Estudio de datos temporales con R

Trabajo Fin de Grado

Autor:
Irene Capel Sanz

Tutor:
Fernando Reche Lorite

Grado en Matemáticas

Julio, 2016
Universidad de Almería

Índice general

Abstract in English

The present work tries to do a study on a small scale about the past, current and future situation of a company which is situated in Almería (there are more branches in other cities of Andalusia, though). This company was founded in 1953 and it is dedicated to selling replacement parts for motorbikes, bycicles and cars, mainly. For this study, the period that goes from the year 2012 to 2015 has been chosen as a reference. We have taken into consideration the monthly profits, as well as the profits made by families of products during these years, and we have obtained 29 variables and 48 data for each one. We have suggested a series of goals which are described as follows:

1. To do a descriptive study about some of these variables (families) in which we would include whether there are seasonal variation or not, trend, outliers,...
2. To apply statistical models according to the stationary to obtain more information about each of the families.
3. To do forecasts to the future about the profits of certain variables.

To get all of these objetives, we will use a tool called **RStudio** to work with **R** software which is an enviroment and programming language with an approach to the statistical analysis. Due to the fact that we have time datas (it is a sequence of N observations (data) about one or more variables,ordered and chronological equidistant (normally), in our case monthly), we will have to use a specific language for them, which we will explain briefly in the next section.

Time series are analyzed to understand the past and to predict the future. The time series analysis stand out the best features in the data. These reasons together with the computing potential make the methods to time series applicable in industry, governments and business. Occasionally, it is useful to do simulations to different strategies upon planning possible future decisions.

Resumen en español

El presente trabajo intenta hacer un estudio a pequeña escala de la situación pasada, actual y futura, de una empresa asentada en Almería (aunque con más tiendas por varias ciudades andaluzas). Dicha empresa tuvo sus orígenes en el año 1953 y se dedica a la venta de recambios de motos, bicicletas y automóviles, principalmente.

Para el estudio descrito con anterioridad, se han tomado como años de referencia los comprendidos entre 2012 y 2015, inclusive estos. Se han considerado las ganancias mensuales y por familias de productos de la empresa durante estos años, obteniendo así 29 variables y 48 datos para cada una de ellas.

Hemos planteado una serie de objetivos los cuales se describen a continuación:

1. Hacer un estudio descriptivo de algunas variables (familias) en el que se incluya la observación de estacionalidad o ausencia de la misma, tendencia, outliers,...

2. Aplicar modelos estadísticos según la estacionaridad para obtener más información de cada una de las familias.

3. Predicciones para el futuro sobre las ganancias de determinadas variables.

Para obtener todas estos objetivos, usaremos una herramienta llamada **RStudio** para trabajar con **R** el cuál es un entorno y lenguaje de programación con un enfoque al análisis estadístico. Al ser nuestros datos temporales (secuencia de N observaciones (datos) de una o más variables, ordenadas y equidistantes (normalmente) cronológicamente, en nuestro caso mensualmente), tendremos que usar un lenguaje específico para ellos, el cuál explicaremos brevemente en el siguiente apartado.

Las *series temporales* son analizadas para entender el pasado y predecir el futuro. El análisis de *series temporales* destaca las mayores características en los datos. Estas razones junto con la potencia computacional, hacen que los métodos para series temporales sean aplicables en industria, gobiernos y comercio. En ocasiones, se suele acudir a simulaciones de los datos para diferentes estrategias a la hora de planear posibles decisiones futuras.

1 Introducción

El problema que se nos plantea para el estudio de las series temporales que tratamos en este trabajo, es el poder conocer la evolución de una empresa almeriense dedicada, principalmente, a la venta de bicicletas, motos y recambios de motos, bicicletas y automóviles y conocer todas las características de dichas series. Para ello se consideran los años comprendidos entre el año 2012 y 2015.

Los datos me fueron proporcionados gracias a la colaboración del responsable de la empresa. Estos datos tratan de las ventas de todos sus productos, agrupados en familias, en las fechas descritas anteriormente. Estas familias, que consideraremos como nuestras variables expresan las ventas de:

1. Bicicletas
2. Recambios de bicicletas
3. Recambios de automóvil
4. Carrocería
5. Recambios de motos
6. Recambios Yamaha (origen)
7. Recambios Peugeot (origen)
8. Krafft
9. Baterías de automóvil e industria
10. Material eléctrico automóvil
11. Neumáticos de motos y bicicletas
12. Pintura
13. Aceites
14. Rodamientos
15. Suministros y herramientas
16. Material deportivo
17. Neumáticos de automóvil
18. Suzuki
19. Material eléctrico moto
20. Rieju
21. Microcar

22. Motorhispania
23. Malaguti
24. Citroën
25. Piaggio
26. Recambios quad varios
27. Recambios auto origen
28. Grupo 50
29. Varios

Sin embargo, en este trabajo se estudiarán aquellas variables que se consideren más características y de las que podemos obtener más información.

Los objetivos principales que nos planteamos para cada una de las variables estudiadas son las siguientes:

- Construir un modelo univariante.
- Hacer predicciones.

Para ello utilizaremos la herramienta **RStudio**, que es un lenguaje de programación, que incluye una consola, editor de sintaxis que apoya la ejecución de código, herramientas para el trazado, la depuración y la gestión del espacio de trabajo. Nos permite poder analizar datos con **R** ya que nos proporciona el entorno informático estadístico.

Utilizaremos en el transcurso del trabajo los siguientes paquetes:

- *xts*: es uno de los paquetes de **R** más flexibles para el manejo de datos dependientes del tiempo.
- *lattice*: paquete que es muy útil para describir gráficamente datos multivariantes.
- *psych*: paquete que dispone de funciones muy útiles en el análisis estadístico, siendo una de ella la función `describe` para hacer un análisis descriptivo.
- *forecast*: incluye métodos y herramientas para analizar series univariantes para predicciones incluyendo la función `auto.arima`, que usaremos en el estudio.
- *xtable*: este paquete permite generar el código para tablas, marcos de datos y matrices.

2

Software R

Como hemos dicho anteriormente, **R** es un entorno y lenguaje de programación con un enfoque al análisis estadístico. Fue desarrollado inicialmente por Robert Gentleman (1959) y Ross Ihaka (1954) del Departamento de Estadística de la Universidad de Auckland en 1993 y es una implementación de software libre del lenguaje **S** y es uno de los lenguajes más utilizados en investigación por la comunidad estadística.

R proporciona un amplio abanico de herramientas estadísticas (modelos lineales y no lineales, tests estadísticos, análisis de series temporales, algoritmos de clasificación y agrupamiento,...) además de la posibilidad de diseñar gráficas. También puede usarse como herramienta de cálculo numérico.

A esto se le suma la posibilidad de cargar diferentes bibliotecas o paquetes con funcionalidades de cálculo o graficación. En el repositorio oficial los paquetes se agrupan por naturaleza y función. Por ejemplo hay grupos de paquetes relacionados con estadística bayesiana, econometría, series temporales,...

Es de enorme flexibilidad, guarda los resultados como un «objeto», de tal manera que se puede hacer un análisis sin necesidad de mostrar su resultado inmediatamente. El usuario puede extraer solo aquella parte de los resultados que le interesa. Es un lenguaje *Orientado a Objetos*. Esto significa que las variables, datos, funciones, etc., se guardan en la memoria del ordenador en forma de objetos con un nombre específico. El usuario puede modificar o manipular estos objetos con operadores (aritméticos, lógicos, y comparativos) y funciones (que a su vez son objetos).

Los argumentos pueden ser objetos (como fórmulas, datos,...), algunos de los cuales pueden ser definidos por defecto en la función. Una función en **R** puede carecer de argumentos, ya sea porque todos están definidos por defecto o porque la función no tiene argumentos.

Todas las acciones en **R** se realizan con objetos que son guardados en la memoria activa del ordenador. La lectura y escritura de archivos solo se realiza para la entrada y salida de datos y resultados. El usuario ejecuta funciones gracias a comandos definidos. Los resultados se pueden visualizar en pantalla, guardarlos en el disco o guardar en un objeto. Los resultados, al ser objetos, pueden ser considerados como datos y analizados como tal.

A modo de resumen, **R** es un conjunto de programas integrados para manipular datos, realizar cálculos y construir gráficos y que tiene las siguinetes características:

1. Las acciones se ralizan mediante órdenes.
2. **R** distingue entre mayúsculas y minúsculas.
3. Es un lenguaje basado en funciones.
4. Los argumentos de las funciones aparecen entre paréntesis.
5. El separador entre argumentos es la coma.
6. El separador entre funciones es un salto de línea o un punto y coma.

3 Definiciones y modelos matemáticos

Durante este trabajo se van a utilizar varios conceptos y modelos estadísticos que requiren ser definidos y explicados previamente para la mejor comprensión durante el estudio.

Definición 3.1. *Un proceso estocástico es un conjunto de variables aleatorias $\{X_t\}$ donde el índice t toma valores en un cierto conjunto C. En nuestro caso, este conjunto es ordenado y coresponde a los instantes temporales meses. Para cada valor t del conjunto C (para cada instante temporal) está definida una variable aleatoria, X_t, y los valores observados de la variable aleatoria en distintos instantes forman una serie temporal.*

Definición 3.2. *Una serie temporal se define como una colección de observaciones de una variable recogidas secuencialmente en el tiempo. Estas observaciones se suelen recoger en instantes de tiempo equiespaciados de manera uniforme, así los datos usualmente son dependientes entre sí. Una serie temporal puede ser discreta o continua dependiendo de cómo sean las observaciones.*

Podemos denotar a $\{X_t\}_{t=1,\dots,n}$ como la serie temporal de longitud n tomada en n tiempos.

Algunas de las principales características de una serie temporal son la tendencia y las variaciones estacionales.

La *tendencia* de una serie temporal es la trayectoria a largo plazo de la misma, haciendo abstracción de las fluctuaciones que se producen a intervalos más breves de tiempo. Este movimiento puede ser ascendente, descendente, estable o combinación de éstos, pero siempre ha de observarse un periodo de tiempo muy amplio para poder captar dicha componente.

Si se pueden predecir exactamente los valores de la serie, se dice que la serie es *determinística*. Por otro lado, las series que evolucionan sin seguir aparentemente un patrón en concreto son definidas como aquellas que presentan una tendencia *estocástica* (aleatoria) en su evolución temporal. El futuro sólo se puede determinar de modo parcial por las observaciones pasadas y no se pueden determinar exactamente en el futuro.

Los *ciclos* son movimientos a medio plazo (superior a un año) en torno a la tendencia cuyo período y amplitud pueden presentar cierta regularidad. Refleja comportamientos recurrentes, aunque no tienen por qué ser exactamente periódicos.

Definición 3.3. *El periodo estacional, S, se define como el número de observaciones que forman el ciclo estacional (s=12 para series mensuales, s=4 para series trimestrales,...). Supondremos que el valor de s es fijo en la serie.*

Definición 3.4. *Diremos que una serie es estacional cuando su valor esperado no es constante, pero varía con una pauta cíclica. En concreto, si:*

$$E(X_t) = E(x_{t+s}) \tag{3.1}$$

diremos que la serie tiene estacionalidad de periodo s. Una serie mensual, por ejemplo tiene estacionalidad si los valores esperados en distintos meses del año son distintos, pero el valor esperado en el mismo mes en distintos años es el mismo [4].

Estos tipos de efectos hay que determinarlos y se pueden medir explícitamente o incluso se pueden eliminar del conjunto de los datos, desestacionalizando la serie original.

Definición 3.5. *Se denominan series estacionarias o estrictamente estacionarias a aquellas series cuya tendencia es constante a lo largo del tiempo y que no presentan tendencias obvias o ciclos estacionarios. En ellas,*

- *la media y la varianza (las distribuciones marginales) se mantienen constantes a lo largo del tiempo y la autocovarianza $Cov(X_t, X_s)$ sólo depende del retardo $m = |t - s|$.*
- *la dependencia entre variables sólo depende de sus retardos.*

Ambas condiciones se resumen en que la distribución conjunta de $(X_{t_1}, ..., X_{t_n})$ es la misma a la de $(X_{t_1+m}, ..., X_{t_n+m})$ $\forall t_1, ..., t_n$ y m, siendo m el retardo (esto es, la distribución es invariante si trasladamos las variables en el tiempo).

Pueden presentar además, variaciones accidenteles o irregulares que son fluctuaciones producidas por factores eventuales, esporádicos e imprevisibles, que no muestran una periocidad reconocible.

Definición 3.6. *Una serie temporal es estacionaria en sentido débil si la media y la varianza permanecen constantes en el tiempo y la autocovarianza sólo depende sólo de su separación.*

Definición 3.7. *Un proceso de ruido blanco es un caso simple de proceso estacionario (se denota por ε_t), donde los valores son independientes e idénticamente distribuidos a lo largo del tiempo con*

1. $E[\varepsilon_t] = 0$, $\forall t$.
2. $Var(\varepsilon_t) = \sigma^2$, $\forall t$.
3. $Cov(\varepsilon_t, \varepsilon_{t-m}) = 0$, $\forall k = \pm 1, \pm 2, ...$

La primera condición establece que la esperanza es siempre constante y nula, la segunda condición indica que la varianza es constante y la tercera, que las variables del proceso están incorreladas para todos los retardos.

En este proceso conocer los valores pasados no proporciona ninguna información sobre el futuro, ya que el proceso «no tiene memoria».

En contaposición a las series estacionarias, las *series no estacionarias* son aquellas en las cuales la media y/o variabilidad cambian en el tiempo. Los cambios en la media determinan una tendencia a crecer o decrecer a largo plazo, por lo que la serie no oscila alrededor de un valor constante. En resumen, muchas series no son estacionarias bien porque:

- Presentan ciertas tendencias a lo largo del tiempo (la media crece o baja a lo largo del tiempo).
- Su dispersión no es constante (su varianza no es constante).
- Son estacionales y poseen patrones periódicos.

Definición 3.8. *Un paseo aleatorio suele proporcionar un buen modelo a datos con tendencia estocástica. Así, la serie temporal X_t es un paseo aleatorio si:*

$$X_t = X_{t-1} + \varepsilon_t \tag{3.2}$$

donde $\{\varepsilon_t\}$ es una serie de ruido blanco.

Definición 3.9. *Un proceso es integrado de orden $h \geq 0$, y lo representaremos por $I(h)$, cuando al diferenciarlo h veces se obtiene un proceso estacionario. Un proceso estacionario es, por tanto, siempre $I(0)$.*

Una vez que hemos representado gráficamente nuestra serie, podemos buscar en ella características como tendencias y variaciones estacionales.

Una tarea que podemos pasar a realizar es la descomposición de la serie en diferentes elementos para analizarla desde el punto de vista de sus componentes estructurales.

Descomponer una serie significa separarla en sus componentes, normalmente la tendencia y la componente aleatoria. Además, si la serie es estacional, se puede añadir una componente estacional.

Si el modelo de descomposición es aditivo, podemos formularlo de la forma:

$$X_t = m_t + s_t + z_t \tag{3.3}$$

donde en el tiempo t, X_t es la serie temporal observada, m_t es la tendencia, s_t es el efecto estacional y z_t el término error (o componente aleatoria), que es en general, una secuencia de variables aleatorias correladas con media cero (la serie se muestra como resultado de factores fortuitos que le inciden de forma aislada. Esta componente es de naturaleza aleatoria, ya que su movimientos no vienen definidos es decir son irregulares).

Esta descomposición se basa en métodos elementales: la tendencia se calcula con una media móvil, el efecto estacional se calcula promediando los valores de cada unidad de tiempo para todos los periodos y luego centrando el resultado. Finalmente, los residuos se obtienen restando a la serie observada las dos componentes anteriores.

Cuando el efecto estacional tiende a incrementar a la par que la tendencia, el modelo multiplicativo es el más apropiado:

$$X_t = m_t \cdot s_t \cdot z_t \tag{3.4}$$

En este caso, si el término error tiene un incremento de varianza y la variable es positiva, es indicativo que una transformación logarítmica puede ser la más adecuada [1, 4].

Una vez que hemos identificado cualquier tendencia y efecto estacional, estacionarizamos una serie no estacionaria para eliminar así estas características. A la resultante podremos aplicar los modelos adecuados para su estudio.

1. **Estabilización de la varianza:** Para estabilizar la variabilidad se suelen tomar logaritmos (hace que la dispersión sea más o menos constante a medida que crece la media).

2. **Eliminación de tendencia:** Una forma sencilla de eliminar una tendencia aproximadamente lineal es diferenciar la serie, es decir, consiste en suponer que la

tendencia evoluciona lentamente en el tiempo, de manera que en el instante t la tendencia debe estar próxima a la tendencia en el instante $t-1$. De esta forma, si restamos a cada valor de la serie el valor anterior, la serie resultante estará aproximadamente libre de tendencia. Dicho de otra forma, calcularemos $\nabla X_t = x_t - x_{t-1}$.

3. **Eliminación de estacionalidad:** Para eliminar la estacionalidad de una serie para convertirla en estacionaria, aplicaremos una *diferencia estacional*.

Definición 3.10. *Definimos el operador de retardo, $\boldsymbol{B}$, como un operador lineal que aplicado a una serie temporal, obtendremos la misma serie temporal retardada un período,es decir,*

$$\boldsymbol{B}x_t = x_{t-1} \tag{3.5}$$

A veces es llamado operador lag, aplicándolo repetidamente:

$$\boldsymbol{B}^n x_t = x_{t-n} \tag{3.6}$$

Definición 3.11. *Definimos el operador diferencia estacional de período s como:*

$$\nabla_s = 1 - \boldsymbol{B}^s \tag{3.7}$$

Aplicándolo a una serie temporal, obtenemos otra, cuyo valor en cada instante t es la diferencia entre el valor de la serie original en t, y el valor en $t-s$. Así,

$$\nabla_s X_t = (1 - \boldsymbol{B}^s)X_t = x_t - x_{t-s} \tag{3.8}$$

Una vez hemos transformado la serie en estacionaria para poder aplicar modelos acordes a esa situación, obtendremos las funciones de autocovarianza y autocorrelación.

La función de autocorrelación nos servirá para poder analizar la estacionalidad de una serie. Esta función mide la correlación entre los valores de la serie en un lapso de tiempo m. El correlograma es una representación gráfica de las autocorrelaciones $\rho(m)$, es decir, las correlaciones entre x_t y x_{t+m}, en función de m.

En el gráfico que obtenemos, siempre se tiene que $\rho(0) = 1$. Las líneas discontinuas representan las bandas de confianza de $\rho(m)$ de nivel 95% bajo la hipótesis de que la serie es un ruido blanco (incorrelada).

De igual forma, dada una secuencia temporal de n observaciones $x_1, ..., x_n$, podemos formar $n-1$ parejas de observaciones contiguas $(x_1, x_2), (x_2, x_3), ..., (x_{n-1}, x_n)$ y calcular el coeficiente de correlación de estas parejas. A este coeficiente lo denominaremos coeficiente de autocorrelación de orden 1 y lo denotamos como r_1. Análogamente se pueden formar parejas con puntos separados por una distancia 2, es decir $(x_1, x_3), (x_2, x_4)$,... y calcular el nuevo coeficiente de autocorrelación de orden 2. De forma general, si preparamos parejas con puntos separados una distancia m, calcularemos el coeficiente de autocorrelación de orden m.

Al igual que para el coeficiente de correlación lineal simple, se puede calcular un error estándar y por tanto un intervalo de confianza para el coeficiente de autocorrelación.

Definición 3.12. *Siendo X_t una serie, definimos la función de autocorrelación entre los instantes t y s como:*

$$z = \frac{\gamma(s,t)}{\sqrt{\gamma(s,s)\gamma(t,t)}} \tag{3.9}$$

donde $\gamma(s,t)$ es la covarianza. Este coeficiente mide la capacidad de predecir la serie en un instante t a partir de un valor del instante s.

La función de autocorrelación es el conjunto de coeficientes de autocorrelación r_k desde 1 hasta un máximo que no puede exceder la mitad de los valores observados, y es de gran importancia para estudiar la estacionalidad de la serie, ya que si ésta existe, los valores separados entre sí por intervalos iguales al periodo estacional deben estar correlacionados de alguna forma. Es decir que el coeficiente de autocorrelación para un retardo igual al periodo estacional debe ser significativamente diferente de 0.

Relacionada con la función de autocorrelación nos encontramos con la función de autocorrelación parcial. En el coeficiente de autocorrelación parcial de orden m, se calcula la correlación entre parejas de valores separados esa distancia pero eliminando el efecto debido a la correlación producida por retardos anteriores a m.

Una perturbacin transitoria sobre una variable estacionaria tiene efectos puramente transitorios; pueden durar varios períodos, pero sus efectos terminan desapareciendo. Los valores sucesivos de su función de auocorrelación convergen rápidamente hacia cero, excepto quizá en los retardos de carácter estacional. La serie temporal correspondiente a una variable estacionaria no deambula durante períodos largos de tiempo a un mismo lado de su media muestral, sino que cruza frecuentemente dicho nivel medio.

Por el contrario, una perturbación de carácter transitorio sobre una variable no estacionaria tiene efectos permanentes. La función de autocorrelación de una variable no estacionaria converge a cero muy lentamente, y su serie temporal muestra claramente largos períodos de tiempo en que deambula sin cruzar su nivel medio.

A continuación, vamos a iniciar el estudio de **modelos de procesos estacionarios** que son útiles para representar la dependencia de los valores de una serie temporal de su pasado. Los modelos más simples son los autorregresivos, que generalizan la idea de regresión para representar la dependencia lineal entre dos variables aleatorias.

Procesos autorregresivos

El primer proceso estocástico lineal, son los modelos autorregresivos de orden p conocidos como AR(p). Estos modelos parten del supuesto de que el valor presente de la serie X_t, se explica en función de p valores previos, así $X_{t-1}, X_{t-2}, ..., X_{t-p}$ siendo p el número de retardos necesarios para pronosticar $\hat{X}_t$.

El proceso general de AR(p) se puede modelizar bajo la siguiente ecuación:

$$X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + ... + \alpha_p X_{t-p} + \varepsilon_t \tag{3.10}$$

donde $X_t, X_{t-1}, ...$ representan las variables aleatorias concebidas como realizaciones de un proceso estocástico en los momentos de tiempo $t, t-1, ...$, ε_t es ruido blanco y los α_i son los parámetros del modelo con $\alpha_p \neq 0$ para un proceso de orden p. La ecuación (1.5) puede ser expresada como un polinomio de orden p en términos del operador de retardo:

$$\theta_p(\mathbf{B})X_t = (1 - \alpha_1 \mathbf{B} - \alpha_2 \mathbf{B}^2 - ... - \alpha_p \mathbf{B}^p)X_t = \varepsilon_t \tag{3.11}$$

Una predicción en el tiempo t viene dada por:

$$\hat{X}_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + ... + \alpha_p X_{t-p} \tag{3.12}$$

Procesos autorregresivos estacionarios y no estacionarios

La ecuación $\theta_p(\mathbf{B}) = 0$, donde $\mathbf{B}$ es tratada como un número (real o complejo), es llamada ecuación característica. Las raíces de esta ecuación (la (1.6)) deben exceder, sin excepción, de la unidad en valor absoluto para que el proceso sea estacionario. Notar que el paseo aleatorio tiene $\theta = 1 - \mathbf{B}$ con raíz $\mathbf{B} = 1$ (el paseo aleatorio es el caso especial AR(1) con $\alpha_1 = 1$) es no estacionario.

Por ejemplo, si aplicamos una estructura de dependecia a las observaciones contiguas en una serie temporal, X_t, y tomamos $y_t = x_t$ y $z_t = x_{t-1}$, obtenemos el modelo de dependecia que llamamos proceso autorregresivo de primer orden y que denotamos como AR(1). Es un proceso donde el valor presente de la serie sólo depende de forma lineal del último valor observado. El proceso AR(1) viene dado por:

$$X_t = \alpha x_{t-1} + \varepsilon_t + c \tag{3.13}$$

donde $\{\varepsilon_t\}$ es una serie de ruido blanco con media cero y varianza σ^2. También, c y α son constantes a determinar y tenemos que $-1 < \alpha < 1$, lo cuál es necesario para que el proceso sea estacionario. Utilizando la notación del operador retardo, $\mathbf{B}$, tenemos que:

$$\mathbf{B}x_t = x_{t-1} \tag{3.14}$$

En este caso, si el parámetro AR es positivo ($\alpha > 0$), la dependecia lineal del presente de los valores pasados es siempre positiva, mientras que si el parámetro es negativo, esta dependecia es positiva para los retardos pares y negativa para los impares. En el caso de un modelo AR(1) con $\alpha_1 = 1$, no es estacionario.

En general, en un proceso AR(p) las autocorrelaciones del correlograma parcial son nulas después del retardo p. Estos procesos se caracterizan por tener muchos coeficientes de autocorrelación distintos de cero y que decrecen con el retardo. El valor actual está correlado con todos los anteriores, aunque con coeficientes decrecientes y por ellos son procesos con memoria relativamente larga.

Procesos de media móvil

Una familia de procesos que tienen memoria muy corta, son la de media móvil, o procesos MA. Estos procesos son función de un número finito, y generalmente pequeño de las innovaciones pasadas.

Los procesos MA(q) son aquellos cuyo valor actual depende de las q últimas innovaciones (los ε_t). Su fórmula general dada la serie X_t es:

$$X_t = \varepsilon_t + \beta_1 \varepsilon_{t-1} + \ldots + \beta_q \varepsilon_{t-q} \tag{3.15}$$

donde $\{\varepsilon_t\}$ es ruido blanco con media cero y varianza σ_{ε^2}. Si lo queremos representar utilizando el operador retardo,

$$X_t = (1 + \beta_1 \mathbf{B} + \beta_2 \mathbf{B}^2 + \ldots + \beta_q \mathbf{B}^q)\varepsilon_t = \phi_q(B)\varepsilon_t \tag{3.16}$$

donde ϕ_q es un polinomio de orden q. Como los procesos MA consisten en una suma finita de términos estacionarios de ruido blanco, son siempre estacionarios y tienen una media y autovarianza invariantes en el tiempo. El proceso es *invertible* si las raíces del operador $\phi_q = 0$ son, en módulo, mayores que la unidad.

Procesos ARMA

Combinando las propiedades de los procesos AR(p) y MA(q), podemos definir los procesos ARMA(p,q), dando lugar a una familia de procesos estocásticos estacionarios. Matemáticamente, los procesos ARMA resultan de añadir estructura MA a un proceso AR o viceversa. En estos modelos supondremos siempre que no hay raíces comunes en los operadores AR y MA.

La fórmula general de un modelo autorregresivo de media móvil de orden (p,q), ARMA(p,q), queda:

$$X_t = \alpha_1 x_{t-1} + \alpha_2 x_{t-2} + \ldots + \alpha_p x_{t-p} + \varepsilon_t + \beta_1 \varepsilon_{t-1} + \beta_2 \varepsilon_{t-2} + \ldots + \beta_q \varepsilon_{t-q} \tag{3.17}$$

donde $\{\varepsilon_t\}$ es un proceso de ruido blanco. En términos del operador retardo,

$$\theta_p(B) X_t = \phi_q(B) \varepsilon_t \tag{3.18}$$

Como características de este modelo, tenemos:

1. El proceso es estacionario cuando todas las raíces de θ exceden la unidad en valor absoluto.
2. El proceso es invertible cuando todas las raíces de ϕ exceden la unidad en valor absoluto.
3. El modelo AR(p) es el caso especial de ARMA(p,0).
4. El modelo MA(q) es el caso especial de ARMA(0,q).
5. Un modelo ARMA será siempre más eficiente que AR o MA por requerir menos parámetros.

Las autocorrelaciones disminuyen geométricamente, y se hacen prácticamente cero a los pocos retardos.

Una vez examinados los modelos disponibles para series estacionarias, nos queda conocer cuáles son los modelos no estacionarios para series temporales, los modelos ARIMA y SARIMA.

Procesos ARIMA

Primero extendemos el modelo paseo aleatorio para incluir autorregresividad y términos de media móvil. Como las series diferenciadas necesitan ser «integradas» para recuperar la serie original, el proceso estocástico es el llamado autorregresivo integrado de media móvil.

Los modelos ARIMA si tenemos una serie no estacionaria, como primer paso tenemos que diferenciarla hasta que obtengamos una serie estacionaria. Si tenemos que diferenciarla d veces para obtener una serie estacionaria, entonces tenemos un modelo ARIMA(p,d,q), donde d es el orden de integración del proceso (es el número de raíces unitarias), p es el orden de la parte autorregresiva estacionaria y q es el orden de la parte media móvil.

La fórmula general del ARIMA(p,d,q) es:

$$(1-\alpha_1\mathbf{B}-...-\alpha_p\mathbf{B}^p)(1-\mathbf{B})^d X_t = c+(1-\beta_1\mathbf{B}-...-\beta_q\mathbf{B}^q)\varepsilon_t \tag{3.19}$$

En términos del operador diferencia:

$$\theta_p(\mathbf{B})\bigtriangledown^d X_t = \phi_q(\mathbf{B})\varepsilon_t \tag{3.20}$$

donde $\bigtriangledown = (1-\mathbf{B})$ y θ_p y ϕ_q son polinomios de órdenes p y q respectivamente.

Procesos ARIMA estacionales

Finalizamos este apartado con el modelo SARIMA o ARIMA estacional. Al proceso ARIMA le podemos incluir términos estacionales de manera multiplicativa para dar lugar a un proceso no estacionario estacional para dar lugar al proceso SARIMA. Surge del hecho de que podemos convertir series no estacionarias en estacionarias tomando diferencias regulares (entre períodos consecutivos) y podemos eliminar la estacionalidad mediante diferencias estacionales.

Lo podemos expresar como $ARIMA(p,d,q)(P,D,Q)_m$ donde la primera parte corresponde con la parte no estacional del modelo, y la segunda, con la estacional. Además m es el número de periodos por estación. Esta serie es estacionaria y la función de autocorrelación será siempre nula salvo en los retardos estacionales $s, 2s, 3s, ..., Qs$ donde Q es el orden del modelo.

En términos del operador retardo:

$$\Theta_P(\mathbf{B}^s)\theta_p(\mathbf{B})(1-\mathbf{B}^s)^D(1-\mathbf{B})^d X_t = \Phi_Q(\mathbf{B}^s)\phi_q(\mathbf{B})\varepsilon_t \tag{3.21}$$

donde Θ_P, θ_p, Φ_Q y ϕ_q son polinomios de órdenes P,p,Q y q respectivamente. En general, el modelo es no estacionario salvo que $D = d = 0$ y todas las raíces de la ecuación característica excedan la unidad en valor absoluto.

La función de autocorrelación simple de este proceso es una mezcla de las funciones de autocorrelación correspondientes a las partes regular y estacional.

4 Análisis de las series temporales

4.1 Introducción de los datos en R

Ilustraremos el análisis de series temporales con los datos que nos ha proporcionado la empresa de sus ventas mensuales desde 2012 hasta 2015.

Importamos los datos del fichero `trabajo_TFG_nuevo.csv` en un `data frame` llamado `datos1`. Este fichero consta de 28 variables y 48 datos. En código,

```
datos.origen <- read.table("trabajo_TFG_nuevo.csv", header=F,sep=";",
                           dec=",")
datos1 <- t(datos.origen[,-1])
datos1 <- as.data.frame(datos1)
colnames(datos1) <- datos.origen[,1]
```

Para realizar un estudio descriptivo clásico de series temporales, separando el componente estacional, el componente tendencia-ciclo así como los residuos, es necesario constituir una serie cronológica, especificando en qué unidad de tiempo empezamos y con qué frecuencia tenemos observaciones.

En nuestro caso, empezamos en el año 2012 y al tratarse de datos mensuales, en cada unidad de tiempo (años) tenemos 12 datos. Utilizaremos la instrucción `ts` (time series), para definir el objeto serie temporal. Así,

```
datos <- ts(datos1, frequency=12, start=2012)
```

Entre los paquetes de **R** más flexibles para el manejo de datos dependientes del tiempo es `xts`. Por tanto, pasamos nuestro dato a este formato ya que para algunos casos, nos será de más utilidad. Para ello, primero instalamos el paquete `xts` en nuestro programa.

```
library(xts)
datos.xts <- as.xts(datos)
```

Si quisiéramos obtener más información acerca de los datos, podemos escribir la función `summary`, la cuál permite calcular algunas medidas descriptivas elementales de todas las variables del fichero simultáneamente (tales como el valor mínimo, máximo, media,...).Por ejemplo, si queremos obtener una estadística descriptiva básica de las 3 primeras variables, lo haremos de la siguiente forma:

```
summary(datos[,1:3])

##    Bicicletas    Recambios bicicletas Recambios automovil
##  Min.   :15902   Min.   : 4745        Min.   :30289
##  1st Qu.:28410   1st Qu.: 7085        1st Qu.:41836
##  Median :33812   Median : 8313        Median :45012
##  Mean   :36123   Mean   : 8749        Mean   :44642
```

```
##  3rd Qu.:45632   3rd Qu.:10363        3rd Qu.:46340
##  Max.   :68076   Max.   :14848        Max.   :59005
```

Finalmente, podemos calcular algunos datos numéricos globales, como por ejemplo, las medias anuales y cuatrimestrales para después graficarlas si deseamos. Para obtener estos resultados, nos será de utilidad las datos en formato `xts` ya que así podremos usar las funciones `apply.year` y `apply.quarterly` (son funciones del paquete `xts` para aplicar funciones a periodos concretos).

```
medias.anuales <- apply.yearly(datos.xts,FUN=mean)
medias.cuatrimestre <- apply.quarterly(datos.xts,FUN=mean)
medias.anuales[,"Bicicletas"]

##           Bicicletas
## dic. 2012   41109.36
## dic. 2013   37866.15
## dic. 2014   32780.42
## dic. 2015   32737.62
```

Otra función que es útil si queremos hacer un resumen estadístico más amplio que el que suministra la función básica `summary`, es la función `describe` del paquete `psych`. Como ejemplo podemos ver una estadística descriptiva de las ventas mensuales de 15 variables:

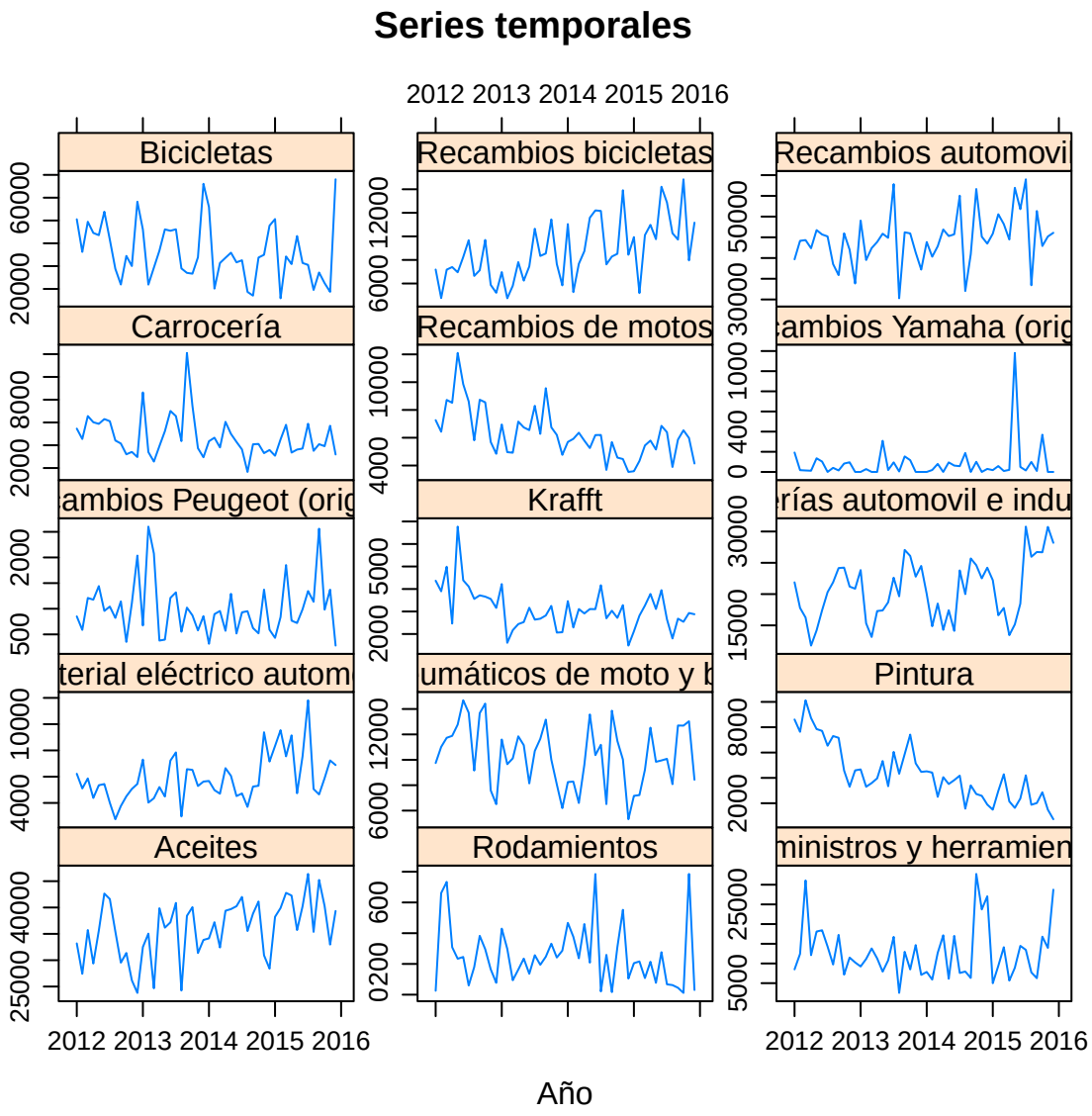
		n	$\overline{x}$	σ	$se = \sigma/\sqrt{n}$
Bicicletas	1	48	36123.39	12662.49	1827.67
Recambios bicicletas	2	48	8748.89	2557.95	369.21
Recambios automovil	3	48	44641.91	6306.94	910.33
Carrocería	4	48	4707.00	1782.00	257.21
Recambios de motos	5	48	6283.49	1769.41	255.39
Recambios Yamaha (origen)	6	48	81.23	180.38	26.04
Recambios Peugeot (origen)	7	48	1001.87	531.39	76.70
Krafft	8	48	3129.40	931.61	134.47
Baterías automovil e industria	9	48	20688.64	4793.11	691.83
Material eléctrico automovil	10	48	5871.11	1800.35	259.86
Neumáticos de moto y bici	11	48	10316.28	2425.24	350.05
Pintura	12	48	4352.65	2227.47	321.51
Aceites	13	48	35931.73	5554.12	801.67
Rodamientos	14	48	253.70	196.90	28.42
Suministros y herramientas	15	48	12798.80	6774.98	977.88

Cuadro 4.1: Estadística descriptiva

Es necesario instalar el paquete `lattice` para realizar gráficos ya que es un paquete de visualización de datos. El gráfico `xyplot` se utiliza para representar un gráfico bivariante pero en este caso nos permite ver simultáneamente el comportamiento de varias variables en su sencuencia temporal.

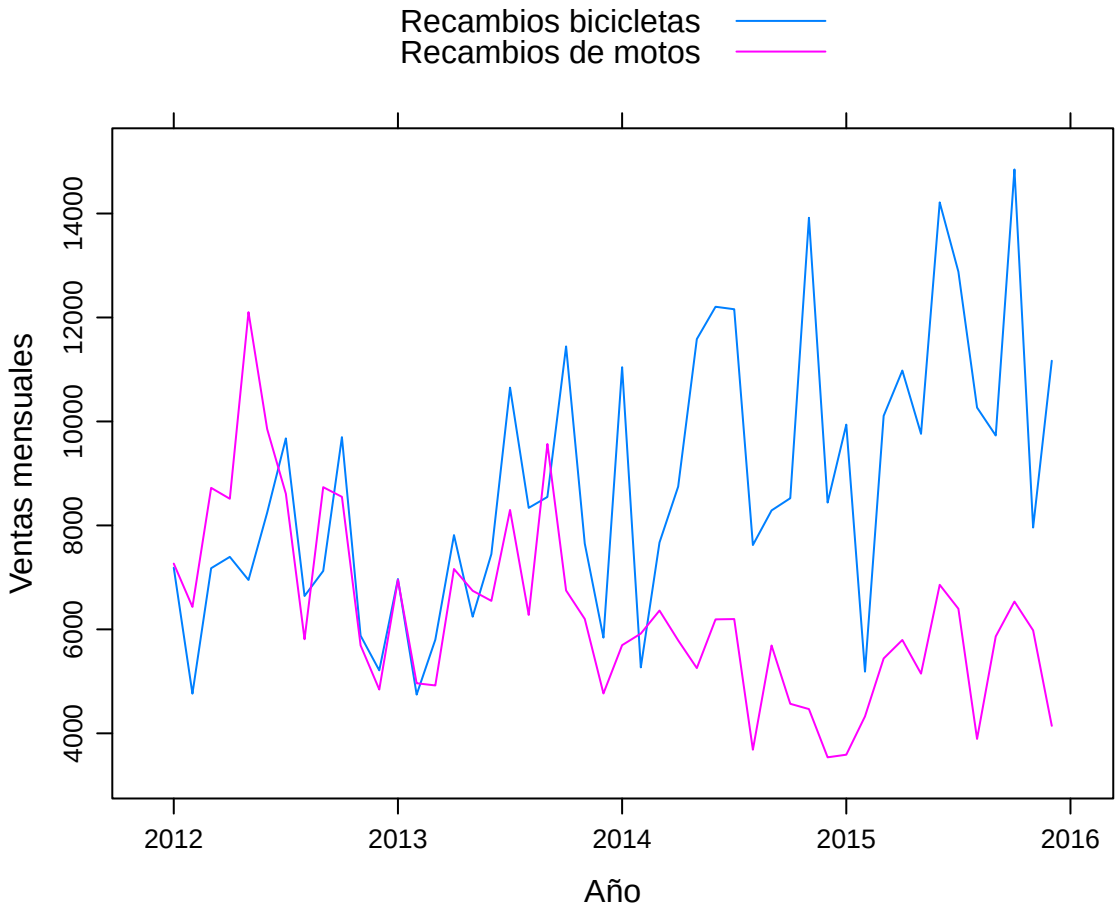
Apoyándonos en este paquete ya instalado, graficaremos todas las variables en un solo gráfico, a modo de visión general de las mismas. El argumento `main` se utiliza para poner un título al gráfico que queremos representar.

```
library(lattice)
xyplot(datos[,1:15], main="Series temporales",xlab="Año")
```



Además, si disponemos de variables con escalas similares que puedan ser comparadas, podemos elaborar gráficos superpuestos. Por ejemplo, podemos comparar como se han comportado la venta de recambios de bicicletas y de motos:

```
xyplot(datos[,c(2,5)],superpose=T,xlab="Año",ylab="Ventas mensuales")
```



Este tipo de gráficos son muy ilustrativos y nos proporcionan bastante información sobre la evolución temporal de nuestras variables. Además nos permiten una visión en conjunto del problema.

4.2 Análisis de las variables

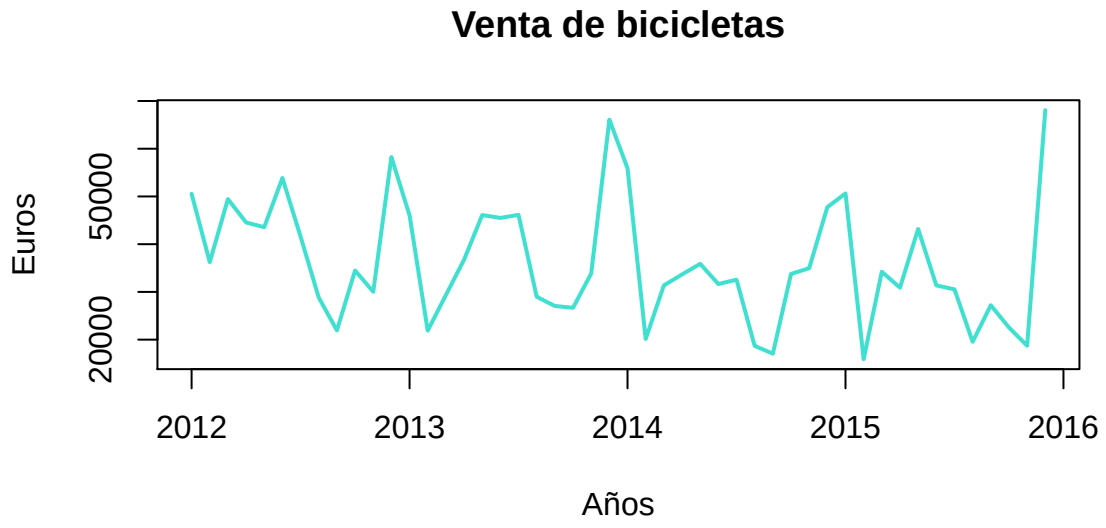
En esta sección se estudiarán las variables más significativas del total. Se les aplicará un análisis descriptivo para concretar sus principales características y se les aplicará después el corespondiente modelo dependiendo de ellas.

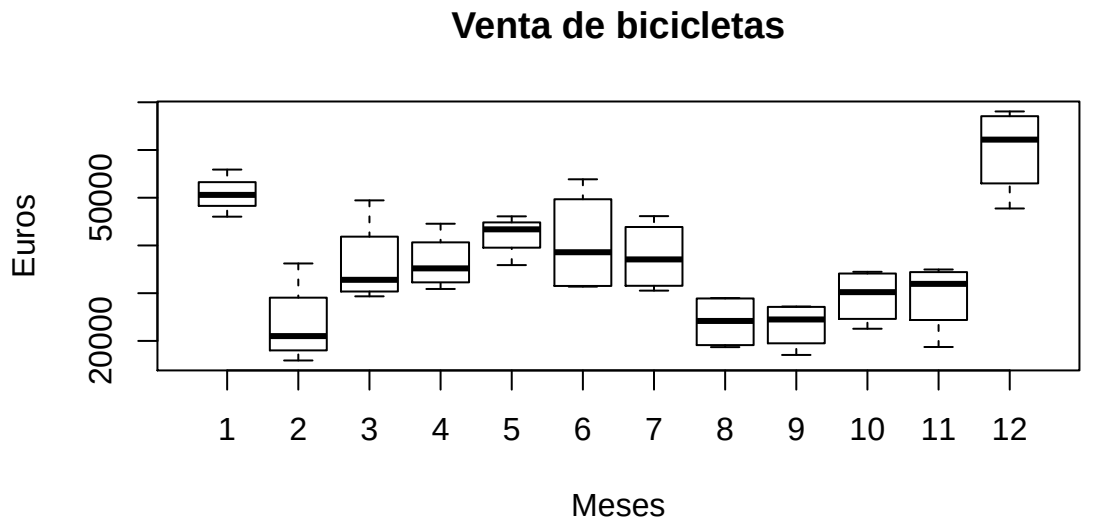
Bicicletas

Una vez que hayamos leido los datos en **R**, el siguiente paso normalmente es hacer un gráfico de los datos de la serie temporal, el cuál lo podemos hacer con la función `plot`. En primer lugar lo haremos para la variable bicicletas, la cuál (como todas las demás) consta de 48 datos.

Usaremos pues, el siguiente código para mostrar el gráfico de la serie temporal:

```
par(mfrow=c(2,1),mar=c(4,4,4,2),oma=c(0,0,0,0))
plot(datos[,1], xlab="Años", ylab="Euros",
     main="Venta de bicicletas", col="turquoise", lwd=2)
boxplot(datos[,1]~cycle(datos[,1]), main="Venta de bicicletas",
        xlab="Meses", ylab="Euros")
```





El primer componente de la función `plot`, `datos[,1]`, denota a la variable bicicletas, que es el primer dato del fichero. Los argumentos `xlab` e `ylab` nos permiten escribir las etiquetas de los ejes. El argumento `col` nos da la opción de dar color a nuestra serie temporal, y con `lwd` controlamos el grosor de la línea.

A simple vista, se puede observar cómo parece haber una componente estacional ya que se intuye un patrón anual en Navidad mayormente, y en verano. También puede apreciarse una tendencia ligeramente decreciente.

El siguiente paso, si queremos comparar la distribución de las ventas para cada mes, haremos un gráfico Box-Plot. En él, podemos ver las variaciones estacionales, es decir, si existen meses que en comparación con otros, resultan ser más productivos en cuanto a ventas. El argumento `cycle` extrae las estaciones para cada dato.

En este gráfico podemos ver cómo se cumplen nuestras sospechas y se aprecia perfectmanente como sobre todo en Navidad (Enero y Diciembre), las ventas son muy superiores a las que hay el resto del año. Por otro lado, se aprecia también, aunque más ligeramente, que en Junio las ventas son algo superiores que en los 9 meses restantes. Parece haber pues, estacionalidad en nuestra serie temporal.

Si queremos hacer un análisis descriptivo de nuestra variable, podemos usar el `psych` junto con la función `describe`. No es necesario que volvamos a incluir el paquete, pues en la sección anterior ya lo cargamos. Así,

		n	$\overline{x}$	σ	min	max	rango	$se = \sigma/\sqrt{n}$	NA	NA	
Bicicletas	1	48	36123.39	12662.49	33811.57	35380.56	14178.34	15902.41	68076.15	52173.74	18

Cuadro 4.2: Estadística descriptiva

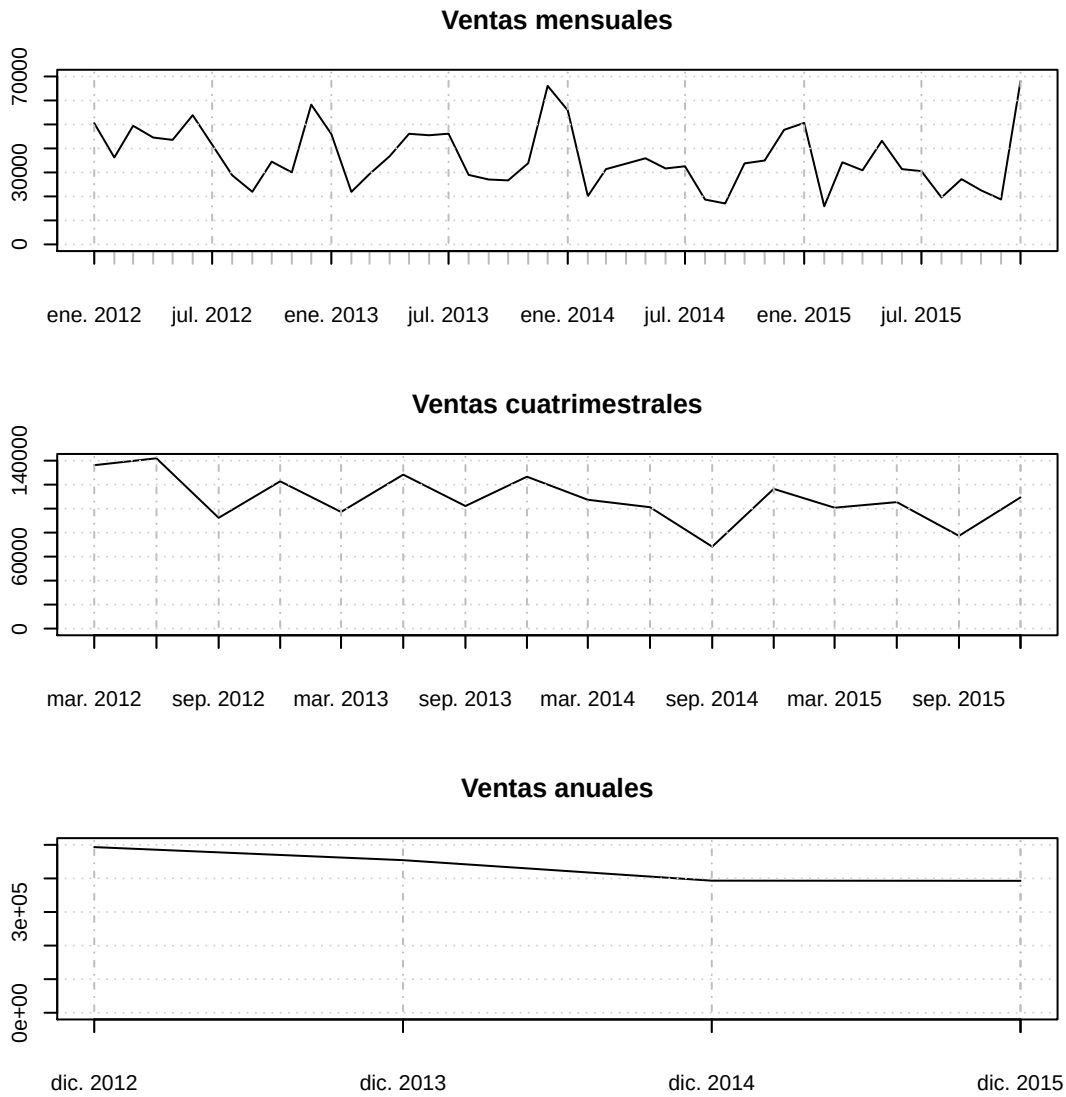
En este análisis vemos cómo esta variable posee 48 observaciones, la media de ventas de bicicletas es de 36123.39 euros, la desviación estándar es 12662.49 (promedio de las desviaciones individuales de cada observación con respecto a la media de una distribución. Mide el grado de dispersión o variabilidad) y la mediana es la cantidad de 33811.57 euros. Los valores superior e inferior en ventas son respectivamente, 68076.15 y 15902.41 euros.

Veamos ahora unos gráficos que aunque no son comparativos entre sí, nos pueden dar algo más de información acerca de la venta de biciletas. Para ello, vamos a calcular en primer lugar, las ventas cuatrimestrales y anuales de la variable.

```
datos.cuatrimestrales.bici <- apply.quarterly(datos.xts[,1],FUN=sum)
datos.anuales.bici <- apply.yearly(datos.xts[,1],FUN=sum)
datos.anuales.bici

##           Bicicletas
## dic. 2012   493312.3
## dic. 2013   454393.8
## dic. 2014   393365.1
## dic. 2015   392851.4

datos.cuatrimestrales.bici

##           Bicicletas
## mar. 2012  136229.60
## jun. 2012  141975.05
## sep. 2012   92343.39
## dic. 2012  122764.30
## mar. 2013   97259.80
## jun. 2013  128342.70
## sep. 2013  102163.84
## dic. 2013  126627.45
## mar. 2014  107389.77
## jun. 2014  101189.98
## sep. 2014   68312.88
```

```
## dic. 2014  116472.45
## mar. 2015  100771.39
## jun. 2015  105444.90
## sep. 2015   77288.96
## dic. 2015  109346.14
```

```
par(mfrow=c(3,1),mar=c(4,4,4,2),oma=c(0,0,0,0))
plot(datos.xts[,1],main="Ventas mensuales",ylim=c(0,70000))
plot(datos.cuatrimestrales.bici,main="Ventas cuatrimestrales",
     ylim=c(0,140000))
plot(datos.anuales.bici,main="Ventas anuales",ylim=c(0,500000))
```



La primera línea de este código, la escribimos para que se nos permita en un mismo renglón, visualizar las tres gráficas. En los tres gráficos, hemos añadido el argumento `ylim` para especificar hasta qué valor queremos que se nos muestre la variable (ventas mínimas y máximas).

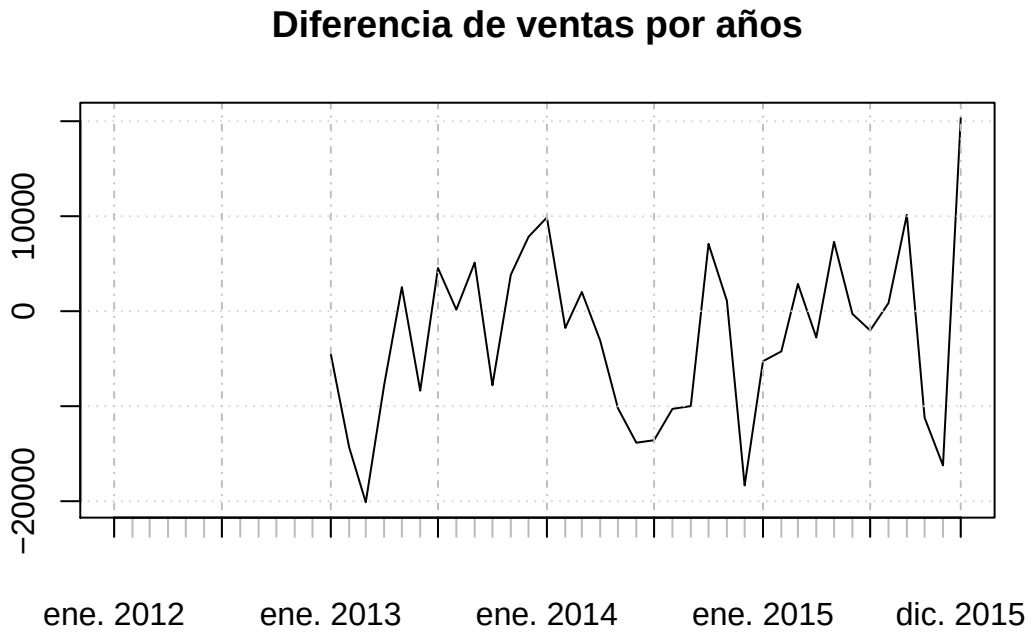
Volvemos a apreciar una ligera tendencia decreciente en los tres gráficos. Sin embargo no lo consideraremos tendencia pues una diferencia de unos 100000 euros con ganancias de 500000 euros, no es una tendencia considerable.

Si nos interesara tener algún dato cuantitativo, podemos por ejemplo calcular todas las ventas de bicicletas del año 2012 con la primera línea de código. Si queremos concretar algo más, podemos hallar lo mismo pero de un mes en concreto. Lo haremos por ejemplo, como muestra la segunda línea, del mes de Enero de 2012. También podemos selecionar periodos concretos como de junio de 2012 a marzo de 2013.

```
datos.xts['2012',1]

##          Bicicletas
## ene. 2012   50553.71
## feb. 2012   36226.87
## mar. 2012   49449.02
## abr. 2012   44548.58
## may. 2012   43562.18
## jun. 2012   53864.29
## jul. 2012   41597.10
## ago. 2012   28812.56
## sep. 2012   21933.73
## oct. 2012   34474.76
## nov. 2012   30034.11
## dic. 2012   58255.43

datos.xts['2012-01',1]

##          Bicicletas
## ene. 2012   50553.71

datos.xts['2012-06/2013-03',1]

##          Bicicletas
## jun. 2012   53864.29
## jul. 2012   41597.10
## ago. 2012   28812.56
## sep. 2012   21933.73
## oct. 2012   34474.76
## nov. 2012   30034.11
## dic. 2012   58255.43
## ene. 2013   46029.06
## feb. 2013   21894.25
## mar. 2013   29336.49
```

Con esto, se nos puede ocurrir querer comparar las ventas en varios años diferentes. Podemos entonces usar la función `diff` para obtener los resultados. Si escribimos,

```r
plot(diff(datos.xts[,1],lag=12), main="Diferencia de ventas por años")
```



Obtenemos la diferencia de ventas de una año respecto a otro. Por eso escribimos el argumento `lag` igual a 12 (anual). Para verlo de una manera un poco más clara, lo haremos de manera cuantitativa, con los dos primeros años:

```r
diff(datos.xts[1:24,1],lag=12)
```

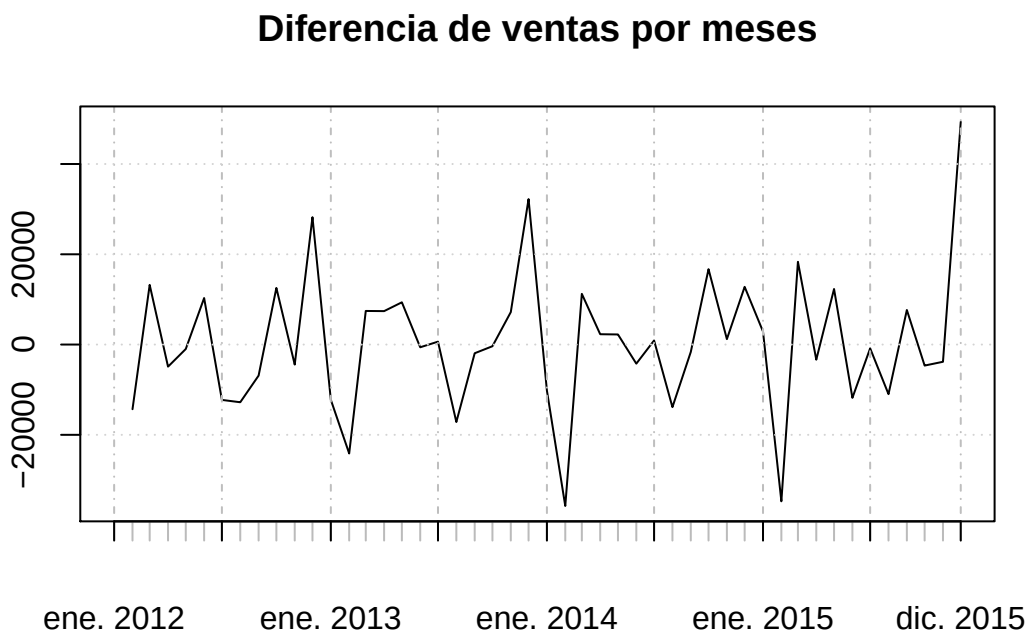
```
##           Bicicletas
## ene. 2012         NA
## feb. 2012         NA
## mar. 2012         NA
## abr. 2012         NA
## may. 2012         NA
## jun. 2012         NA
## jul. 2012         NA
## ago. 2012         NA
## sep. 2012         NA
## oct. 2012         NA
## nov. 2012         NA
## dic. 2012         NA
## ene. 2013   -4524.65
## feb. 2013  -14332.62
## mar. 2013  -20112.53
## abr. 2013   -7805.10
## may. 2013    2537.91
## jun. 2013   -8365.16
```

```
## jul. 2013    4542.86
## ago. 2013     163.35
## sep. 2013    5114.24
## oct. 2013   -7799.62
## nov. 2013    3827.55
## dic. 2013    7835.22
```

El año 2012 nos aparece como **NA** ya que lo que estamos calculando es la diferencia, como hemos dicho, anual y así por ejemplo los resultados que aparecen en los meses del año 2013, son la diferencia de ventas del año 2012 con 2013. Así por ejemplo, aparece en enero de 2013, -4524,65. Esto es porque en Enero de 2013, se obtuvieron 4524,65 euros menos de beneficio que en Enero de 2012. Podemos por tanto hacernos una idea de cómo han ido las ventas evolucionando a lo largo de estos cuatro años.

De la misma manera podemos ahora escribir,

```
diff(datos.xts[,1])
plot(diff(datos.xts[,1]), main="Diferencia de ventas por meses")
```

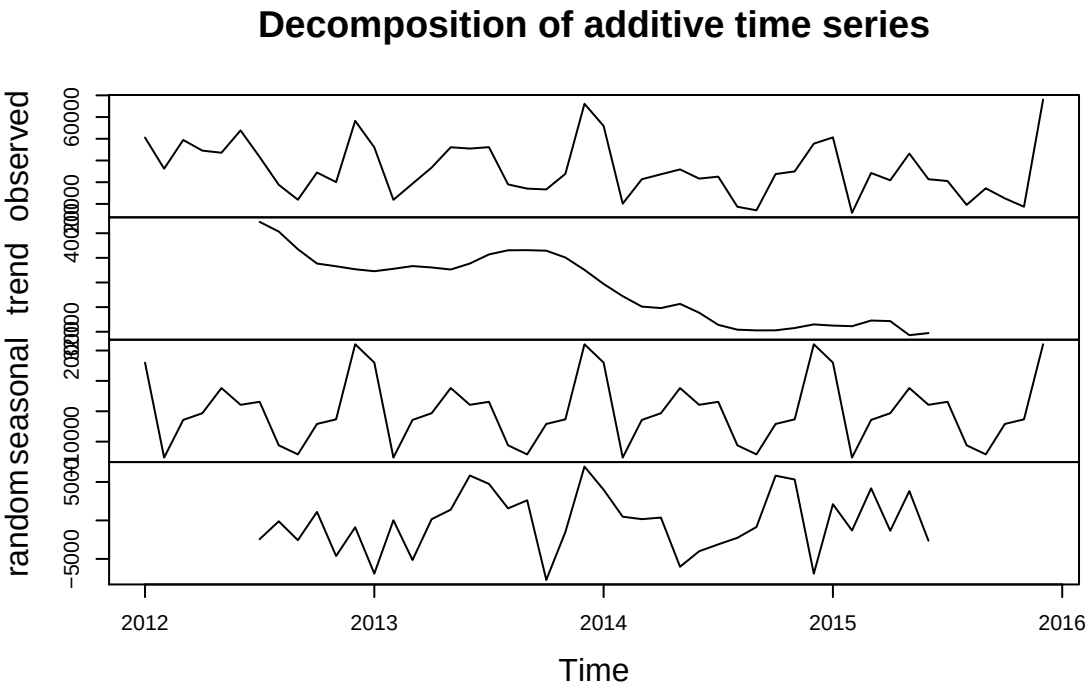


donde al no especificar el argumento `lag`, la función tiene como valor por defecto 1 y lo que nos calcula entonces son las diferencias mensuales. Esto es, obtenemos por ejemplo, como pirmer dato, la diferencia del mes de Enero de 2012 con Febrero del mismo año, obteniendo que en Enero de 2012 tenemos **NA** y en Febrero de 2012, -14326,84. Vuelve a ser porque en Febrero se ingresaron 14326,84 euros menos que el mes anterior.

Como se puede apreciar en el primer gráfico que hemos hecho, debido a que el tamaño de las fluctuaciones estacionales no varían demasiado de un año para otro y no

se aprecia una tendencia fuerte, parece conveniente describir la serie con un modelo aditivo. Escribiremos entonces la siguiente línea de código:

```
plot(decompose(datos[,1]))
```

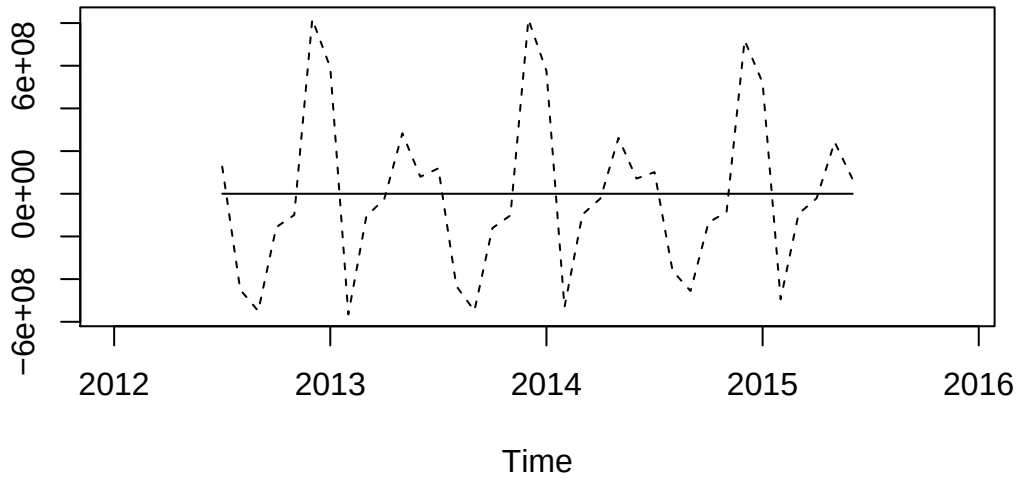


En **R**, la función `decompose` estima la tendencia y la componente estacional de una serie temporal. La función `plot`, produce una única figura donde se muestra la serie original, x_t, y las series descompuestas m_t, s_t y z_t. Como se dijo anteriormente, m_t representa la tendencia, ligeramente decreciente, s_t es la componente estacional (se aprecia que hay dos diferentes) y finalmente, z_t que corresponde a la componente aleatoria. Esta última es una estimación debido a que es obtenida de la serie original usando estimaciones de la tendencia y los efectos estacionales. Esta estimación de la realización de un proceso aleatorio es error residual. Sin embargo, nosotros lo tratamos como la realización de un proceso aleatorio.

En la gráfica podemos observar, en la componente `trend`, que los valores decrecen ligeramente de un valor de algo más de 40000 a estar ligeramente por debajo de los 32000.

Otro gráfico interesante que podemos mostar, es el siguiente:

```
Bicis.decom=decompose(datos[,1])
Trend=Bicis.decom$trend
Seasonal=Bicis.decom$seasonal
ts.plot(cbind(Trend,Trend*Seasonal),lty=1:2)
```



Este último gráfico usa `lty` para poder apreciar dos tipos de líneas, como son la superposición de la tendencia (línea continua) con la variación estacional (línea discontinua). Vemos así como esta serie no posee tendencia alguna.

Al no haber tendencia ni diferencias en la variabilidad, parece que esta serie de venta de bicicletas es una serie estacionaria. La media y la variabilidad se mantienen bastante estables a lo largo del tiempo. Es importante saberlo pues una serie estacionaria siempre será preferible en un estudio frente a una no estacionaria por tres razones principalmente:

1. Con series estacionarias podemos obtener predicciones fácilmente.

2. Como la media es constante, podemos estimarla con todos los datos, y utilizar este valor para predecir una nueva observación.

3. También se pueden obtener intervalos de predicción (confianza) para las predicciones asumiendo que x_t sigue una distribución conocida, por ejemplo, la normal.

Los valores estimados de la estacionalidad, la tendencia y la componente irregular son almacenadas en las variables `Bicis.decom$seasonal`, `Bicis.decom$trend` y `Bicis.decom$random`.

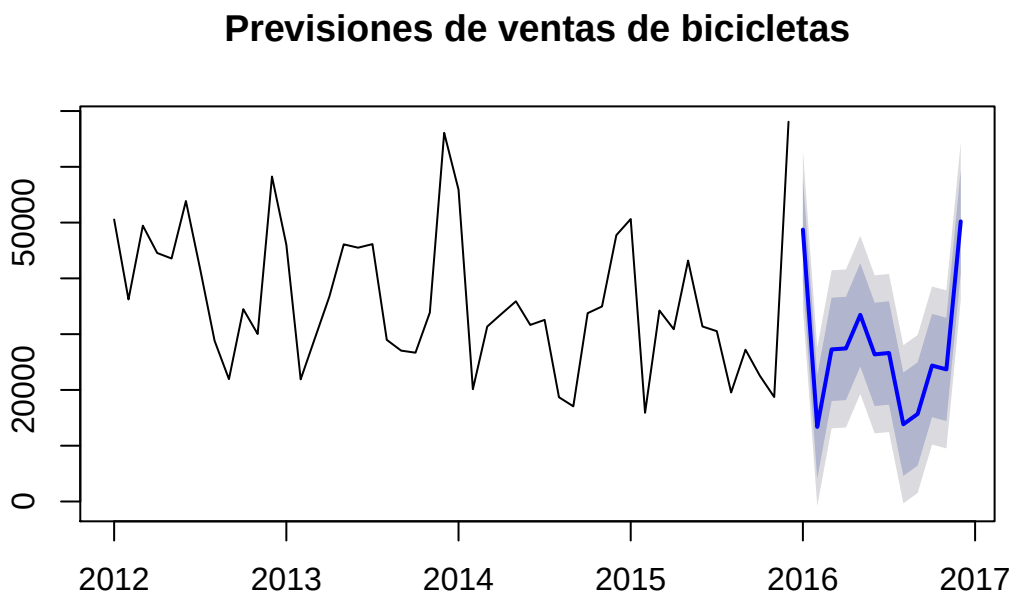
Estos valores vienen dados de Enero a Diciembre y son los mismos cada año. El valor mayor corresponde al mes de Diciembre (22067,9347) y el menor a Febrero (-15239,6321). Esto nos quiere decir que el mayor pico de ventas es en el mes de Diciembre y el menor en Febrero y esto es una situación constante cada año.

Para poder estimar un modelo ARIMA(p,d,q) que explique nuestra serie temporal, necesitamos transformarla en estacionaria (en este caso ya es estacionaria), es decir, tenemos que eliminar la varianza, la tendencia y la componente estacional. En nuestro caso, al tener una serie temporal estacional que puede ser descrita usando un modelo aditivo, podemos ajustar estacionalmente la serie estimando la componente estacional y eliminándola de la serie original.

Sin embargo, nosotros utilizaremos la función `auto.arima` para encontrar ese modelo. Esta función nos convierte sin necesidad de que diferenciemos la serie manualmente, el mejor ARIMA(p,d,q) para nuestro modelo, diferenciándola automáticamente.

Por tanto, usaremos esta función que nos ayudará a hacer un ajuste predictivo de la serie para después poder graficar nuestra predicción.

```
library(forecast)
(ajuste <- auto.arima(datos[,1]))

## Series: datos[, 1]
## ARIMA(0,0,0)(1,1,0)[12] with drift
##
## Coefficients:
##          sar1      drift
##       -0.6241  -265.4843
## s.e.   0.1354    69.8947
##
## sigma^2 estimated as 52234216:  log likelihood=-372.9
## AIC=751.79   AICc=752.54   BIC=756.54

plot(forecast(ajuste,h=12),main="Previsiones de ventas de bicicletas")
```



Podemos en este punto ver varias cosas importantes.

En primer lugar, si nos fijamos en el gráfico, vemos que con bastante detalle, nos ha dibujado una predicción para el año siguiente, 2016 (ya que hemos especificado `h=12`).

Vemos que aparece una línea central en la predicción con más grosor, y unas bandas a ambos lados de color más claro. Estas bandas son el intervalo de predicción donde

las ventas se pueden mover en el año 2016 (con más probabilidad cuanto más cerca se está de la línea central de la predicción, con el 80 % y el 95 %).

Por otro lado hemos obtenido un resultado numérico. Vemos cómo nos aparece que esta serie se comporta según un modelo SARIMA o ARIMA estacional.

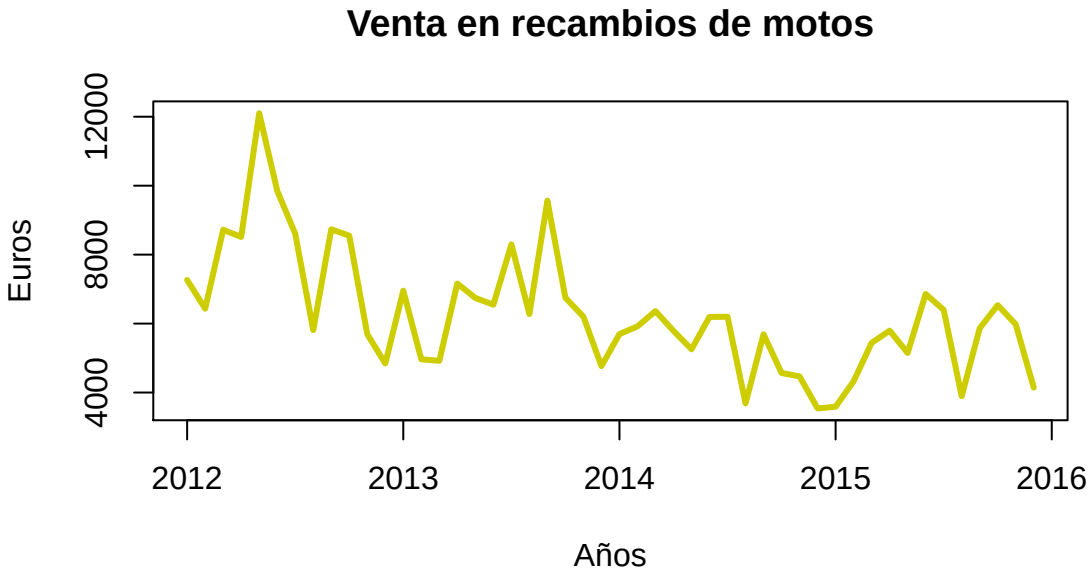
Que $d = 0$ quiere decir que no hemos tenido que diferenciar la serie pues ya era estacionaria y los meses no tienen relación con los inmediatamente anteriores a él (el modelo no tiene relación con el mes anterior, pues obtenemos un ARIMA(0,0,0)).

Por otro lado, el que $D = 1$, la diferenciación con respecto al año anterior, lo que hacer es eliminar la estacionalidad de la serie (la parte estacional, $(1,1,0)_{12}$, indica que es un AR(1) respecto del año anterior).

En este caso, hemos obtenido un modelo $ARIMA(0,0,0)(1,1,0)_{12}$. El primer paréntesis indica la parte no estacional de la serie y el segundo, la estacional. Por otro lado, el primer paréntesis nos indica la relación con los meses inmediatamente anteriores (en esta caso, no tienen relación) y la segunda parte, la relación con el mes del año anterior (estacionalidad, en un año en nuestro caso).

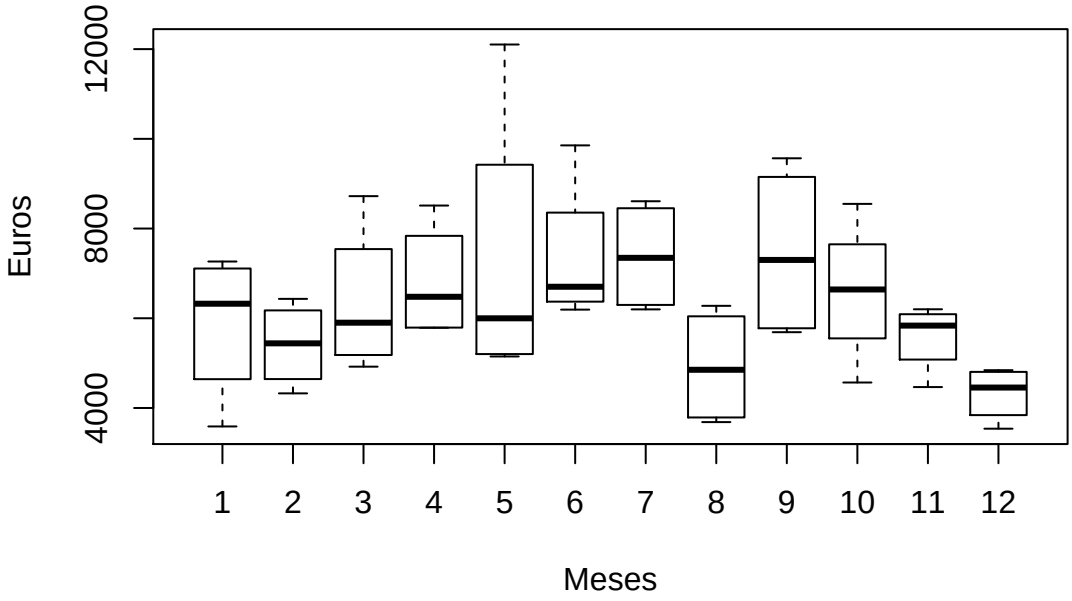
Recambios de motos

Como en el caso anterior, nuestro primer paso será representar el gráfico de la variable «Recambios de motos». Su representación gráfica es:



A simple vista, podríamos decir que esta serie parece que es no estacionaria por varias características apreciables. En primer lugar, esta serie no es estable en el tiempo ya que no oscila alrededor de ningún nivel fijo (tiene un nivel variable en el tiempo). Es no lineal pues combina periodos de crecimiento con otros de decrecimiento. Además, vemos que muestra una tendencia no lineal decreciente y es un ejemplo de serie estacional. Se observa que algunos meses tienen sistemáticamente más ventas que otros. Por ejemplo, los meses de antes y después de verano son más productivos. Además hay variabilidad.

Nuestro siguiente paso será hacer el gráfico de cajas donde podremos observar qué meses son respecto a los otros, los más rentables en cuanto a ganancias en recambios de motos.



Podemos confirmar que los meses más representativos son Mayo y Septiembre. Además podemos ver como en el mes de Diciembre, las ventas caen desde unos 12000 euros a estar cercano a 5000 euros, siendo uno de los meses con menos ganancias.

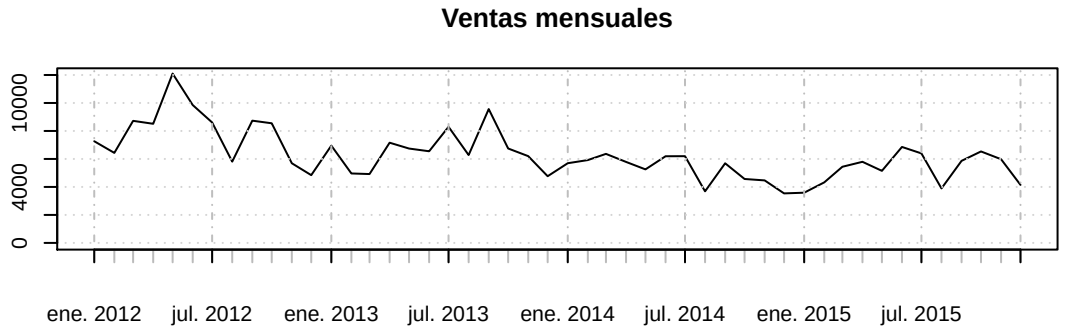
Para hacer un análisis descriptivo de la serie de ventas en recambios de motos,

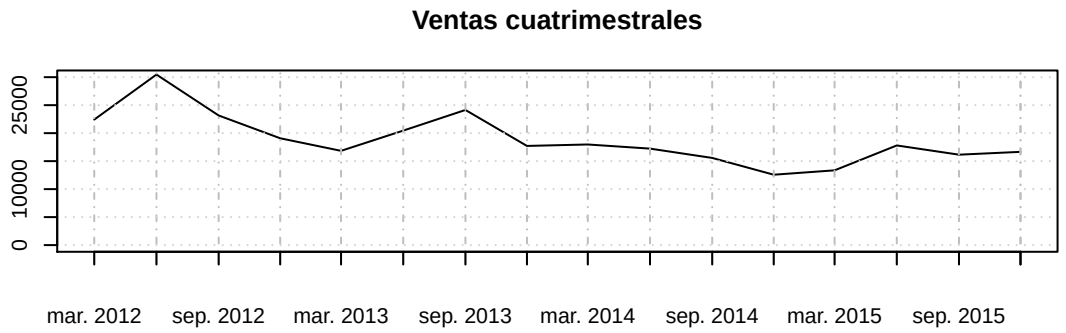
		n	$\overline{x}$	σ	min	max	rango	$se = \sigma/\sqrt{n}$	NA	NA	NA
Recambios de motos	1	48	6283.49	1769.41	6086.94	6166.09	1337.56	3538.57	12102.79	8564.22	255.39

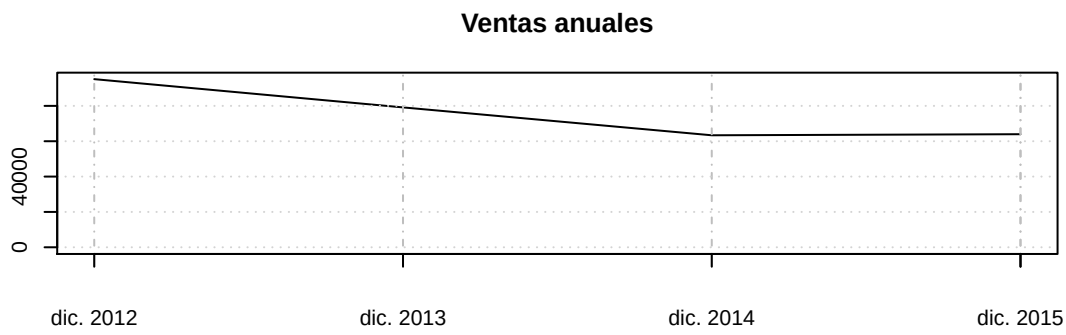
Cuadro 4.3: Estadística descriptiva

Podemos apreciar cómo esta variable posee 48 observaciones, la venta media es de 6283.49 euros, la desviación estándar es 1769.41 y la mediana es 6086.94 euros. Los valores superior e inferior en ventas son respectivamente, 12102.79 y 3538.57 euros.

Veamos cómo quedan las gráficas mensuales, cuatrimestrales y anuales de nuestra serie temporal.



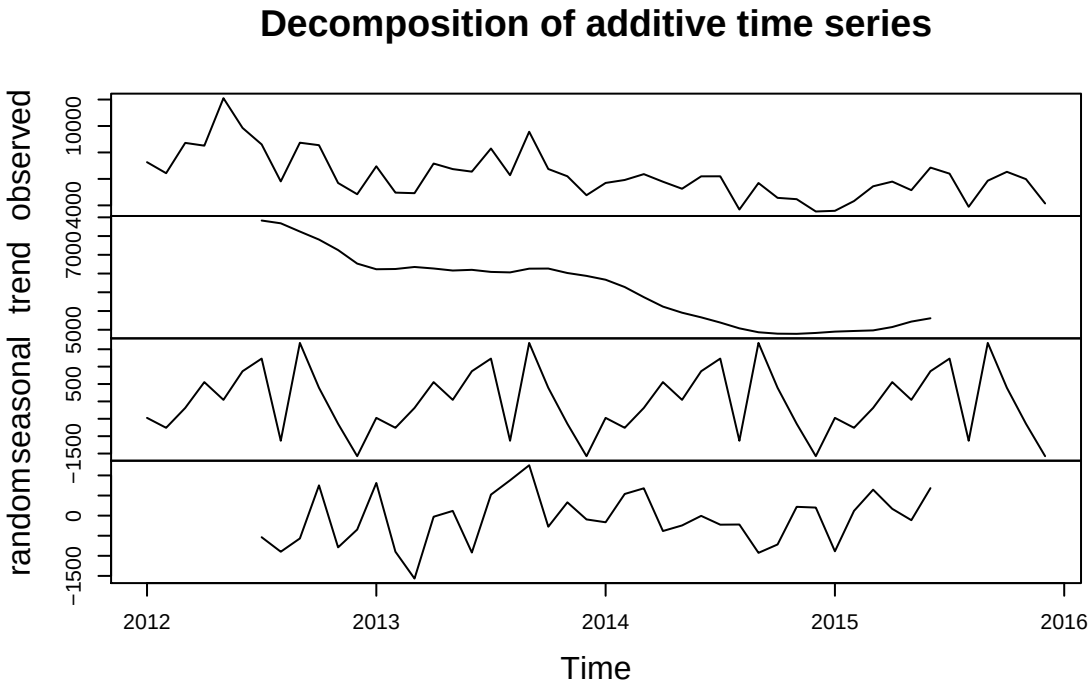




En el gráfico de ventas anuales podemos ver cómo las ventas caen de los 95000 euros a algo más de 60000 a finales del año 2014. A partir de ahí, parece que las ventas empiezan a ascender muy ligeramente.

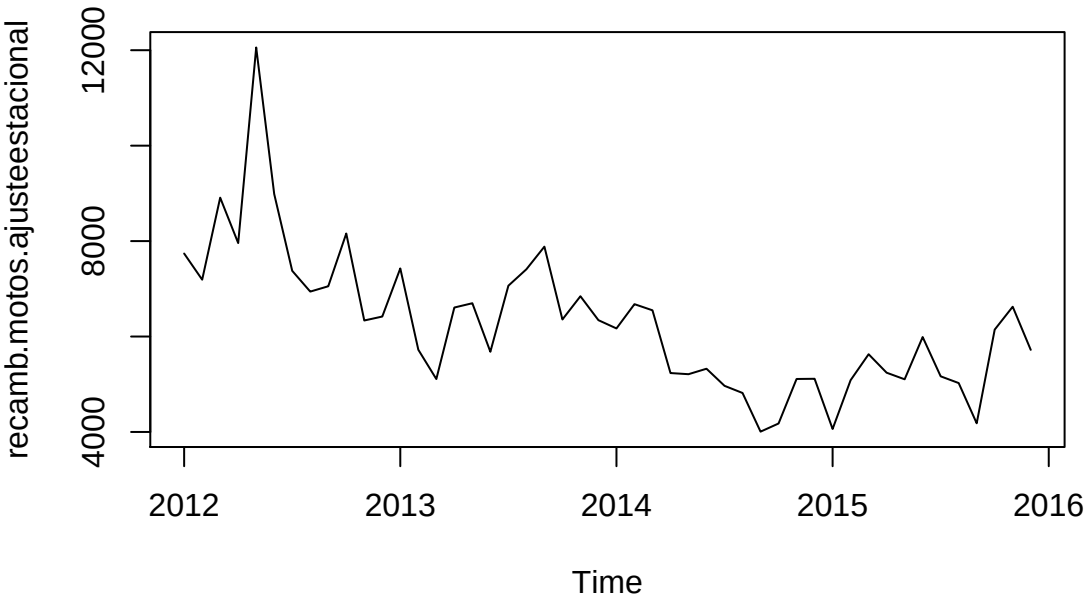
En el gráfico de las ventas cuatrimestrales, vemos que las ventas son irregulares en el tiempo ya que hay épocas de creciemiento con otras de decrecimiento. Por ejemplo, hasta Abril de 2012 las ventas crecen para después decrecer hasta mediados de 2013, y así a lo largo del tiempo.

Al igual que hicimos anteriormente veamos como quedaría gráficamente la descomposición de la serie.



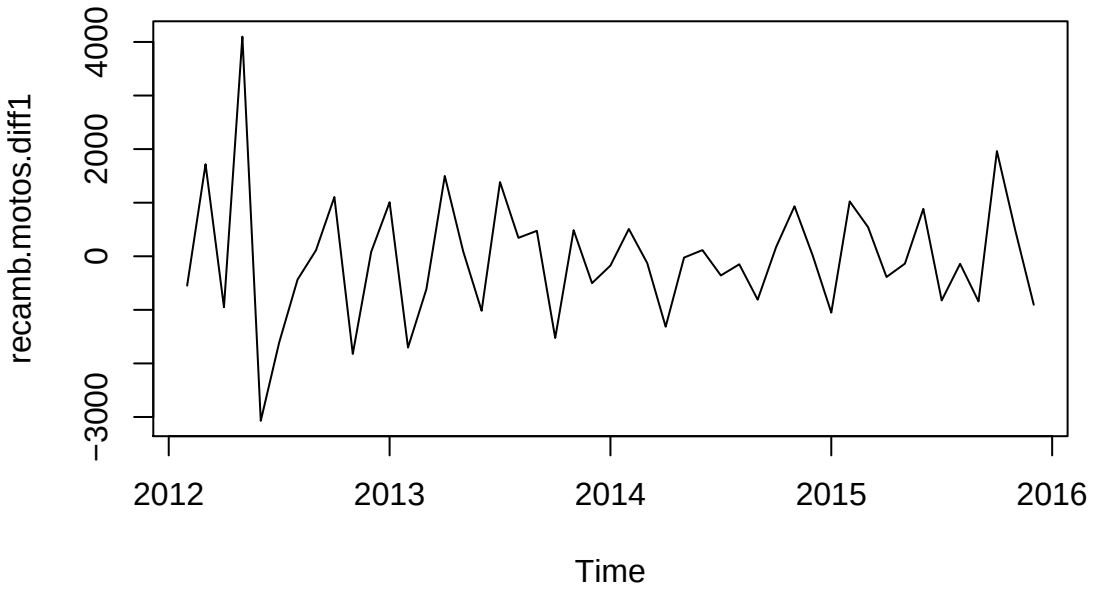
De esta manera podemos visualizar cada una de las componentes por separado. Vemos como por ejemplo la tendencia, desciende desde un valor cercano a 8000, a 5000 aunque parece que al final, asciende ligeramente hasta alcanzar los 5500.

Si tenemos una serie temporal estacional que puede ser descrita usando un modelo aditivo, podemos ajustar estacionalmente la serie estimando la componente estacional y quitándola después de la serie original. Si realizamos ese proceso y quitamos la componente estacional de la serie original obtenemos:



Podemos ver como la variación estacional ha sido eliminada de la serie temporal. Este ajuste estacional ahora sólo contiene las componente tendencia e irregular.

Si diferenciamos una vez la serie temporal no estacionaria para transformarla en estacionaria y lo aplicamos a la serie desestacionalizada obtenemos:

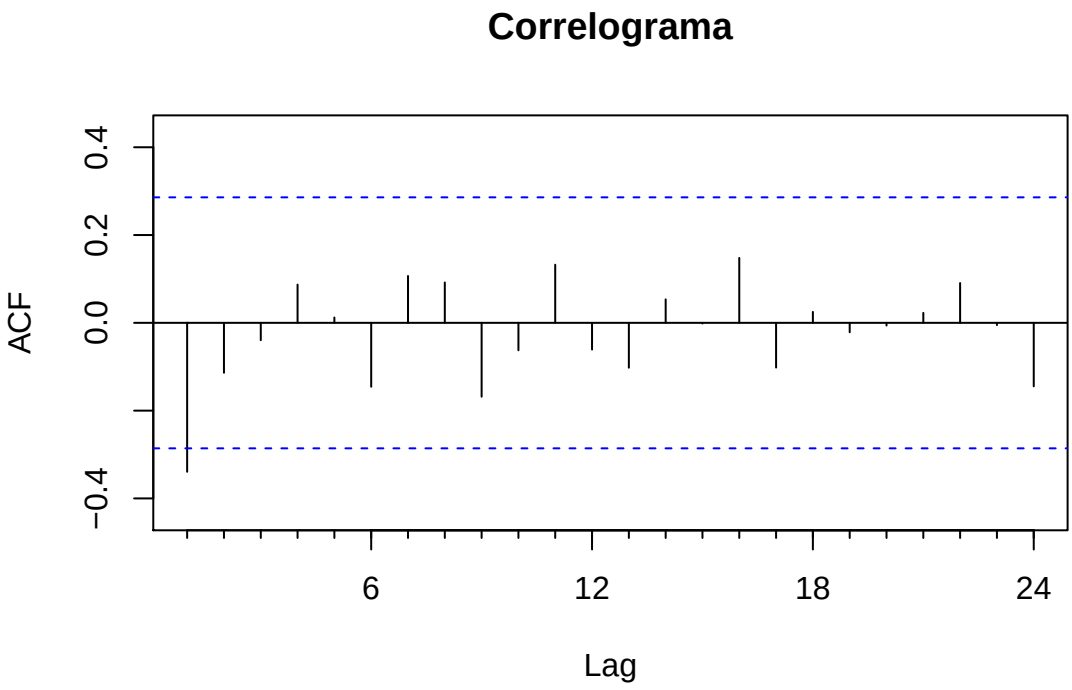


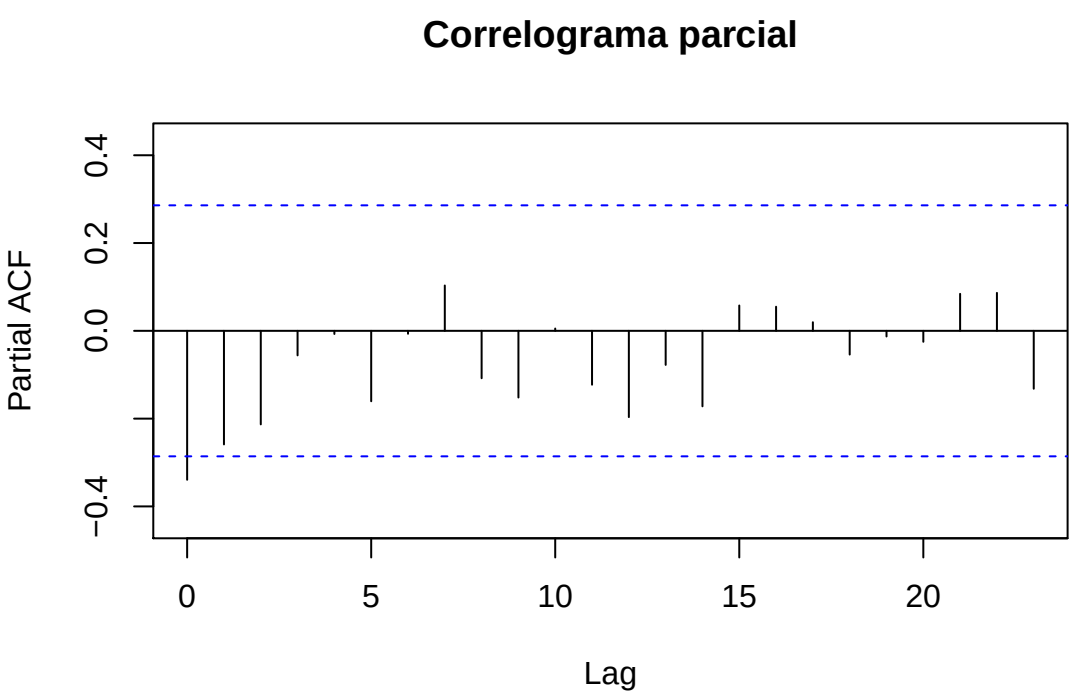
La serie con la primera diferencia parece ya ser estacionaria en media y varianza. El nivel de la serie permanece bastante constante a lo largo del tiempo, y la varianza también. En consecuencia, parece que necesitamos diferenciar la serie una sola vez para poder conseguir una serie estacionaria.

Por todo esto, en nuestra serie de ventas en recambios de motos, el orden de diferenciación (d) es 1. Nuestra serie es por tanto integrada de orden 1, $I(1)$. Esto significa que podemos usar un modelo ARIMA(p,1,q) para nuestra serie, ya que parece ser el más adecuado. El siguiente paso es descifrar los valores de p y q para nuestro modelo ARIMA.

Por tomar la primera diferenciación, hemos eliminado la componente tendencia y ya sólo nos queda la componente irregular (pues previamente hemos eliminado la estacionalidad). Podríamos por tanto examinar si hay correlaciones entre los términos sucesivos de esta componente irregular. Si es así, esto podría ayudarnos a realizar un modelo predictivo para nuestra serie temporal.

Seleccionemos ahora, el mejor modelo ARIMA para la serie. Esto significa encontrar los mejor valores e p y q para nuestro modelo ARIMA(p,1,q). Para hacer esto, observaremos el correlograma y el correlograma parcial de la serie estacionaria.



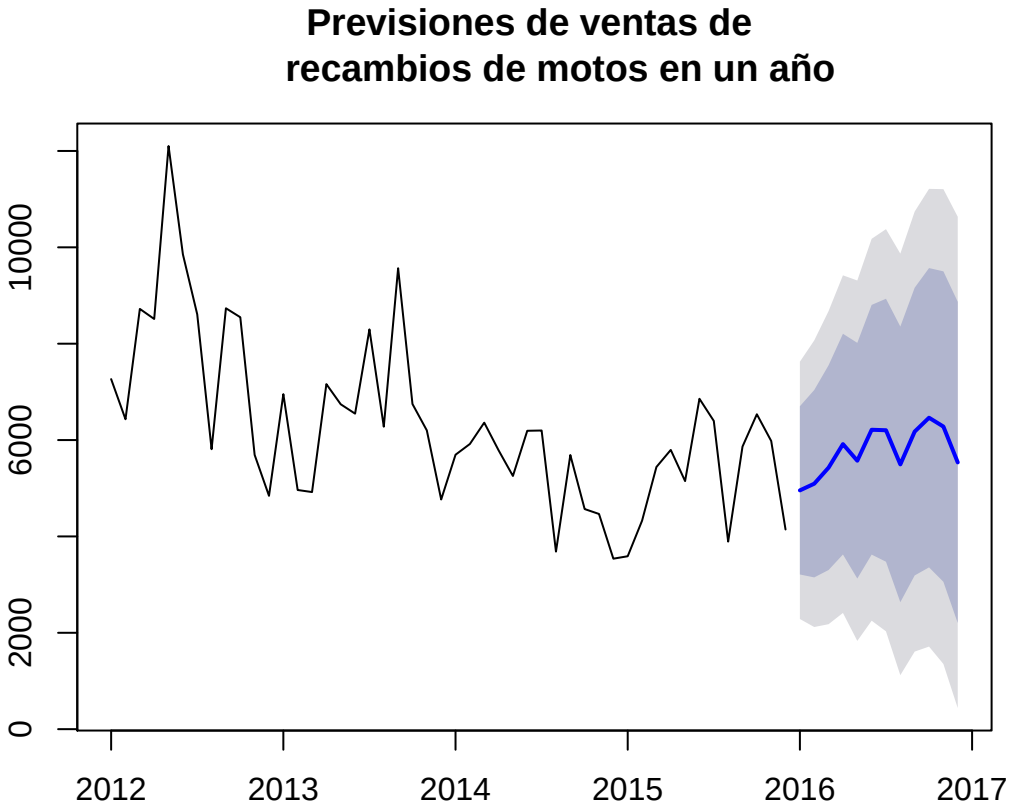


Podemos ver en el correlograma cómo la autocorrelación en lag=1 excede el límite de confianza tomando un valor de -0.339. Por tanto, esto sugiere un término $MA(1)$ no estacional. Como además, hemos tenido que diferenciar una vez y la serie ya no es estacional, el mejor modelo ARIMA que podemos encontrar para nuestra serie es ARIMA(0,1,1). En este caso no se trata de un modelo SARIMA, pues hemos eliminado la parte estacional de la variable.

Si lo queremos hacer con la función `auto.arima` y además encontrar una predicción

de la serie, obtenemos:

```
## Series: datos[, 5]
## ARIMA(0,1,1)(0,0,1)[12]
##
## Coefficients:
##           ma1    sma1
##       -0.5093  0.4548
## s.e.   0.1587  0.1935
##
## sigma^2 estimated as 1853477:  log likelihood=-406.37
## AIC=818.74   AICc=819.3   BIC=824.29
```



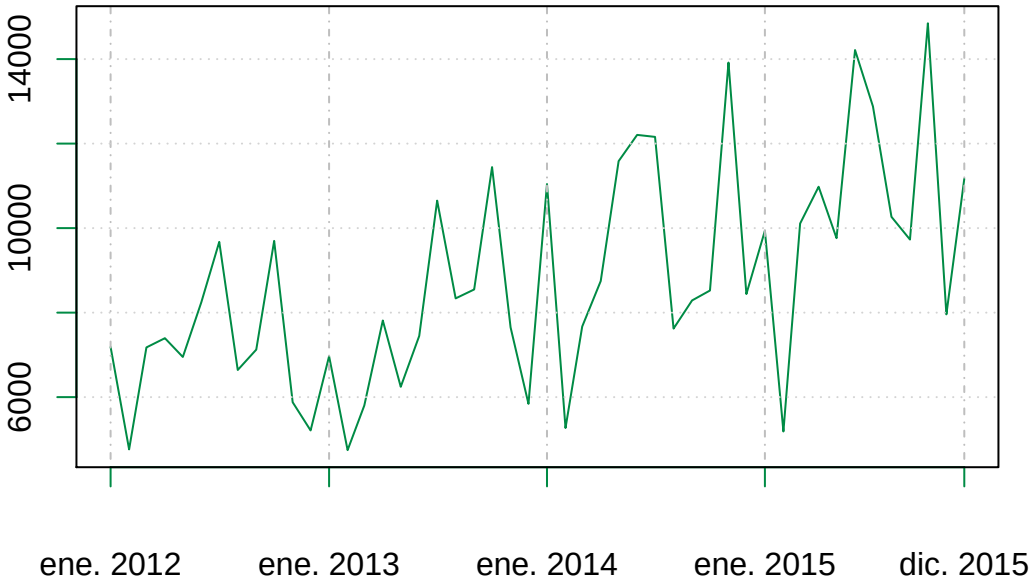
No tiene por qué coincidir el modelo que nosotros hemos creido más adecuado para la variable ya que usando la función, se nos muestra el más adecuado. En este caso si nos aparece un modelo SARIMA pues con la función utilizada, no es necesario que la convirtamos en estacionaria y eliminemos las componentes estacionales. En este resultado podemos ver por ejemplo,que la serie es estacionaria y la hemos tenido que diferenciar una vez, pues $d = 1$.

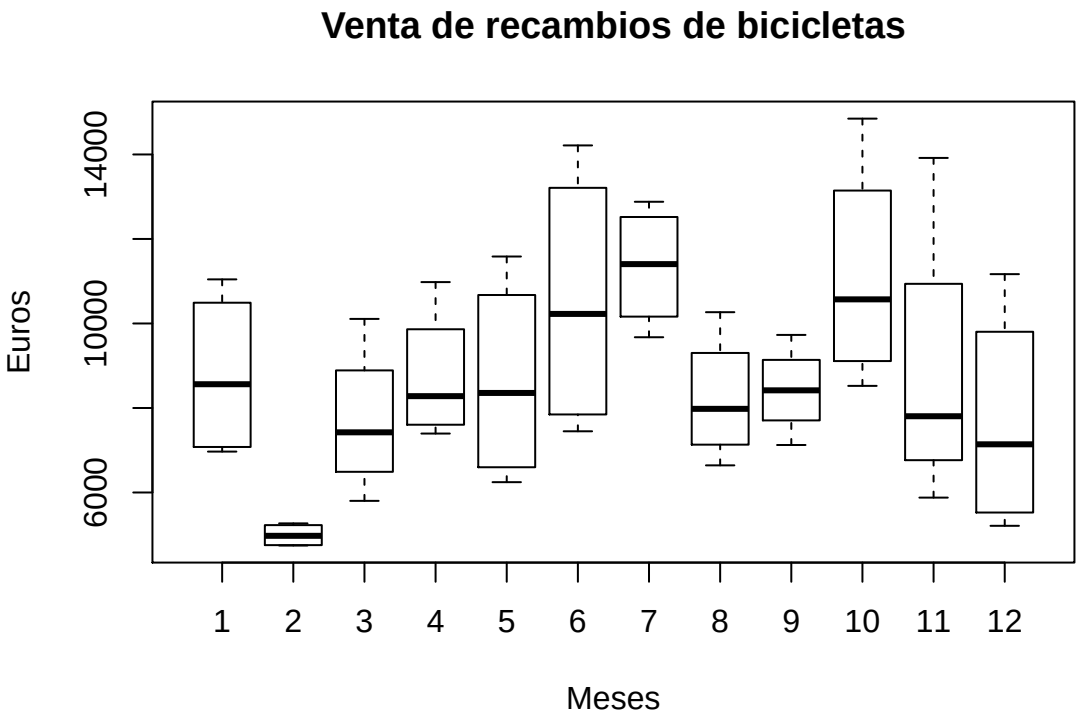
Como conclusión, obtenemos que esta segunda serie es no estacionaria, tiene tendencia estocástica, variabilidad, inestable respecto a la media y el mejor modelo que modela esta serie es $ARIMA(0,1,1)(0,0,1)_{12}$.

Recambios de bicicletas

Estudiemos en tercer lugar el análisis correspondiente a la serie temporal de las ventas en recambios de bicicletas.

Como tenemos por costumbre, vamos a ver el gráfico de las ventas mensuales de nuestra serie para poder empezar a sacar conclusiones.





Parece, a primera vista, que se trata de una serie no estacionaria pues parece poseer

una tendencia creciente en el tiempo, variación estacional (en los meses previos y siguientes a verano), variabilidad que parece ir creciendo a medida que pasa el tiempo y tendencia creciente. Por tanto, podría ser descrita por un modelo multiplicativo.

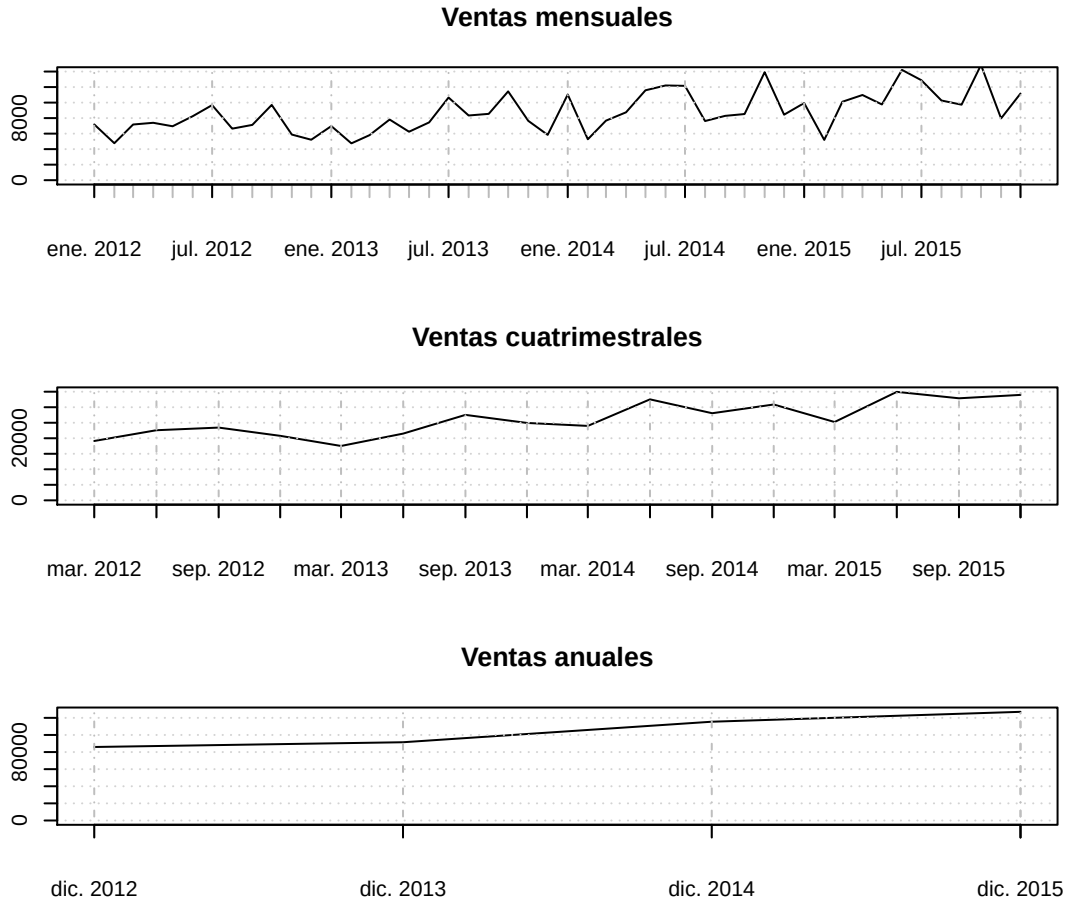
Gracias al gráfico de Box-Plot, confirmamos nuestras sospechas ya que parece que en los meses de Junio, Julio y Octubre, las ventas son superiores a los demás meses. Podemos apreciar además, cómo las ventas en el mes de Febrero caen por debajo de los 6000 euros cuando por ejemplo en Octubre, las ventas son superiores a los 14000 euros, convirtiéndose con diferencia, en el mes con menos ganancias de todo el año.

A continuación se nos muestra un análisis descriptivo de la serie de ventas en recambios de bicicletas:

		n	$\overline{x}$	σ	min	max	rango	$se = \sigma/\sqrt{n}$	NA	NA	N
Recambios bicicletas	1	48	8748.89	2557.95	8312.58	8604.50	2443.59	4745.47	14847.84	10102.37	369.2

Cuadro 4.4: Estadística descriptiva

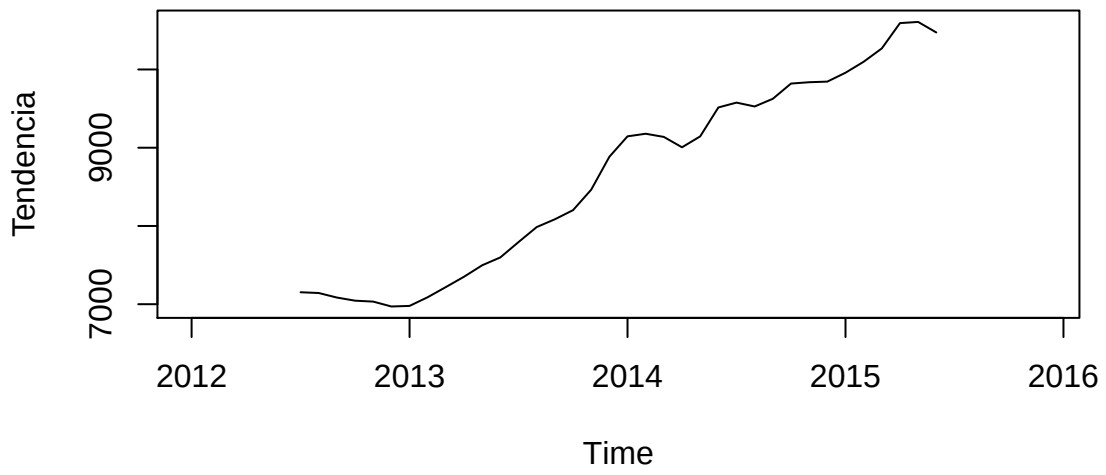
Calculemos ahora las ventas tanto cuatrimestrales como anuales de esta serie y hagamos los correspondientes gráficos, donde se puede apreciar la tendencia.



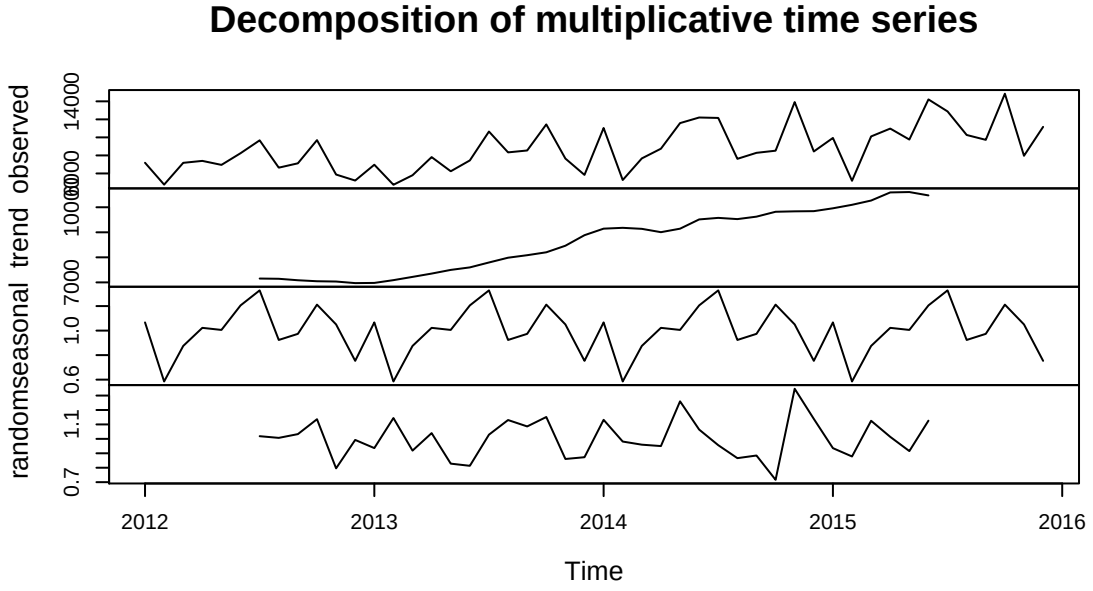
Podemos sacar varias conclusiones. En primer luegar, vemos por los datos de ventas anuales que hemos obtenido, que las ventas han ido creciendo a lo largo de los años, pasando de unas ventas iniciales de 85944.90 euros, a obtener de beneficio en el año 2015, 127039.19 euros. Este dato se refleja con la gráfica correspondiente a las ventas anuales donde vemos una clara tendencia creciente donde parece haber tres

situaciones diferenciadas. Por otro lado, en el gráfico de ventas cuatrimestrales podemos observar que las ventas comienzan por unos 20000 euros y finalizan cercanas a las 35000 euros donde se aprecian subidas y bajadas y donde además vemos que las variaciones van aumentando en el transcurso del tiempo.

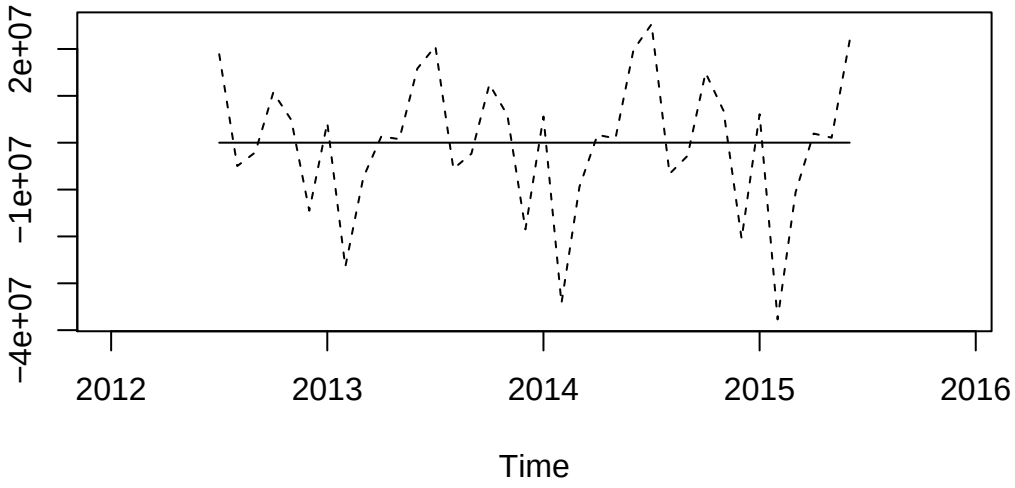
Podemos visualizar exclusivamente la tendencia de la variable y observamos que ésta es creciente.



Pasemos a ver la descomposición de la serie temporal. Debido a lo que hemos argumentado antes, dado que la tendencia y la varianza de la serie original son crecientes, el modelo de descomposición más adecuado parece ser el multiplicativo.



Mostramos ahora la imagen del efecto estacional (con línea discontinua) superpuesta con la tendencia (línea continua).



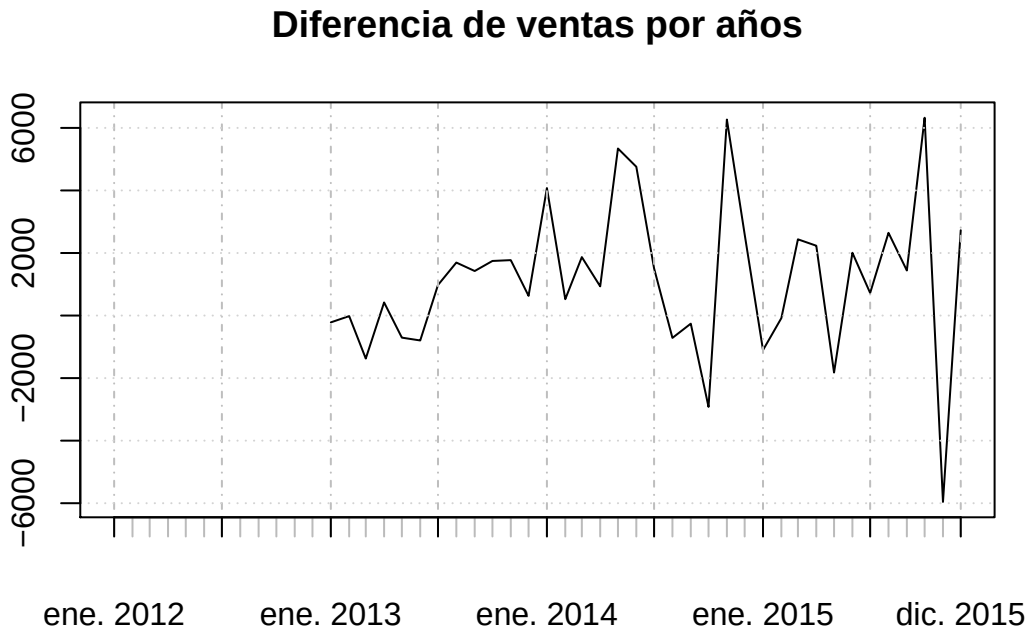
Vemos cómo la pauta estacional se superpone a la tendencia global creciente (aunque no se aprecia por la escala), produciendo un comportamiento cíclico que se repite en los distintos años de la muestra.

Podemos sacar más datos interesantes de esta serie. Mostremos entonces ahora, por ejemplo,las ventas de Junio (el mes con más ventas) de cada uno de los cuatro años para ver cómo han ido evolucionando las ventas en ese mes:

```
##          Recambios bicicletas
## jun. 2012              8245.58
##          Recambios bicicletas
## jun. 2013              7448.59
##          Recambios bicicletas
## jun. 2014             12206.37
##          Recambios bicicletas
## jun. 2015             14214.03
```

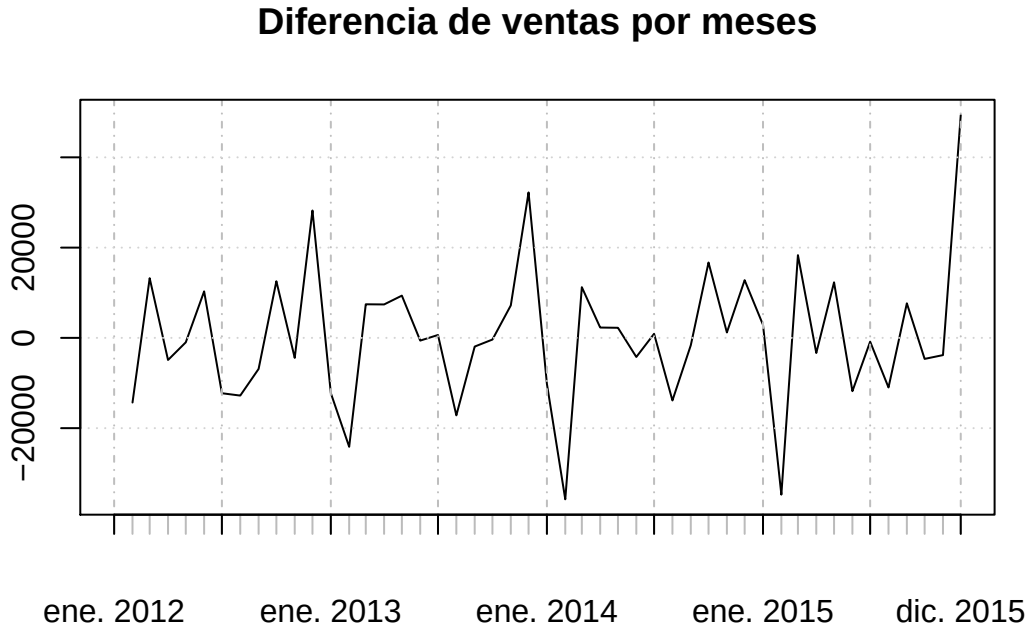
Vemos que tras una disminución del año 2012 a 2013, las ventas comienzan después a crecer año tras año. Podemos calcular que la diferencia de ganacias del año 2015 y 2012 es de 5968,45 euros más en el último año.

Otros resultados que podemos sacar, son la diferencia de ventas de un mes de un año con respecto al del año anterior:



Aquí lo que estamos haciendo es comparar cada mes con el mismo del año anterior. Por ejemplo vemos que en Enero de 2013 se ganaron 218,21 euros menos que en Enero de 2012. Sin embargo el mismo mes del año 2013 se ganan 4075.82 más que en el mismo mes de 2013.

También podemos visualizar la diferencia de ventas con respecto al mes anterior:



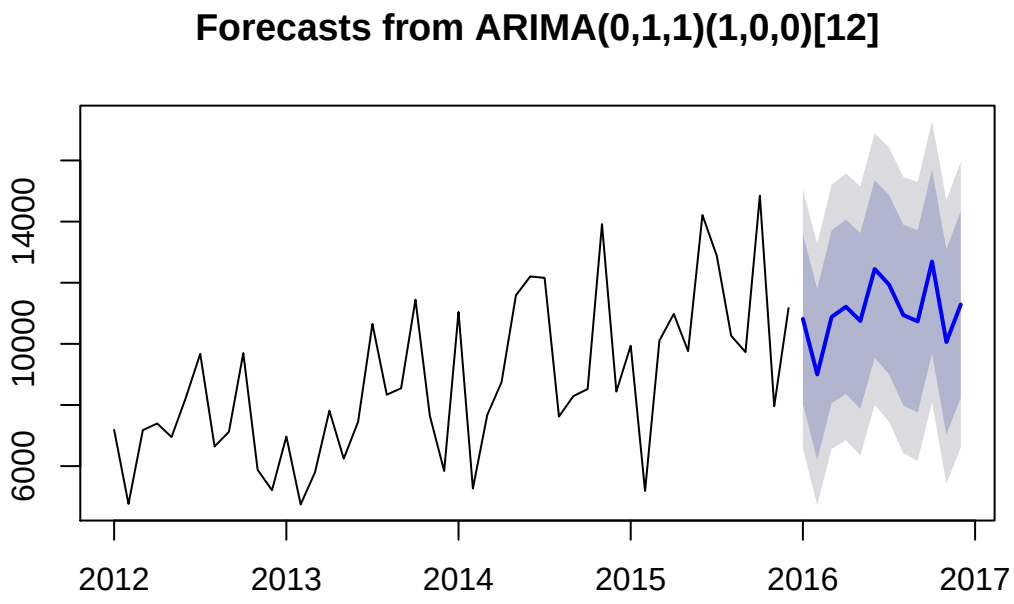
En este gráfico vemos cómo el resultado es bastante homogéneo en media y varianza durante todo el transcurso del tiempo.

A continuación, veamos mediante la función `auto.arima`, qué modelo es el más adecuado para nuestra serie temporal.

```
## Series: datos[, 2]
## ARIMA(0,1,1)(1,0,0)[12]
##
## Coefficients:
##          ma1    sar1
##       -0.8593  0.3820
## s.e.   0.0654  0.1499
##
## sigma^2 estimated as 4674343:  log likelihood=-428.12
## AIC=862.25   AICc=862.81   BIC=867.8
```

Obtenemos que el modelo $ARIMA(0,1,1)(1,0,0)_{12}$ sería el que mejor describe la serie de ventas en recambios de bicicletas lo que implica que la parte estacional es anual y sigue un modelo AR estacionario y que la parte no estacional es no estacionaria, que hemos tenido que diferenciar una vez y que el modelo es de media móvil con respecto al mes anterior.

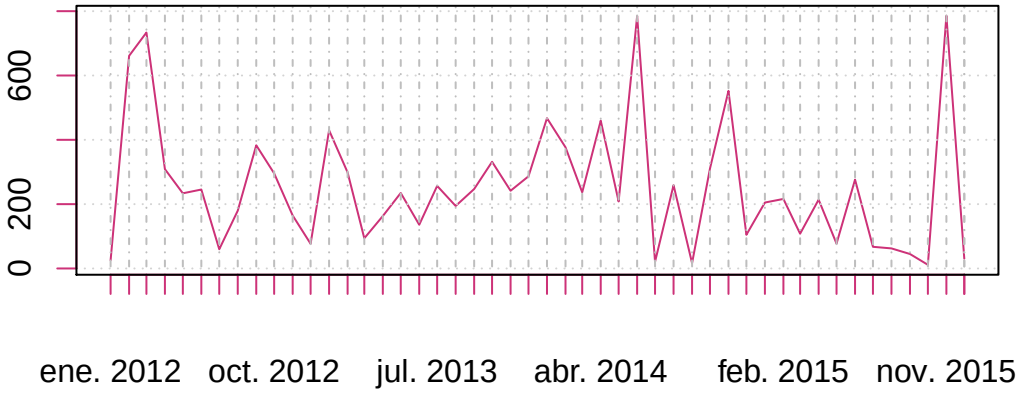
Por otro lado, gracias al ajuste que hemos hecho antes, podemos predecir los valores en ventas para el año 2016, los cuáles parecen razonables por el transcurso de los años. Siempre obtenemos este resultado con unos intervalos de predicción, los cuales también se ven reflejados en el gráfico, de otros colores.

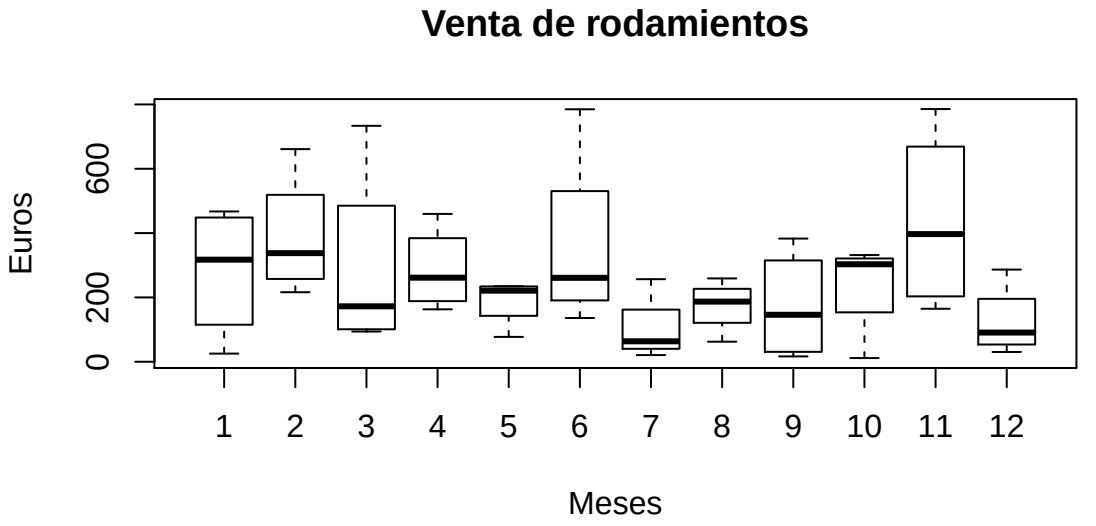


Después del estudio, hemos concluido que la serie es no estacionaria, tiene tendencia creciente, componente estacional, variabilidad y el modelo que mejor la describe según la función `auto.arima` es $ARIMA(0,1,1)(1,0,0)_{12}$.

Rodamientos

Nuestro estudio ahora se basará en la serie temporal de «Ventas en rodamientos». Siguiendo la metodología realizada durante el trabajo, vemos inicialmente la representación gráfica de la serie:





En esta ocasión, podemos observar que no hay una tendencia clara pues hasta Junio de 2014 la tendencia parece ser creciente aunque a partir de ahí, las ventas comienzan a descender, dibujando una tendencia decreciente. Se puede decir que se trata de una serie estocástica, pues no sigue un único patrón. Por otro lado, parece haber una componente estacional que se repite a final de cada año. Se trata de una serie no estacionaria.

Observamos, fijándonos en el gráfico de Box-Plot, que aunque no con mucha diferencia, el mes de Noviembre parece destacar entre los demás, pasando de una media de 253,70 euros, a casi alcanzar los 800 euros.

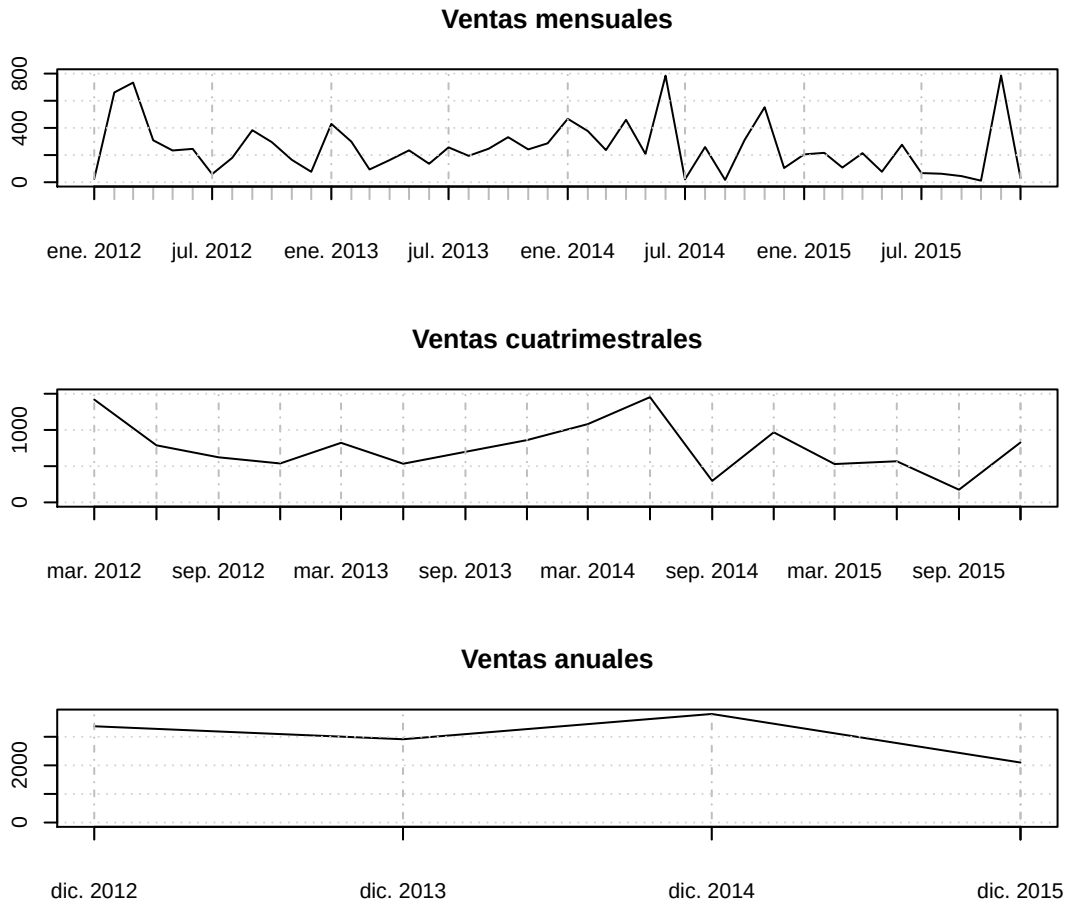
El análisis descriptivo de la serie de ventas en rodamientos nos muestra los siguientes datos:

Calculemos las ventas medias cuatrimestrales y anuales, para después hacer sus

		n	$\overline{x}$	σ	min	max	rango	$se = \sigma/\sqrt{n}$	NA	NA	NA
Rodamientos	1	48	253.70	196.90	234.19	228.46	166.33	11.46	785.71	774.25	28.42

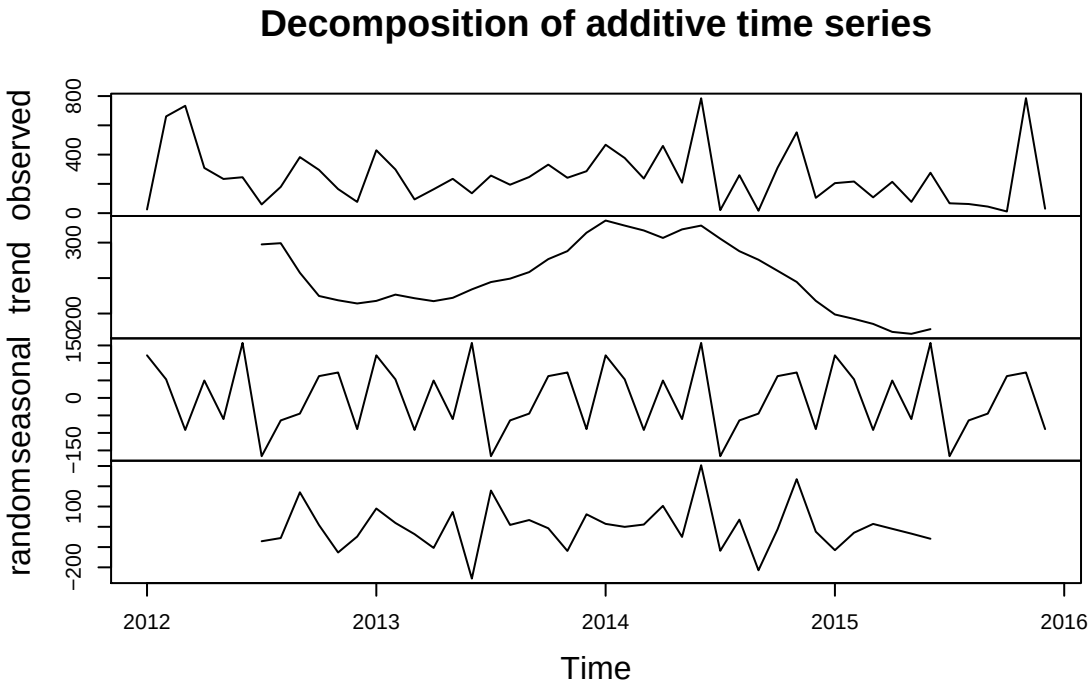
Cuadro 4.5: Estadística descriptiva

correspondientes gráficos:



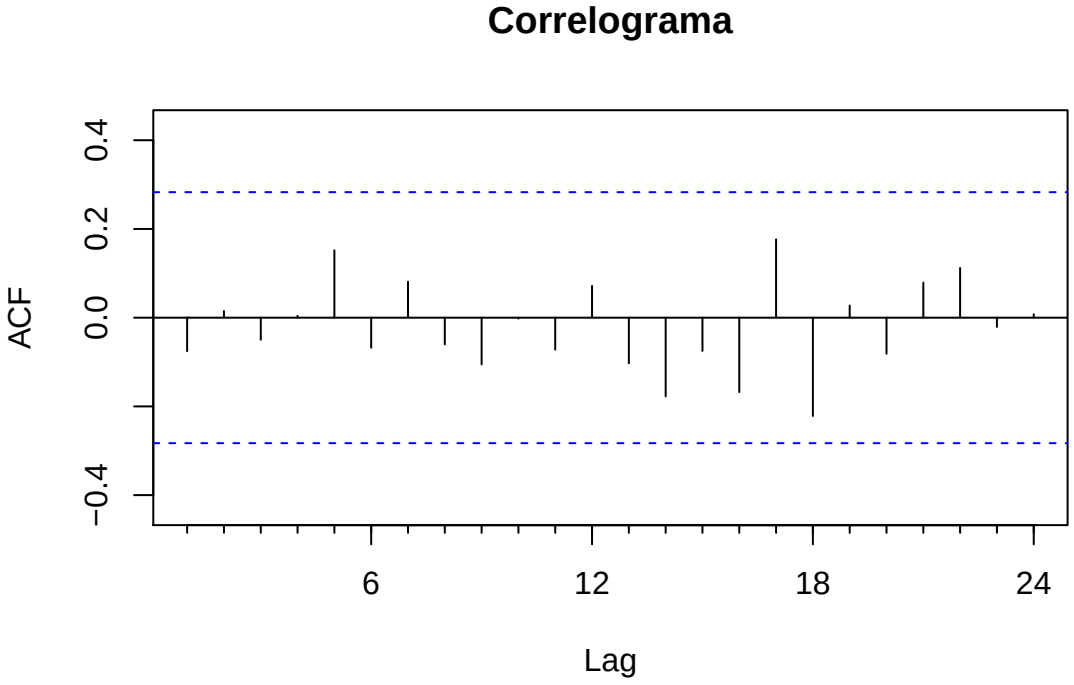
En el gráfico de ventas anuales, podemos ver varias fases en la tendencia. Hay dos periodos de decrecimiento (de Enero de 2012 hasta Diciembre de 2013 y de Diciembre de 2014 a Diciembre de 2015) y uno de crecimiento (de Diciembre de 2013 a Diciembre de 2014). Parece complicado hacer una predicción de los datos pues la tendencia no es clara.

Pasemos a ver la descomposición de la serie. Parece más adecuado descomponerla con un modelo aditivo, pues las fluctuaciones aleatorias en los datos son bastante constantes a lo largo del tiempo. Además no hay tendencia, luego haremos una descomposición para una serie sin tendencia y explicada con un modelo aditivo.

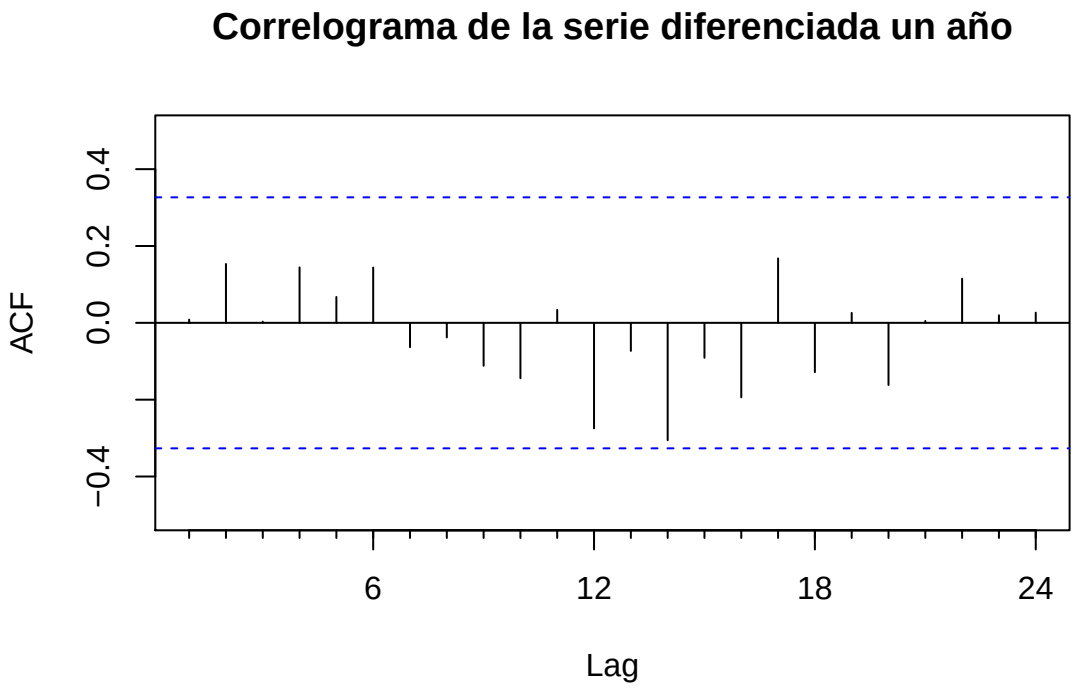


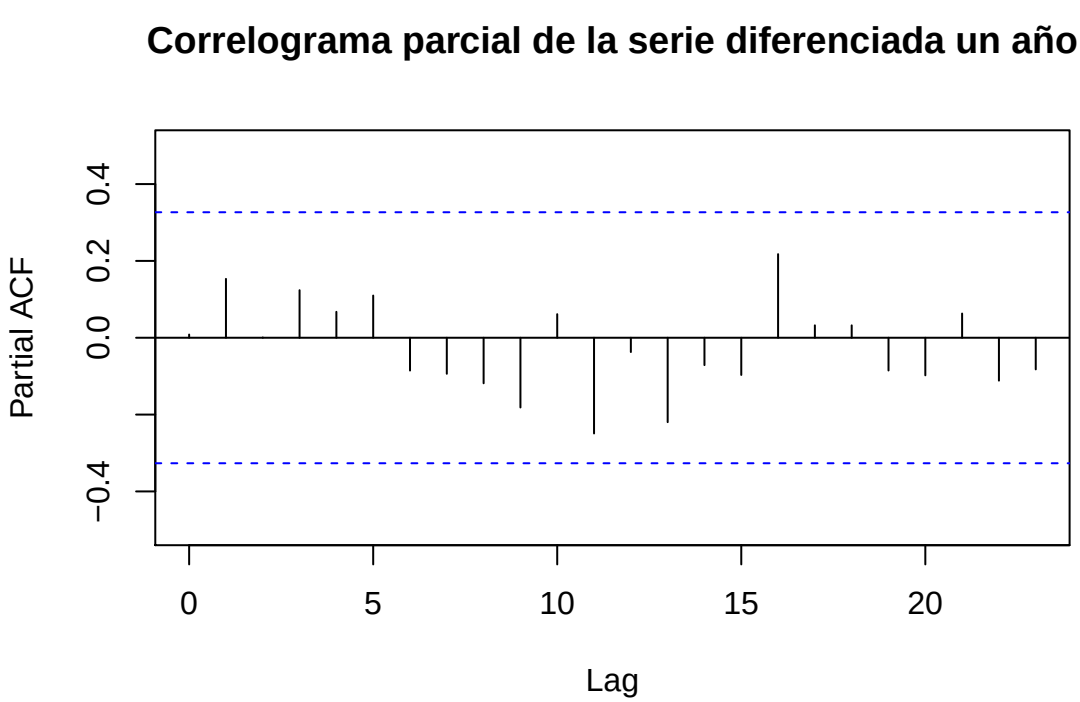
De estos gráficos parece desprenderse que no hay una tendencia apreciable pero sí una cierta estacionalidad.

Si dibujamos el correlograma parece que nos econtramos ante un ruido blanco ya que no aparece ningún valor fuera de la banda de confianza.



En consecuencia parece que el modelo es estacionario y además un ruido blanco pero con algún comportamiento estacional (sospechamos que anual). Si dibujamos ahora el correlograma de la serie diferenciada con respecto al año anterior obtenemos





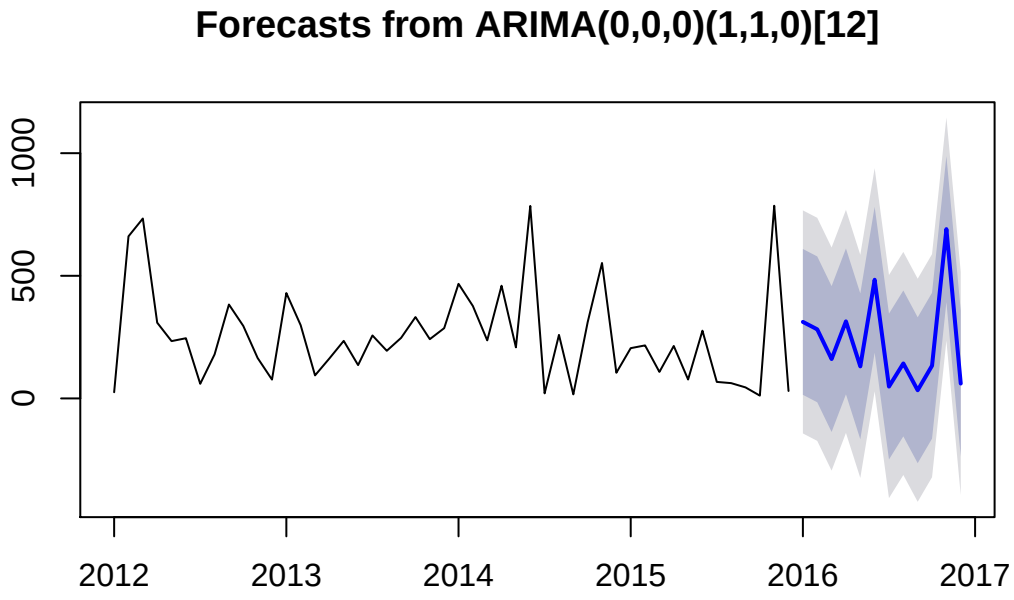
En este caso, tampoco identificamos un modelo autorregresivo o de media móvil. Si recurrimos a la función `auto.arima`, obtenemos uno modelo $ARIMA(0,0,0)(1,1,0)_{12}$, es decir un «ruido blanco» (abusando un poco de la expresión porque realmente no lo es) con una componente estacional anual autorregresivo de orden 1 no estacionaria.

```
## Series: datos[, 14]
## ARIMA(0,0,0)(1,1,0)[12]
```

```
##
## Coefficients:
##          sar1
##       -0.4082
## s.e.   0.1737
##
## sigma^2 estimated as 53837:  log likelihood=-247.76
## AIC=499.51   AICc=499.87   BIC=502.68
```

Realmente la función `auto.arima` ha buscado el mejor modelo (dentro de los que dispone) que mejor se ajusta a los datos, pero que realmente no tiene porqué ser un buen modelo.

Si representamos la posibles ventas que nos proporciona este modelo para el año que viene, obtenemos



que parece una predicción razonable si nos basamos en un periodo corto anterior (quizás los dos últimos años) ya que el periodo anterior a 2014 parece bastante anárquico. Esto nos da una pequeña muestra de lo que la capacidad computacional y la posibilidad de probar muchos modelos (de forma automática) nos puede ayudar a modelar situaciones que en otras épocas, con cálculos a mano hubiera sido imposible.

5 Conclusiones

Durante este trabajo han sido analizadas exclusivamente 4 de las 29 variables, pues se ha considerado, después de estudiar el resto, que son las que presentan los resultados más interesantes y, además tienen un comportamiento diferente.

Con estas variables hemos podido construir diferentes modelos de series temporales, pudiendo ver la mayoría de las posibilidades que este tipo de datos ofrecen.

En este trabajo hemos podido exponer de una forma práctica con datos reales como se comportan series estacionarias y no estacionarias, además de estudiar el concepto de estacionalidad o la ausencia de la misma así como la existencia de diferentes tipos de tendencias (decrecientes, crecientes y estables).

Se han modelado las series utilizando la conocida metodología ARIMA, y hemos podido observar y analizar de primera mano algunos de las características más importantes de datos temporales reales, pues los que se han utilizado en esta memoria nos lo ha proporcionado una empresa almeriense procedentes de sus balances de ventas. Eso ha hecho que nos aproximemos al mundo de los datos temporales de una forma directa aplicando los conceptos teóricos conocidos desde hace bastante tiempo al mundo real. En este caso la experiencia ha sido muy rica pues he podido apreciar la dificultad existente a la hora de trasladar la teoría a los datos reales.

Un hecho a resaltar ha sido el modelaje de la última serie temporal que hemos considerando, la de «ventas de rodamientos». Al intentar construir un ajuste sin la función `auto.arima` han aparecido algunas cuestiones interesantes.

A primera vista esta serie parece que tiene una componente estacional. Sin embargo el estudio de los correlogramas no podemos interpretarlos con facilidad como para poder elegir un modelo acorde a la serie.

En este caso, si aplicamos la función `auto.arima` mencionada anteriormente, sí que hemos obtenido un modelo que se ajusta a los datos de esta variable (cosa que es de gran utilidad pues a mano nos hubiera sido imposible llegar a una conclusión). Esta función prueba entre muchos modelos posibles y de entre ellos, escoge aquél que mejor encaje con la serie, según criterios propios. De esta forma, la capacidad de cálculo y los métodos de computación nos permiten construir modelos que, en otras épocas, donde los cálculos habían de hacerse a mano, hubiera sido imposible o muy difícil de realizar.

Hemos por tanto, podido llevar a cabo aquellos objetivos que nos propusimos en un principio: describir y construir modelos univariantes con datos temporales así como realizar predicciones de su comportamiento futuro. Ello nos ha permitido, aparte de tener que conocer alguno los modelos teórico de series existentes, adquirir destreza en el manejo de herramientas estadísticas (**R** en este caso) que nos ha permitido construir los mejores modelos adaptados a nuestros datos.

Finalmente, el uso de LaTeX y el paquete *knitr* de **R** ha permitido elaborar esta memoria incluyendo tanto el código utilizado como los resultados obtenidos directamente en el informe sin tener que «copiar y pegar» resultados lo que redunda en la buena presentación y uniformidad del texto.

Bibliografía

[1] Avril Coghlan, *A Little Book of R For Time Series*, 2015.

[2] Paul S.P. Cowpertwait, Andrew V. Metcalfe, *Introductory Time Series with R*, Springer, 2009.

[3] Jonathan D. Cryer,Kung-Sik Chan, *Time Series Analysis with Applications in R*, Springer, 2010.

[4] Daniel Peña, *Análisis de series temporales*, Alianza Editorial, 2010.

[5] Bernhard Pfaff, *Analysis of Integrated Series with R and Cointegrated Time*, Springer, 2008.

[6] Robert H. Shumway,David S. Stoffer, *Time Series Analysis and its Applications*, Springer, 2010.

[7] Eric Zivot and Jiahui Wang, *Modelling Financial Time Series with S-PLUS*, New York Springer, 2005.

[8] https://www.otexts.org/fpp/8/9.

[9] https://rpubs.com/joser/SeriesTemporalesBasicas.

[10] https://www.otexts.org/fpp/8/7.

[11] http://halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/EDescrip/tema7.pdf.