# *Series Temporales y Minería de Flujo de Datos*

Máster en Ciencia de Datos e Ingeniería de Computadores

**Universidad de Granada**

Trabajo autónomo I: Series Temporales

19 Mayo 2017

Alberto Castillo Lamas
alcasla90@gmail.com

# 1 Teoría

## 1.1 Serie temporal. Definición

Se define como una colección de observaciones de una variable recogidas secuencialmente en el tiempo u ordenadas cronológicamente. Las observaciones usualmente se recogen en instantes de tiempo equiespaciados. Su análisis se realiza mediante métodos que ayudan a interpretarlas y que permiten extraer información representativa sobre las relaciones subyacentes entre los datos de la serie o diversas series que permiten en diferente medida y con distinta confianza extrapolar o interpolar los datos y así predecir el comportamiento de la serie en momento futuros o no observados. Por eso uno de los usos más habituales de las series de datos temporales es su análisis para predicción o pronóstico.

## 1.2 Metodología Box-Jenkins

Se aplica a los modelos autorregresivos y de media móvil (ARMA), o a los modelos autorregresivos integrados de media móvil (ARIMA). Para encontrar el mejor ajuste de una serie temporal de valores, con el fin de que los valores pronosticados sean certeros. El modelo original utiliza un enfoque de modelado iterativo en tres etapas. Las etapas son las siguiente:

- Identificación y selección del modelo. Asegurarse de que las variables son estacionarias, identificación de la estacionalidad, y el uso de las funciones de autocorrelación y de autocorrelación parcial de la serie de tiempo para deducir que componentes del modelo ARIMA seleccionar.
- Estimación de parámetros usando algoritmos de cálculo para tener los coeficientes que mejor se ajusten al modelo ARIMA.
- Comprobar el modelo mediante el ensayo, si el modelo estimado se ajusta a las especificaciones de un proceso univariado estacionario. Siendo así los residuos independientes unos de otros, además la media y varianza deben ser constantes en el tiempo.

## 1.3 Modelado de tendencia

- Estimación funcional. Aproximar la tendencia de la serie como una función. Hacer hipótesis sobre el modelo que rige la tendencia. Por ejemplo se puede usar un modelo lineal, logarítmico, polinomial,...
- Filtrado. Aplicar un filtro de medias móviles para estimar la tendencia. Se puede variar el número de muestras usadas en la aplicación del filtro provocando un mayor suavizado de la serie con su aumento.
- Diferenciación. Diferenciar la serie hasta que desaparezca la tendencia.

En mi caso lo he realizado mediante una aproximación funcional lineal y con un filtro de medias móviles, pero se ha descartado ambas hipótesis debido a la ausencia de tendencia en la serie.

## 1.4 Modelado de estacionalidad

- Diferenciación. Utilizando un desfase igual al periodo de la estacionalidad. Nos podemos apoyar con la visualización de la autocorrelación.
- Transformada de Fourier
- Método de Hols-Winter

En este caso práctico si existe estacionalidad con un periodo anual y se ha eliminado utilizando diferenciación.

## 1.5 Hipótesis ARIMA

Para realizar una hipótesis de este tipo de modelo hay que estimar tres valores fundamentales que definen como tal al modelo. El grado de AR (modelo autorregresivo), I (diferenciación), y MA (modelo de medias móviles).

- I. El más simple, grado de diferenciación. Tras el proceso de eliminación de estacionalidad necesitamos que la serie se estacionaria, y para ello en caso de no serla se procede a su diferenciación para hacerla estacionaria.

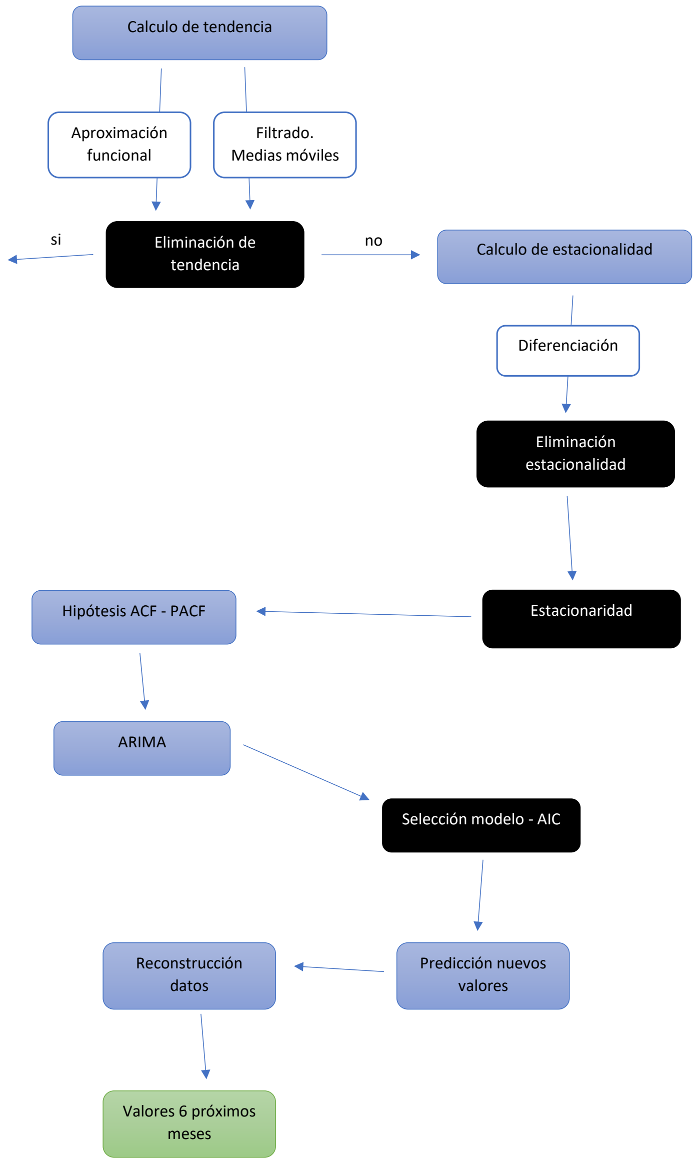
- AR. Observado la forma de la gráfica de autocorrelación junto al número de coeficientes significativos, mayores y/o entorno al umbral, en la autocorrelación parcial, empezando desde cero.

- MA. Al contrario que AR, número de coeficientes significativos en la autocorrelación junto a la tendencia de los coeficientes de la autocorrelación parcial.

Hay que tener ciudado para no seleccionar un modelo que ajuste demasiado la señal, ya que estaríamos modelando *ruido blanco*, lo cual no nos interesa. El objetivo es obtener un modelo que generalice adecuadamente, por eso a veces el modelo que se obtiene es la media formando así una linea sobre la media en la predicción.

## 2 Práctica

Figure 1: Pasos seguidos durante la realización de la práctica



En la figura 1 se observa el esquema de los pasos llevados a cabo durante el análisis, descomposición y predicción para la serie temporal dada.

Los valores en bruto leídos del fichero proporcionado para práctica *SerieTrabajoPractico.dat* se distribuyen como muestra la gráfica de la figura 2. Contiene valores entre 0 y 1.3, así el primer paso va a ser normalizar los datos, como resultado la serie queda como se muestra en la figura 3.

Figure 2: Distribución de valores de la serie en bruto

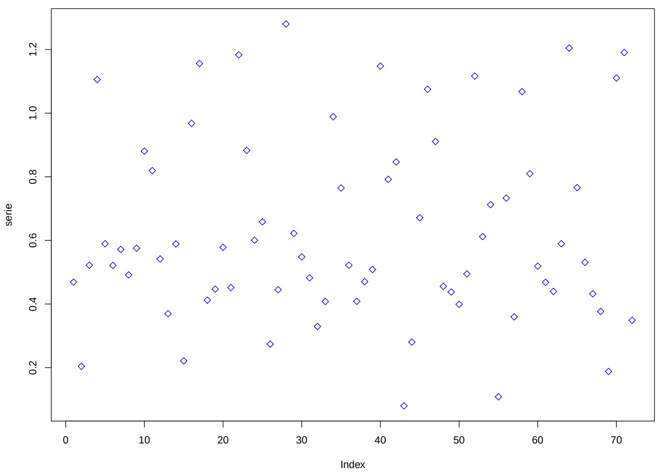
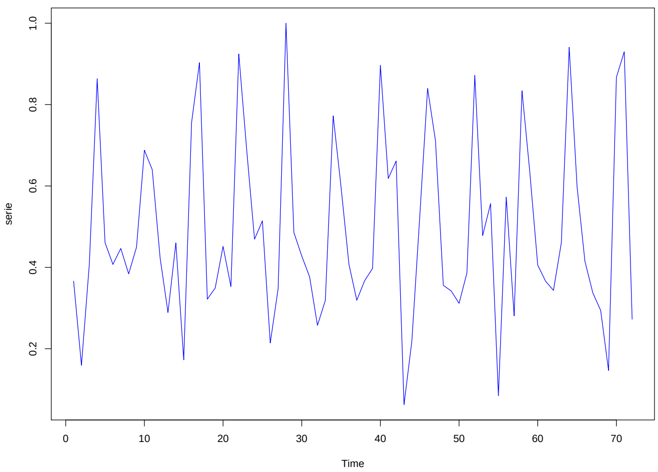



Figure 3: Serie tras normalizacion ([0,1])




Continua con el análisis de la serie. Primero, se calcula la tendencia, que se llevará a cabo de dos formas. Primero por una *aproximación funcional lineal*, y la segunda mediante filtro de medias móviles tomando *k=5* muestras, cuya tendencia (línea roja) se muestra en las figuras 4 y 5 respectivamente.

Como se observa ambas tendencias son planas, literalmente en la lineal y también en el segundo caso en general. Además la serie tras eliminar tendencia por cualquier de las dos formas se mantiene igual, por tanto confirma la no existencia de tendencia.

Figure 4: Tendencia mediante aproximación funcional lineal

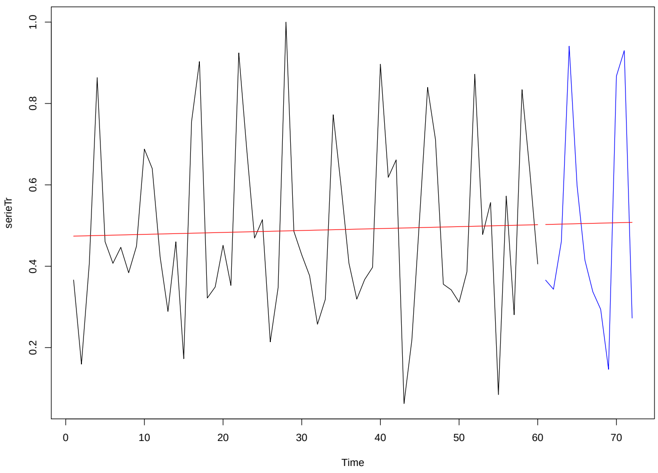
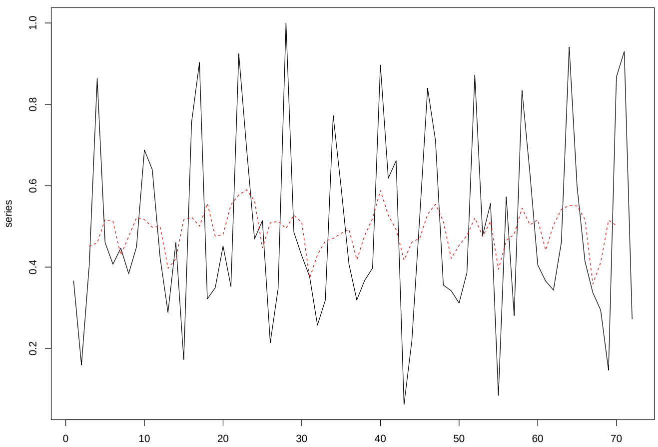



Figure 5: Tendencia mediante filtro de medias móviles k=5 muestras

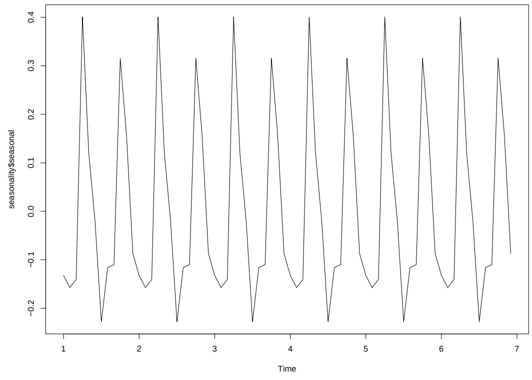



Se continua con la serie normalizada sin eliminar tendencia. El segundo paso será eliminar estacionalidad, mediante la descomposición de la serie por una frecuencia de 12 muestras (meses de año) se observa como el componente estacional tiene un claro patrón, figura 6 superior. Tras eliminar este componente estacional de periodo 12 nos queda la señal como muestra la figura 6 (inferior).

A continuación se chequea la estacionaridad de la serie mediante el test, y con el apoyo de la autocorrelación. Pasa el test y la autocorrelación se muestra como la figura 7, claramente estacionaria. Por tanto se puede proceder a la construcción del modelo ARIMA, o en este caso ARMA al no haber sido diferenciada la serie (d=0).

Según observo ambas gráficas que muestra la figura 7 se plantean las siguiente hipótesis.

Figure 6: Componente estacional de la serie tras su descomposición, y el resultado tras eliminar estacionalidad.



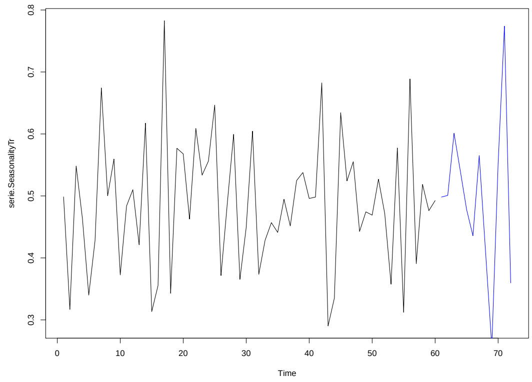
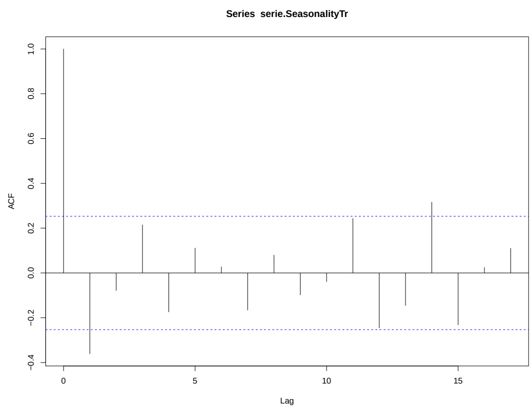
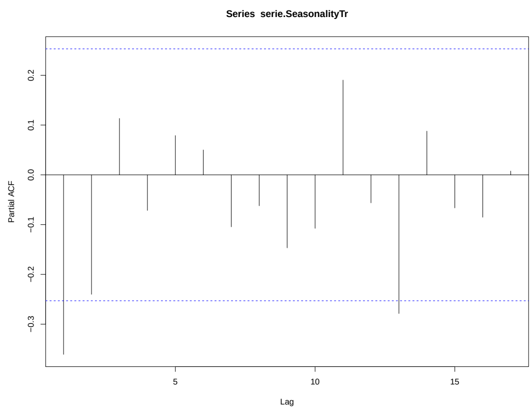




Figure 7: Autocorrelación (acf) y autocorrelación parcial (pacf)

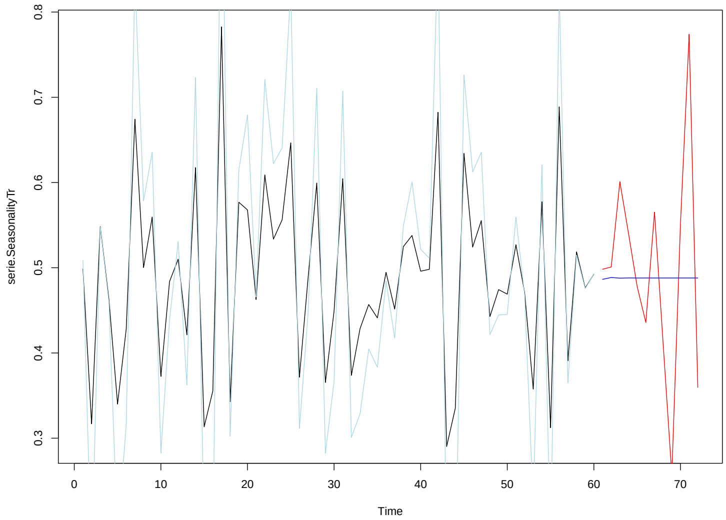


**ARIMA 1,0,0**

Observando la oscilación de la autocorrelación y dependiendo de la interpretación en la autocorrelación parcial podemos tomar el coeficiente AR como 0 o 1. En mi caso tomo el coeficiente cercano al umbral como significativo.

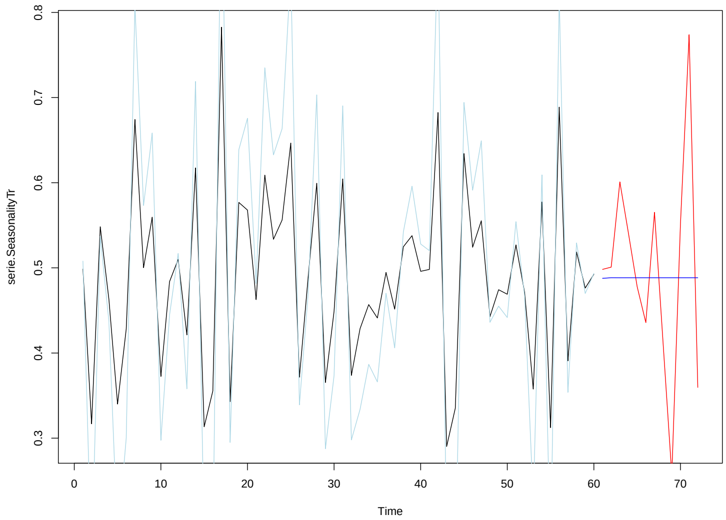
Figure 8: Modelo ARIMA para la primera hipótesis. Predicción realizada por el modelo como línea azul.




**ARIMA 1,0,1**

Al igual que en la hipótesis anterior, si se observa la función de autocorrelación parcial por esa oscilación y coeficientes significativos en la autocorrelación diría que el coeficiente MA es 2, aunque en este caso pruebo primero con 1, es dudoso el segundo coeficiente en la autocorrelación.

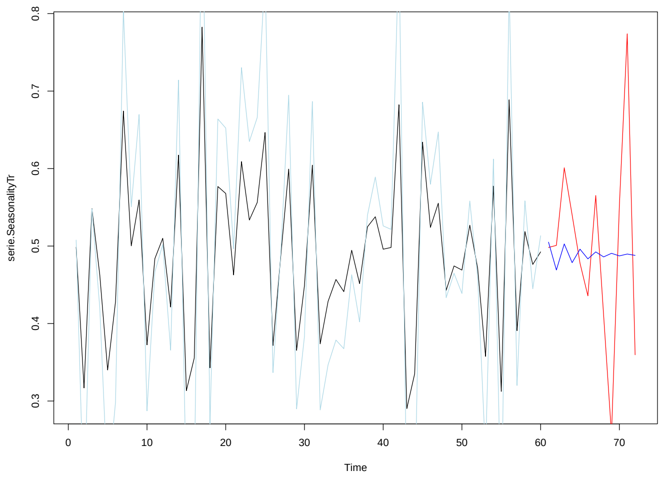
Figure 9: Modelo ARIMA para la segunda hipótesis. Predicción realizada por el modelo como línea azul.




**ARIMA 1,0,2**

Ahora si, aunando intuiciones de ambas hipótesis anteriores se genera este modelo más cercano a mi interpretación de ACF y PACF.

Figure 10: Modelo ARIMA para la tercera hipótesis. Predicción realizada por el modelo como línea azul.




# 3 Elección de modelo y predicción

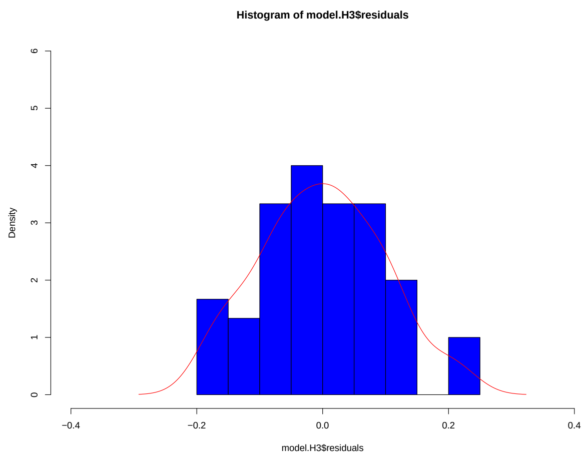
Para cada modelo se ha calculado el error cuadrático acumulado para ambos conjuntos en los que se ha dividido la serie para poder testearlo. Además se calcula el AIC, medida a la que se otorgará mayor importancia para la selección del modelo más adecuado.

Table 1: Estadísticos sobre cada modelo obtenido. En negrita el modelo seleccionado.

| Modelo | Error cuadrático acumulado Train | Error cuadrático acumulado Test | AIC |
|---|---|---|---|
| ARIMA 1,0,0 | 0.5837419 | 0.1864003 | 101.5511 |
| ARIMA 1,0,1 | 0.5631754 | 0.1862647 | 101.6426 |
| **ARIMA 1,0,2** | 0.555204 | 0.1842127 | 100.4651 |

Se observa como poseen valores de AIC similares, aunque el más reducido y cuyo error cuadrático acumulado en el test ha sido más bajo es el modelo ARIMA 1,0,2. Dicho modelo muestra un histograma y función de densidad mostrado en la figura 11.

Figure 11: ARIMA 1,0,2 histograma y función de densidad




**Predicción** Con este modelo se predice los valores de ventas para los próximos 6 meses, estos datos hay que reconstruirlos para la serie inicial aplicándoles la estacionalidad correspondiente a la primera mitad del ciclo, y al no tener tendencia nada más. El resultado se muestra en la figura 12. La línea azul representa los valores de ventas de los 6 próximos meses teniendo en cuenta los datos de ventas de los últimos 6 años.

Figure 12: ARIMA 1,0,2 predicción próximos 6 valores