Trabajo autónomo I:

Series Temporales

Nombre: Francisco Pérez Hernández

E-mail: [herpefran92@gmail.com](mailto:herpefran92@gmail.com)

Asignatura: Series temporales y minería de flujos de datos

Máster: Ciencia de Datos e Ingeniería de Computadores

Fecha de entrega: 19/05/2017

# 1 Parte teórica

## Serie temporal

Una serie temporal es cualquier magnitud observada a lo largo del tiempo, en un intervalo regular ya sea cada minuto, hora, día, semana, etc. Además, tenemos series estacionarias o no estacionarias, en las que, se dice estacionaria cuando su media y varianza no varía con el tiempo. La estacionaridad indica que las propiedades estadísticas de la serie no varían en el tiempo y, por tanto, los datos pueden estudiarse bajo un mismo modelo paramétrico independiente del tiempo. Además, que sea estacionaria es un requisito para poder aplicar modelos paramétricos de análisis y predicción de series de datos. Las series no estacionarias pueden tener o no tener tendencia o estacionalidad.

## Metodología Box-Jenkins para predicción de series temporales

La metodología Box-Jenkins para predicción de series temporales conlleva:

* Dividir la serie X(t) en tendencia T(t), estacionalidad S(t) y una componente irregular (E(t)). Enfoque aditivo: X(t) = T(t) + S(t) + E(t)
* **Cálculo de la estacionalidad**: valor medio de los valores de la serie dentro del mismo punto estacional.
* **Cálculo de la tendencia**: regresiones, filtrado de la señal, diferenciación, etc.
* **Metodología**:
  + Eliminar tendencia y estacionalidad de la serie.
  + Hacer E(t) estacionaria y aplicar métodos paramétricos.

## Técnicas de modelado de tendencia

La tendencia es el incremento o decremento a largo plazo de los datos. Para modelarla tenemos diferentes técnicas como pueden ser:

* **Estimación funcional**: Aproximar la tendencia de la serie como una función. Requiere realizar hipótesis sobre el modelo que rige la tendencia como los modelos lineales, polinómicos, etc.
* **Filtrado**: Esta opción consiste en aplicar un filtro de medias móviles para estimar la tendencia.
* **Diferenciación**: Se diferencia la señal hasta que desaparezca la tendencia.

## Técnicas de modelado de estacionalidad

La estacionalidad son los datos afectados por un patrón estacional tal como el día del año o el día de la semana. Para modelarla necesitamos:

* Encontrar el periodo de la estacionalidad. Para ello podemos ayudarnos de la ACF para calcular el periodo.
* Se calcula el valor promedio de cada punto de la estación (si es mensual, la media para cada mes; si es semanal, la media para cada día de la semana, etc.).
* Otro método consiste en utilizar la diferenciación, con un desfase d igual al periodo de la estacionalidad: X’(t) = X(t) – X(t-d)
* Existen métodos como la transformada de Fourier o el método de Holts-Winter de alisado exponencial para estacionalidad.

## Proceso para obtener los parámetros de un modelo ARIMA

Un modelo ARIMA está compuesto de 3 componentes p, d y q: ARIMA(p,d,q). Lo primero que tenemos que hacer es saber si el modelo es AR o MA y en qué orden:

* Los modelos AR tienen un ACF que decrece a 0 (con diferentes posibles formas: regulares, sinusoidales, alternando +/-). El número del orden “p” (AR(p)) es tantos valores “distintos de 0 como haya en el PACF”.
* Los modelos MA tienen un PACF que decrece a 0 (con diferentes posibles formas: regulares, sinusoidales, alternando +/-). El número del orden “q” (MA(q)) en tantos “valores distintos de 0” como haya en el ACF.
* Un valor se considera “distinto de cero” si no está en el rango (-2/sqrt(N), 2/sqrt(N)), con N=longitud de la serie.

También tenemos que hallar el parámetro d de la siguiente forma:

* Cuando la serie no es estacionaria, el ACF decrece lentamente a 0.
* La parte integrada es necesaria normalmente para corregir la estacionaridad en la varianza.
* Si la serie presenta tendencia lineal, normalmente con d=1 es suficiente. Si la tendencia no es lineal, puede ser necesario usar d>1.
* Si la serie presenta estacionalidad, puede ser necesario un d=periodo de estacionalidad.

# 2 Parte Práctica

## 2.1 Pasos seguidos

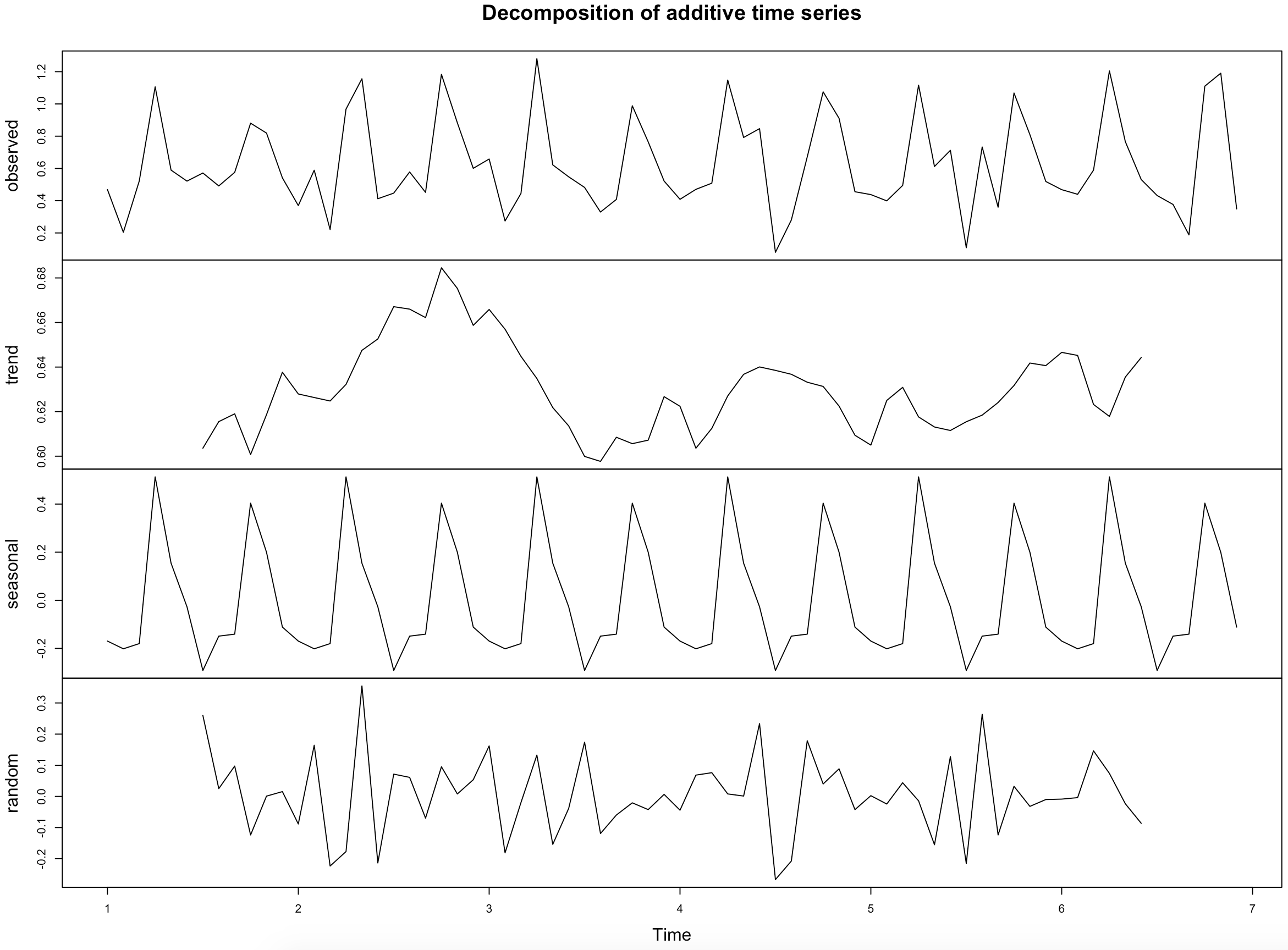
Los pasos que se van a seguir durante esta parte práctica serán:

1. Analizar el problema
2. Pre-procesamiento
3. Eliminación de tendencia
4. Eliminación de estacionalidad
5. Hacer la serie estacionaria
6. Obtención de los parámetros del modelo ARIMA
7. Selección del mejor modelo
8. Realizar predicción

## 2.2 Necesidad o no de pre-procesamiento

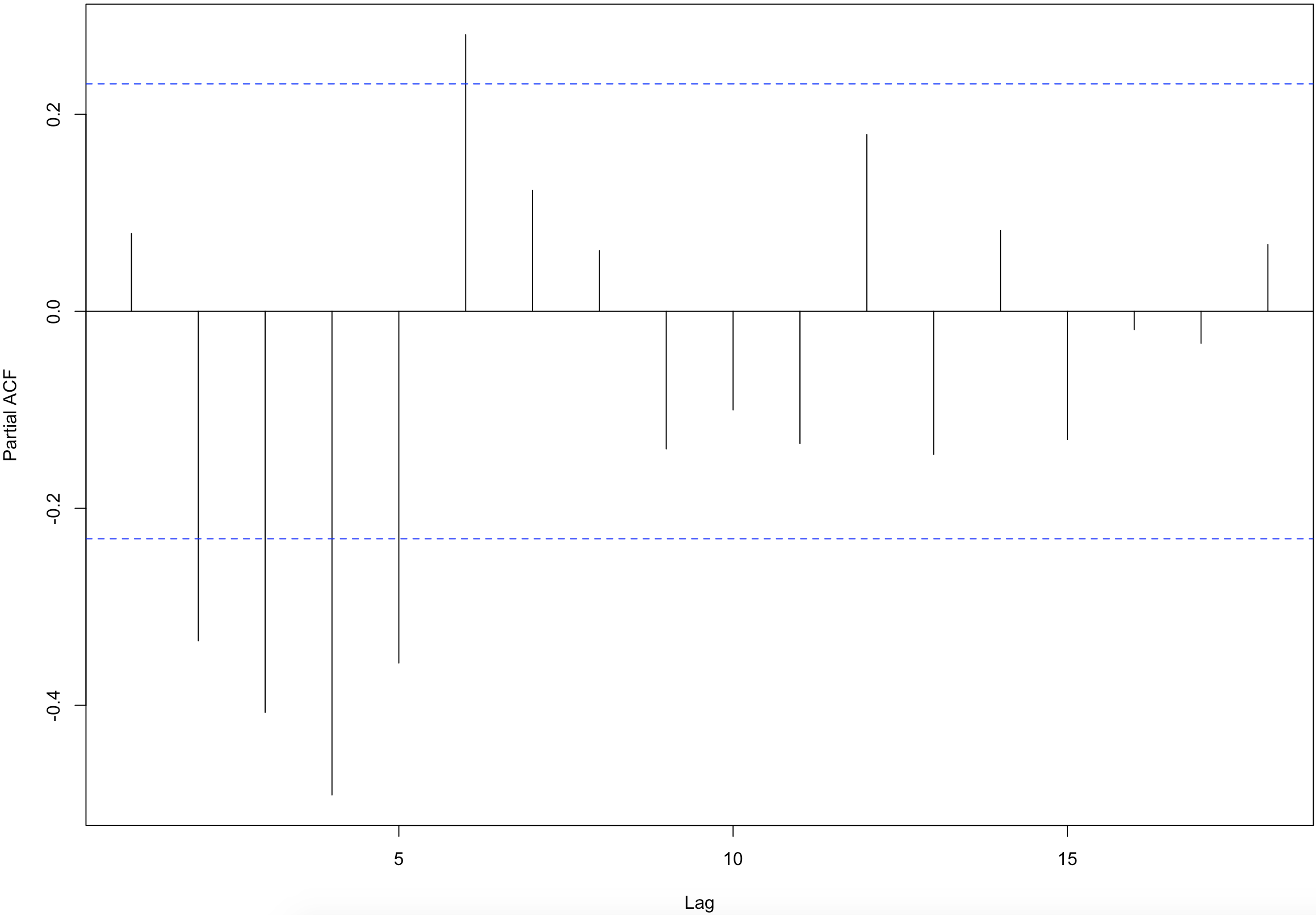
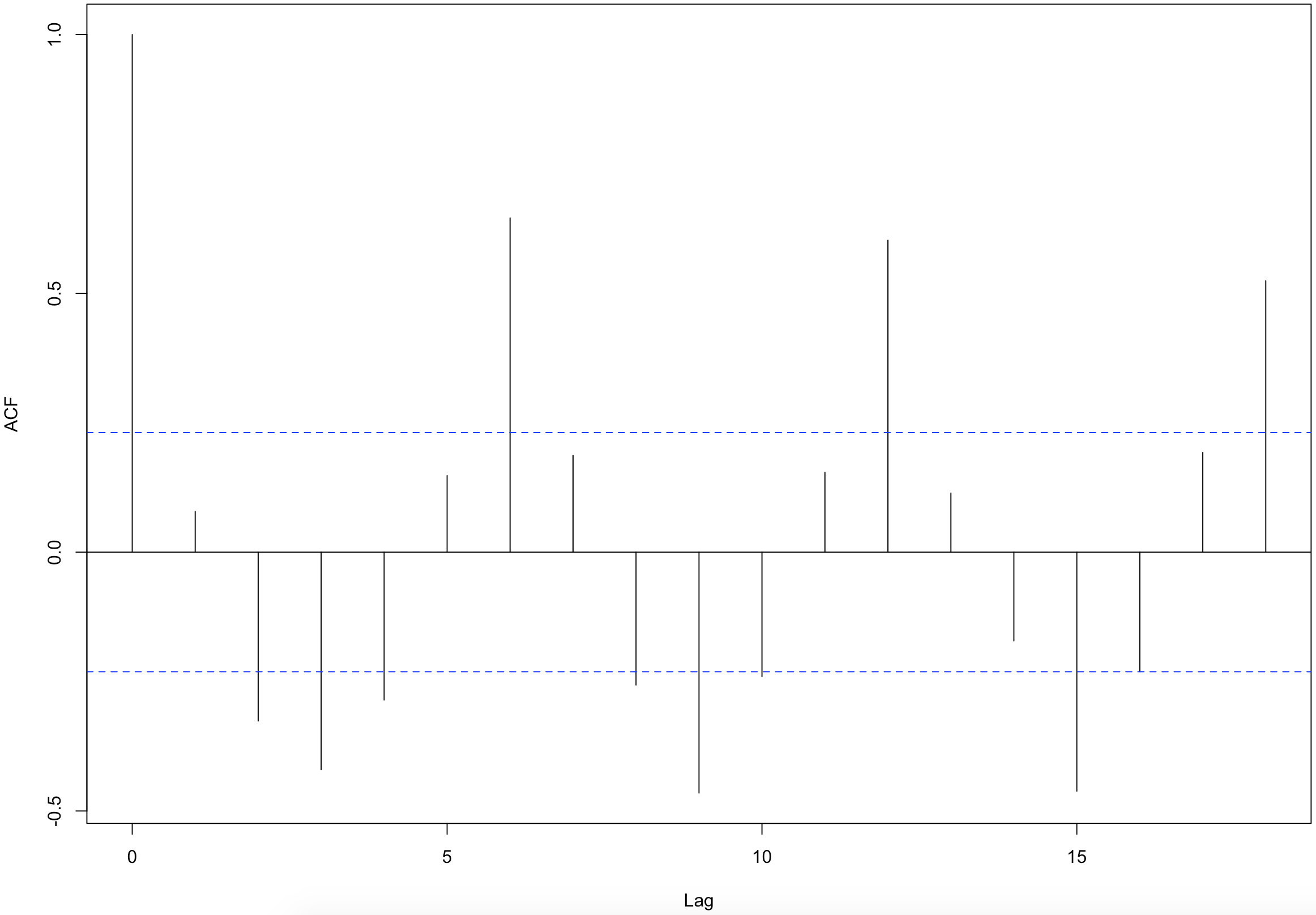
El fichero de datos que vamos a analizar es un fichero en el que tenemos los datos sobre el número de ventas (en miles) de un producto durante 6 años en los conocidos almacenes Guardo-To-Íto, desde enero de 2010 hasta 2015. Lo que queremos realizar, por tanto, es predecir las ventas para el año 2016.

Lo primero que hacemos es cargar la serie como una serie temporal con una frecuencia de 12 ya que estamos tratando las ventas anuales. Pero cuando vemos la primera gráfica, vemos como la frecuencia puede ser de 6 meses por el comportamiento que se visualiza. Seguidamente hacemos una descomposición para ver más información:



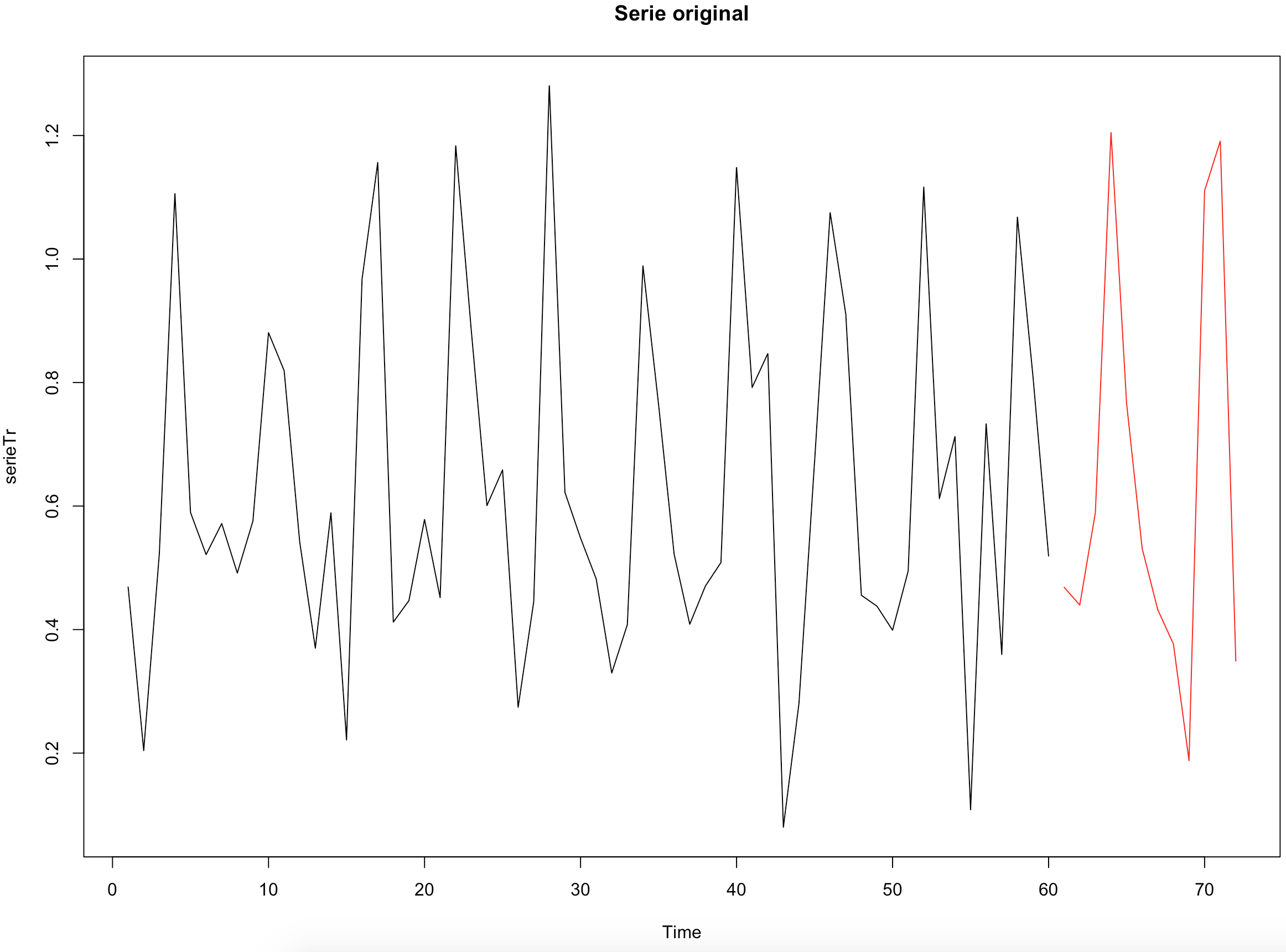
Podemos ver 3 aspectos: 1 La variabilidad de la estacionalidad no aumenta ni decrece. 2 Puede que no haya tendencia en la serie. 3 Por la gráfica de estacionalidad, se puede ver que habrá estacionalidad en la serie.

Mostramos además el gráfico ACF y PACF:



Si además, hacemos el test de Dickey-Fuller aumentado, nos da un valor de p-value = 0.01 por lo que podemos decir que la serie es Estacionaria con un nivel de confianza del 99%. Además, en el gráfico ACF vemos que el Lag es 6.

Vamos a dividir la serie original en train y test, dejando un total de un año, 12 elementos, para la parte de test.



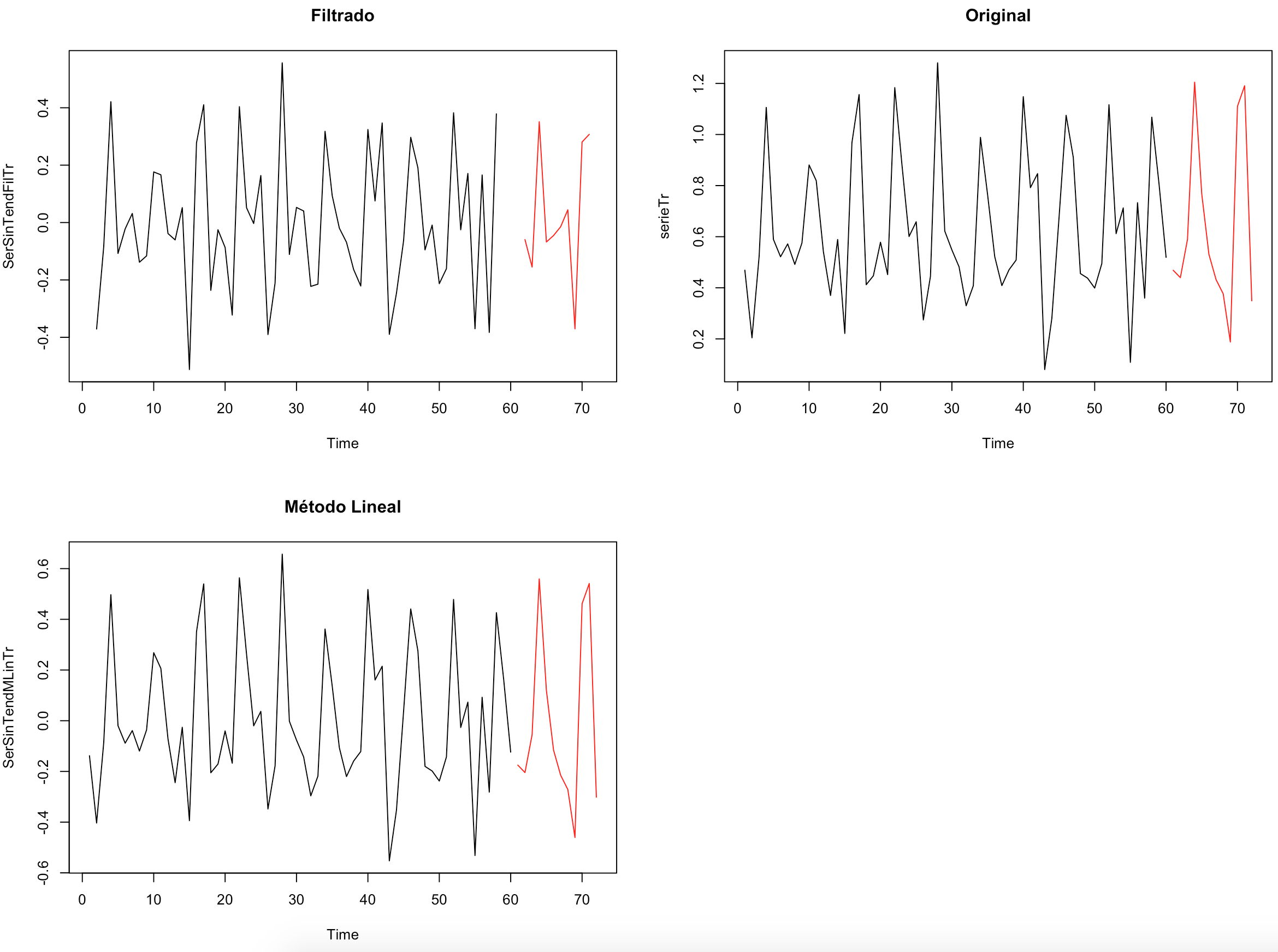
## 2.3 Necesidad o no de eliminación de tendencia

Para modelar la tendencia he decidido realizar dos modelos, el lineal y el filtrado.

El modelo lineal nos ha dado una suma de residuos RSS igual a 4,717. Además, pasa los test de jarque.bera y el t.test, por lo que podemos decir que la aproximación lineal es factible.

El modelo basado en filtrado, se ha quedado con unos valores de k para train de 4 y de 3 para test.

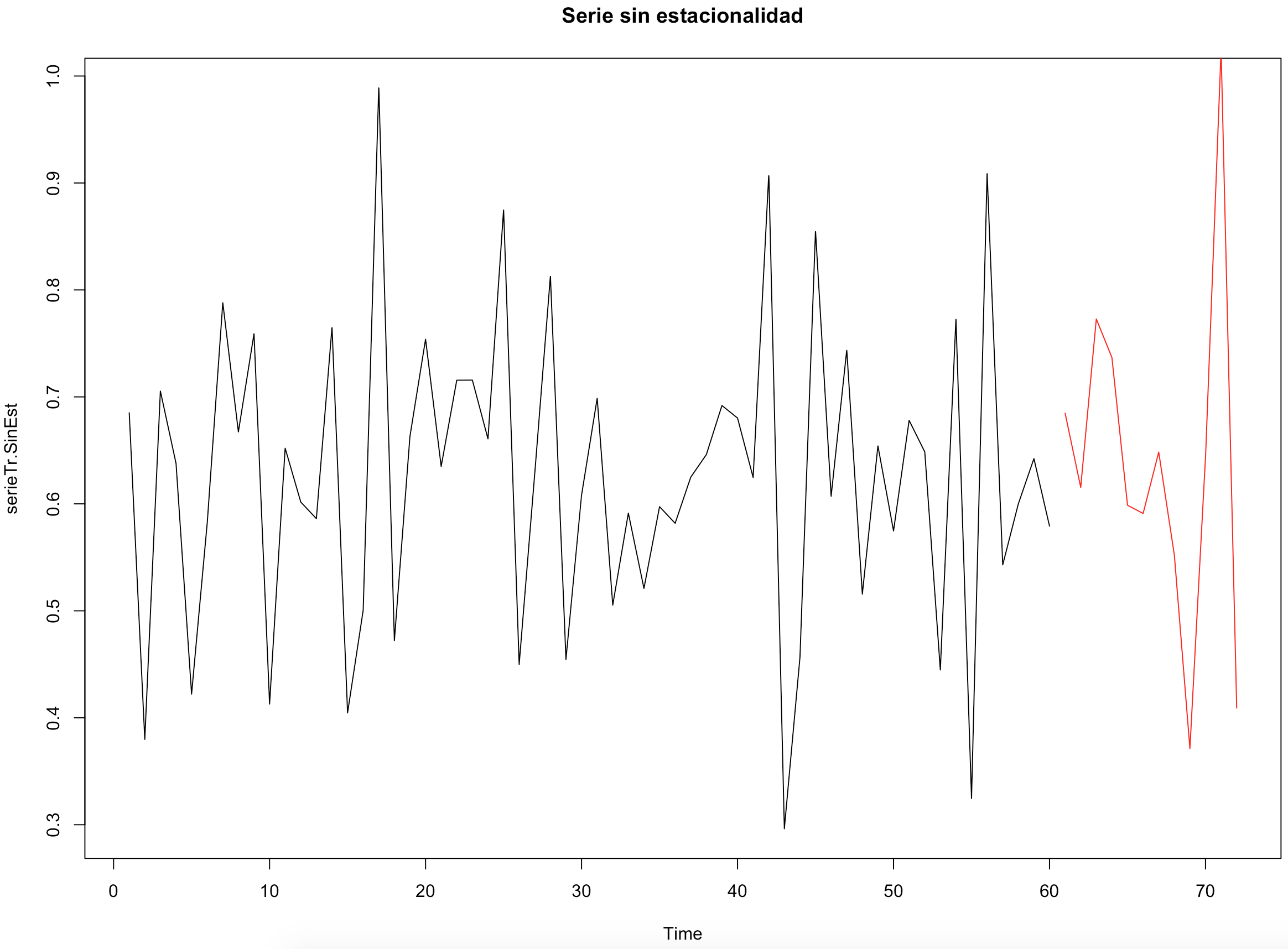
Los resultados de la serie sin tendencia serían:



En base a estos resultados, he decidido no eliminar tendencia en la serie, ya que como habíamos dicho en el pre-procesamiento, la serie no presenta tendencia.

## 2.4 Necesidad o no de eliminación de estacionalidad

Como habíamos visto en el apartado 2.2, la serie presenta una estacionalidad que se repite cada 6 meses. Primero habíamos pensado en 12 al ser una serie de datos anual, pero hemos visto en las gráficas que se trataba de una de 6 meses. Por lo tanto, si eliminamos estacionalidad, sabiendo el periodo, quedando la serie de la siguiente forma:



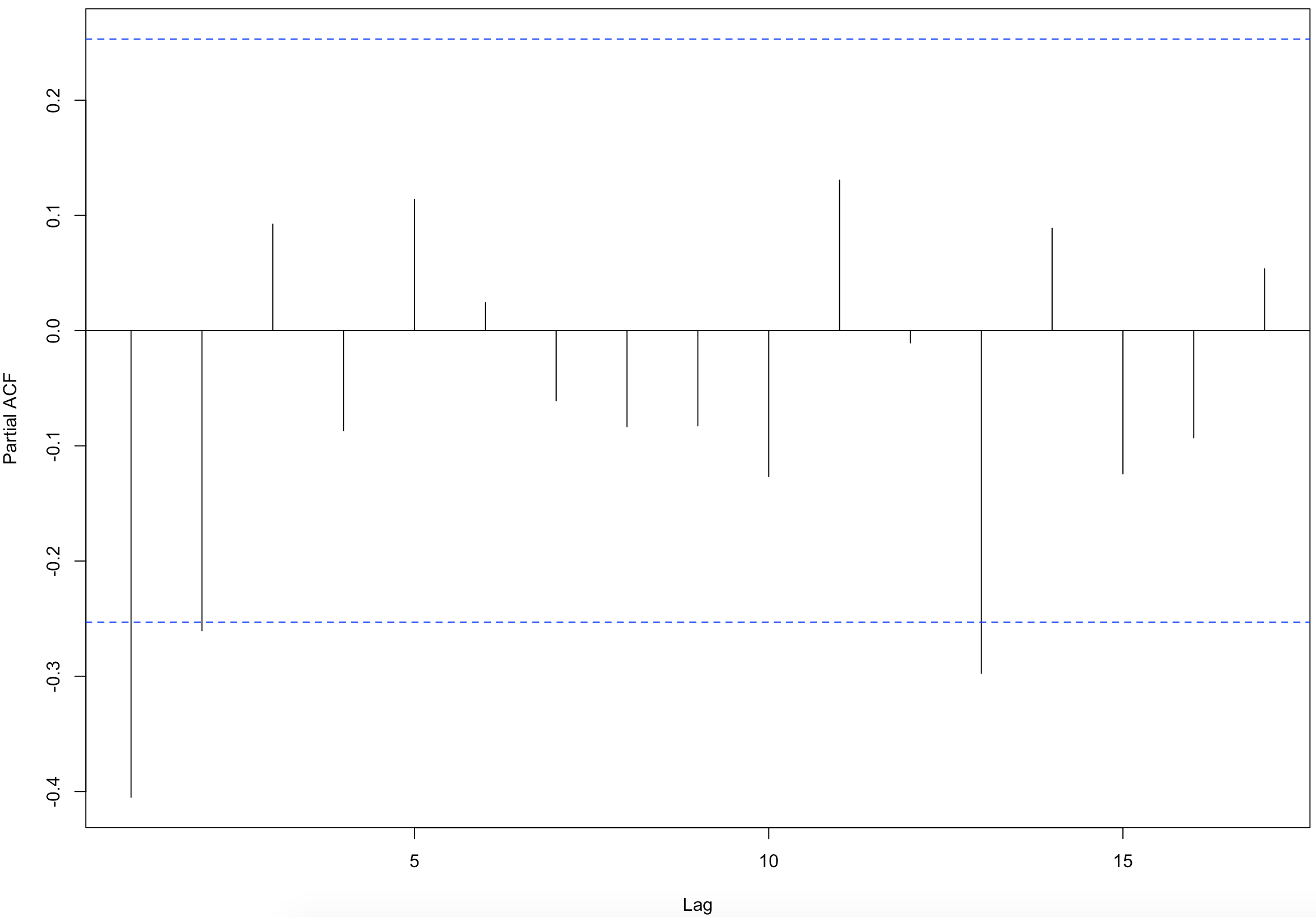
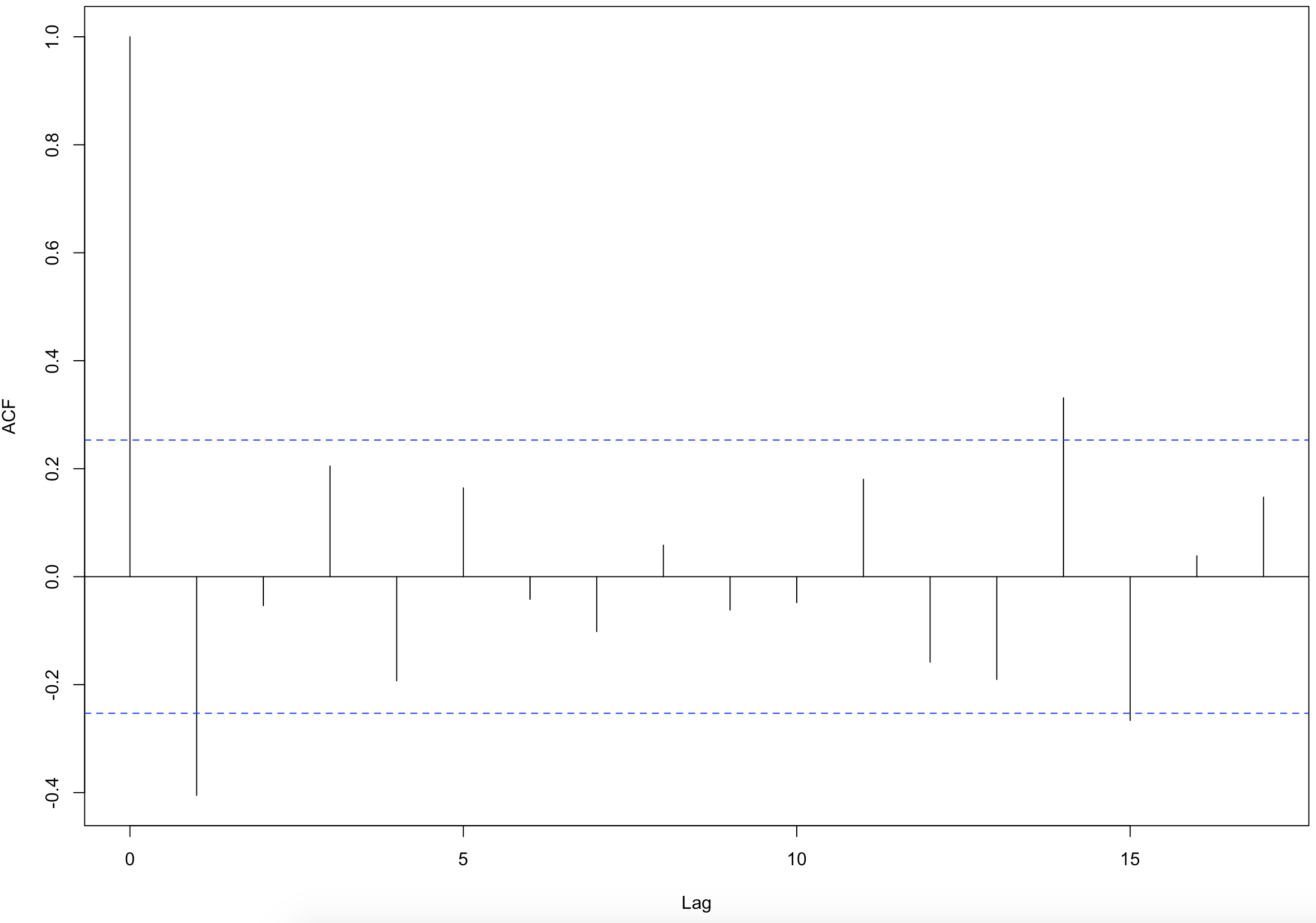
## 2.5 Necesidad o no de hacer la serie estacionaria

Para saber si la serie es estacionaria o no, como ya hemos visto en el apartado 2.2, realizamos el test de Dickey-Fuller aumentado, con el que hemos visto con un 99% de confianza, que la serie presenta estacionariedad. Por lo tanto, como ya sabemos que es estacionaria, no es necesario diferenciarla.

Por lo tanto, vamos a pasar a realizar nuestro modelo ARIMA.

## 2.6 Obtención de los parámetros del modelo ARIMA

Veamos para comenzar los gráficos ACF y PACF:



Viendo estos gráficos podemos decidir si tenemos modelos AR o MA y en qué orden.

De forma que, para tener un AR, el gráfico ACF tiene que decrecer a 0, condición que se cumple. Además, su orden “p” será 2, ya que el gráfico PACF tiene dos valores distintos a 0.

Para tener un MA, el gráfico PACF decrece a 0, con un orden “q” igual a 2 ya que tiene 2 valores distintos a 0 en el ACF.

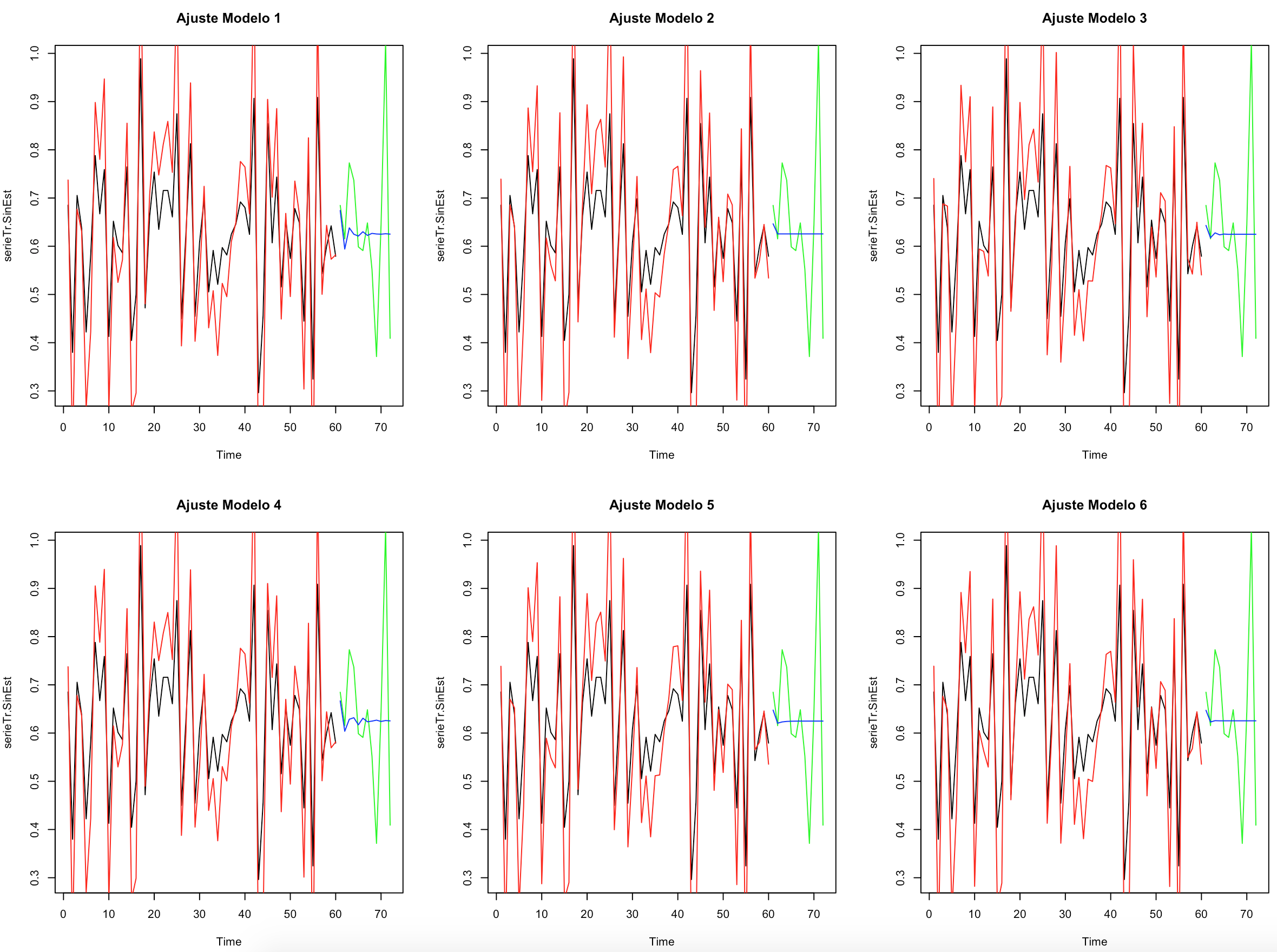
Ya solo falta el valor “d”, pero como no hemos realizado diferenciación, su valor será 0.

Por lo tanto vamos a probar con los modelos ARIMA(2,0,2), ARIMA(0,0,1), ARIMA(1,0,0), ARIMA(2,0,1), ARIMA(1,0,2) y ARIMA(1,0,1) para poder tener una comparación amplia.

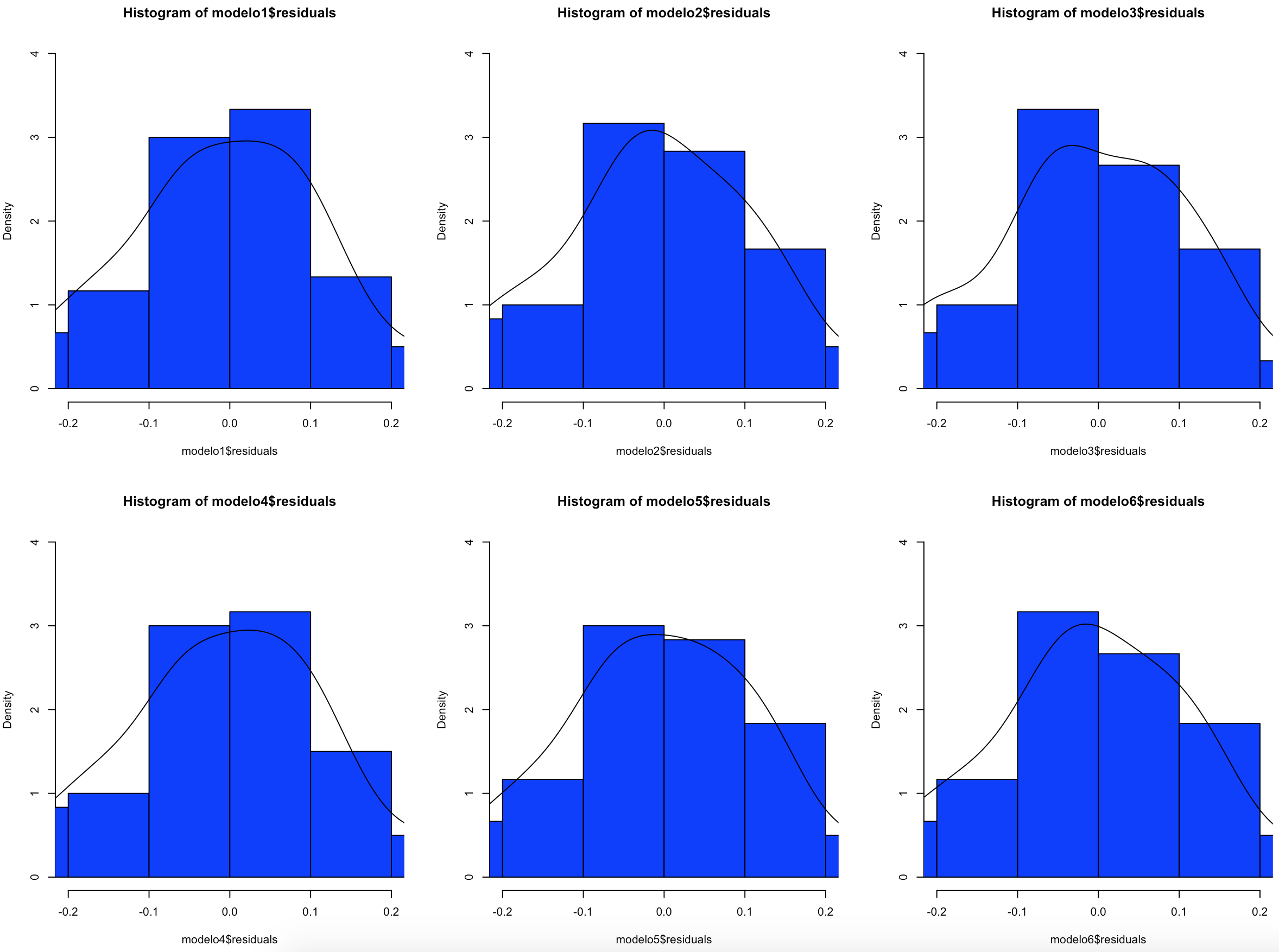
## 2.7 Selección del mejor modelo

Para cada uno de los 6 modelos que voy a comparar, le calculo los errores de ajuste en train, los errores de predicción de test, mostrando sus gráficas de ajuste y predicción en test. Además, realizamos los test Box-Pierce, Jarque Bera y el Shapiro-Wilk. Estos test son pasados por los 6 modelos en todos los test. Además, se muestra los histogramas de los residuos.

Si mostramos las gráficas de los 6 modelos tenemos:



Además, podemos ver los histogramas de los residuos de los 6 modelos:



Para seleccionar el mejor modelo, nos vamos a basar en el criterio de Akaike (AIC), ya que los 3 test, que hemos comentado anteriormente, han pasado los resultados satisfactoriamente.

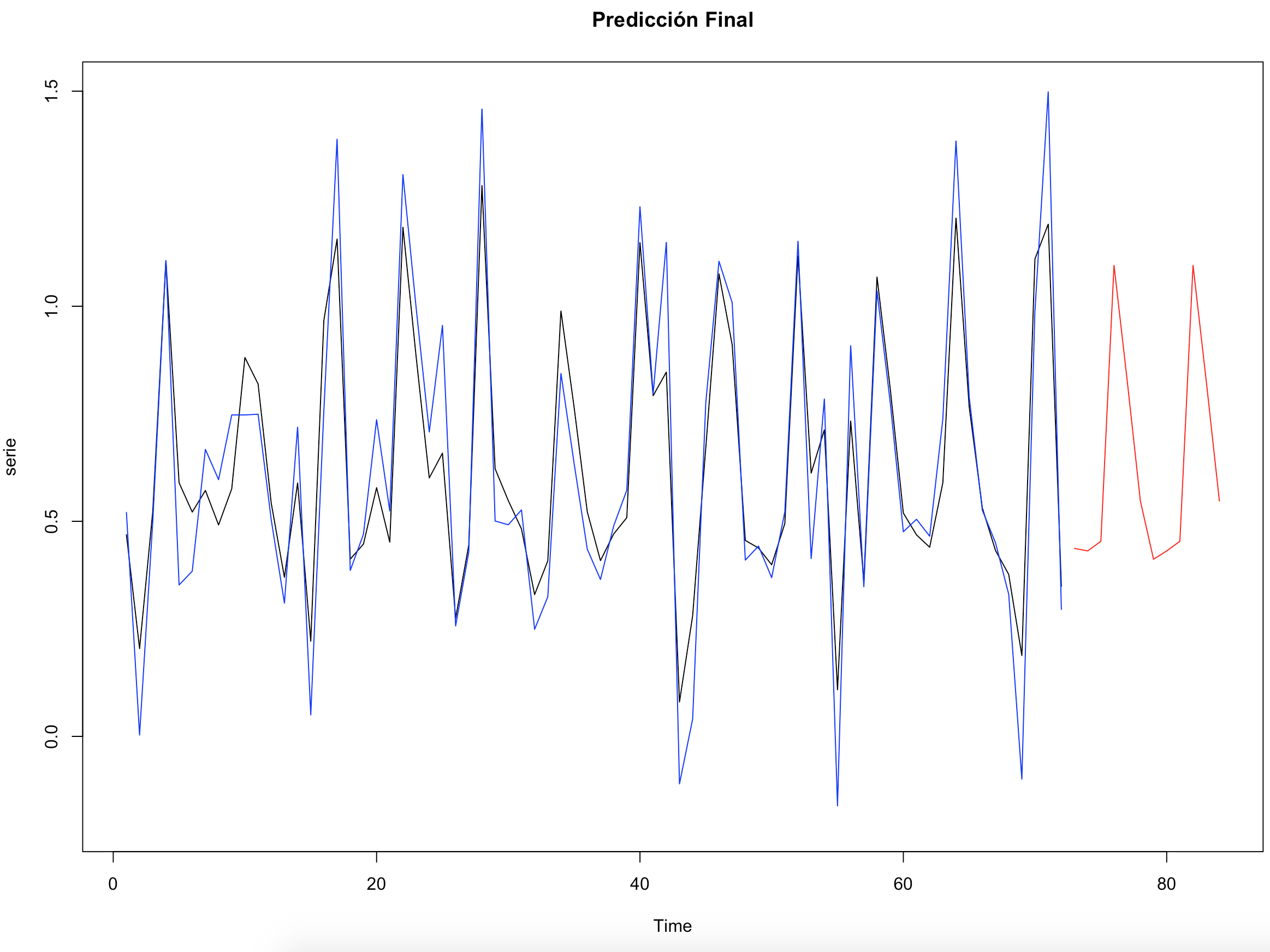
El resultado del criterio de Akaike es:

|  |  |  |
| --- | --- | --- |
| Modelo | df | AIC |
| ARIMA(2,0,2) | 6 | - 67.67963 |
| ARIMA(0,0,1) | 3 | - 70.55199 |
| ARIMA(1,0,0) | 3 | - 68.12978 |
| ARIMA(2,0,1) | 5 | - 69.54970 |
| ARIMA(1,0,2) | 5 | - 67.26716 |
| ARIMA(1,0,1) | 4 | - 68.76011 |

Por lo que nos vamos a quedar con el modelo ARIMA(0,0,1) al presentar menos errores, con un valor de -70.55.

## 2.8 Obtención de los valores predichos para la serie

Para predecir la serie lo primero que vamos a hacer es aplicar lo hecho durante la práctica, es decir, vamos a quitarle la estacionalidad, vamos a aplicar el modelo ARIMA(0,0,1) y vamos a predecir las siguientes 12 resultados de estas ventas de las que trata el conjunto de datos. Seguidamente deshacemos los cambios e imprimimos la serie final con su predicción, quedando:



En esta gráfica vemos, en negro la serie original, en azul los valores ajustados y en rojo, los valores predichos para los siguientes 12 meses de esta serie.

\*El código estará adjunto en la entrega de la práctica.