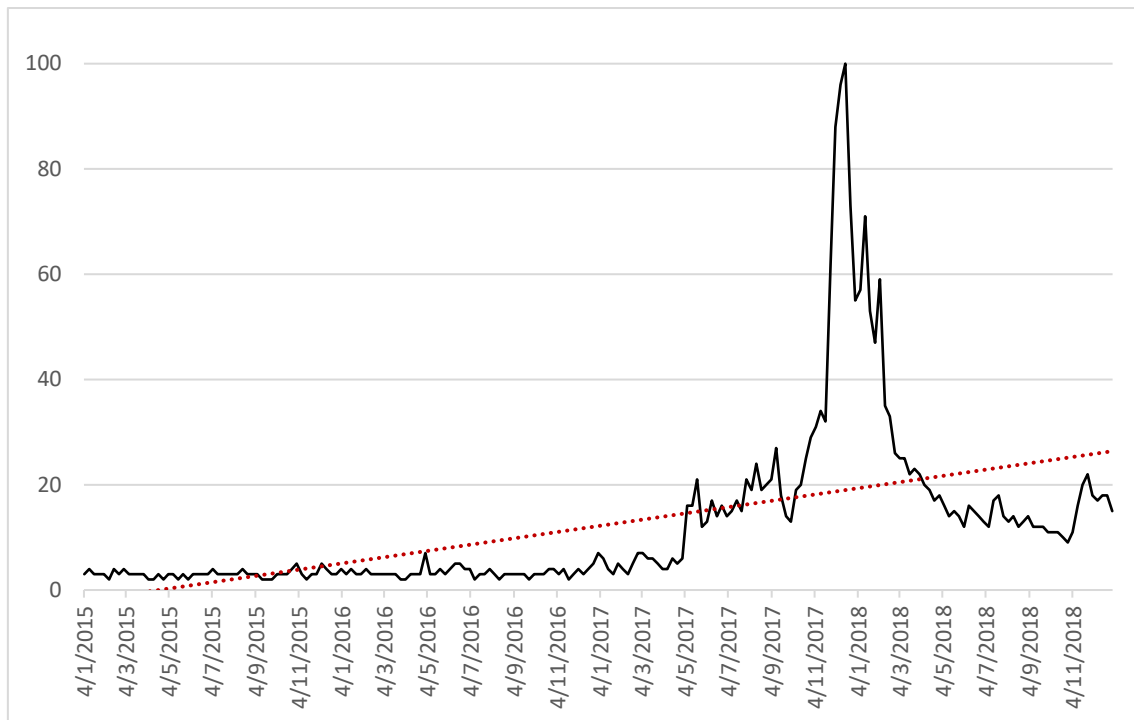


Figura 4.8. Gràfic de la sèrie bitcoin amb Excel amb la línia de la tendència.

A la Figura 4.8, el més destacable d'aquesta sèrie és el bruscanvi en la sèrie a partir de la meitat de l'any 2017. S'observa un moviment sense tendència i no molt elevat al principi, però llavors comença a incrementar-se fins que a principis del 2018 es produeix un pic molt elevat. Després, la sèrie torna a disminuir en valors, però segueixen sent més elevats que al principi. Pel que fa l'estacionalitat, no sembla que pugui apreciar-se'n. Amb la línia de tendència afegida sembla que sí que hi hagi una certa tendència creixent.

Amb el contrast de Daniel obtinc un valor de l'estadístic Z de 12.04, motiu pel qual hi ha suficient evidència estadística per a rebutjar la hipòtesi nul·la. S'accepta que la sèrie té tendència.

Pel que fa el Test de Kruskal-Wallis, s'ha hagut de fer una petita modificació. Al tractar-se d'una sèrie setmanal, per avaluar l'estacionalitat, s'ha hagut de calcular una mitjana de cada conjunt de valors corresponents al mateix mes, per a poder comparar-lo amb la mitjana del conjunt de valors del mes següent. El problema és que en els mateixos mesos, alguns anys tenen 4 valors i d'altres en tenen 5. Aplicant el Test de Kruskal-Wallis a la sèrie modificada d'aquesta manera, obtinc un valor KW = 2.35, que queda molt lluny de la zona de rebuig de la hipòtesi nul·la. Per tant, no es pot provar que la sèrie presenti estacionalitat.

Tal i com s'ha avaluat de manera gràfica, la sèrie presenta tendència creixent i no estacionalitat. Per tant, es tracta d'una sèrie de tipus III.

4.2 Predicció determinista

La sèrie ja classificada segons la tendència i l'estacionalitat serveix per a poder aplicar-hi uns mètodes de predicció o uns altres. Les prediccions serveixen per a poder estimar els futurs valors de la sèrie, que ara per ara són desconeguts. Sempre s'ha de tenir un error de predicció que acompanya les estimacions futures. Aquest error es pot tenir més o menys controlat, tot i que també es poden produir canvis estructurals que poden ser detectats a posteriori.

En aquest apartat només es tractaran mètodes de predicció per a les sèries tipus III i les sèries tipus IV, ja que són el tipus que s'ha trobat.

4.2.1 Sèries tipus III

Per a les sèries de tipus III es proposa que les sèries tinguin la següent estructura:

$$Y_t = T_t + U_t = \beta_0 + \beta_1 t + U_t$$

On es suposa linealitat per part de la tendència, T_t , i per tant, segueix la fórmula d'una recta amb el pendent β_1 que acompanya a t , i un valor β_0 com a terme d'intersecció amb l'eix d'ordenades.

Per a les sèries tipus III s'utilitza el Mètode de la tendència lineal. Aquest es tracta de calcular una simple recta de regressió amb un valor β_0 i un valor β_1 al llarg de tota la sèrie.

$$\widehat{\beta}_1 = \frac{\text{covar}(y_t, t)}{\text{var}(t)}, \widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 * \bar{t}$$

El següent mètode és el de les dobles mitjanes mòbils. Aquest mètode suposa que la sèrie segueix una tendència localment lineal. És a dir, no té en compte tot el conjunt de dades per a predir la tendència de cada punt, sinó que només té en compte uns quants punts anteriors a aquest. Aquesta quantitat depèn d'un valor de k , que com més elevat més memòria presenta la predicció de la tendència local.

En primer lloc calculem una mitjana de longitud k (MM) dels valors anteriors al punt t en que ens trobem. En aquest cas s'ha escollit $k = 2$. A continuació, es torna a fer una mitjana dels valors MM anteriors, també de longitud k , anomenada MM'. És a dir, es calculen les mitjanes mòbils de les mitjanes mòbils calculades en el pas anterior. Seguidament, s'estima la tendència i el pendent per cada punt:

$$\widehat{T}_t = 2MM_t - MM'_t$$

$$\widehat{\beta}_1(t) = \frac{2}{k-1} (MM_t - MM'_t) \quad , \text{ amb } k = 2$$

Per últim, per al mètode de l'allisament exponencial lineal de Holt (AEH) també se suposa que la tendència és localment lineal, i per tant, la finalitat és calcular el valor de la tendència i del pendent localment a cada punt. La diferència és que es tenen en compte tots els valors anteriors al punt donat amb més o menys intensitat, segons els paràmetre α i γ que es pot canviar.

Per a poder inicialitzar l'algorisme, s'inicialitza la variable T (la tendència local) amb Y_t , i β_1 amb el valor zero.

Els valors de de T són una combinació lineal del valor y_t i la seva predicció en aquell moment t . Depèn d'un paràmetre $\alpha \in [0,1]$. La tendència es va recalculant amb la següent equació d'actualització:

$$\widehat{T}_t = \alpha y_t + (1 - \alpha) \widehat{y}_t(1)$$

Les estimacions de β_1 són una combinació lineal entre la diferència entre la tendència en aquell moment i just l'anterior. L'estimació de β_1 en el moment anterior. Aquesta combinació depèn d'un paràmetre $\gamma \in [0,1]$. El pendent es va recalculant amb la següent equació d'actualització:

$$\widehat{\beta}_1(t) = \gamma [\widehat{T}_t - \widehat{T}_{t-1}] + (1 - \gamma) \widehat{\beta}_1(t-1)$$

Per últim, la predicció consisteix en la suma del valor estimat de la tendència i del pendent.

$$\widehat{y}_T(m) = \widehat{T}_t + \widehat{\beta}_1(T) * m$$

On m és l'índex del valor que ocupa l'observació en la sèrie.

A continuació, hi ha les prediccions de la sèrie de la paraula "manel". En el següent gràfic es mostra com evolucionen les prediccions a través del temps.

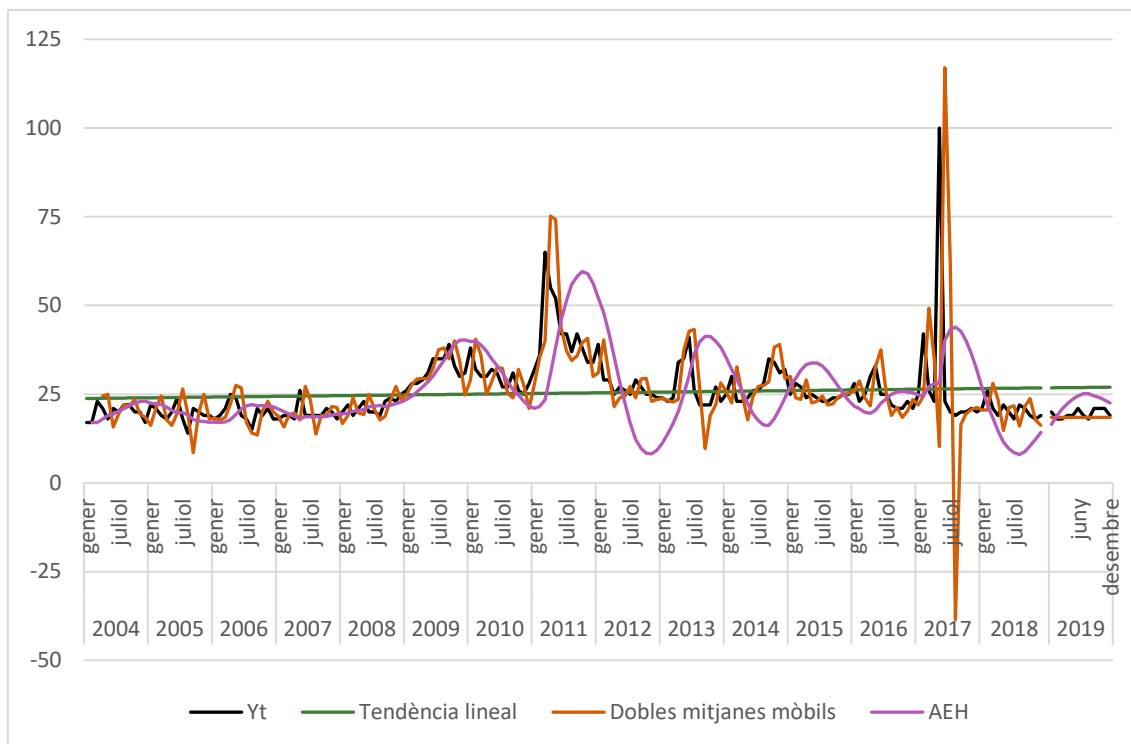


Figura 4.9. Prediccions per a la sèrie manel.

En la Figura 4.9 s'aprecia de color negre la sèrie original i en colors diferents la distribució que segueixen les prediccions que s'han fet amb els mètodes de la Tendència Lineal (verda), el mètode de les Dobles mitjanes mòbils (taronja) i l'Allisament Exponencial de Holt (lila). Per a aquest últim s'ha utilitzat una $\gamma = 0.79$ i una $\alpha = 0.09$, ja que són els que menys error mostraven.

També es pot apreciar com del 2018 al 2019 es produeix un salt. Això és degut a que totes les dades anteriors fins al desembre del 2018 és el període mostral, i el 2019 s'ha utilitzat com a període extra-mostrat.

Per a poder comparar entre mètodes, es necessita un indicador adimensional. S'utilitza l'Error Percentual Absolut Mitjà (EPAM) i a la Taula 4.1 hi ha els resultats obtinguts:

| EPAM mostral | | | EPAM extra-mostrat | | |
|--------------|-----------|-------|--------------------|-----------|-------|
| T. Lineal | Dobles MM | AEH | T. Lineal | Dobles MM | AEH |
| 0.204 | 0.282 | 0.305 | 0.384 | 0.432 | 0.221 |

Taula 4.1. Errors dels mètodes de Tendència Lineal, Dobles Mitjanes Mòbils i d'AEH.

El mètode que té menys error extra-mostrat. és l'AEH amb un 22.1%. Tot i així, és bastant elevat. Per a la resta de sèries temporals que anteriorment s'han classificat com a tipus III es seguiria el mateix procediment de mètodes de predicció.

A través d'aquestes prediccions he pogut veure que les bombolles són uns períodes que desestabilitzen el períodes posteriors a elles i conseqüentment, alteren les prediccions i fa que s'obtinguin errors molt elevats.

4.2.2 Sèries tipus IV

Per a les sèries tipus IV, les quals tenen tendència i component estacional, se suposa que segueixen el següent esquema:

$$Y_t = T_t + S_t + U_t$$

On la component T_t és la tendència i S_t és la component estacional. Per a aquest tipus de sèries es presenten dos mètode de predicció.

El primer és el mètode de descomposició. Aquest consisteix en fer una primera estimació de la tendència, $T_t^{(1)}$, a través d'una mitjana mòbil de longitud k ; assignant a cada valor de la mitjana mòbil el període que és el centre de la mateixa. Si k és un nombre parell es calcula una mitjana mòbil centrada, que és la mitjana mòbil de longitud 2 de la mitjana mòbil calculada anteriorment. Un cop tenim la primera estimació de la tendència, es calculen les components estacionals. Aquestes es calculen fent la diferència del valor real de la sèrie menys la primera estimació de la tendència.

$$S_t + U_t = Y_t - T_t^{(1)}$$

A continuació, es fa una mitjana per obtenir una component estacional mitjana per a cada estació (en aquest cas n'hi ha 12).

$$\overline{S}_i = \frac{\sum_{t \in T_i} S_t + U_t}{T_i}, \quad i = 1, 2, \dots, s$$

Seguidament es calcula la mitjana, \bar{S} , d'aquests valors per restar-la de cada \overline{S}_i i d'aquesta manera, eliminar les components irregulars. Finalment obtinc una component estacional per a cada estació que compon un període sencer, \widehat{S}_i . És a dir, s'obtenen 12 valors que són estimacions de les 12 components estacionals.

Lavors, es fa la segona estimació de la tendència restant el valor de les component estacional corresponent al valor real: $T_t^{(2)} = y_t - \widehat{S}_i$

A continuació, es segueix el mateix mètode de l'estimació de la tendència lineal que una sèrie tipus III.

Com a segon i últim mètode de predicció per a les sèries tipus IV hi ha l'allisament exponencial de Holt-Winters. Aquest mètode està basat en l'estimació de la tendència, del pendent i de la component estacional per a cada moment del temps. Tot això es fa a través d'uns paràmetres de pes: α , γ i δ . Tots aquests paràmetres $\in [0,1]$

S'ha utilitzat la sèrie de la paraula "lithium" per a les prediccions d'una sèrie tipus IV.

A continuació, el gràfic de la sèrie temporal de la paraula "lithium" amb les prediccions fetes a través dels dos mètodes de predicció per a sèries tipus IV explicats anteriorment.

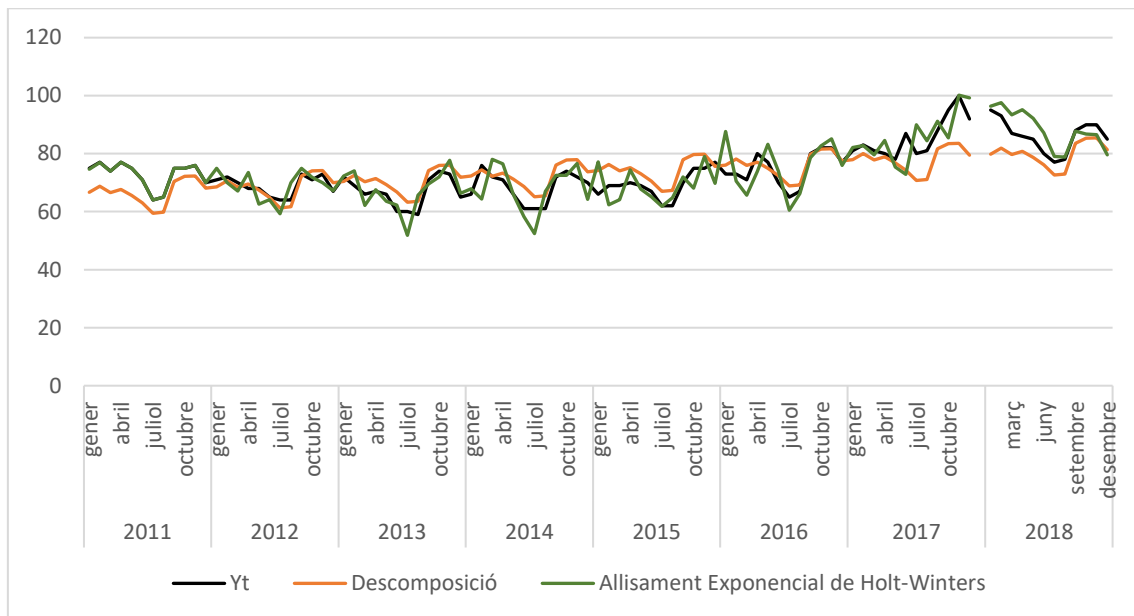


Figura 4.10. Gràfic de les prediccions de la sèrie lithium

A la Figura 4.10 s'aprecia en color negre els valors reals de la sèrie; en color taronja el mètode de Descomposició, i en verd el mètode d'Allisament Exponencial de Holt-Winters. A la Taula 4.2 es mostren els errors obtinguts.

| EPAM mostral | | EPAM extra-mostrat | |
|---------------|-------|--------------------|-------|
| Descomposició | AEHW | Descomposició | AEHW |
| 4,415 | 3,483 | 6,284 | 4,264 |

Taula 4.2. Errors dels mètodes de Descomposició i d'AEHW.

En ambdós casos dels mètodes de predicció, l'error de predicció per al període extra-mostrat és major que per al període mostral. Obtinc un error menor amb el mètode d'Allisament Exponencial de Holt-Winters que amb el mètode de Descomposició. Per a l'AEHW s'ha determinat els valors $\alpha = 0.9$, $\beta = 0.95$ i $\delta = 0.05$.

4.3 Estocàstic

En l'enfoc estocàstic de l'anàlisi de sèries temporals, es modelitza un conjunt de variables aleatòries. Un procés estocàstic està definit com a una successió infinita de variables aleatòries, cada una referenciada a un instant del temps i que estan ordenades cronològicament. Per tant, es defineix la sèrie temporal com una realització finita (de tamany u) d'un procés estocàstic.

Un procés estocàstic és estacionari en sentit estricte si al realitzar un desplaçament en el temps de totes les variables aleatòries de la sèrie, la distribució no canvia. Un procés estocàstic és estacionari en sentit dèbil si els moments de primer i segon ordre (la mitjana, la variància i la covariància) no depenen de t .

Per a l'anàlisi estocàstica s'utilitza la funció d'autocovariància i la funció d'autocorrelació parcial per a poder identificar quin procés estocàstic és el que millor defineix el comportament de la sèrie. Aquestes funcions es defineixen de la següent manera:

Definició. La funció d'autocovariància:

$$\gamma_{t,s} = E[(y_t - \mu_t)(y_s - \mu_s)], \forall t,s \in \mathbb{Z}$$

Definició. La funció d'autocorrelació parcial:

$$\rho_s = \frac{\gamma_s}{\gamma_0}, \quad s = 1, \dots, n \text{ i sent } \gamma_0 \text{ és la variància.}$$

Definició. Un soroll blanc $\{\varepsilon_t, t \in \mathbb{Z}\}$ és un procés estocàstic que compleix les següents característiques:

$$\begin{aligned} E[\varepsilon_t] &= 0 & \forall t \in \mathbb{Z} \\ \text{Var}[\varepsilon_t] &= \sigma_\varepsilon^2 & \forall t \in \mathbb{Z} \\ \text{Covar}[\varepsilon_t] &= 0 & \forall t \neq s \in \mathbb{Z} \end{aligned}$$

Es denota com a $\varepsilon \sim SB(0, \sigma_\varepsilon^2)$.

Definició. Un procés estocàstic és lineal quan es pot reescriure com al resultat d'una transformació lineal d'un procés de soroll blanc, de la següent manera:

$$y_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j}, \quad \text{amb } \varepsilon \sim SB(0, \sigma_\varepsilon^2)$$

On la sèrie ψ_0, ψ_1, \dots , són coeficients fixos.

Definició. Models autoregressius - AR(p)

Un procés estocàstic és un procés autoregressiu d'ordre p si existeixen els coeficients $\delta, \phi_1, \dots, \phi_p$, tals que

$$y_t = \delta + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \dots + \phi_p y_{t-p} + \varepsilon_t, \{\varepsilon_t\} \sim SB(0, \sigma_\varepsilon^2)$$

Definició. Models mitjana mòbil – MA(q)

Un procés estocàstic és un procés de mitjana mòbil d'ordre q si existeixen els coeficients $\theta_1, \dots, \theta_q$, tals que

$$y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \dots + \theta_q \varepsilon_{t-q} = \sum_{j=0}^q \theta_j \varepsilon_{t-j}, \theta_0 = 1, \{\varepsilon_t\} \sim SB(0, \sigma_\varepsilon^2)$$

Definició. Models autoregressius de mitjana mòbil – ARMA(p,q)

Un procés estocàstic és anomenat autoregressiu de mitjana mòbil d'ordre p i q si per a cada t:

$$y_t + \phi_1 y_{t-1} + \dots + \phi_p y_{t-p} = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \dots + \theta_q \varepsilon_{t-q}, \{\varepsilon_t\} \sim SB(0, \sigma_\varepsilon^2)$$

Tal i com s'observa amb la fórmula, els models ARMA(p,q) són models que contenen una part autoregressiva i una de mitjana mòbil.

Definició. Models autoregressius integrats de mitjana mòbil – ARIMA(p,d,q)

Y_t correspon a un model ARIMA(p,d,q) si s'hi ha aplicat la transformació següent,

$$Y_t = (1-B)^d Y_t, d \geq 0$$

B és l'*operator retard*, i.e. $B Y_t = Y_{t-1}$

Definició. Models estacionals – SARIMA(p,d,q)(P,D,Q)_s

Un model SARIMA és un model ARIMA que conté una part regular i una part estacional. Si per exemple, el nostre període sencer és un any i tenim dades mensuals, la part regular correspon a com es comporta el procés dins de l'any. En canvi, la part estacional, amb $s = 12$, és com es comporta la sèrie d'any en any, a través de diferències estacionals

4.4 Predicció estocàstica

Per a fer una anàlisi estocàstica de les sèries que he estat utilitzant fins aquí, es fa servir l'R i les funcions del paquet *stats*, ja integrat dins l'R.

La finalitat és altre cop, poder predir els valors futurs a través dels de la mateixa sèrie. Quan he anat fent aquesta anàlisi he vist que just després d'un període explosiu, com per exemple a la sèrie *spread*, és molt complicat trobar un model que defineixi tot el conjunt de valors del procés estocàstic i que serveixi per fer unes bones prediccions. No obstant, amb una sèrie com *pants*, on sembla que aquesta segueixi un patró més repetitiu (com hem vist a l'anàlisi determinista, és una sèrie tipus IV: tendència i estacionalitat) i sense tants canvis bruscos com una bombolla, l'anàlisi estocàstica sembla que funcioni millor per a fer prediccions.

Així doncs, a continuació presento com he fet l'anàlisi estocàstica d'aquestes dues sèries esmentades.

4.4.1 Predicció estocàstica de la sèrie *spread*

He començat per la sèrie *spread*. La seva gràfica es pot veure a la Figura 4.2.

En primer lloc fem el test de Normalitat de Jarque-Bera. En aquest test es plantegen les següents hipòtesis,

$$\begin{cases} H_0: \text{les dades segueixen una distribució Normal} \\ H_1: \text{les dades no segueixen una distribució Normal} \end{cases}$$

L'apliquem i obtenim el següent resultat, a la Figura 4.11.

```
Title:
Jarque - Bera Normality Test

Test Results:
STATISTIC:
X-squared: 167.7957
P VALUE:
Asymptotic p value: < 2.2e-16
```

Figura 4.11. Test de Jarque-Bera per a la sèrie *spread*.

Com es pot comprovar, es pot rebutjar la hipòtesi nul·la que ens diu que les dades segueixen una distribució Normal. Tot i així per l'anàlisi que hem de fer no és necessari que ho siguin, però simplement és interessant saber si ho són o no. També per a l'hora de veure com es comporten els residus del model ajustat.

A continuació a la Figura 4.12 es mostren els gràfics de les funcions d'autocovariància i la d'autocorrelació parcial.

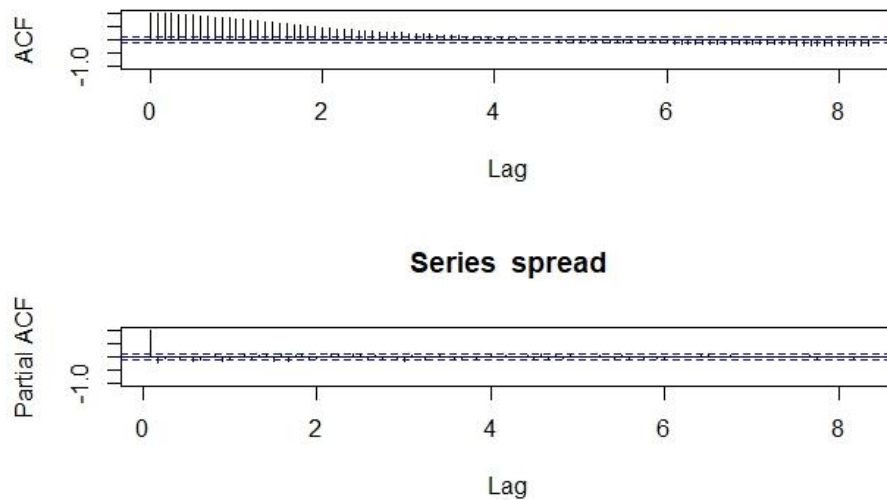


Figura 4.12. Gràfics de la funció d'autocovariància i d'autocorrelació parcial de la sèrie spread.

L'objectiu és que aquests dos gràfics tinguin el comportament com si fossin un soroll blanc. Interessa veure la funció ACF només amb el primer valor significativament diferent de zero (és a dir, fora de l'interval) i la funció Partial ACF amb els valors tots zero.

Després, es tracta d'anar-hi fent transformacions fins a aconseguir-ho. En aquesta sèrie l'he transformant primer utilitzant la funció logaritme i després he aplicat diferències regulars. Finalment es veu com a la Figura 4.13 la sèrie té la següent forma i es veu que presenta un caràcter més estacionari que abans.

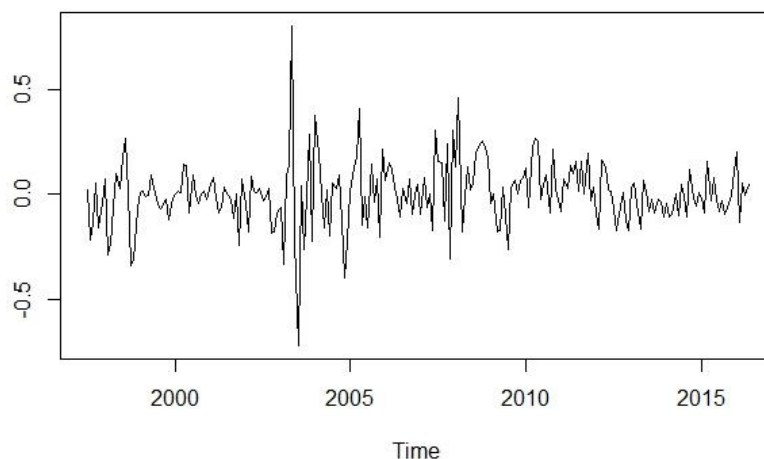


Figura 4.13. Gràfics de la funció d'autocovariància i d'autocorrelació parcial de la sèrie spread.

Després plantejo varis model SARIMA, anar canviant els paràmetres i comprovant la significació d'aquests fins que he acabat escollint un model SARIMA(0,1,1)(0,0,0)₁₂. La significació del paràmetre del model està a la Taula 4.3.

| Paràmetre | Significació |
|-------------------|--------------|
| MA (part regular) | 0.00106 |

Taula 4.3. Significació dels paràmetres del model proposat per a la sèrie spread.

Per fer una anàlisi dels residus, utilitzo la funció *tsdiag* que mostra el següent gràfic, a la Figura 4.14. Es pot veure com entre el 2010 i el 2015 hi ha un augment dels valors dels residus estandarditzats. Això correspon amb el període on, mirant la Figura 4.2 i la Taula 3.6 es detecta una bombolla. Per la resta, el gràfic de la funció ACF és correcte ja que s'espera que els residus es comportin com un soroll blanc, i l'ACF d'aquests

Per últim, hi ha els p-valors del Test de Ljung-Box. Aquest test serveix per mirar si els residus són independents els uns dels altres. Té el següent planteig d'hipòtesis,

$$\begin{cases} H_0: \text{les dades estan independentment distribuïdes.} \\ H_1: \text{les dades no estan independentment distribuïdes.} \end{cases}$$

L'ideal és que tots els valors estiguin per sobre la línia blava, és a dir, està fora la regió de rebuig. D'aquesta manera, es pot dir que els residus són independents. Tot i que l'últim valor sembla estar una mica al límit, està per sobre, i és el model amb el qual he obtingut millors p-valors.

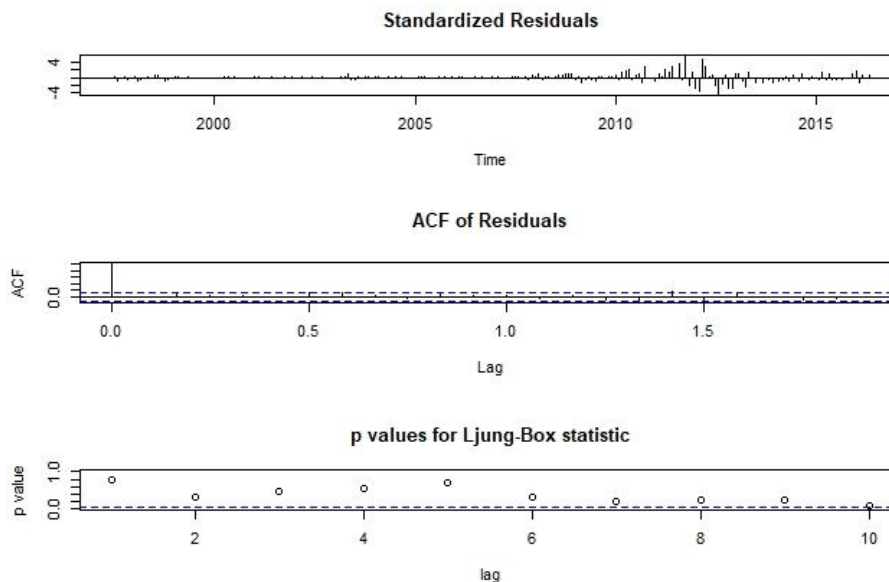


Figura 4.14. Anàlisi de residus del model proposat per a la sèrie spread.

Per últim, faig les prediccions amb el model proposat anteriorment i en calculo els errors. En primer lloc calculo les prediccions *ex-post*, les quals serveixen per calcular els errors de predicció del model. Aquestes prediccions són de les últimes dades que ja tinc, per a poder comparar. Són com unes prediccions de validació creuada, ja que el valor

real és conegut. Després faig les prediccions *ex-ante*, les quals són “reals” ja que no en conec el valor real. A la Figura 4.15 es mostren dos gràfics amb les prediccions *ex-post* i *ex-ante*.

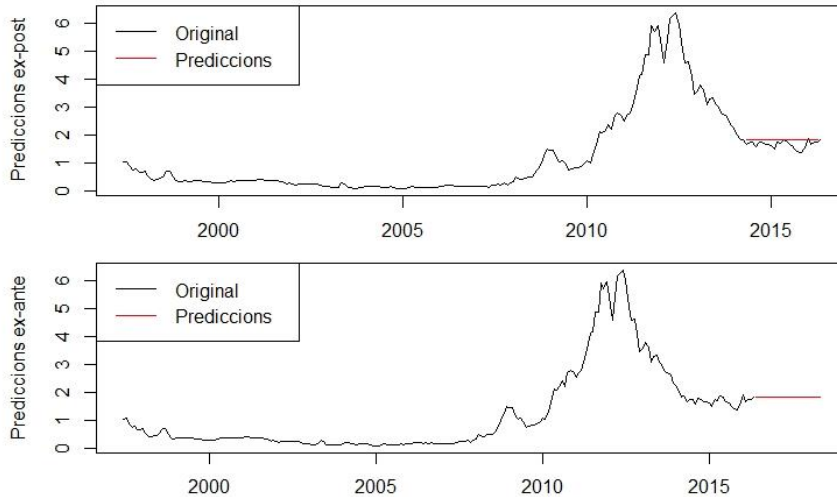


Figura 4.15. Prediccions *ex-post* de la sèrie *spread* (dalt). Prediccions *ex-ante* de la sèrie *spread* (baix).

Com es pot veure, en ambdós casos les prediccions semblen bastant dolentes. El model proposat, amb una diferència regular ($d = 1$) i un model MA(1) a la part regular. Tot i així, a la Taula 4.4 hi ha el càlcul dels errors de les prediccions *ex-post*.

| Error | Valor |
|--|---------------|
| Error quadràtic mitjà (EQM) | 0.053 |
| Error de l'arrel quadrada de la mitjana (REQM) | 0.223 |
| Error absolut mitjà (EAM) | 0.193 |
| Error percentual absolut mitjà (EPAM) | 0.124 (12.4%) |

Taula 4.4. Errors de les prediccions *ex-post* del model de la sèrie *spread*.

Si l'EPAM és superior al 5% ja es considera una predicció bastant dolenta, i això és el que també es pot concloure amb el gràfic.

La següent sèrie analitzada és la sèrie *pants*.

4.4.2 Predicció estocàstica de la sèrie *pants*

El motiu pel qual he analitzat aquesta sèrie és perquè com s'aprecia a la Figura X aquesta presenta un caràcter molt diferent de la sèrie *spread*. A l'apartat anterior, en l'anàlisi determinista l'he classificada com a sèrie de tipus IV, amb tendència i estacionalitat. Observant el gràfic ja es pot veure que sembla que tingui en general un caràcter bastant similar sempre, sense cap pic relativament diferent de la resta.

En primer lloc faig el test de Jarque-Bera, i com s'observa a la Figura 4.16, amb un p-valor de $0.254 > 0.05$, no puc rebutjar la hipòtesi nul·la i per tant, accepto la normalitat de la sèrie.

```

Title:
Jarque - Bera Normalality Test

Test Results:
STATISTIC:
X-squared: 2.7373
P VALUE:
Asymptotic p value: 0.2545
    
```

Figura 4.16. Test de Jarque-Bera per a la sèrie pants.

Altre cop, la finalitat és aconseguir que la sèrie sigui estacionària. Per a aconseguir-ho, en aquest cas he d'aplicar diferències regulars i diferències estacionals ($s = 12$). A través del gràfic de la funció ACF i PACF, provo varis models i miro la significació dels paràmetres fins que finalment escullo un model SARIMA(0,1,2)(0,1,1)₁₂.

A la Figura 4.17 mostro els diagnòstics d'aquest model i la significació dels paràmetres d'aquest. Com es pot comprovar, els residus estandarditzats semblen estar homogèniament distribuïts, tot i haver-hi alguna observació que el seu valor absolut és major que 3.

D'altra banda, tots els p-valors associats al test de Ljung-Box són majors que 0.05 i per tant, no hi ha evidència per dir que els residus estan relacionats.

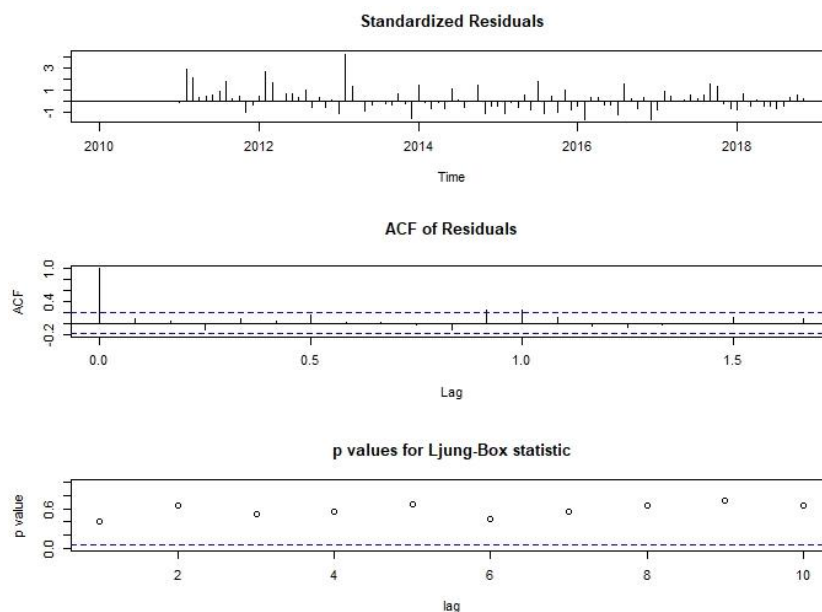


Figura 4.17. Anàlisi de residus del model proposat per a la sèrie pants.

Com es pot comprovar, a la Taula 4.5, els paràmetres són significatius, ja que tots els p-valors són menors que 0.05 i per tant, es pot rebutjar la hipòtesi nul·la que ens diu que no són rellevants per a l'explicació del comportament de la sèrie.

| Paràmetre | Significació |
|-------------------------|--------------|
| MA(1) (part estacional) | 0.0011 |
| MA(2) (part estacional) | 0.0242 |
| MA(1) (part estacional) | 4.85e-05 |

Taula 4.5. Significació dels paràmetres del model proposat per a la sèrie pants.

Per últim, una vegada ja tinc escollit el model em dispo a fer les prediccions. Igual que amb la sèrie spread, faig prediccions ex-post i ex-ante. Amb les prediccions “falses”, ex-post, en calculo els errors, mostrats a la Taula 4.6.

| Error | Valor |
|--|----------------|
| Error quadràtic mitjà (EQM) | 9.100 |
| Error de l'arrel quadrada de la mitjana (REQM) | 3.017 |
| Error absolut mitjà (EAM) | 2.273 |
| Error percentual absolut mitjà (EPAM) | 0.0273 (2.73%) |

Taula 4.6. Errors de les prediccions ex-post del model de la sèrie pants.

A la Figura 4.18 es mostren els gràfics de la sèrie i les prediccions, en color vermell. Aquesta vegada, per a les prediccions *ex-ante* també hi he afegit els intervals de confiança.

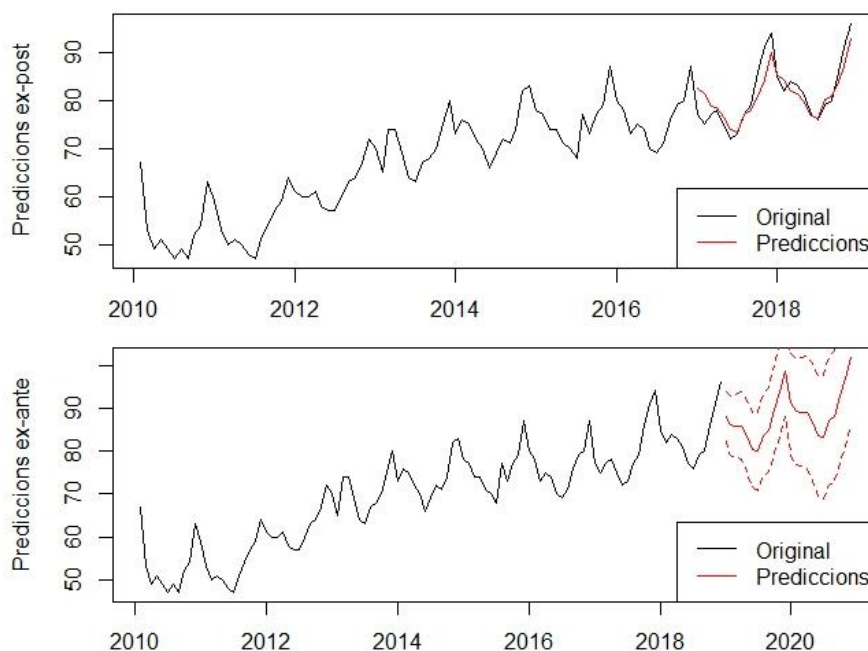


Figura 4.18. Prediccions ex-post de la sèrie pants (dalt). Prediccions ex-ante de la sèrie pants (baix).

Com es pot comprovar, i basant-nos en que l'error percentual absolut mitjà és major del 5%, les prediccions (tant les *ex-post* com les *ex-ante*) són força més acurades que les de

la sèrie anterior *spread*. Això és probablement degut a que el comportament d'aquesta sèrie és força més estable i similar al llarg del temps.

Respecte les prediccions *ex-ante* també podem dir que és un bon resultat, ja que els intervals de confiança segueixen força bé la forma del comportaments de les prediccions, però tampoc són molt amples.

4.5 Conclusió de l'anàlisi clàssica

Pel que fa a l'anàlisi determinista, vull destacar que la detecció de tendència a la sèrie es veu fàcilment afectada per al mínim detall o comportament explosiu, encara que sigui breu. És a dir, encara que la sèrie pugui semblar que tingui un comportament més o menys similar i aleatori sense cap pic o moment a destacar, a la que apareix almenys un període de més variància o de comportament explosiu, el Contrast de Daniel ja detecta tendència. D'altra banda, amb el Test de Kruskal-Wallis sí que sembla que detecti amb més eficiència i robusitat l'estacionalitat de la sèrie, tot i haver-hi algun període diferent de la resta.

Respecte l'anàlisi estocàstica, pot semblar fàcil pensar que la determinació d'un model per a l'ajust dels valors de les taules vagi millor que no pas basant-se en només dos criteris: l'estacionalitat i la tendència.

El moment de determinar el model és força complicat ja que hi ha força criteris en els quals basar-se a l'hora d'escollir els paràmetres; l'AIC, la significació d'aquests, els p-valors del test de Ljung-Box, la forma de la funció d'autocovariància, etc. Tot i així, en els dos exemples (*spread* i *pants*) que explico anteriorment he intentat aconseguir sobretot que tots els paràmetres del model fossin significatius a un nivell de significació del 5%, com sempre, i en cas de dubte també utilitzava els valors AIC per comparar i em quedava amb el mínim.

Amb la sèrie *spread*, la qual té un clar caràcter explosiu en un moment donat, i que tal i com he analitzat a l'apartat de **3.6** és un clar caràcter de bombolla/tendència viral, és més complicat ajustar un model amb unes bones prediccions. Potser seria més interessant analitzar just quan la sèrie comença a mostrar uns patrons on sembla que comença a fer-se viral.

Sembla ser que tant sigui amb un enfoc determinista o un enfoc estocàstic, és bastant complicat quan una sèrie està a punt d'entrar en un període de crisi/de canvi.

Això denota que aquestes bombolles, les quals he estudiat i analitzat la seva presència o no són més fàcils de detectar i en conseqüència estudiar una vegada ja han passat.

Tampoc seria mala idea estudiar com es comencena comportar els valors de la sèrie en un principi de període d'aquests d'exhuberància.

V. SHINY

5.1 Metodologia

Com ja he explicat a la introducció, un objectiu és aconseguir crear una aplicació Shiny d'R que serveixi per a mostrar el que he fet amb el test. En primer lloc, he creat una Shiny anomenada "Visualització" on carrego totes les dades i els resultats de les set sèries analitzades.

En segon lloc he creat una Shiny "interactiva" on la finalitat és poder aplicar el test PSY (Phillips, Shi i Yu, 2015) a una sèrie de dades de Google Trends, però poden escollir la paraula al moment.

5.2 Visualització

Com acabo de dir, en aquesta Shiny presento les sèries que he analitzat i tractat individualment. Els resultats i gràfics que es mostren són iguals que les de l'apartat III. Al següent enllaç hi ha pujada la Shiny que es pot visitar a través del navegador.

<https://sabinacriballes.shinyapps.io/aplicacioTFG/>

En primer lloc, a la pestanya de "Presentació" hi ha el botó a l'esquerra que permet canviar la sèrie que es té pensat visualitzar. A la dreta es mostra el gràfic de la sèrie; l'evolució del nombre de cerques de la paraula en qüestió. A la Figura 5.1 es veu quin format té aquesta primera pestanya.

Botó per escollir la paraula de la qual se'n vol mostrar l'anàlisi

Gràfic de la sèrie escollida



Figura 5.1. Impressió de pantalla de la Shiny de visualització, apartat de Presentació.

A continuació, a la Figura 5.2 es mostra la segona pestanya. Aquí es visualitza el gràfic de l'evolució dels estadístics al llarg de la sèrie. A la dreta es mostra quina sèrie és l'analitzada.

Com ja he comentat a l'apartat III , en aquests gràfics els primers valors no existeixen ja que no es poden calcular. Això és degut a que l'algorisme necessita uns quants valors previs al valor del qual se'n vol calcular l'estadístic del test.

Aquesta visualització facilita poder veure el valor de l'estadístic i com es comporta la sèrie en aquest moment, ja que els gràfics estan a la mateixa posició en diferents pestanyes.

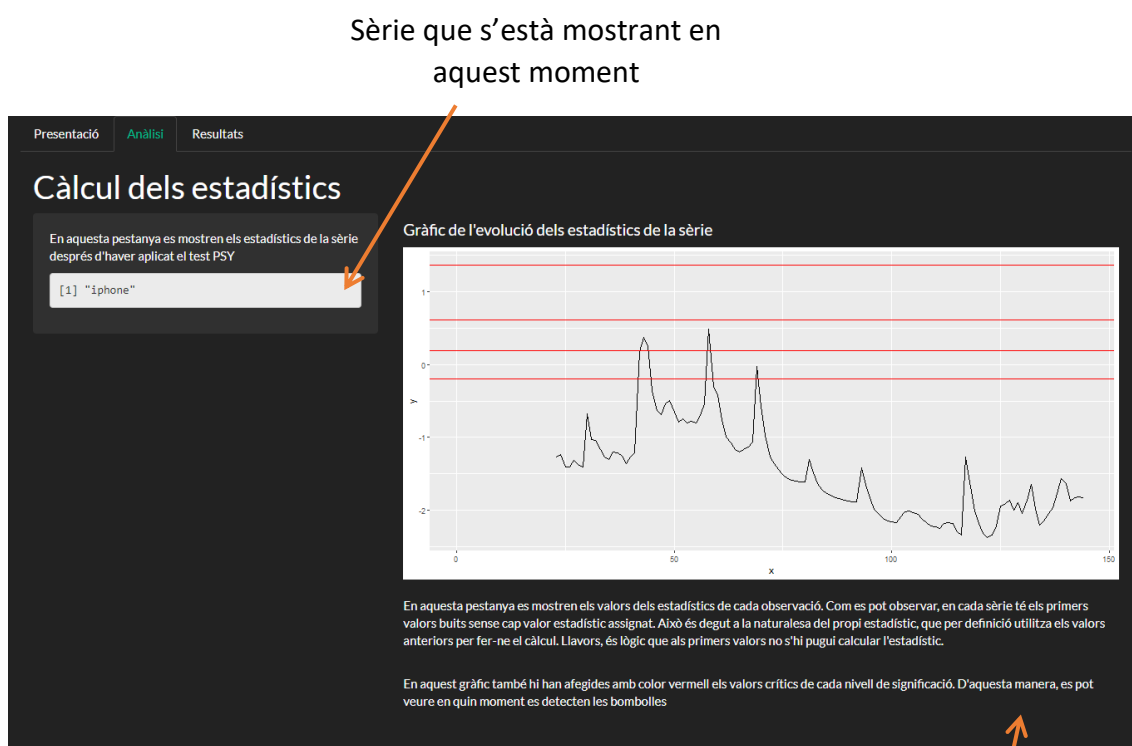


Figura 5.1. Impressió de pantalla de la Shiny de visualització, apartat d'Anàlisi.

Per últim, a la Figura 5.3, es mostra la tercera i última pestanya d'aquesta Shiny. En aquesta es mostra altre cop quina és la sèrie que s'està mostrant i a sota, la taula amb els períodes de crisi trobats. Igual que abans, es mostren els períodes ordenats de menys a més tolerància i es mostra: o bé un missatge que diu que no s'han trobat bombolles amb aquest nivell de significació, o el període amb data d'inici (start) i de finalització (end) de la bombolla que s'ha trobat.

A més, sota la taula hi ha una breu explicació del perquè de les bombolles trobades a la sèrie corresponent.

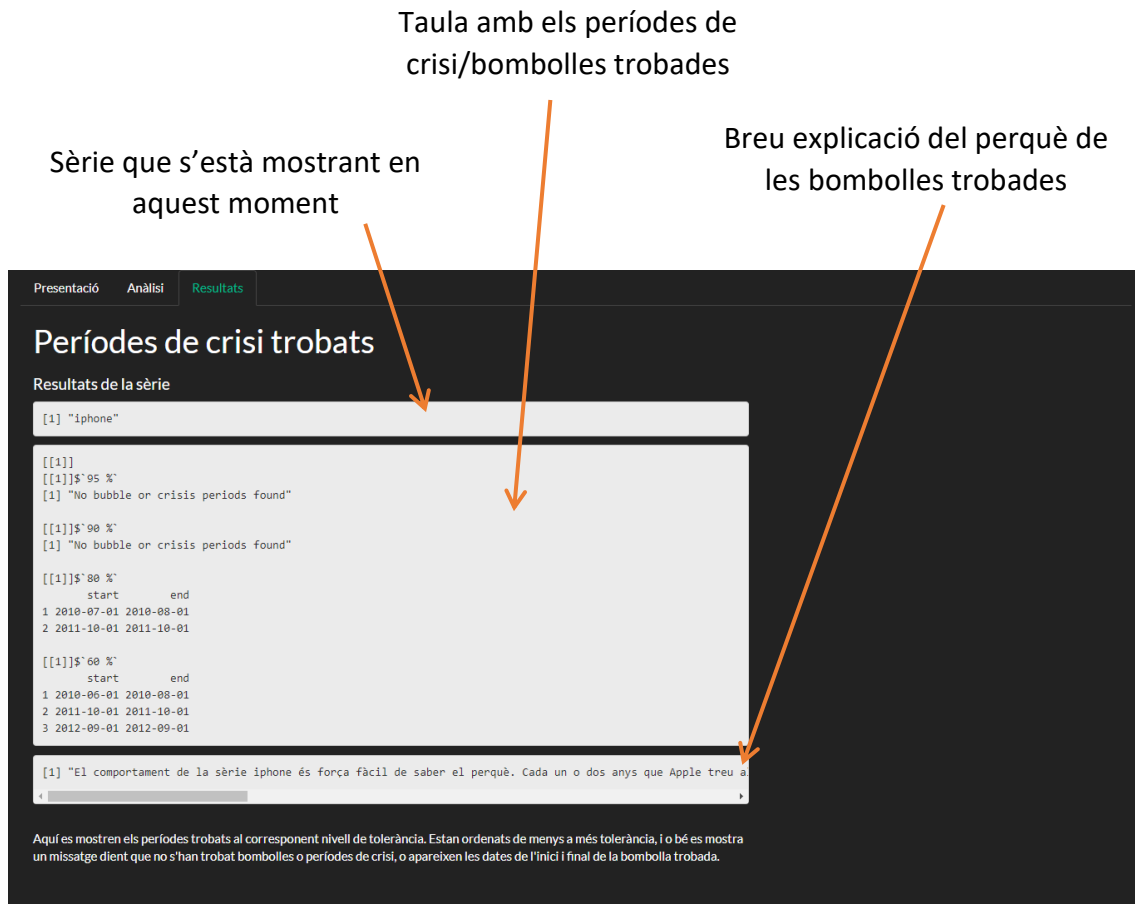


Figura 5.3. Impressió de pantalla de la Shiny de visualització, apartat de Resultats.

5.3 Interactiva

Com ja he dit anteriorment, la segona Shiny consisteix en una aplicació on es pot interaccionar i modificar la paraula de la qual se'n vol fer l'anàlisi al moment. Al contrari de la sèrie de visualització, en aquesta no hi ha dades carregades i es necessita connexió a Internet per a poder descarregar del Google Trends la sèrie de la paraula escollida. Al següent enllaç hi ha pujada la Shiny que es pot visitar a través del navegador.

<https://sabinacriballes.shinyapps.io/interactiuTFG/>

Per a facilitar la cerca i la rapidesa del test, tots els paràmetres ja els he escollit amb antelació, i igual que els nivells de significació. *IC* pren el valor 0, *adflag* és 2, i *yr* és 1. Com ja he comentat abans a l'apartat II, *yr* influeix molt en la detecció de bombolles, ja que serveix per a calcular l'amplada de la finestra gran de l'algorisme i com a més petit, més tolerància té. He escollit 1 perquè penso que és preferible a que el test detecti tots els caràcters explosius com una bombolla i que després l'investigador pugui decidir si tenir en compte tots els períodes trobats o no. A més a més, els quantils que es calculen

són per defecte el 60, 80, 90 i 95%. Com he vist a l'apartat III, on analitzo les set sèries, algunes sèries necessiten un nivell de tolerància més elevat per a poder detectar bombolles.

Consisteix en dues pestanyes, que es poden veure una impressió de pantalla de cada una a la Figura 5.4 i a la Figura 5.5.

En la primera pestanya, "Inici", hi ha la finestra per a poder escriure-hi la paraula a analitzar. A la dreta, es mostra el gràfic de la sèrie.

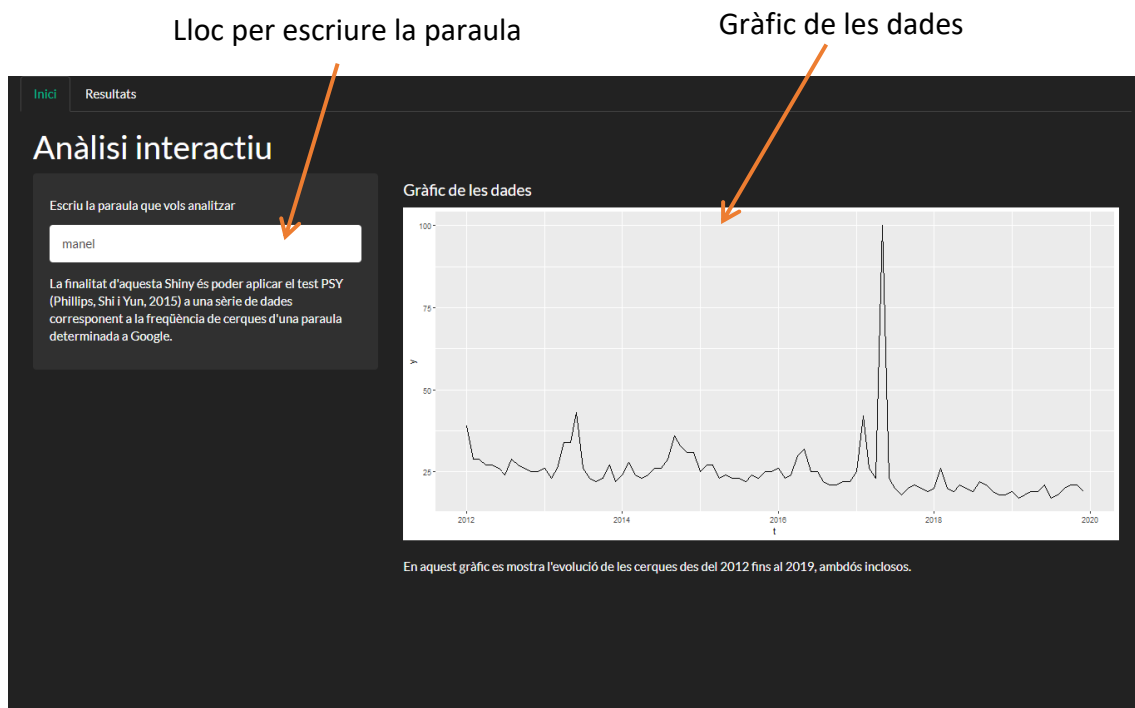


Figura 5.4. Impressió de pantalla de la Shiny interactiva, la pestanya inicial.

En la segona pestanya és on es mostren les bombolles, si és que n'hi ha, detectades en el test. També es mostra el gràfic de l'evolució dels estadístics i afegit amb unes línies horitzontals vermelles, els valors crítics corresponents al 60, 80, 90 i 95%.

Primer, dins cada apartat del nivell de significació corresponent, es mostra un missatge amb "Bubble found" o "No bubble or crisis periods found", segons s'hagi detectat bombolla o no amb aquesta tolerància, respectivament. Si s'ha trobat, es mostra una taula amb *start* i *end* amb les dates d'inici i finalització de la/les bombolla/es. Si no s'ha trobat, apareix un error que simplement s'ha d'ignorar.

Missatge amb si s'ha trobat o no bombolla

Períodes de bombolla trobats al nivell de tolerància corresponent

Gràfic amb l'evolució dels estadístics

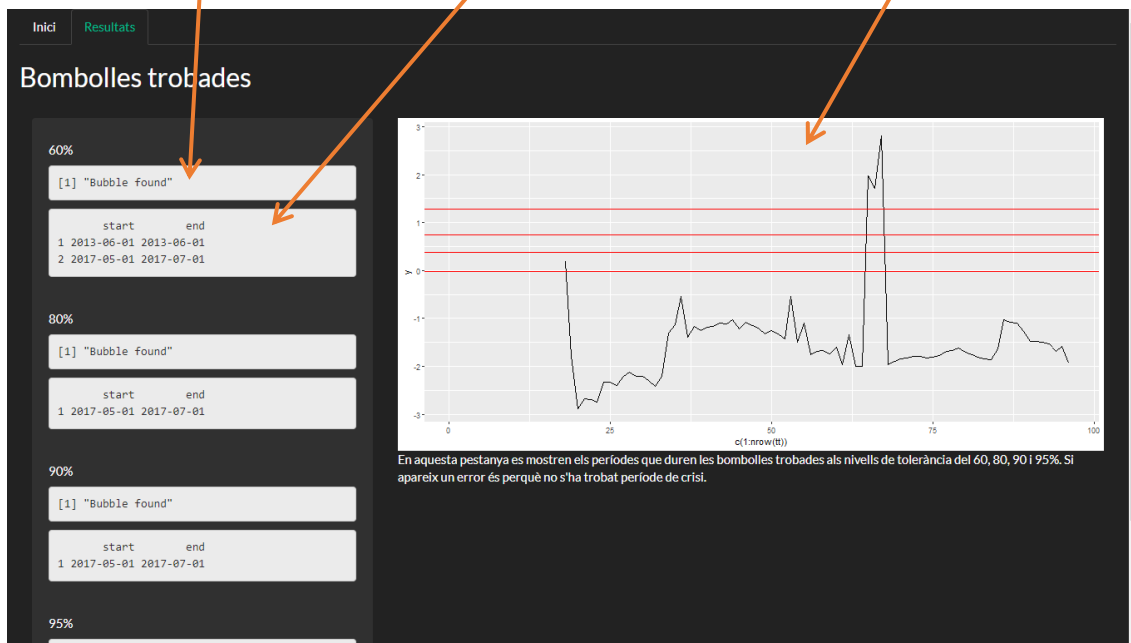


Figura 5.5. Impressió de pantalla de la Shiny interactiva, la pestanya de resultats.

VI. CONCLUSIONS

Per a la finalització d'aquest treball, faré un petit repàs a els resultats obtinguts amb el test, les dues tipus d'anàlisi realitzades, tenint en compte el objectius plantejats a l'inici.

A través de les set sèries escollides he pogut veure quan el test PSY és l'encarregat de detectar i determinar què és i què no és una bombolla. He vist que modificant els paràmetres de les funcions, els quals estan explicats a la metodologia, aquesta detecció pot variar en més o menys mesura. És a dir, que canviant els valors *adflag* o *Tb*, en una sèrie es pot passar de detectar-ne més d'una a no detectar cap bombolla. Podria veure's que el test és una mica sensible als canvis en el valor dels paràmetres.

En primer lloc, pel que fa al comportament de les sèries, hi ha una gran diversitat en el seu comportament. Per exemple, a la sèrie *spread* es pot veure que hi ha sobretot una bombolla molt senyalada i que aquestes dades s'utilitzen com a exemple en el paquet *psymonitor* mateix. Llavors és fàcil veure que el test detecta sempre aquesta període de bombolla. D'altra banda, també hi ha sèries que tenen punts o moments que no dona a pensar que s'hagi produït cap bombolla, i que consegüentment no hagi sigut trending topic en aquell moment. Però tot i així el test en detecta. Quan això succeeix és quan és important fer un anàlisi amb profunditat del perquè de cada moment i què va passar en aquelles dates per a que el test pugui detectar-hi una bombolla. És imprescindible no separar les bombolles detectades de la naturalesa de les dades; complementar els resultats obtinguts del test amb la informació respectiva de la sèrie.

En segon lloc, és important comentar que al principi, a l'utilitzar les funcions del paquet *psymonitor*, moltes vegades el test PSY no trobava bombolles. Al canviar i poder rebaixar el nivell de tolerància, aquest problema l'he pogut resoldre, però amb aquest canvi en el nivell de tolerància a vegades succeeix que la bombolla varia de llargada, variant les dates d'inici i/o finalització. Això produeix que la llargada de les bombolles no sigui sempre del tot homogènia. El més pràctic potser és, un cop detectada la bombolla, declarar un nivell de tolerància màxim on es delimiti l'inici i el final de la bombolla.

Amb l'anàlisi clàssica he vist que la component de la tendència es veu altament afectada quan hi ha períodes bombolla (com a l'apartat **4.4.1**). Com ja he comentat a l'apartat **4.5**, el Contrast de Daniel és molt sensible a aquests canvis bruscos i per tant la tendència es veu afectada.

D'això se'n deriva una mala predicció ja que és complicat determinar un model que serveixi per a definir el comportament al llarg de tota la sèrie. Una proposta per a solucionar aquest problema en el model clàssic, seria afegir variables dummies. Aquestes amortiguarien els períodes de bombolla i es podria analitzar la sèrie amb més

facilitat. D'aquesta manera, els errors de predicció extra-mostrals millorarien molt. El més important és veure que les bombolles desestabilitzen les prediccions immediatament posteriors, on representa que la bombolla ha evolucionat a crisi, però les prediccions no.

L'anàlisi estocàstica també es veu afectada per als períodes aquests. Tot i que podria semblar que com que al determinar el model pots variar si té més o menys memòria, hauria d'anar millor, no és així. Una possible solució podria ser dividir la sèrie i ajustar uns models diferents quan es produeixin aquests períodes.

En tot cas, el més important és veure que és important detectar les bombolles, determinar el seu inici i la finalització i després tractar-les separatament del model.

Hi ha algunes sèries, com per exemple la de la paraula *iphone* (apartat **4.1.4**) on es comencen a produir uns pics amb certa estacionalitat, els quals al principi són detectats com a bombolla i els seus successors el test ja no els detecta. Això també és curiós ja que es pot veure com un terme pot ser "normal" que esdevingui *trending topic* cada cert temps. Tot i així, el test detecta bé les primeres bombolles, per tant, podria dir-se que el test funciona i és aplicable quan es presenten múltiples bombolles.

Les aplicacions Shiny realitzades són fàcils de manejar, ja que no tenen masses opcions. La de visualització és simplement escollir la paraula en concret a mostrar i ja està. L'aplicació interactiva pot presentar més problemes ja que el paquet Google Trends es renova molt sovint i si ha algun canvi molt rellevant en la forma d'emmagatzemar les dades pot ser que es produeixi algun problema d'incompatibilitat i no funcioni del tot bé. Tot i així, puc dir que l'objectiu de poder analitzar al moment qualsevol paraula i saber si presenta o no bombolla, ha estat assolit.

VII. ÍNDEX DE FIGURES

| | | |
|------|---|----|
| 1.1 | Cerques del terme “coronavirus” segons Google a Xina (superior) i a Espanya (inferior). | 7 |
| 2.1 | Part superior del data frame d'exemple. | 12 |
| 2.2 | Gràfic de la sèrie temporal d'exemple. | 13 |
| 2.3 | Esquema de com funcionen les submostres del test PSY (Phillips, Wu i Yu (2015)). | 16 |
| 2.4 | Evolució dels estadístics d'exemple amb els punts crítics al 95, 90, 80 i 60% .. | 20 |
| 2.5 | Gràfic de la sèrie manel amb períodes comentats. | 22 |
| 3.1 | Gràfic de la sèrie iphone. | 24 |
| 3.2 | Gràfic de la sèrie pants. | 25 |
| 3.3 | Gràfic de la sèrie lithium. | 26 |
| 3.4 | Gràfic de la sèrie obama. | 27 |
| 3.5 | Gràfic de la sèrie bitcoin. | 28 |
| 3.6 | Gràfic de la sèrie spread. | 30 |
| 4.1 | Gràfic de la sèrie manel amb Excel amb línia de la tendència. | 33 |
| 4.2 | Gràfic de la sèrie spread amb Excel amb línia de la tendència. | 35 |
| 4.3 | Gràfic de la sèrie pants amb Excel amb línia de la tendència. | 36 |
| 4.4 | Gràfic de la sèrie obama amb Excel amb línia de la tendència. | 37 |
| 4.5 | Gràfic de la sèrie iphone amb Excel amb la línia de la tendència. | 38 |
| 4.6 | Gràfic de la sèrie lithium amb Excel amb la línia de la tendència. | 39 |
| 4.7 | Gràfic de la sèrie lithium canviat d'escala amb Excel amb la línia de la tendència. | 40 |
| 4.8 | Gràfic de la sèrie bitcoin amb Excel amb la línia de la tendència. | 41 |
| 4.9 | Prediccions per a la sèrie manel. | 44 |
| 4.10 | Gràfic de les prediccions de la sèrie lithium. | 46 |

| | | |
|------|--|----|
| 4.11 | Test de Jarque-Bera per a la sèrie spread. | 49 |
| 4.12 | Gràfic de la funció d'autocovariància i d'autocorrelació parcial de la sèrie spread. | 50 |
| 4.13 | Gràfic de la funció d'autocovariància i d'autocorrelació parcial de la sèrie spread. | 50 |
| 4.14 | Anàlisi de residus del model proposat per a la sèrie spread. | 51 |
| 4.15 | Prediccions ex-post de la sèrie spread (dalt). Prediccions ex-ante de la sèrie spread (baix). | 52 |
| 4.16 | Test de Jarque-Bera per a la sèrie pants. | 53 |
| 4.17 | Anàlisi de residus del model proposat per a la sèrie pants. | 53 |
| 4.18 | Prediccions ex-post de la sèrie pants (dalt). Prediccions ex-ante de la sèrie pants (baix). | 54 |
| 5.1 | Impressió de pantalla de la Shiny de visualització, apartat de Presentació. | 57 |
| 5.2 | Impressió de pantalla de la Shiny de visualització, apartat d'Anàlisi. | 58 |
| 5.3 | Impressió de pantalla de la Shiny de visualització, apartat de Resultats. | 59 |
| 5.4 | Impressió de pantalla de la Shiny interactiva, la pestanya inicial. | 60 |
| 5.5 | Impressió de pantalla de la Shiny interactiva, la pestanya de resultats. | 61 |

VIII. ÍNDEX DE TAULES

| | | |
|-----|---|----|
| 2.1 | Exemple del funcionament de l'algorisme. | 16 |
| 2.2 | Evolució dels estadístics de la sèrie d'exemple (manel) amb els punts crítics al 95, 90, 80 i 60%. | 21 |
| 3.1 | Períodes detectats pels estadístics de la sèrie iphone amb els punts crítics al 95, 90, 80 i 60%. | 24 |
| 3.2 | Períodes detectats pels estadístics de la sèrie pants amb els punts crítics al 60, 80, 90 i 95%. | 25 |
| 3.3 | Períodes detectats pels estadístics de la sèrie lithium amb els punts crítics al 90, 80 i 60%. | 26 |
| 3.4 | Períodes detectats pels estadístics de la sèrie obama amb els punts crítics al 90, 80 i 60%. | 28 |
| 3.5 | Períodes detectats pels estadístics de la sèrie bitcoin amb els punts crítics al 95, 90, 80 i 60%. | 29 |
| 3.6 | Períodes detectats pels estadístics de la sèrie spread amb els punts crítics al 99, 95 i 90%. | 30 |
| 4.1 | Error dels mètodes de Tendència Lineal, Dobles Mitjanes Mòbils i d'AEH. | 44 |
| 4.2 | Error dels mètodes de Descomposició i d'AEHW. | 46 |
| 4.3 | Significació dels paràmetres del model proposat per a la sèrie spread. | 51 |
| 4.4 | Error de les prediccions ex-post del model de la sèrie spread. | 52 |
| 4.5 | Significació dels paràmetres del model proposat per a la sèrie pants. | 54 |
| 4.6 | Error de les prediccions ex-post del model de la sèrie pants. | 54 |

IX. REFERÈNCIES

8.1 Articles

- [1] Peter C. B. Phillips; Shu-Ping Shi i Jun Yu. (2015). *Testing for multiple bubbles: historical episodis of exuberance and collapse in the S&P 500*. Recuperat de <https://onlinelibrary.wiley.com/doi/abs/10.1111/iere.12132>
- [2] Peter C. B. Phillips; Shu-Ping Shi i Jun Yu. (2011). *Testing for multiple bubbles*. Recuperat de https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1981976
- [3] Thomas Quistgaard Pedersen i Erik Christian Montser Schütte. (2017). *Testing for Explosive Bubbles in Presence of Autocorrelated Innovations*. Recuperat de https://pure.au.dk/portal/files/109652663/rp17_09.pdf

8.2 Articles electrònics

- [4] Bego Romero. (2018). ¿Qué es y cómo usar Google Trends?. *Raiola Networks*. Recuperat de <https://raiolanetworks.es/blog/google-trends-tendencias-palabra-clave/>
- [5] Enderrock.cat. (2011). Manel, el tema del día a Twitter. *Enderrock*. Recuperat de <http://www.enderrock.cat/noticia/4713/manel/tema/dia/twitter>
- [6] Aragón, Laura. (2020). Coronavirus: radiografía de la alarma social según nuestras búsquedas en internet. *La Vanguardia*. Recuperat de https://www.lavanguardia.com/vida/20200313/474097773348/coronavirus-radiografia-alarma-social-busquedas-internet.html?utm_term=botones_sociales_app&utm_source=social-otros&utm_medium=social
- [7] Jaume V. Aroca. (2020). Sí, la pandemia remite (Eso indica Google). *La Vanguardia*. Recuperat de https://www.lavanguardia.com/vida/20200409/48393852483/pandemia-coronavirus-remite-google-fact-check.html?utm_term=botones_sociales_app&utm_source=social-otros&utm_medium=social
- [8] Philippe Massicotte i Dirk Eddelbuettel. (May 2, 2019). Package 'gtrendsR'. *CRAN*. Recuperat de <https://cran.r-project.org/web/packages/gtrendsR/gtrendsR.pdf>

- [9] Peter C.B. Phillips; Shuping Shi i Itamar Caspi. (March 20, 2019). Package 'psymonitor'. CRAN. Recuperat de <https://cran.r-project.org/web/packages/psymonitor/psymonitor.pdf>

8.3 Apunts

- [10] Oleg Seleznev. *Time series analysis*. Department of Mathematics and Mathematical Statistics, Umeå University.
- [11] Helena Chuliá. *Análisis de series temporales. 1era Parte: ANÁLISIS DETERMINISTA*. School of Economics, Universitat de Barcelona.
- [12] Ernest Pons i Fanals. *Análisis de series temporales. (Análisis estocástico de series temporales; Modelos lineales de series temporales i Metodología Box-Jenkins)*. Departament d'Econometria, Estadística i Economia Espanyola, Universitat de Barcelona.

8.4 Llibres

- [13] Brockwell, P.J. i Davis, R.A. (2002). *Introduction to Time Series and Forecasting* (2na edició). Springer.

8.5 Pàgines web

- [14] Per a l'obtenció de les dades per a les anàlisis s'ha utilitzat el següent enllaç, <https://trends.google.com/trends/?geo=US>

IX. ANNEX

9.1 Script Shiny “Visualització”

```
library(shiny)
library(ggplot2)
library(shinyWidgets)
library(shinythemes)
# Càrrega de les dades
setwd("C:/Users/sabin/Desktop/TFG/Data")
load("./TOT.RData")

data_sets = list(dfm_bitcoin=dfm_bitcoin, dfm_iphone=dfm_iphone, dfm_lithium=
dfm_lithium, dfm_manel=dfm_manel, dfm_obama=dfm_obama, dfm_pants=dfm_pants, dfm_spread=dfm_s
pread)
data_bsadf = list(bsadf_bitcoin=bsadf_bitcoin,
bsadf_iphone=bsadf_iphone, bsadf_lithium=
bsadf_lithium, bsadf_manel=bsadf_manel, bsadf_obama=bsadf_obama, bsadf_pants=bsadf_pants, bs
adf_spread=bsadf_spread)
data_quantiles95 = list(quantile95_bitcoin=quantile95_bitcoin,
quantile95_iphone=quantile95_iphone, quantile95_lithium=
quantile95_lithium, quantile95_manel=quantile95_manel, quantile95_obama=quantile95_obama, q
uantile95_pants=quantile95_pants, quantile95_spread=quantile95_spread)
data_quantilesBsadf2 = list(quantilesBsadf2_bitcoin=quantilesBsadf2_bitcoin,
quantilesBsadf2_iphone=quantilesBsadf2_iphone, quantilesBsadf2_lithium=
quantilesBsadf2_lithium, quantilesBsadf2_manel=quantilesBsadf2_manel, quantilesBsadf2_obam
a=quantilesBsadf2_obama, quantilesBsadf2_pants=quantilesBsadf2_pants, quantilesBsadf2_spre
ad=quantilesBsadf2_spread)
data_tablecrisis = list(tablecrisis_bitcoin=tablecrisis_bitcoin,
tablecrisis_iphone=tablecrisis_iphone, tablecrisis_lithium=
tablecrisis_lithium, tablecrisis_manel=tablecrisis_manel, tablecrisis_obama=tablecrisis_ob
ama, tablecrisis_pants=tablecrisis_pants, tablecrisis_spread=tablecrisis_spread)
data_bsadfNA =
list(bsadfNA_bitcoin=bsadfNA_bitcoin, bsadfNA_iphone=bsadfNA_iphone, bsadfNA_lithium=bsadf
NA_lithium, bsadfNA_manel=bsadfNA_manel, bsadfNA_obama=bsadfNA_obama, bsadfNA_pants=bsadfNA
_pants, bsadfNA_spread=bsadfNA_spread)

ui <- fluidPage(theme = shinytheme("darkly"),

tabsetPanel(

  tabPanel("Presentació", fluid=TRUE,

    headerPanel("Sèrie escollida"),

    sidebarPanel(
      selectInput(inputId = "Paraula", label = "Selecciona la paraula que vols
analitzar", multiple = FALSE,
        choices =
list("obama", "manel", "spread", "pants", "lithium", "bitcoin", "iphone"), selected =
"lithium")
    ),

    mainPanel(
      h4("Gràfic de les dades"),
      plotOutput("Grafic"), # Gràfic de les dades
      br(),
      p("En aquesta Shiny es presenten les set sèries analitzades a les quals s'hi
ha aplicat el test PSY(Phillips, Wu i Yu (2015)).
      En aquesta pestanya es mostra un gràfic de la sèrie que escollida en aquest
moment. És important fer una primera visualització
      del comportament de la sèrie, per veure i poder situar els períodes
d'exhuberència i poder identificar el perquè
      de les bombolles (si és que n'hi ha) trobades.")
    )
  ),
),
```

```

tabPanel("Anàlisi",fluid=TRUE,

  headerPanel("Càlcul dels estadístics"),

  sidebarPanel(
    p("En aquesta pestanya es mostren els estadístics de la sèrie després d'haver
aplicat el test PSY"),
    verbatimTextOutput("Nom1")
  ),
  mainPanel(
    h4("Gràfic de l'evolució dels estadístics de la sèrie"),
    plotOutput("GraficEst"),
    br(),
    p("En aquesta pestanya es mostren els valors dels estadístics de cada
observació. Com es pot observar, en cada sèrie té els primers valors
buits sense cap valor estadístic assignat. Això és degut a la naturalesa
del propi estadístic, que per definició utilitza els valors anteriors
per fer-ne el càlcul. Llavors, és lògic que als primers valors no s'hi
pugui calcular l'estadístic."),
    br(),
    p("En aquest gràfic també hi han afegides amb color vermell els valors crítics
de cada nivell de significació. D'aquesta manera,
es pot veure en quin moment es detecten les bombolles")
  )
),

tabPanel("Resultats",fluid=TRUE,

  headerPanel("Períodes de crisi trobats"),

  mainPanel(
    h4("Resultats de la sèrie"),
    verbatimTextOutput("Nom2"),
    verbatimTextOutput("Taules"), # Taula amb els resultats
    verbatimTextOutput("Desc"),
    br(),
    p("Aquí es mostren els períodes trobats al corresponent nivell de tolerància.
Estan ordenats de menys a més tolerància, i o bé es mostra un
missatge dient que no s'han trobat bombolles o períodes de crisi, o
apareixen les dates de l'inici i final de la bombolla trobada.")
  )
)
))

server <- function(input, output) {

  dfm <- reactive({
    data.frame(data_sets[[paste("dfm_",input$Paraula,sep="")]])
  })

  bsadf <- reactive({
    data.frame(data_bsadf[[paste("bsadf_",input$Paraula,sep="")]])
  })

  bsadfNA <- reactive({
    data.frame(data_bsadfNA[[paste("bsadfNA_",input$Paraula,sep="")]])
  })

  quantilesBsadf2 <- reactive({
    data.frame(data_quantilesBsadf2[[paste("quantilesBsadf2_",input$Paraula,sep="")]])
  })

  tablecrisis = reactive({
    list(data_tablecrisis[[paste("tablecrisis_",input$Paraula,sep="")]])
  })

  output$Desc <- renderPrint({
    if(input$Paraula == "bitcoin"){ nomr <- 1 }
    if(input$Paraula == "iphone") { nomr <- 2 }
  })
}

```

```

    if(input$Paraula == "lithium"){ nomr <- 3 }
    if(input$Paraula == "manel") { nomr <- 4 }
    if(input$Paraula == "obama") { nomr <- 5 }
    if(input$Paraula == "pants") { nomr <- 6 }
    if(input$Paraula == "spread") { nomr <- 7 }
    data_explicacio[nomr,2]
    as.character(data_explicacio[nomr,2])
  })

nom_serieanalitzada = reactive({
  as.character(input$Paraula)
})

output$Grafic <- renderPlot({
  ss=dfm()
  ggplot(data = ss, aes(x = t, y = y)) + geom_line(color = "black", size = 0.4))

output$GraficEst <- renderPlot({
  tt=bsadfNA()
  zz=quantilesBsadf2()
  ggplot(data = tt,aes(x=x,y=y)) + geom_line(color = "black", size = 0.4) +
  geom_hline(yintercept = as.numeric(zz[,1]), color = "red"))

output$Taules <- renderPrint({
  rr=tablecrisis()
  rr})

output$Nom1 <- renderPrint({
  nn=nom_serieanalitzada()
  nn
})
output$Nom2 <- renderPrint({
  n2=nom_serieanalitzada()
  n2
})

}

shinyApp(ui = ui, server = server)

```

9.2 Script Shiny “Interactiu”

```

# Càrrega de les funcions i paquets

library(shiny)
library(gtrendsR)
library(psymonitor)
library(ggplot2)
library(psymonitor)
library(shinyWidgets)
library(shinythemes)

setwd("C:/Users/sabin/Desktop/TFG/")
source("ADFres.R")
source("quantilsPSY.R")
source("PSY2.R")
source("locate2.R")

ui <- fluidPage(theme = shinytheme("darkly"),

tabsetPanel(

  tabPanel("Inici", fluid=TRUE,

    headerPanel("Anàlisi interactiu"),
    br(),
    sidebarPanel(

```



```

      textInput("Paraula", label = h5("Escriu la paraula que vols analitzar"),
value = "manel"),
      p("La finalitat d'aquesta Shiny és poder aplicar el test PSY (Phillips,
Shi i Yun, 2015) a una sèrie
de dades corresponent a la freqüència de cerques d'una paraula determinada
a Google.")
    ),
    mainPanel(
      h4("Gràfic de les dades"),
      plotOutput("Grafic"), # Gràfic de les dades
      br(),
      p("En aquest gràfic es mostra l'evolució de les cerques des del 2012 fins al
2019, ambdós inclosos.")
    )
  ),
  tabPanel("Resultats", fluid=TRUE,
    h2("Bombolles trobades"),
    br(),
    sidebarPanel(
      h5("60%"),
      verbatimTextOutput("resultat60"),
      verbatimTextOutput("resultat60_si"),
      br(),
      h5("80%"),
      verbatimTextOutput("resultat80"),
      verbatimTextOutput("resultat80_si"),
      br(),
      h5("90%"),
      verbatimTextOutput("resultat90"),
      verbatimTextOutput("resultat90_si"),
      br(),
      h5("95%"),
      verbatimTextOutput("resultat95"),
      verbatimTextOutput("resultat95_si"),
      br()
    ),
    mainPanel(
      plotOutput("GraficEst"),
      p("En aquesta pestanya es mostren els períodes que duren les bombolles
trobades als nivells de tolerància del
60, 80, 90 i 95%. Si apareix un error és perquè no s'ha trobat període de
crisi.")
    )
  )
)
)

server <- function(input, output) {

  paraula <- reactive({input$Paraula})

  resm <- reactive(gtrends(keyword = paraula(),
time = "2012-01-01 2019-12-31",
hl="es"))

  dfm <- reactive({
    resmact=resm()
    data.frame(t=resmact[["interest_over_time"]]$date,
y=(resmact[["interest_over_time"]]$hits))
  })

  output$Grafic <- renderPlot({
    ggplot(data = dfm(), aes(x = t, y = y)) + geom_line(color = "black", size = 0.4)
  })

# Paràmetres
y <- reactive({
  dfmact=dfm()
  dfmact$y
})
}

```

```

obs      <- reactive({length(y())})
swindow0 <- reactive({floor(obs() * (0.01 + 1.8 / sqrt(obs())))})
IC       <- 0
adflag   <- 2
yr       <- 1
Tb       <- reactive({12*yr + swindow0() - 1})
nboot    <- 99

# Càlcul dels estadístics
bsadf2   <- reactive({PSY2(y(), swindow0 = swindow0(), IC = IC, adflag = adflag)})
output$bsadf <- renderText({
  bb = bsadf2()
  bb
})

dif <- reactive({nrow(dfm()) - length(bsadf2())})
x1 <- reactive({data.frame(y=rep(NA,dif())) })
x2 <- reactive({data.frame(y=bsadf2())})
bsadf2NA <- reactive({rbind(x1(),x2())})

# Càlcul dels valors crítics
val.quantils <- c(0.6,0.8,0.9,0.95)

quantilesBsadf2 <- reactive({
  quantilsPSY(y(), swindow0 = swindow0(), IC = IC, adflag = adflag, Tb = Tb(), nboot
= 99, nCores = 2, val.quantils=val.quantils)
})

# Gràfic estadístics i valors crítics
output$GraficEst <- renderPlot({
  tt=bsadf2NA()
  zz=quantilesBsadf2()
  ggplot(data = tt,aes(x=c(1:nrow(tt)),y=y)) + geom_line(color = "black", size = 0.4)
+ geom_hline(yintercept = as.numeric(zz[,1]), color = "red")
})

# Dates
dim <- reactive({obs() - swindow0() + 1})

monitorDates <- reactive({
  dfmact=dfm()
  as.Date(dfmact[swindow0():obs(),1])
})

quantile95 <- reactive({
  qq = quantilesBsadf2()
  quantile95 = qq %%% matrix(1, nrow = 1, ncol = dim())
})

#----- 60
# 60 %
ind60 <- reactive({
  bb = bsadf2()
  qq60 = quantile95()
  ind60 <- bb > t(qq60[1, ]) * 1
})
periods60 <- reactive({
  ii = ind60()
  mm = monitorDates()
  periods60 <- locate2(ii, mm)
})

logical_foundperiods60 <- reactive({is.null(periods60())})
# output$prova <- renderPrint({
#   dd = logical_foundperiods60()
#   dd
# })
res60 <- reactive({
  if(as.logical(logical_foundperiods60())){           res60 <- "No bubble or crisis
periods found"
} else{                                               res60 <- "Bubble found"   }
}

```

```

})
res60_si <- reactive({
  pp60 <- as.POSIXlt(as.POSIXct( periods60() ), tz="GMT")
  long60 <- length(pp60)
  res60_si <- data.frame(start = pp60[1:(long60/2)],
                        end = pp60[(long60/2 + 1):long60])
})
output$resultat60_si <- renderPrint({
  resultat60_si = res60_si()
  resultat60_si
})
output$resultat60 <- renderPrint({
  resultat60 = res60()
  resultat60
})

#----- 80
# 80 %
ind80 <- reactive({
  bb = bsadf2()
  qq80 = quantile95()
  ind80 <- bb > t(qq80[2, ]) * 1
})
periods80 <- reactive({
  ii = ind80()
  mm = monitorDates()
  periods80 <- locate2(ii, mm)
})

logical_foundperiods80 <- reactive({is.null(periods80())})
output$prova <- renderPrint({
  dd = logical_foundperiods80()
  dd
})
res80 <- reactive({
  if(as.logical(logical_foundperiods80())){ res80 <- "No bubble or crisis
periods found"
} else{ res80 <- "Bubble found" }
})
res80_si <- reactive({
  pp80 <- as.POSIXlt(as.POSIXct( periods80() ), tz="GMT")
  long80 <- length(pp80)
  res80_si <- data.frame(start = pp80[1:(long80/2)],
                        end = pp80[(long80/2 + 1):long80])
})
output$resultat80_si <- renderPrint({
  resultat80_si = res80_si()
  resultat80_si
})
output$resultat80 <- renderPrint({
  resultat80 = res80()
  resultat80
})

#----- 90
# 90 %
ind90 <- reactive({
  bb = bsadf2()
  qq90 = quantile95()
  ind90 <- bb > t(qq90[2, ]) * 1
})
periods90 <- reactive({
  ii = ind90()
  mm = monitorDates()
  periods90 <- locate2(ii, mm)
})

logical_foundperiods90 <- reactive({is.null(periods90())})

```

```

output$prova <- renderPrint({
  dd = logical_foundperiods90()
  dd
})
res90 <- reactive({
  if(as.logical(logical_foundperiods90())){           res90 <- "No bubble or crisis
periods found"
  }else{           res90 <- "Bubble found"   }
})
res90_si <- reactive({
  pp90 <- as.POSIXlt(as.POSIXct(periods90()),tz="GMT")
  long90 <- length(pp90)
  res90_si <- data.frame(start = pp90[1:(long90/2)],
                        end   = pp90[(long90/2 +1):long90])

})
output$resultat90_si <- renderPrint({
  resultat90_si = res90_si()
  resultat90_si
})
output$resultat90 <- renderPrint({
  resultat90 = res90()
  resultat90
})

#----- 95
# 95 %
ind95 <- reactive({
  bb = bsadf2()
  qq95 = quantile95()
  ind95 <- bb > t(qq95[2, ]) * 1
})
periods95 <- reactive({
  ii = ind95()
  mm = monitorDates()
  periods95 <- locate2(ii, mm)
})

logical_foundperiods95 <- reactive({is.null(periods95())})
output$prova <- renderPrint({
  dd = logical_foundperiods95()
  dd
})
res95 <- reactive({
  if(as.logical(logical_foundperiods95())){           res95 <- "No bubble or crisis
periods found"
  }else{           res95 <- "Bubble found"   }
})
res95_si <- reactive({
  pp95 <- as.POSIXlt(as.POSIXct(periods95()),tz="GMT")
  long95 <- length(pp95)
  res95_si <- data.frame(start = pp95[1:(long95/2)],
                        end   = pp95[(long95/2 +1):long95])

})
output$resultat95_si <- renderPrint({
  resultat95_si = res95_si()
  resultat95_si
})
output$resultat95 <- renderPrint({
  resultat95 = res95()
  resultat95
})

}

shinyApp(ui = ui, server = server)

```