Estudio estadístico de un caso real. Series temporales.

Trabajo Fin de Grado

Autor:
Ana Bonilla Ruiz

Tutor:
Fernando Reche Lorite

Grado en Matemáticas

Septiembre, 2019
Universidad de Almería

Índice general

Abstract in English

In a real field, especially from an economic point of view, is common to deal with data which take place at different time. Time series are so useful in this context. Studing those series, univariate and multivariate cases, is our main goal in this research.

This document will be divided mainly in two different parts. Firstly we will introduce some theoretical results which are needed in order to understanding our methods and useful to manage time series. Furthermore, we will be provided by useful tools to estimate models which give us predictions about the future behaviour of those time series.

During the second part of this text we will work with all the tools introduced before. Focusing in a real data provided by an international hotel company, located in Roquetas de Mar (Almería). The analysis of the data have been done with **R** software, mainly with its Rstudio tool, where we have managed temporal formats.

Resumen en español

En un ámbito real, y sobre todo si hablamos desde un punto de vista económico, es común trabajar con datos que tienen lugar en distintos instantes de tiempo. Por ello, resulta de interés estudiar *series temporales*. El estudio de estas series es el principal objetivo, tanto en su forma univariante como en la multivariante.

El documento estará dividido principalmente en dos partes. En la primera se hará un recorrido teórico dando definiciones adecuadas para el estudio de dichas series y se introducirán recursos estadísticos que nos ayudaran a estudiar su comportamiento. Además, nos proporcionarán herramientas para la estimación de modelos óptimos que se adapten a las condiciones de los diferentes tipos de series, los cuales, no solo nos dan información sobre su comportamineto pasado, si no que también nos serán de utilidad para hacer previsiones futuras.

En la segunda parte, llevaremos a la práctica los anteriores conceptos teóricos. Para ello, haremos uso de unos datos reales proporcionados por una cadena hotelera, la cual se expande en territorio nacional e internacional, teniendo su sede en Roquetas de Mar (Almería). Para el análisis de estos datos se hará uso del software **R**, en particular de su entorno *Rstudio*, en el cual se trabajará en un formato temporal.

1 Introducción

Para hacer el estudio de series temporales hemos recurrido a unos datos proporcionados por la empresa *Senator Hotels & Resort*, comúnmente conocida como *Grupo Hoteles Playa* (primera cadena de hoteles de Andalucía). Es una empresa cuya sede se encuentra Roquetas de Mar (Almería), aunque tiene comercios de sector turístico situados por varios puntos turísticos nacionales e internacionales. Nosotros nos centraremos principalmente en el sector hotelero nacional.
Tenemos datos desde el año 2010 (aunque algunos hoteles con los que trabajamos tuvieron fecha de apertura posterior) hasta la actualidad (entendiendo por actualidad el momento en el que actualicé los datos, 28 de Marzo de 2019). Los hoteles con los que trabajamos tienen asociado un código de identificación, y son los siguientes:

- **10**: PLAYACAPRICHO HOTEL (Roquetas de mar, Almería).
- **20**: SENATOR CASTELLANA (Madrid, Madrid).
- **30**: PLAYADULCE HOTEL (Aguadulce, Almería).
- **40**: DIVERHOTEL AGUADULCE (Aguadulce,Almería).
- **50**: PLAYASOL SPA HOTEL (Roquetas de Mar, Almería).
- **60**: DIVERHOTEL ROQUETAS (Roquetas de Mar,Almería).
- **80**: VERA PLAYA CLUB HOTEL (Vera, Almería).
- **90**: ZIMBALI PLAYA SPA HOTEL (Vera, Almería).
- **130**: SENATOR MAR MENOR GOLF & SPA (Los Alcázares,Murcia).
- **140**: HOTEL CABO DE GATA (Retamar (El Toyo), Almería).
- **150**: SENATOR BARAJAS HOTEL(Madrid, Madrid).
- **180**: PLAYABALLENA SPA HOTEL (Costaballena (Rota), Cádiz).
- **190**: PLAYACÁLIDA SPA HOTEL (Almuñecar, Granada).
- **200**: PLAYALINDA HOTEL (Roquetas de Mar, Almería).
- **210**: PLAYABONITA HOTEL (Benalmádena, Málaga).
- **220**: MARBELLAPLAYA (Marbella, Málaga).
- **250**: DIVERHOTEL MARBELLA (Marbella, Málaga).
- **270**: PLAYACANELA HOTEL (Ayamonte (Isla Canela), Huelva).
- **280**: SENATOR HUELVA HOTEL (Huelva, Huelva).
- **290**: SUITES PUERTO MARINA (Mojácar, Almería).
- **340**: HOTEL VIRGEN DE LOS REYES (Sevilla, Sevilla).

- **350**: SENATOR MARBELLA SPA HOTEL (Marbella, Málaga).
- **360**: SENATOR PARQUE CENTRAL HOTEL (Valencia, Valencia).
- **370**: SENATOR BARCELONA SPA HOTEL (Barcelona, Barcelona).
- **400**: SENATOR GRAN VÍA 70 SPA HOTEL (Madrid, Madrid).
- **410**: PLAYAMARINA SPA HOTEL (Ayamonte (Isla Canela), Huelva).
- **420**: PLAYACARTAYA SPA HOTEL (Cartaya, Huelva).
- **430**: APARTAMENTOS PLAYAMARINA (Ayamonte(Isla Canela), Huelva).
- **440**: SENATOR CÁDIZ SPA HOTEL (Cádiz, Cádiz).
- **450**: SENATOR GRANADA SPA HOTEL (Granada, Granada).
- **460**: CALEIA TALAYOT SPA HOTEL (Cala Millor, Mallorca).
- **470**: SENATOR BANUS SPA HOTEL (Puerto Banus- Estepona, Málaga).
- **480**: PARAÍSO PLAYA (Vera, Almería).
- **500**: ALMUÑECAR PLAYA SPA HOTEL.(Almuñecar, Granada).

Durante el transcurso del trabajo analizaremos variables que tienen que ver con dichos hoteles. Se mostrarán los resultados más interesantes.

Una vez tenemos las variables planteamos los siguientes objetivos:

- **Estudio de series univariantes:**
  - Estudio de estacionalidad y estacionariedad.
  - Construcción de un modelo.
  - Selección del modelo.
  - Realización de predicciones.
  - Validación del modelo.
- **Estudio de series multivariantes**.
  - Construcción de un modelo VAR.
  - Estudio de cointegración.
  - Construcción de un modelo VEC.

Para el cumplimiento de estos objetivos, junto a una base teórica, la herramienta **R** toma un papel fundamental, así como lo hacen paquetes muy concretos que nos permiten trabajar con series temporales, los cuales describiremos durante el desarrollo del trabajo.

2 Fundamentos matemáticos

Durante el transcurso de este informe vamos a usar conceptos y modelos estadísticos que necesitamos definir para un mejor entendimiento de nuestro análisis.

2.1 Definiciones de interés

El desarrollo de nuestro trabajo se basa en el estudio de *series temporales*, así pues empezaremos por dar una definición de este término.

Definición 2.1. *Se llama* ***serie temporal*** *a una variable estadística cuyos valores varían a lo largo del tiempo de forma equiespaciada. La podemos expresar de la siguiente forma:*

$$Y_t, \qquad t = 1, 2, \ldots, T,$$

donde el subíndice t indica el tiempo en que se observa el dato Y_t.

Por un lado tenemos las series temporales univariantes en las que sólo se analiza una serie temporal en función de su propio pasado. Por otro, tenemos las series multivariantes, donde se analizan varias series a la vez.

Desde el punto de vista clásico una serie está formada por cuatro componentes:

- **Tendencia:** refleja la evolución de la serie a largo plazo. Esta componente puede ser de carácter estacionario (gráficamente se representa por una recta paralela al eje de abscisas), lineal (creciente o decreciente), parabólica, exponencial, etc. Para captar esta componente es necesario observar un periodo de tiempo amplio.
- **Ciclo:** se caracteriza por oscilar al rededor de la tendencia, estas oscilaciones no son regulares. Como resulta complicado separar la tendencia y el ciclo, normalmente se trabaja con una mezcla de ambas que se denomina *ciclo-tendencia*.
- **Componente estacional:** esta componente recoge las oscilaciones que se producen durante periodos de la serie y que se repiten regularmente en los diferentes periodos.
- **Componente aleatoria:** Es comunmente conocida como «ruido». Son movimientos imprevisibles sin pauta periódica ni tendencia reconocible. Actúa en cualquier serie temporal, ya sea en mayor o menor medida.

En cuanto a *series temporales* tienen gran importancia las *series estacionarias*, las cuales pueden ser *fuertemente estacionarias* o *débilmente estacionarias* (cuando se hable de estacionaridad nos referiremos a este último). Otro concepto destacable es el de *serie estacional*. Veamos una definición formal de estos conceptos.

Definición 2.2. *Se dice que la serie* Y_t *es* ***fuertemente estacionaria*** *cuando cualquier colección finita* $\{Y_{t_1}, Y_{t_2}, \ldots, Y_{t_k}\}$ *con* $k \geq 1$ *tiene el mismo comportamiento probabilístico que* $\{Y_{t_1+h}, Y_{t_1+h}, \ldots, Y_{t_k+h}\}$ *para cualquier entero* $h = \pm 1, \pm 2 \ldots$, *es decir,*

$$F(y_{t_1}, y_{t_2}, \ldots, y_{t_k}) = F(y_{t_1+h}, y_{t_2+h}, \ldots, y_{t_k+h})$$

siendo F la función de distribución.

Definición 2.3. *Decimos que Y_t es **débilmente estacionaria** si se cumple que, para todo $t \neq s$:*

1. $E[y_t] = E[y_s] = \mu$.
2. $Var(y_t) = Var(y_s) = \sigma^2$.
3. $Cov(y_t, y_{t+k}) = Cov(y_s, y_{s+k})$.

Si se cumple la condición 1 diremos que la serie es **estacionaria en media**, si se cumple también la 2, la serie será **estacionaria en varianza**.

Definición 2.4. *Una serie Y_t es **estacional** si su valor esperado no es constante, pero varía con una pauta cíclica. Más concretamente, si:*

$$E[y_t] = E[y_{t+s}]$$

diremos que la serie tiene estacionalidad de periodo s.

En una serie temporal Y_t cabe analizar como afectan las observaciones del pasado en las observaciones del futuro. Para identificar esta dependencia vamos a definir algunas funciones.

La *función de autocorrelación* tiene gran utilidad para estudiar la estacionalidad de la serie, ya que si los valores estan separados entre sí por un cierto periodo de tiempo, estos deben estar correlacionados.

Definición 2.5. *Sea la serie temporal Y_t, entonces la **función de autocorrelación** entre el instante t y el instante s viene dada por*

$$\rho(s,t) = \frac{\gamma(s,t)}{\sqrt{\gamma(s,s)\gamma(t,t)}},$$

donde $\gamma(t,s)$ es la covarianza.

Este coeficiente mide la capacidad de predecir la serie en el instante t a partir de un valor del instante s.

Existirá correlación si el coeficiente de autocorrelación para un retardo igual al periodo estacional es significativamente distinto de cero.

Ahora bien, la función de autocorrelación parcial mide la correlación entre dos variables separadas por h periodos cuando no se considera la dependencia creada por los retardos intermedios existentes entre ambas.

Definición 2.6. *La **función de autocorrelación parcial** (PACF) de una serie Y_t , ϕ_{hh} para $h = 1, 2, \ldots$ se define como*

$$\phi_{11} = corr(Y_{t+1}, Y_t) = \rho(1)$$

$$\phi_{hh} = corr(Y_{t+1} - \hat{Y}_{t+1}, Y_t - \hat{Y}_t)$$

La interpretación gráfica de estas funciones nos da gran información sobre el modelo que puede seguir nuestra serie temporal.

No obstante, existen contrastes de hipótesis que nos permiten ver si los coeficientes de correlación son simultáneamente iguales a cero, es decir, ausencia de correlación. Nosotros vamos a definir el *Test de Box-Pierce* y el *Test de Ljung-Box*.

Test de Box-Pierce. Este test parte de que $Y_t \sim \mathcal{N}(0, \sigma)$ y que son independientes, es decir,

$$\begin{aligned} H_o &: \rho_k = 0 \quad k = 1, 2, \ldots, p \\ H_1 &: \text{algún} \quad \rho_k \neq 0 \end{aligned}$$

El estadístico que se usa es

$$Q = T \sum_{k=1}^{p} \hat{\rho}_k^2$$

que, bajo la hipótesis nula, asintóticamente se distribuye como una χ_p^2.

Test de Ljung-Box. Las hipótesis son las mismas salvo que el estadístico, distribuido asintóticamente como anteriormente, es

$$Q_1 = T(T-2) \sum_{i=1}^{p} \frac{\hat{\rho}_k^2}{T-k}$$

Esta forma tiene un mejor funcionamiento para muestras pequeñas y es más robusto frente a la falta de normalidad.

Existe un modelo estadístico simple en el que hay ausencia de correlación en el tiempo entre sus observaciones, este es el conocido *ruido blanco*, procedemos a dar una definición de este.

Definición 2.7. *Un **ruido blanco** es una serie, el cual habitualmente notamos por a_t, $t = \pm 1, \pm 2, \ldots$, tal que su media es cero, la varianza es constante y es incorrelada. Es decir:*

- $E[a_t] = 0$.
- $Var(a_t) = \sigma_a^2$.
- $Cov(a_t, a_{t+k}) = 0$.

Se trata de un proceso en el que todas sus variables son independientes.

Existen modelos que estudian tanto series estacionarias (modelos ARMA) como no estacionarias (modelos ARIMA). Estos fueron estudiados por Box y Jenkins en los años 70. A estos modelos les dedicaremos las siguientes secciones.

2.2 Series estacionarias

En esta sección vamos a analizar los modelos estacionarios. Estos, además de ser útiles para modelar series estacionarias, pueden ser aplicados a series no estacionarias inicialmente, pero que, tras transformaciones sencillas lo serán, como veremos posteriormente.

Procesos Autorregrevos: AR(p)

Diremos que un proceso es **autorregresivo de orden p** o **AR(p)** si el valor de variable temporal Y_t depende de sus valores pasados y del ruido blanco, a_t , a tiempo t, donde $t = 1, 2, \ldots, T$. Es decir,

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \ldots + \phi_p Y_{t-p} + a_t = \sum_{j \geq 0}^{p} \phi_j a_{t-j} \quad p = 1, 2, \ldots, k$$

donde $\phi_1, \phi_2, \ldots, \phi_p$ son coeficientes no nulos y p el número de retardos necesarios para pronosticar la serie.

Podemos definir, de forma alternativa a la ya dada, el modelo haciando uso de un operador de retardo [1], el cual definimos de la siguiente forma,

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p,$$

una vez definido esto podemos construir el modelo AR(p) como sigue

$$\phi_p(B) Y_t = a_t$$

Estos procesos se usan para definir fenómenos donde los eventos producen un efecto inmediato durante un corto periodo de tiempo.

En estos modelos nos encontramos con el problema que no en todos los casos son estacionarios, vemos el comportamiento del modelo **AR(1)**, el cual para ciertas restricciones de los parámetros no es estacionario.

Definimos el modelo AR(1) de la siguiente forma:

$$Y_t = \phi_1 Y_{t-1} + a_t$$

Para ver cuales son estos parametros vamos a comprobar las condiciones de estacionaridad para ver cuales son estas restricciones.

1. Estacionariedad en media.

$$E[Y_t] = E[Y_t = \phi_1 Y_{t-1} + a_t]$$

[1] Definimos el operador retardo B como

$$B Y_t = Y_{t-1},$$

si aplicamos este operador k veces, se puede comprobar que $B^k Y_t = Y_{t-k}$

Ahora bien, haciendo uso de la linealidad de la esperanza y sabiendo que, por definición $E[a_t] = 0$, se tiene:

$$E[Y_t = \phi_1 Y_{t-1} + a_t] = E[\phi_1 Y_{t-1}] + E[a_t] = \phi_1 E[Y_{t-1}]$$

Como la media debe ser constante y finita en el tiempo, por lo que,

$$E[Y_t] = \phi_1 E[Y_t]$$

$$(1 - \phi)E[Y_t] = 0$$

$$E[Y_t] = \frac{0}{1 - \phi_1} = 0$$

Entonces, para que el proceso sea estacionario $\phi_1 \neq 1$.

2. Estacionario en covarianza.

$$\gamma_0 = E[Y_t - E[Y_t]]^2 = E[\phi_1 Y_{t-1} + a_t - 0]^2 = \phi^2 Var(Y_{t-1}) + \sigma^2$$

Dada la autocorrelación del proceso

$$E[Y_{t-1} a_t] = E[(Y_{t-1} - 0)(a_t - 0)] = Cov(Y_{t-1} a_t) = 0$$

Suponiendo que el proceso es estacionario ,

$$E[Y_{t-1}]^2 = Var(Y_{t-1}) = Var(Y_t) = \gamma_0$$

Por tanto, $\gamma_0 = \phi_1^2 \gamma_0 + \sigma^2$, enotnces $\gamma_0 = \frac{\sigma^2}{1-\phi_1^2}$
Para que un proceso sea estacionario, es necesario que $|\phi_1| < 1$.

Procesos de Medias Móviles: MA (q)

El proceso de **medias móviles de orden q** o **MA(q)** son procesos compuestos por una combinación lineal de ruido blanco, a_t, con media 0 y varianza contante σ_a^2. Es decir, una serie temporal Y_t sigue dicho modelo si

$$Y_t = a_t + \theta_1 a_{t-1} + \theta_2 a_{t-2} + \ldots + \theta_q a_{t-q} = \sum_{j \geq 0}^{q} \theta_j a_{t-j} \quad q = 1, 2, \ldots, k$$

donde hay q retardos y $\theta_1, \theta_2, \ldots, \theta_q$ son coeficientes no nulos.

Como en el modelo anterior, en este también podemos hacer uso del operador retardo para definirlo:

$$Y_t = \theta(B) a_t,$$

siendo $\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \ldots + \theta_q B^q$.

A diferencia de los modelos AR(p), estos son siempre estacionarios.

Procesos Autorregresivos de Medias Móviles : ARMA (p,q)

El proceso **autorregresivo de Medias Móviles** o **ARMA(p,q)** es un proceso mixto en el que se combinan los modelos AR(p) y MA(q) vistos anteriormente. Así pues, lo definimos de la siguiente forma:

Una serie Y_t con $t = 1, 2, \ldots, T$ es un modelo $ARMA(p,q)$ si es estacionario, y se verifica que:

$$Y_t = \phi_1 Y_{t-1} + \ldots + \phi_p Y_{t-p} + a_t + \theta_1 a_{t-1} + \ldots + \theta_q a_{t-q} \quad p, q = 1, 2, \ldots, k$$

donde ϕ_p y θ_q y $\sigma_w^2 > 0$. Y a_t es ruido blanco.

Como es de esperar, este modelo también lo podemos definir en función del operador retardo :

$$\phi(B) Y_t = \theta(B) a_t,$$

donde $\phi(B)$ y $\theta(B)$ son los operadores definidos anteriormente.

Como este modelo contiene una parte autorregresiva, y en dichos modelos se requeria un control de los parámetros para determinar si el modelo era estacionario o no, es de esperar que en este modelo también existan ciertas condiciones, estas vienen dada por el siguiente teorema:

Teorema 2.1. *Un proceso autorregresivo de medias móviles finito ARMA(p,q) es estacionario sí y solo sí el módulo de las raíces del polinomio autorregresivo $\phi(B)$ está fuera del disco unidad.*

Tras un pequeño recorrido por los modelos estacionarios, damos pie a los no estacionarios.

2.3 Series no estacionarias

Los modelos propuestos anteriormente, tienen como característica común la *estacionariedad*. Pero, en numerosas ocasiones, las series temporales presentan una *tendencia determinista* o la varianza no permanece constante, *varianza heterocedástica*. Según en que caso nos encontremos la serie tendrá propiedades distintas, y la forma de transformación en estacionaria será distinta. En los siguiente apartados lo vemos con más detalle.

No estacionariedad en media

Una de las características más notables en una serie temporal es la presencia tendencia. Cuando haya presencia de dicho comportamiento la serie no será estacionaria, puesto que no evolucionan a un nivel constante, es decir, su valor medio cambia con el tiempo.

Existen diferentes formas de modelar una serie no estacionaria en media. La forma más sencilla de modelarla es:

$$Y_t = \mu_t + Z_t$$

donde μ_t modela la tendencia y Z_t es una serie estacionaria. Podemos estimar la tendencia siguiendo el siguiente modelo,

$$\mu_t = \alpha_0 + \alpha_1 t$$

y usar un modelo para la serie temporal de la siguiente forma:

$$Z_t = \alpha_0 + \alpha_1 t + a_t,$$

donde a_t es ruido blanco y α_0 y α_1 los podemos determinar usando mínimos cuadrados.

En este modelo se puede decir que la tendencia es **determinista** y Z_t podría ser un modelo ARMA.

Otro modelo se da cuando la tendencia es **estocástica**, en este caso podemos modelar la tendencia de la forma:

$$\mu_t = \delta + \mu_{t-1} + a_t,$$

Este modelo es conocido como **paseo aleatorio**. Es realmente un modelo AR(1) con $\phi_1 = 1$ (anteriormente vimos que no era estacionario).

Si el modelo es correcto , diferenciando los datos Y_t , obtenemos una serie estacionaria, es decir,

$$Y_t - Y_{t-1} = n(\mu_t + Z_t) - (\mu_{t-1} - Y_{t-1}) = \delta + X_t + a_t$$

donde $X_t = Z_t - Z_{t-1}$ es estacionaria.

Como hemos visto antes, podemos modelar la serie dentro de la clase de los modelos ARMA, por tanto, para determinar que la serie sea estacionaria, podemos usar los criterios vistos en la sección anterior.

Para determinar si una serie es estacionaria en media podemos hacer uso del **test ADF (test de Dickey- Fuller aumentado)**, donde se tiene como hipótesis nula que la serie tiene raices unitarias (es decir $\phi < 1$, lo que implica que no hay estacionaridad, y lo denotaremos como **I(1)**) y como hipótesis alternativa que no las hay (denotamos por **I(0)**, se da la estacionaridad).

Teniendo en cuenta estos tipos de no estacionariedad podemos definir los modelos ARIMA (p,d,q).

Modelos ARIMA(p,d,q): Dada una diferenciación de nivel d, se denomina *ARIMA(p,d,q)* al modelo definido como se sigue :

$$\phi_p(B)(1-B)^d Z_t = \theta_0 + \theta_q(B) a_t$$

donde ϕ_p es el operador estacionario AR y $\theta_q(B)$ es el operador MA.

No estacionariedad en varianza

Hay ocasiones en la que la varianza no se mantiene estable en toda la serie y crece con el tiempo, cuando esto ocurre se dice que la serie es **no estacionaria en varianza**. Para determinar si la serie es o no estacionaria en varianza tenemos el *test de Kwiatkowski, Philips, Schmidth, Shin (KPSS)*.

Contraste KPSS: Este contraste plantea como hipótesis nula que la serie no es estacionaria. El método se basa en estudiar la varianza de u_t en los siguiente modelos no observables :

- Modelo 1:

$$Y_t = \psi t + r_t + a_t$$
$$r_t = r_{t-1} + a_t$$

donde el valor inicial r_0 es fijo, a_t estacionario y u_t es $iid(0, \sigma_u)$.

Si la serie es estacionaria en torno a una tendencia determinista, $\sigma_u = 0$, en caso contrario sigue un paseo aleatorio.

- Modelo 2:

$$Y_t = r_t + w_t$$
$$r_t = r_{t-1} + u_t$$

En este caso la hipótesis nula es estacionariedad frente a paseo aleatorio.

Normalmente para estabilizar la varianza se usan las transformaciones de Box-Cox:

$$Y_t^{(\lambda)} = \begin{cases} \frac{Y_t^{\lambda}-1}{\lambda} & \lambda \neq 0 \\ ln(Y_t) & \lambda = 0 \end{cases} \tag{2.1}$$

Series estacionales

Anteriormente, hemos definido el concepto de *serie estacional*. Ahora bien, en esta sección describiremos un modelo adecuado para este tipo de series.

Así pues una serie describe un comportamiento estacional de periodo s si existen patrones en la serie estudiada cada s intervalos.

Para la explicación y la determinación de un modelo con estas características tendrá gran importancia el operador retaro B^s, siendo este el resultado de aplicar el operador retardo B s veces, es decir, $B^s Y_t = Y_{t-s}$.

En este tipo de series las observaciones dependen de las observaciones anteriores y de las observaciones en el periodo anterior (si por ejemplo tomamos datos mensuales, si tomamos el dato de un mes, tendremos en cuenta los meses anteriores y las observaciones de ese mes en años anteriores).

Luego para relacionar estas observaciones proponemos un modelo de la siguiente forma:

$$\Phi_P(B^s)(1-B^s)^D Z_t = \Theta_Q(B^s)\alpha_t$$

Como los α_t pueden estar correlaccionados, se introduce un segundo modelo de la forma

$$\phi(B)\alpha_t = \theta(B)a_t$$

siendo a_t ruido blanco. Ahora, sustituimos la segunda ecuación en la primera dando lugar a un modelo multiplicativo $ARIMA(p,d,q) \times ARIMA(P,D,Q)_s$

$$\Phi_(B^s)\phi_p(B)(1-B^s)^D Z_t = \Theta_Q(B^s)\theta_q(B)a_t.$$

Así pues tenemos:

- La componente ARIMA(p,d,q) modela la dependencia regular,que es la dependencia asociada a observacones consecutivas.
- La componenete ARIMA(P,D,Q) modela de dependencia estacional, que está asociada a observaciones asociadas a s periodos.

2.4 Series multivariantes

Como hemos dicho anteriormente, las series multivariantes estudian el comportamiento de varias series a la vez, por tanto tenemos series *k-dimensionales*, así pues procedemos a definir dicho concepto.

Definición 2.8. *Una **serie k-dimensional** se define como una serie de la forma* $\boldsymbol{Y}_t = (Y_{1t},\dots,Y_{kt})'$*, en la que cada* Y_{it}*, con* $i = 1,2,\dots,k$*, es una serie temporal unidimensional.*

Como en el caso univariante, tiene gran interés el comportamiento de las series, por lo que pasamos a dar alguna definición.

Definición 2.9. *Sea* $\boldsymbol{Y}_t$ *una serie k-dimensional de la forma definida anteriormente, se dice que esta es **débilmente estacionaria** si cumple que* $E[\boldsymbol{Y}_t] = \mu$ *y* $Cov[\boldsymbol{Y}_t] = \sum_z$ *es una matriz contante* $k \times k$ *definida positiva.*

Definición 2.10. *Una serie k-dimensional* $\boldsymbol{Y}_t$ *se dice **fuertemente estacionaria** si la distribución conjunta de un conjunto* $(Y_{t_1},\dots,Y_{t_m})$ *es la misma que la de* $(Y_{t_{1+j}},\dots,Y_{t_{m+j}})$ *donde* m*,* j *y* $t_1,\dots,t_m$ *son enteros positivos.*

Modelos VAR

Los *modelos de autorregresión vectorial* son una extensión del modelo autorregresivo univariante a series multivariantes. Este es un modelo estacionario, no obstante también tiene su utilidad para la no estacionariedad, puesto que podemos usar modelos VAR que incorporan relaciones de cointegración, en los cuales profundizaremos más adelante.

Tomamos $\mathbf{Y}_t = (Y_{1t},\dots,Y_{kt})$ una serie k-variante, se dice que esta sigue un modelo **VAR(p)** si la podemos expresar de la siguiente forma

$$\mathbf{Y}_t = \boldsymbol{\Phi}_1\mathbf{Y}_{t-1} + \dots + \boldsymbol{\Phi}_p\mathbf{Y}_{t-p} + \mathbf{C}\mathbf{D}_t + \mathbf{a}_t \quad con \quad t = 1,2,\dots,T$$

donde $\boldsymbol{\Phi}_i$ para $i = 1,\dots,p$ es una matriz de coeficientes $k \times k$, $\mathbf{a}_t$ es un vector k-dimensional de ruido blanco con matriz de covarianzas $\sum_a$ invariante en el tiempo y definida positiva, $\mathbf{C}$ una matriz con coeficientes de dimensión $k \times m$ que representa la componente determinista y, por último un vector $m \times 1$, $\mathbf{D}_t$ de los regresores deterministas existentes (constante, tendencia o estacionalidad).

Como ocurre en los modelos univariantes, podemos dar una definición desde otro punto de vista, basandonos en un polinomio de retardo. Por lo tanto si consideramos dicho polinomio de la siguiente forma

$$\boldsymbol{\Phi}(B) = \mathbf{I}_k - \boldsymbol{\Phi}_1 B - \dots - \boldsymbol{\Phi}_p B^p,$$

donde $\mathbf{I}_k$ es la matriz identidad de orden k, podemos definir el modelo tratado como se sigue

$$\boldsymbol{\Phi}(B)Y_t = \mathbf{C}\mathbf{D}_t + \mathbf{a}_t$$

Como hemos dicho, los modelos **VAR** son estacionarios, pero estos tienen que cumplir una condicción necesaria y suficiente, esta es que las soluciones de la ecuación

$det(\mathbf{I} - \mathbf{\Phi}B) = 0$ estén fuera del disco unidad, o de forma análoga $det(\mathbf{\Phi}(B)) = 0$ sean en módulo mayor que la unidad.

Cuando se estima un modelo es conveniente realizar una validación del mismo. Para ello, se estudian los residuos del modelo, los cuales denotamos por $\hat{u}_t$. Las características que deben cumplir los residuos para una validación del modelo son, la ausencia de correlación, ver si estos se distribuyen normalmente y la presencia de heterocedasticidad.

Test de Portmanteau: Esta prueba es la que utilizaremos para determinar la existencia de correlación en los residuos .

El estadístico aplicado en el test de Portmanteau se define como

$$Q_h = T \sum_{j=1}^{h} tr(\hat{C}_j' \hat{C}_0^{-1} \hat{C}_j \hat{C}_0^{-1})$$

con $\hat{C}_i = \frac{1}{T} \sum_{t=i+1}^{T} \hat{u}_t \hat{u}_{t-i}'$. El estadśtico se aproxima por una distribución Chi-cuadrado, $\chi^2(K^2h - n^*)$ tal que n^* es el número de coeficientes que excluyen los términos deterministas de un modelo VAR (p).
El planteamiento de contraste de hipótesis es análogo al del *test de Box-Pierce*.

Test multivariante ARCH-LM: En el caso el test es usado para determinar la heterocedasticidad de los residuos. Tenemos como hipótesis nula que las series son homocedásticas, y como hipótesis alternativa que hay heterocedasticidad.

El estadístico para este test es:

$$VARCH_{LM}(q) = \frac{1}{2} TK(K+1)R_m^2$$

definiendo

$$R_m^2 = 1 - \frac{2}{K(K+1)} tr(\hat{\Omega}\hat{\Omega}_0^{-1}),$$

con $\hat{\Omega}$ matriz de covarianza del modelo de regresión dado anteriormente. El estadístico sigue una distribución $\chi^2(\frac{qK^2(K+1)^2}{4})$.

Test de normalidad de Jarque-Bera: Determina la normalidad de los residuos.

Este test se aplica, tanto a casos univariantes, como a casos multivariantes. También se aplican pruebas separadas para la asimetría y los kurtosis multivariadas. El estadístico para el caso multivariante es:

$$JB_{mv} = s_3^2 + s_4^2,$$

con

$$s_3^2 = \frac{T b_1' b_1}{6}$$

$$s_4^2 = \frac{T(b_1 - 3_k)(b_2 - 3_k)}{24}$$

de forma que b_1 y b_2 son vectores son vectores de momentos de orden tres y cuatro no centralizados de los residuos estandarizados $u_t^s = \bar{P} - (\hat{u}_t - \bar{\hat{u}}_t)$ donde $\bar{P}$ es una matriz triangular inferior con diagonal positica tal que $\bar{P}\bar{P}' = \bar{\sum}_u$ (descomposición de Cholesky de la matriz de covarianza residual). El estadśtico definido sigue una distribución $\chi^2(2K)$ y la asimetrá (s_3^2) y los kurtois(s_4^2) una $\chi^2(K)$.

Tenemos como hipótesis nula que los residuos siguen una distribución normal y como alternativa el caso contrario.

Para obtener más información sobre estos contrastes veáse [3, pág 28]

Cointegración

Hay ocasiones en las que las series trabajadas no son estacionarias, pero sí comparten una tendencia estocástica común, es este caso se dice que las series están cointegradas, demos una definición formal de este término.

Definición 2.11. *Sea $\mathbf{Y}_t$ de series no estacionarias en media, es decir son series I(1), entonces $\mathbf{Y}_t$ es **cointegrado** si existe un vector*

$$\boldsymbol{\beta}'\mathbf{Y} = \beta_1 Y_{1,t} + \beta_2 Y_{2,t} + \ldots + \beta_n Y_{n,t} \backsim I(0)$$

*a esta ecuación se le denomina **equilibrio a largo plazo**.*

Este vector de cointegración β no es único puesto que si se hacen combinaciones lineales de este la serie resultante sigue siendo estacionaria.

Una representación normalizada suele ser $\beta = (1, -\beta_2, \ldots, -\beta_n)$, por lo que la relación de cointegración queda de la siguiente forma

$$\beta'\mathbf{Y} = Y_{1,t} - \beta_2 Y_{2,t} - \ldots - \beta_n X_{n,t} \backsim I(0).$$

Así, las relaciones de cointegración pueden ser expresadas de este modo

$$Y_{1,t} = \beta_2 Y_{2,t} + \ldots + \beta_n y_{n,t} + u_t,$$

donde $u_t \backsim I(0)$, al cual llamamos **error de equilibrio** o **residuo de cointegración**.

En un vector $\mathbf{Y}_t$ de dimensión n cointegrado puede haber $0 < r < n$ vectores de cointegración linealmente independientes.

Como hemos dicho al inicio de la sección, aunque las series no sean estacionarias, estas tienen una tendencia común.

Si un vector con dimensión n está cointegrado con $0 < r < n$ vectores de cointegración, entonces hay $n - r$ tendencias estocásticas comunes.

Para hacer un estudio de cointegración existen dos puntos de vista, el dado por Engle y Granger en 1986, que desarrollaron un procedimiento de prueba basado en residuos de dos pasos basado en técnicas de regresión el cual trata de ver si hay al menos un vector de cointegración. Y en segundo lugar, desarrollada por Johansen en 1988, la cual determina el número de vectores de cointegración existentes.

La metodologá de Engle y Granger se basa en calcular el residuo de cointegración $\beta_0 Y_t = u_t$ y determinar si u_t es I(0).

La hipótesis nula de este procedimiento es la no presencia de cointegración frente a la alternativa de cointegración.

Se pueden dar dos posibilidades:

- β **conocido**, es decir, por la estructura del problema está preestablecido, y no necesitamos hacer una estimación. En este caso el contraste de hipótesis planteado es el que sigue:

$$H_o : u_i = \beta' \mathbf{Y}_t \backsim I(1) \quad \text{(no cointegración)}$$
$$H_1 : u_i = \beta' \mathbf{Y}_t \backsim I(0) \quad \text{(cointegración)}$$

  Aplicandolo cualquier test de hipótesis para evaluar las dadas, nosotros trabajamos con el **test ADF**, puesto que es del que hemos hablado durante el transcurso del trabajo.

- β **desconocido** , es necesario hacer una estimación. Primero, es necesario especificar una normalización para que el modelo se identifique de forma única.

  Sea $\mathbf{Y}_t = (Y_{1,t}, \mathbf{Y}_{2,t})$, tal que $\mathbf{Y}_{2,t} = (Y_{2,t}, \ldots, Y_{n,t})'$.

  En este caso el vector de cointegración queda de la siguiente forma $\beta = (1 - \beta_2')$.

  Cuando el β era conocido el contraste era directo, en este caso primero debemos estimar β_2 por mínimos cuadrados de la regresión $Y_{1,t} = c + \beta_2' \mathbf{Y}_{2,1} + u_t$, para luego hacer un contraste de cointegración con un test de raices unitarias del residuo de cointegración $\hat{u}_t = Y_{1,t} - \hat{c} - \hat{\beta}_2 \hat{Y}_{2,t}$.

  En 1990 Phillips y Ouliards encontraron una disteibución apropiada para la realización de este contraste, estos le dieron el nombre a dicho contraste denominandolo **contraste de Phillips- Ouliarts**. En este caso se mantienen las hipótesis del anterior. Podemos encontrar más infomración de este contraste en [5, pág 445]

El metodo de Granger relaciona la cointegración con modelos de corrección de errores. En lo que sigue nuestro objetivo será vincular la cointegración con modelos de correción de errores en un contexto autorregresivo,

Anteriormente hemos definido los modelos **VAR**, y hemos visto que era estable si cierto determinante tenia todas sus raíces fuera del disco unidad, si hubiera alguna de estas (o todas), en este caso podríamos estudiar la cointegración.

En caso de que la serie estuviera cointegrada dicho modelo no sería un modelo adecuado puesto que hay relaciones que no se contemplan. Estas relaciones aparecen cuando transformamos este modelo a un modelo **VECM**, (Modelo de correción de errores vectoriales), el cual viene dado por la siguiente expresión :

$$\Delta \mathbf{Y}_t = \mathbf{C}\mathbf{D}_t + \mathbf{\Phi}\mathbf{Y}_{t-1} + \mathbf{\Gamma}_1 \Delta \mathbf{Y}_{t-1} + \ldots + \mathbf{\Gamma}_{p-1} \Delta \mathbf{Y}_{t-p+1} + a_t,$$

donde

$$\mathbf{\Phi} = \mathbf{\Phi}_1 + \ldots + \mathbf{\Phi}_p - \mathbf{I}_n,$$

$$\mathbf{\Gamma}_k = \sum_{j=1}^{k+1} \mathbf{\Phi}_j, \quad k = 1, 2, \ldots, p-1.$$

La matriz $\mathbf{\Phi}$ se denomina **matriz de impacto a largo plazo** y $\mathbf{\Gamma}$ se le llama **matices de impacto a largo plazo**.

Podemos establecer una relación entre un modelo **VAR** y otro **VECM**, es decir, podemos construir un modelo **VAR** a partir de uno **VECM** estableciendo la siguiente relación:

$$\mathbf{\Phi}_1 = \mathbf{\Gamma}_1 + \mathbf{\Phi} + \mathbf{I}_k$$

$$\mathbf{\Phi}_k = \mathbf{\Gamma}_k - \mathbf{\Gamma}_{k-1}, \quad k = 1, 2, \ldots, p.$$

En este modelo , los $\Delta \mathbf{Y}_t$ y todos sus retardos son estacionarios, es decir I(0), en el término $\mathbf{\Phi Y}_{t-1}$ es en el que se recogen las relaciones de cointegración, siempre que haya presencia de estas.

Si el modelo **VECM** no esta cointegrado este se reduce a un modelo **VAR(p-1)** en primera diferencia.

Anteriormente hemos mencionado que, no lo podemos ver si hay al menos un vector de cointegración, si no que también ver el número de vectores de cointegración que hay, por lo tanto vamos a pasar a hacer un recorrido por la metodologá de Johansen.

Esta metodología sigue los siguientes pasos:

- Especificar y estimar un modelo **VAR(p)**.
- Construir un test de razón de verosimilitudes.
- Imponer condiciones de normalización para poder identificar los vectores de cointegración.
- Estimar el modelo **VECM** cointegrado por máxima verolimilitud.

Para una explicación más extensa de este método consultese [5, pág 445].

3 Software R

R es un entorno de software y programación destinado a la estadística que fue desarrollado por Robert Gentleman y Ross Ihaka del Departamento de Estadística de la Universidad de Auckland (Nueva Zelanda) en 1993. Este es una reimplementación del lenguaje S, con la particularidad de que se distribuye bajo licencia GNU (Licencia Pública General), es decir, es un software libre (los desarrolladores pueden descargar el código de forma gratuita, modificarlo para incluir mejoras, es gratuito,...).

Como ya hemos dicho, **R** tiene su mayor uso en la estadística. Por tanto, posee un gran abanico de herramientas que nos ayudan con esta tarea (estadística descriptiva, test estadísticos, modelos de regresión, distribuciones de probabilidad, algoritmos de clasificación y agrupamiento, herramientas para el análisis de series temporales, ...). También tiene capacidad de generar gráficas de alta calidad, lo cual nos resultará de gran utilidad para el desarrollo de nuestro trabajo.

Puesto que **R** es un lenguaje de programación, los usuarios lo pueden extender con sus propias funciones, aunque también puede extenderse a través de paquetes gratuitos desarrollados por su comunidad de usuarios.

R también posee su propio formato para la documentación basado en LaTeX (haciendo uso del paquete knitr), del cual hemos hecho uso para la realización de la memoria de este trabajo.

Usaremos como Entorno de Desarrollo Integrado *Rstudio*.

3.1 Formato de los datos temporales en R

R tiene unos formatos especiales para manejar información temporal más adecuadamente, ya que el paquete básico de **R** está muy limitado. Para esto existen paquetes y herramientas que nos facilitan el trabajo y nos ayudan a enfrentarnos a situaciones más complejas. Dedicaremos esta sección a introducir estos recursos.

Algunos objetos temporales en R

- **irts** (paquete tseries): Maneja series temporales irregulares.
- **ts** (paquete stats): Trabaja con series temporales regulares.
- **zoo** (paquete zoo): Ordena totalmente los datos temporales, tanto regulares como irregulares.
- **xts** (paquete xts): Tiene como base el objeto **zoo**, por lo que se puede considerar una extensión de este con modificaciones.

  Maneja los datos por un lado, mientras que la información de orientación temporal la codifica como un índice.

Algunas clases

Podemos representar de varias fomas los datos temporales en R, para ello tenemos algunas clases como:

- **Date** (paquete base): Hace una representación de las fechas como el número de días desde 1970-01-01.

  En esta clase no se trabaja con tiempo (horas, minutos, segundos), si no que tan solo se puede especificar hasta el día. Para ello trabajamos con los siguientes códigos de formato:

  - %d, día de mes en formato númerico.
  - %m, mes en formato númerico.
  - %b, mes en fomrato carácter abreviado.
  - %B, mes en formato carácter completo.
  - %y, año en formato númerico con dos dígitos.
  - %Y, año en formato númerico con cuatro dígitos.

- **chron** (paquete chron) : Representa fechas y horas (con signo) como el número de segundos transcurridos desde el inicio de 1970 en un vector. No maneja zona horaria.

- **POSIXct** (paquete base): Similar a chron, pero en este caso si maneja zonas horarias.

- **POSIXlt** (paquete base): Tiene dos subclases:

  - **POSIXct** : Representa las fechas desde la media noche GMT (meridiano de Greenwich) del 01-01-1970 hasta la fecha.
  - **POSIXlt**: Representa las fechas como una lista con elementos para segundos (sec), minutos (min), horas (hour), día del mes (mday), mes (mon), año (yday).

4 Análisis de los datos

4.1 Importación de los datos

Los datos han sido proporcionados a través de tablas de Excel, divididos por años (entendiendo esto como las reservas realizadas durante ciertos años), para importarlos necesitamos el paquete readxl, y usaremos la función read_excel. Vemos como ejemplo en código la importación de un año.

```
library(readxl)
Importes2010<-read_excel("Importes Reservas 2010.xlsx")
```

Lo hacemos de forma análoga para el resto de años, quedándonos sólo con los hoteles nacionales, y los juntamos con la función rbind, la cual combina vectores, matrices o *data.frame* por filas (también existe la función cbind, cuyo uso es el mismo, pero en vez de filas combina columnas). Con esto damos lugar al siguiente *data.frame*, al cual llamaremos Importes2.

```
## # A tibble: 6 x 6
##   RLOCA   RFECGR   RENTRA   RSALID RHOTEL SEL0006
##   <chr>    <dbl>    <dbl>    <dbl>  <dbl>   <dbl>
## 1 M3BU  20110329 20101008 20101009     30       0
## 2 X1VL  20080310 20100921 20100922    370   58000
## 3 1LKT  20080604 20100520 20100523    400   19240
## 4 3D4E  20080814 20100422 20100425    400    3643.
## 5 GCGU  20100518 20100520 20100523    400       0
## 6 GCGV  20100518 20100520 20100523    400       0
```

Tenemos 2.226.108 observaciones y 6 variables, las cuales son :

- **RLOCA**: Localizador de la reserva, el cual se usa como identificador de ésta.
- **FECGR**: Fecha de grabación, fecha en la que se hizo la reserva (si lo combinamos con el localizador, cada reserva quedará identificada de forma única).
- **RENTRA**: Fecha de entrada a la estancia.
- **RSALID**: Fecha de salida de la estancia.
- **RHOTEL**: Códido del hotel en el que el cliente se aloja.
- **SEL0006** : Coste total de la reserva.

En nuestro estudio, el localizador quedará en un segundo plano, puesto que no queremos profundizar en ninguna reserva particular asocida a un cliente, si no el estudio de forma general de los importes.

Como vemos, tenemos un importe por cada reserva asociado a una fecha de entrada y a una de salida, y con ello, este importe está asociado a x días. A nosotros no nos

interesa esto, sino lo que paga el usuario por cada día que está hospedado en el hotel, y puesto que nos es imposible acceder a estos datos, calculamos un importe medio diario. Pero aún así, ahora lo que tenemos es un importe medio asociado a esos x días. Para solventar esto crearemos una secuencia de días que vaya desde el día de entrada al de salida (teniendo en cuenta que el día de salida no cuenta como día de estancia), así tendremos un importe asociado a cada día de estancia, lo que se aproxima más a la realidad.

En primer lugar, creamos una nueva variable, a la que llamamos num_obs, la cual nos contabiliza el número de observaciones que hay, para ello tomamos un vector que vaya desde 1 hasta el número de filas que tiene el **data.frame** con el que estamos trabajando, y lo unimos a éste, creando así un nuevo *data.frame* con 7 varables, al que llamaremos ImportesG .

```
num_obs <- 1:nrow(Importes2)
ImportesG <- cbind(num_obs, Importes2)
```

Ahora bien, calculamos los días de estancia de cada reserva, para ello pasaremos a fecha, con las clases vistas anteriormente, calculamos la diferencia, teniendo así una nueva variable a la que llamamos diasestan, y luego diviendo el importe total de la reserva entre estos días obtenemos otra variable, que es la que finalmente queríamos, a la que llamamos Impmed. De forma que al final el *data.frame* quedará de la siguiente forma:

```
##   num_obs RLOCA   RFECGR   RENTRA   RSALID RHOTEL  SEL0006    ENTRADA
## 1       1  M3BU 20110329 20101008 20101009     30     0.00 2010-10-08
## 2       2  X1VL 20080310 20100921 20100922    370 58000.00 2010-09-21
## 3       3  1LKT 20080604 20100520 20100523    400 19240.00 2010-05-20
## 4       4  3D4E 20080814 20100422 20100425    400  3643.33 2010-04-22
## 5       5  GCGU 20100518 20100520 20100523    400     0.00 2010-05-20
## 6       6  GCGV 20100518 20100520 20100523    400     0.00 2010-05-20
##       SALIDA diasestan    Impmed
## 1 2010-10-09         1     0.000
## 2 2010-09-22         1 58000.000
## 3 2010-05-23         3  6413.333
## 4 2010-04-25         3  1214.443
## 5 2010-05-23         3     0.000
## 6 2010-05-23         3     0.000
```

Una vez tenido este *data.frame*, creamos la secuencia:

```
library(data.table)
ImportesG.F1 <- setDT(ImportesG)[,.(dias=seq(ENTRADA,SALIDA-1,
                                    length.out =diasestan)),
                                 by=num_obs]
```

Como se ve, hemos usado la función setDT, la cual pertenece al paquete data.table. Esta función hace una modificación por referencia, en nuestro caso tomamos como

referencia la variable num_obs y creamos una secuencia, que dará lugar a una variable llamada dias con inicio la fecha de entrada, final la fecha de salida menos uno (puesto que, como hemos mencionado antes, el día de salida no cuenta como día en el que el huésped está hospedado), y tomando como longitud de ésta la variable diasestan. Así pues quedaría de la siguiente forma:

```
##    num_obs       dias
## 1:       1 2010-10-08
## 2:       2 2010-09-21
## 3:       3 2010-05-20
## 4:       3 2010-05-21
## 5:       3 2010-05-22
## 6:       4 2010-04-22
```

Uniéndolo a nuestro *data.frame* original mediante la función merge, la cual fusiona columnas con el mismo argumento. Así pues ya tenemos listo el fichero de datos con el que trabajaremos.

```
ImportesG.F2 <- merge(ImportesG.F1 ,ImportesG)
head(ImportesG.F2)

##    num_obs       dias RLOCA   RFECGR   RENTRA   RSALID RHOTEL  SEL0006
## 1:       1 2010-10-08  M3BU 20110329 20101008 20101009     30     0.00
## 2:       2 2010-09-21  X1VL 20080310 20100921 20100922    370 58000.00
## 3:       3 2010-05-20  1LKT 20080604 20100520 20100523    400 19240.00
## 4:       3 2010-05-21  1LKT 20080604 20100520 20100523    400 19240.00
## 5:       3 2010-05-22  1LKT 20080604 20100520 20100523    400 19240.00
## 6:       4 2010-04-22  3D4E 20080814 20100422 20100425    400  3643.33
##       ENTRADA     SALIDA diasestan    Impmed
## 1: 2010-10-08 2010-10-09         1     0.000
## 2: 2010-09-21 2010-09-22         1 58000.000
## 3: 2010-05-20 2010-05-23         3  6413.333
## 4: 2010-05-20 2010-05-23         3  6413.333
## 5: 2010-05-20 2010-05-23         3  6413.333
## 6: 2010-04-22 2010-04-25         3  1214.443
```

Una vez tenemos los datos de la forma que deseamos los guardamos para su posterior uso con la función save, así podremos cargar el archivo cuando sea necesario (haciendo uso de la función load).

```
save(ImportesG.F2,file="ImportesG.RData")
```

Guardar y cargar datos en **R** puede tener gran utilidad cuando se trabaja con una gran cantidad de datos, puesto que estos se borran de la memoria. Así, aparte de ahorrar memoria, se ahorra tiempo de cómputo, ya que no es necesario hacer todos los cálculos otra vez.

4.2 Análisis univariante

Planteamineto del problema

Como vimos al describir las variables, cada reserva tiene asociada la fecha en la que fue realizada. Posteriormente creamos una variable especificando los días de alojamiento, estas variables tomarán su máxima importancia en nuestro estudio.

Tomaremos las reservas que tuvieron estancia durante cada año, y más tarde nos quedaremos con las que se grabaron hasta cierto día (normalmente para esto usamos el día actual, pero en nuestro caso tomamos hasta el último día de 2019 que actualizamos los datos). Teniendo, por tanto, dos variables, el importe facturado al cerrar un año, y el importe facturado hasta el día actual en otro año. Así podemos ver el incremento de facturación que hubo ese año, teniendo así una serie univariante sobre la que hacer nuestro análisis, que nos permitirá hacer previsiones de cual será ese incremento en este año. Vemos como lo hacemos (cabe destacar que antes de hacer esto nos hemos quedado solo con las variables de interés para este estudio).

```
Imp10 <- filter(ImportesG.F2,dias >= "2010-01-01" & dias <= "2010-12-31")
#Este es para el anio cerrado.
adh10<-filter(Imp10, RFECGR <= "20100328")
# "a dia de hoy.
```

Lo haremos de forma análoga para cada año, para después juntarlos y dar lugar a dos *data.frame*, al de año cerrado, lo llamaremos AC, y al de «hasta la fecha actual» adh.

Una vez llegamos a este punto, vamos a empezar a trabajar con los objetos temporales escritos anteriormente, pasamos ambos archivos a formato xts (puesto que tenemos datos irregulares).

```
library(xts)
adh.xts<-as.xts(adh$Impmed, as.POSIXct(adh$dias,format="%Y-%m-%d"))
Ac.xts <- as.xts(ImportesG.F2$Impmed,
                 as.POSIXct(ImportesG.F2$dias , format="%Y-%m-%d"))
```

Así tenemos el importe medio, asociado a un índice temporal que representa el día al que se asocia dicho importe.

Este formato nos es de utilidad puesto que el paquete xts nos permite usar funciones para la agrupación de datos temporales, tenemos apply.year (agrupa los datos de forma anual), apply.quearterly (por cuatrimestres), apply.monthly (de forma mensual), apply.weekly (semanales), apply.daily (agrupa de forma diaria), se puede agrupar en cuanto a su media en ese periodo, su suma, ...

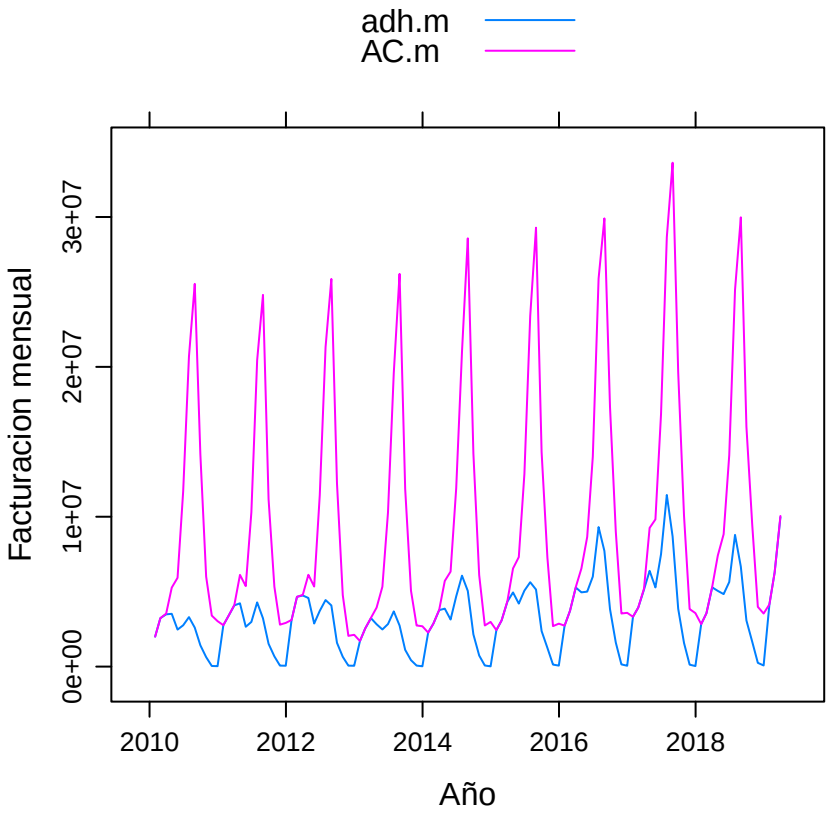
Nosotros, inicialmente trabajaremos con datos mensuales, aplicando su suma, teniendo en cuenta que tenemos flexibilidad para agruparlos como nos interese.

Una vez agrupados, juntamos ambos data.frame con la función merge.xts (también propia del paquete xts), la cual nos permite combinar dos *data.frame* por su fecha.

```
adh.m <- apply.monthly(adh.xts, FUN=sum)
AC.m <- apply.monthly(Ac.xts,FUN=sum)
ACvsADH <- merge.xts(adh.m, AC.m)
```

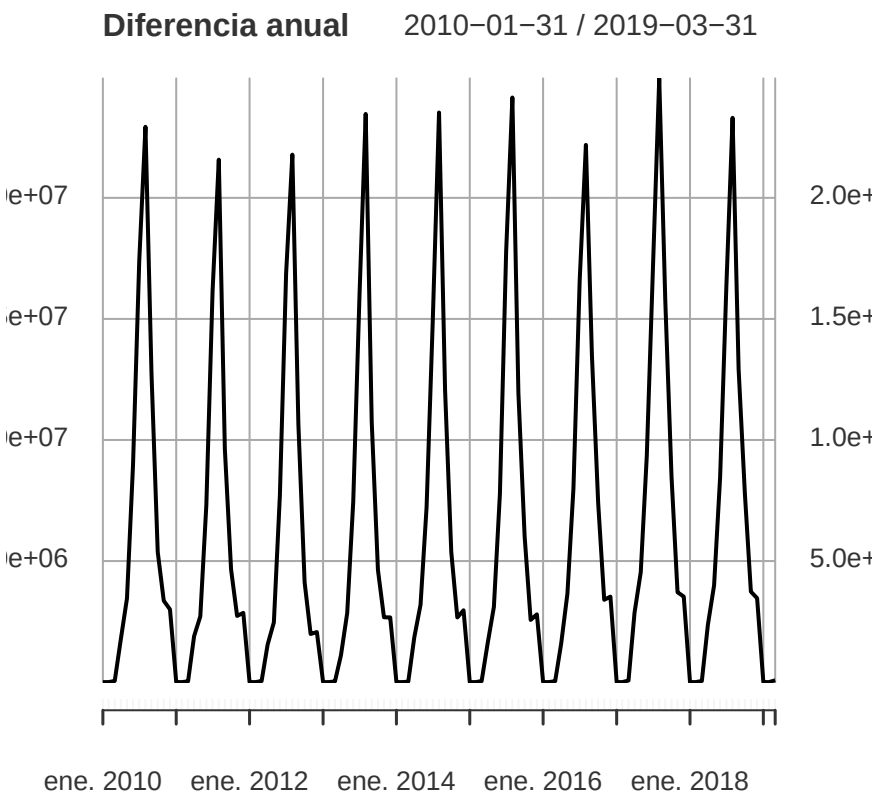
Lo vemos gráficamente, para ello necesitamos el paquete de visualización de datos lattice. Como vamos a representar un gráfico bivariante usaremos la función xyplot, la cual nos permite ver simultáneamente el comportamiento de varias variables en su secuencia temporal.

```
library(lattice)
xyplot(ACvsADH['2010/2019-03'],
       superpose=T,xlab= "Año", ylab="Facturacion mensual")
```



Ahora bien, como lo que queremos estudiar es el incremento de precio, calculamos la diferencia entre ambas series, teniendo así una serie univariante para nuestro estudio. Creamos un nuevo *data.frame* con esta diferencia al que llamamos Dif.anual. Vemos la representación de esta serie.

```
plot(Dif.anual['2010/2019-03'],main="Diferencia anual")
```



Estudio temporal

En esta sección vamos a hacer un estudio temporal completo de los datos en general (incluimos los ingresos totales, sin distinguir cada hotel), para mas tarde usando este mismo código y conceptos particularizar en algunas zonas o en algún hotel en particular, para sacar conclusiones distintas.

Antes de nada, podemos hacer un análisis descriptivo de nuestra variable, para ello usamos la función descr del paquete summarytools. Vemos los resultados de este análisis:

	Dif.anual
Mean	6250629.65
Std.Dev	7102880.44
Min	0.00
Q1	46573.88
Median	3361028.92
Q3	9335320.18
Max	24935522.23
MAD	4982891.35
IQR	9082751.81
CV	1.14
Skewness	1.18
SE.Skewness	0.23
Kurtosis	0.25
N.Valid	111.00
Pct.Valid	100.00

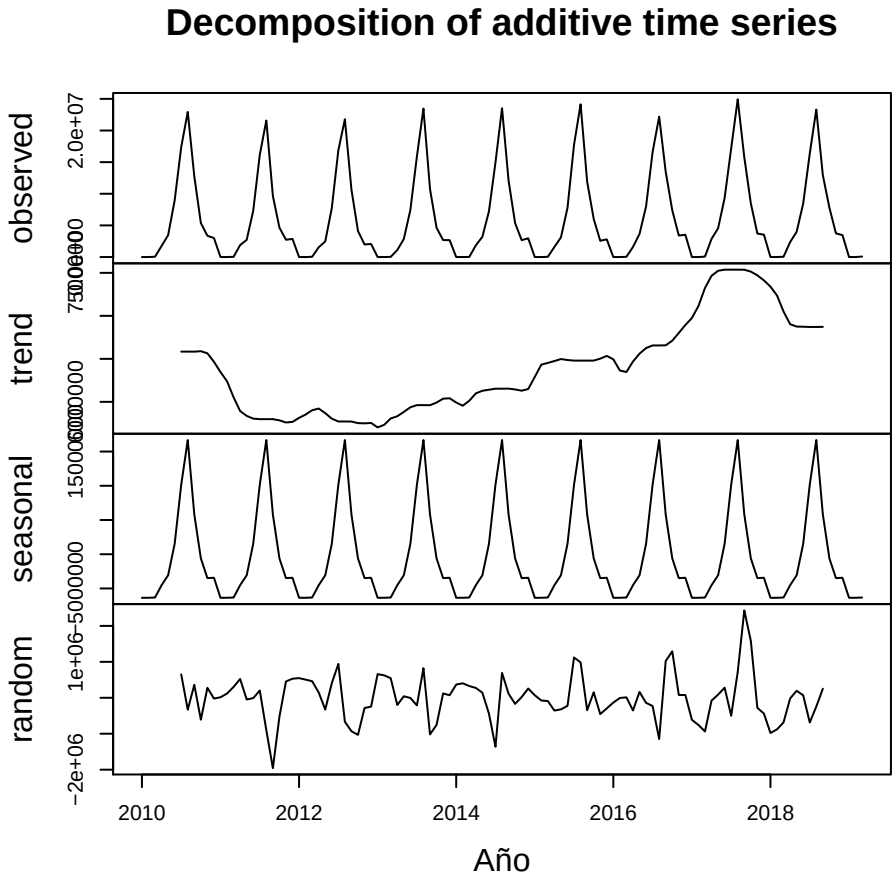
En esta tabla se muestra el número de observaciones (*N.Valid*), el rango de importe (*máx*), el promedio (*mean*), la mediana (*median*), la desviación estándar (*Std.Dev*), el grado de asimetría (*Skewness*) y su error (*SE.Skewness*) , el primer y terrcer cuantil (*Q1*, *Q3*), el coeficiente de variación (*CV*), la desviación media absoluta (*MAD*) y el intervalo intercuadrático (*IQR*).

Una vez realizado esto, pasamos al análisis temporal.

La representación gráfica de la diferencia de importe anual ya nos permite sacar algunas conclusiones. Vemos que ésta tiene un comportamiento similar en cada año, lo que nos da indicios de una regularidad temporal y, además, no parece tener una tendecia fuerte, no obstante, usamos la función descompose, la cual se usa para estimar la tendencia y la componente estacional de una serie temporal. Esta función no tiene buena compatibilidad (al igual que no la tendrá con otras funciones que usaremos posteriormente) con el formato xts. Entonces, puesto que tenemos los datos agrupados mensualmente, los tenemos de forma regular, así pues podemos usar el objeto ts (cuando nos encontramos algún otro caso en el que no haya compatibilidad, haremos el mismo proceso).

```
Dif.anual.desc = decompose(ts(Dif.anual['2010/2019-03'],
                              frequency = 12,start = 2010))
#Vemos que tomamos el comienzo en 2010 con una frecuencia de 12,
#puesto que los datos son mensuales.
```

Usaremos ahora la función plot, la cual produce una única figura donde se muestra la serie original (observed), la tendencia de la serie (trend), la componente estacional (seasonal), y por último la componente aleatoria (random).



Vemos que no hay una tendencia clara, y que hay una estacionalidad de 12 meses (anual), con esto reafirmamos lo que ya podíamos intuir.

Puede observarse, aparentemente, que la serie es estacionaria, para asegurarnos, vamos a aplicar test de estacionariedad, los cuales ya han sido definidos anteriormente. Aplicamos el **test ADF** (Dickey- Fuller) y el **test de KPSS** (Kwiatkowski-Phips-Schmidth-Shin).

En **R** existen varios paquetes que nos calculan estos contrastes, nosotros usaremos el paquete tseries, cada uno de los test con su respentiva función.

```
library(tseries)
adf<-adf.test(Dif.anual['2010/2019-03'])
adf$p.value
```

```
## [1] 0.01
```

Vemos que el p-valor es menor que 0.05, por tanto, se rechaza la hipótesis nula, y con esto la idea de que tenga raíces unitarias. Por tanto, la serie es estacionaria en media.

```
kpss<- kpss.test(Dif.anual['2010/2019-03'])
kpss$p.value

## [1] 0.1
```

El p-valor es mayor que 0.05, por lo que aceptamos la hipótesis nula, por tanto, la serie es estacionaria en varianza.

Una vez visto esto podemos decir que la serie es estacionaria. Trabajar con series estacionarias es interesante, puesto que, nos permiten obtener predicciones fácilmente, por tanto obtener series de este tipo siempre será nuestro objetivo.

En nuestro caso tenemos 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.

Podríamos contruir un conjunto amplio de modelos y quedarnos con el mejor, no obstante la herramienta **R** consta de la función auto.arima (paquete forecast). Esta función nos convierte, sin necesidad de que diferenciemos la serie manualmente, el mejor ARIMA(p,d,q) para nuestro modelo.

Vamos a hacer el ajuste y la predicción (esta se hace con la función forecast).

```
Selec <- Dif.anual['2010/2019-03']  #seleccionamos en base a que fechas
                                    #queremos realizar el modelo
```

```
library(forecast)
(mod<-auto.arima(ts(Selec[,1],frequency = 12,start = 2010)))

## Series: ts(Selec[, 1], frequency = 12, start = 2010)
## ARIMA(1,0,0)(0,1,1)[12]
##
## Coefficients:
##          ar1     sma1
##       0.5615  -0.4899
## s.e.  0.0837   0.1035
##
## sigma^2 estimated as 4.961e+11:  log likelihood=-1474.33
## AIC=2954.66   AICc=2954.92   BIC=2962.45

forecast(mod,h=12)
```

```
##          Point Forecast    Lo 80    Hi 80      Lo 95    Hi 95
## Apr 2019   2.320400e+06  1417786  3223015   939970.3  3700830
## May 2019   3.986672e+06  2951507  5021836  2403524.4  5569819
## Jun 2019   8.526832e+06  7453267  9600397  6884955.4 10168708
## Jul 2019   1.663524e+07 15549845 17720626 14975274.3 18295197
## Aug 2019   2.361846e+07 22529370 24707554 21952839.9 25284084
## Sep 2019   1.356226e+07 12472004 14652516 11894857.3 15229663
## Oct 2019   7.656545e+06  6565922  8747168  5988581.5  9324509
## Nov 2019   3.553587e+06  2462849  4644326  1885447.1  5221728
## Dec 2019   3.410790e+06  2320015  4501565  1742594.2  5078986
## Jan 2020   1.085021e+02 -1090677  1090894 -1668104.0  1668321
## Feb 2020   1.710840e+02 -1090618  1090960 -1668047.0  1668389
## Mar 2020   6.124173e+04 -1029549  1152032 -1606978.1  1729462
```

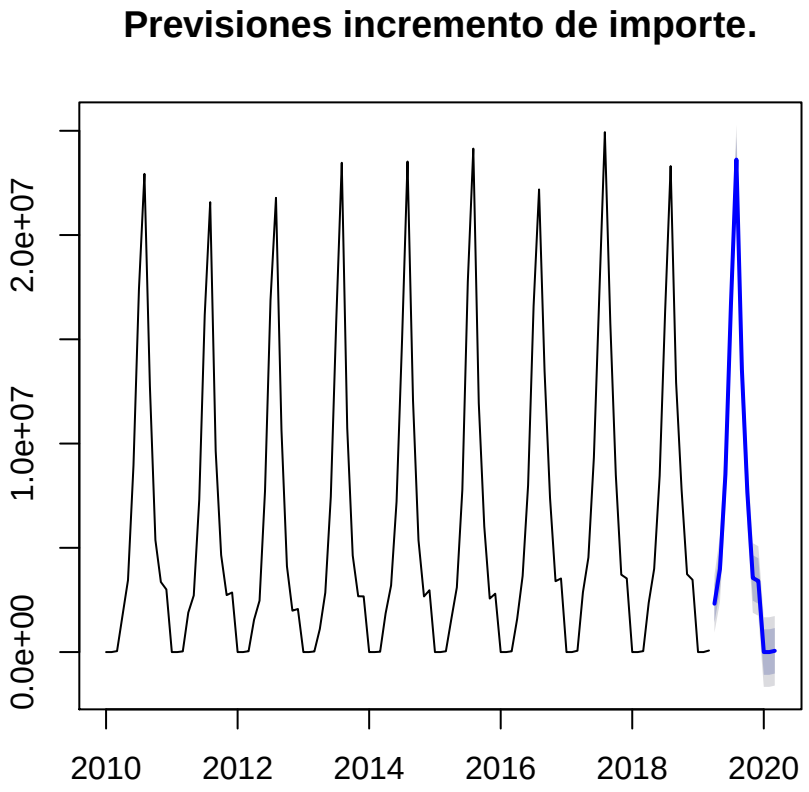
Cabe mencionar algunas observaciones respecto a los resultados.

El modelo propuesto es un $ARIMA(1,0,0)\times(0,1,1)_{12}$.

La primera parte nos muestra la parte no estacional y el segundo la estacional. Ahora bien, que $d = 0$ significa que no hemos tenido que diferenciar la serie puesto que la serie ya era estacionaria (lo hemos visto con un contraste de hipótesis) y que $D = 1$, esto es, la diferenciación respecto al año anterior que elimina la estacionalidad de la serie.

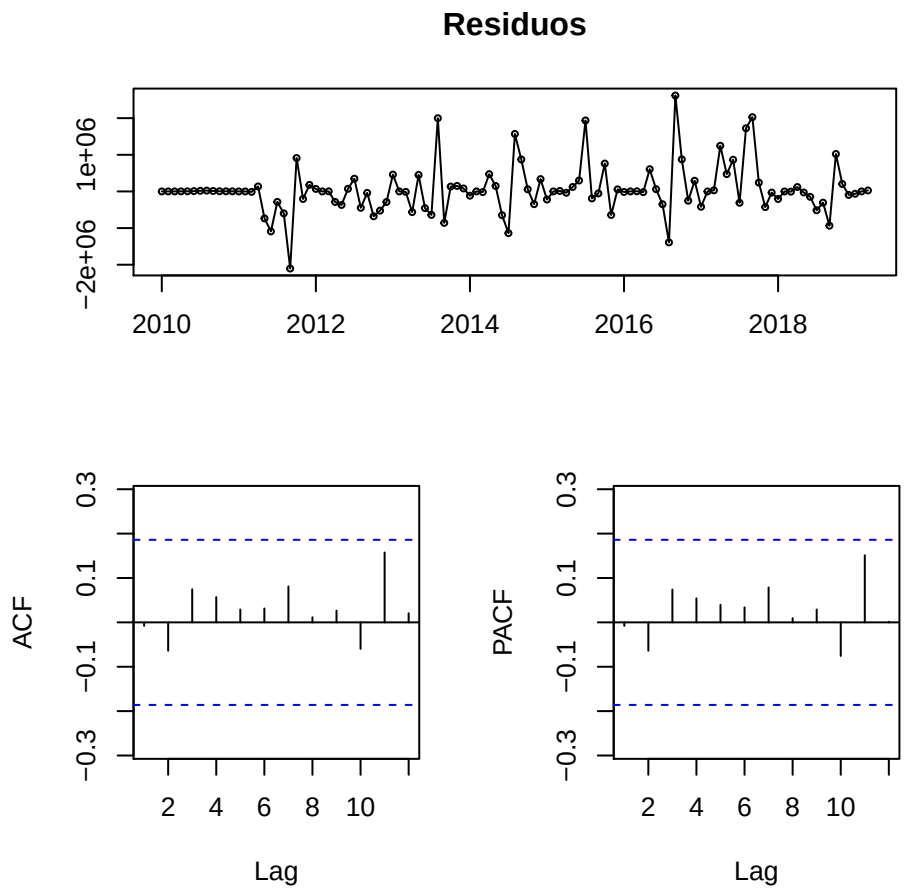
En la primera parte tenemos un AR(1) con respecto al mes anterior, y en la segunda, un MA(1).

La previsión ha sido hecha para los 12 meses siguientes (esto lo hemos impuesto nosotros tomando h=12, podemos tomar cualquier valor según el periodo de tiempo que nos intese preveer). Vemos que tenemos un punto de previsión , y luego unas previsiones entre intervalos de confianza del 80 y el 95 por ciento. Podemos ver esto gáficamente:



En la imagen vemos una linea azul, que representa la predicción, y las bandas de color grisaceo representan los intervalos antes mencionado.

Una vez llegados a este punto, nos faltaría validar el modelo, para ello tenemos que ver que los residuos son ruido blanco.



Se puede apreciar en los correlogramas que no hay ningún retardo significativo que denote ningún tipo de estructura, por lo tanto podemos decir que los residuos son ruido blanco.

También podemos ver esto con algunos contrastes de hipótesis, nosostros usaremos el test de Box- Pierce. Para ello, usaremos la función Box.test.

```
Box.Pierce<-Box.test(residuos_modelo)
Box.Pierce$p.value

## [1] 0.9347248
```

Vemos que el p-valor es mayor que 0.05 por tanto, aceptamos la hipótesis nula. Vemos entonces lo que el gráfico nos anticipaba, los residuos son ruido blanco, y como consecuencia, podemos validar el modelo.

Algunos ejemplos

En el análisis univariante que hemos hecho de forma detallada la serie era estacionaria, así pues vamos a poner algun ejemplos en lo que esto no se de viendo así las transformaciones necesarias que se deben hacer así como sus códigos en **R**. Usaremos

el código que anteriormente hemos explicado con más detenimiento haciando alguna modificación que mencionaremos.

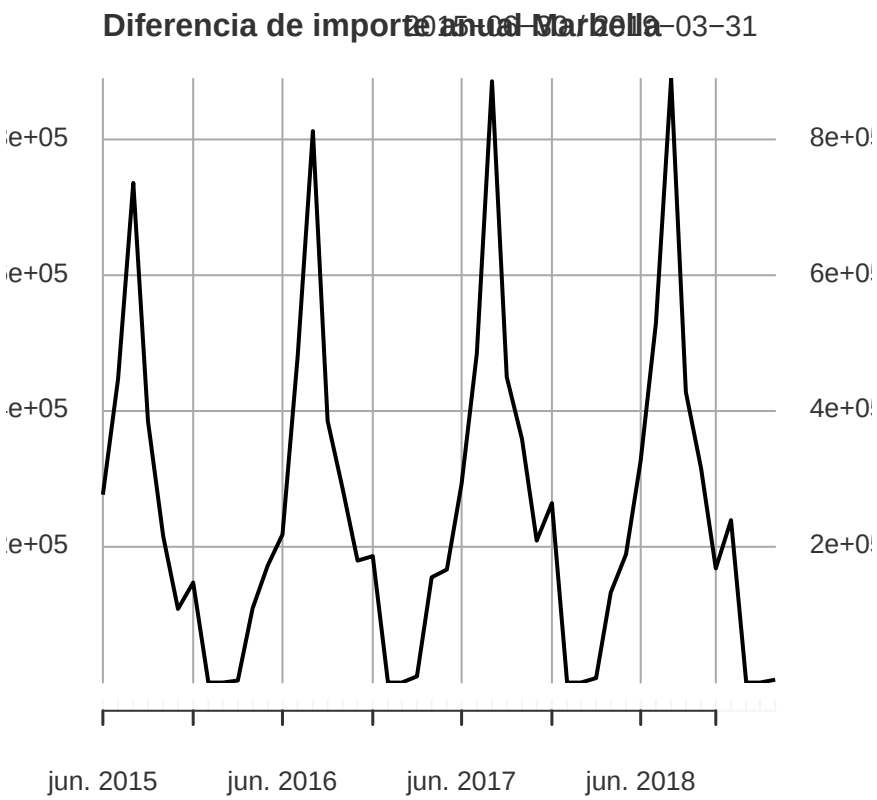
EJEMPLO NO ESTACIONARIDAD EN MEDIA

En este ejemplo vamos a particularizar un hotel, para esto usamos la función filter del paquete dplyr. Seleccionamos el hotel *Senator Marbella Spa Hotel*.En este caso el hotel seleccionado es el que tiene como código **350**, así pues lo filtramos. Vemos en este primer ejemplo como sería en código:

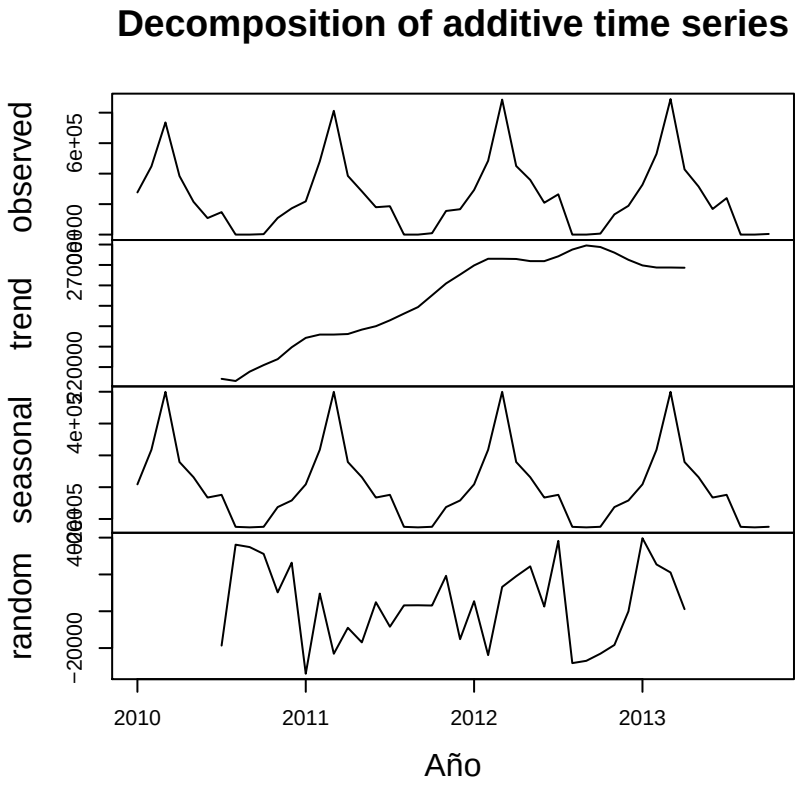
```
library(dplyr)
ImportesG.F2 <- filter(ImportesG.F2 ,RHOTEL=="350")
```

Una vez visto esto empezamos a trabajar a partir de la diferencia, que la serie que vamos a estudiar.

```
plot(Dif.anual.marbella,main="Diferencia de importe anual Marbella")
```



Observamos las componentes de la serie temporal:



En este caso observamos una tendencia ascendente. Aplicamos los contrastes de hipótesis pertinentes.

```
adf.marbella$p.value

## [1] 0.06287113

kpss.marbella$p.value

## [1] 0.1
```

En el **test ADF** el p-valor es mayor que 0.05 por lo que se acepta la hipótesis nula, es decir, la presencia de raíces unitarias y con esto la no estacionaridad es media. En el caso del **test KPSS** también aceptamos la hipótesis nula, pero en esta ocasión esta si que proporciona la estacionaridad de la varianza.

Tenemos una serie no estacionaria en media, pero sí en varianza, por lo que tenemos que diferenciar la serie para transformarla en estacionaria. Para transformar tenemos que diferenciar la serie la opción ndiffs en R nos permite saber cuantas veces es necesario diferenciar la serie para que esta sea estacionaria.

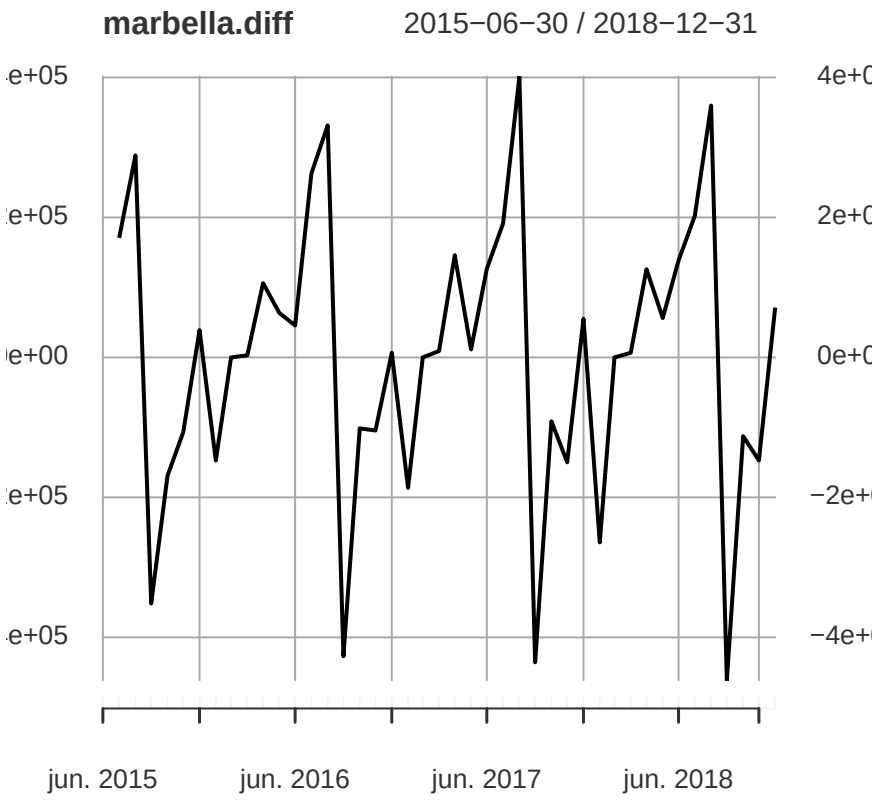
```
ndiffs(Dif.anual.marbella,test = "adf",type="trend")

## [1] 1

#Como lo que queremos eliminar es la tendencia, para que la serie
#sea estacionaria en media, le ponemos type=trend
```

Vemos que hay que hacer una diferencia, puesto que anticipamos que en nuestro modelo ARIMA el parámetro d será 1. Para diferenciar la serie usamos la función diff y la representamos gráficamente:

```
marbella.diff <- diff((Dif.anual.marbella['2015-06/2018-12']))
```



Ahora vamos a ver el modelo que sigue la serie. Como comentamos en el estudio temporal con la función auto.arima se elige un modelo sin necesidad de transformar la serie, entonces se la aplicamos a la serie original y comentamos el resultado.

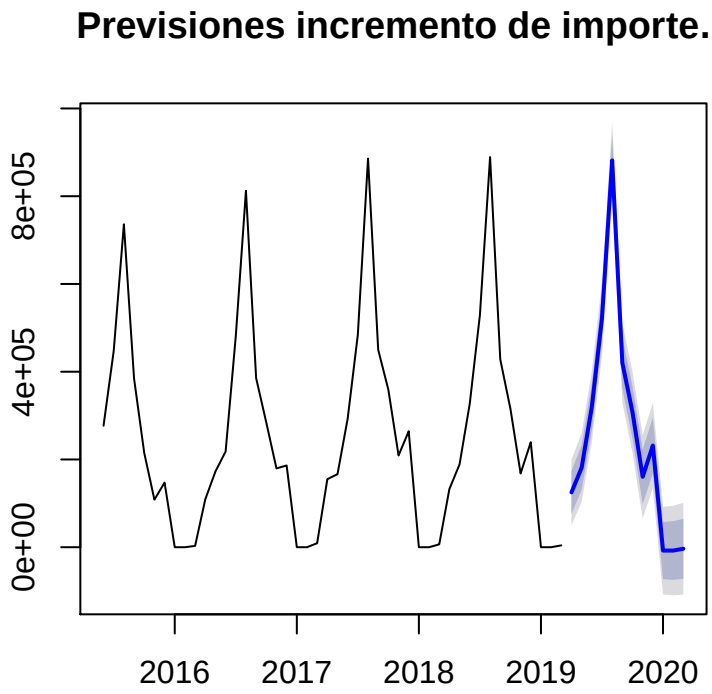
```
library(forecast)
(mod1<-auto.arima(ts(Dif.anual.marbella,frequency = 12,start = c(2015,6))))

## Series: ts(Dif.anual.marbella, frequency = 12, start = c(2015, 6))
## ARIMA(0,1,1)(0,1,0)[12]
##
## Coefficients:
##           ma1
##       -0.7070
## s.e.   0.2255
##
## sigma^2 estimated as 1.462e+09:  log likelihood=-394.86
## AIC=793.72   AICc=794.12   BIC=796.72

forecast(mod1,h=12)
```

```
##          Point Forecast     Lo 80     Hi 80      Lo 95     Hi 95
## Apr 2019     124974.507  75976.34 173972.67   50038.28 199910.73
## May 2019     181254.477 130196.28 232312.67  103167.71 259341.25
## Jun 2019     319583.253 266544.98 372621.53  238468.21 400698.29
## Jul 2019     521851.472 466904.43 576798.52  437817.22 605885.73
## Aug 2019     881659.251 824867.55 938450.95  794803.84 968514.66
## Sep 2019     420339.358 361761.07 478917.65  330751.59 509927.13
## Oct 2019     307745.768 247433.78 368057.75  215506.55 399984.99
## Nov 2019     160540.839  98543.62 222538.06   65724.28 255357.40
## Dec 2019     231715.775 168077.94 295353.61  134390.10 329041.45
## Jan 2020      -7630.620 -72867.83  57606.59 -107402.32  92141.08
## Feb 2020      -7630.620 -74428.92  59167.68 -109789.80  94528.56
## Mar 2020      -3444.596 -71768.32  64879.13 -107936.72 101047.53
```

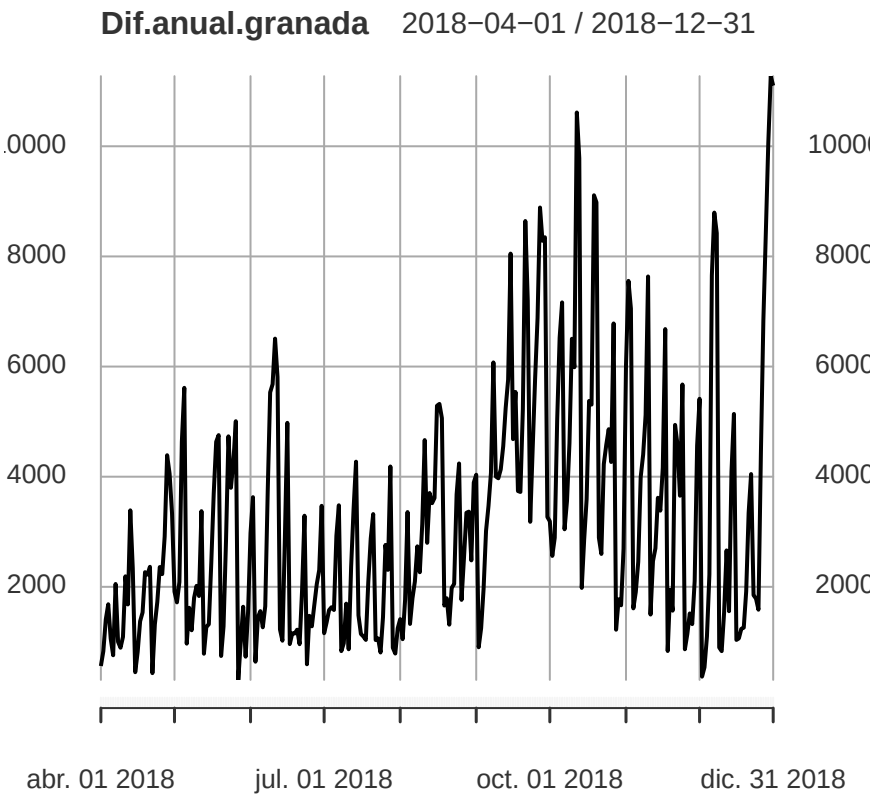
Efectivamente el parámetro d=1 . En la parte estacionaria tenemos un modelo MA de parámetro 1 , y en la estacional se indica D= 1 , esto es, la diferenciación respecto al año anterior que elimina la estacionalidad de la serie. Vemos la previsión de la serie:



EJEMPLO NO ESTACIONARIDAD EN VARIANZA

Para el desarrollo de este ejemplo, a parte de filtrar un hotel (hemos seleccionado el *Senator Granada Spa Hotel*), nos quedamos con los datos de algunos meses del último año (2018) y los agrupamos de forma diaria. Después de hacer esto nos quedará una serie de la siguiente forma:

```
plot(Dif.anual.granada)
```



En este caso solo aplicaremos el **test KPSS**, puesto que, en el caso de que no sea estacionario en varianza, tampoco lo será en media.

```
kpss.granada$p.value

## [1] 0.01
```

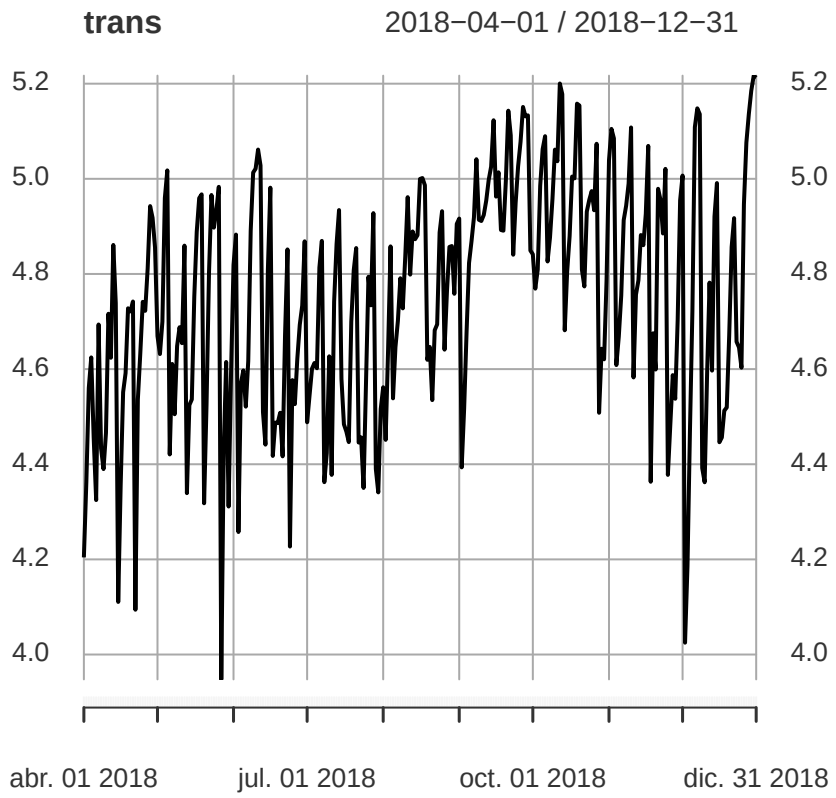
Como vemos se rechaza la estacionaridad de la serie.

Como vimos en secciones anteriores, para estabilizar la varianza podemos recurrir a la transformación de Box-Cox. En **R**, el paquete forecast, el cual ya hemos usado en alguna ocasión, nos facilita el trabajo.

Para determinar el parámetro λ usamos la función BoxCox.lambda y para realizar la tranformación usamos BoxCox:

```
lambda <- BoxCox.lambda(Dif.anual.granada)
trans<- BoxCox(Dif.anual.granada,lambda)
```

```
plot(trans)
```



Una vez hecho esto diferenciamos la serie (puesto que tampoco era estacionaria en media). Le volvemos a aplicar el test KPSS a la serie transformada diferenciada y verficamos que ya la serie es estacionaria en varianza, que es lo que pretendiamos.

```
kpss.transformada<- kpss.test(diff(trans))
kpss.transformada$p.value

## [1] 0.1
```

En este caso, no estimaremos el modelo ni haremos previsiones puesto que carece de interés.

4.3 Estudio Multivariante

Antes de comenzar esta sencción, cabe mencionar que se va a recurrir, a parte de a datos proporcionados por la empresa, a datos del *Instituto Nacional de Estadística*. A los cuales se puede acceder públicamente a través de la siguiente dirección :

https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736177015&menu=resultados&idp=1254735576863

Una vez en la página, y selecionados los datos que nos parezcan interesantes, estos pueden ser descargados en distintos formatos (cvx, Excel, PDF, . . .), en nuestro caso, como nuestros datos particulares han sido proporcionados en Excel, seguiremos trabajando en ese formato.

Planteamiento del problema

Ahora lo que buscamos es ver si el comportamiento del **Porcentaje de ocupación** [1] de los hoteles de la empresa es coherente con la ocupación de la zona turística donde se sitúa.

El Porcentaje de ocupación es un dato que no nos han proporcionado, así pues tenemos que realizar un cálculo de este. Para ello importamos el planinng de la empresa, el cual se proporciona a través de un Excel.

```
Planning<-read_excel("Habitaciones.xlsx")
```

En el archivo Planning tenemos 98.443 y 4 variables (las cuales son, el código del hotel (Hotel), la fecha , de forma diaria, de la que es el planning (Fecha), las habitaciones que había ocupadas (Ocupacion), y las habitaciones disonibles (Cupo)).

Para facilitar el trabajo construiremos un *data.frame* que contemple el Porcentaje de ocupación de las zonas turísticas en formato xts.

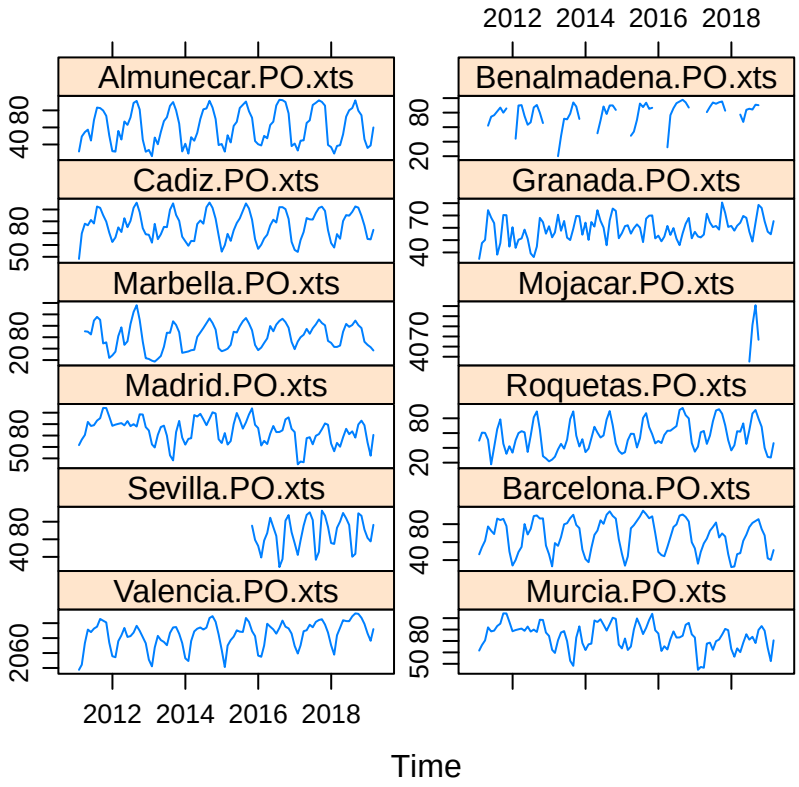
Al final trabajamos con un *data.frame*, al que llamamos PO.particular.xts, que será de la siguiente forma:

```
##            Almunecar.PO.xts Benalmadena.PO.xts Cadiz.PO.xts Granada.PO.xts
## 2011-01-31         31.74789                NA     48.20985       34.74543
## 2011-02-28         49.36645                NA     69.74097       47.63339
## 2011-03-31         54.52399                NA     77.98653       50.17489
## 2011-04-30         57.44147           62.03231     76.66667       74.29719
## 2011-05-31         44.63398           74.39516     81.15050       68.32491
## 2011-06-30         68.53995           76.46825     78.16850       63.81526
##            Marbella.PO.xts Mojacar.PO.xts Madrid.PO.xts Roquetas.PO.xts
## 2011-01-31              NA             NA      61.59215        49.84221
## 2011-02-28              NA             NA      66.41604        60.63612
## 2011-03-31        70.35789             NA      70.51500        60.30169
## 2011-04-30        69.72810             NA      81.94932        50.93587
## 2011-05-31        64.81157             NA      78.36257        17.69157
## 2011-06-30        89.18982             NA      79.06433        40.85532
##            Sevilla.PO.xts Barcelona.PO.xts Valencia.PO.xts Murcia.PO.xts
## 2011-01-31             NA         46.41829        17.54274       61.59215
## 2011-02-28             NA         54.35949        24.36756       66.41604
## 2011-03-31             NA         61.47206        53.24261       70.51500
## 2011-04-30             NA         77.44914        71.44097       81.94932
## 2011-05-31             NA         72.63365        67.99395       78.36257
## 2011-06-30             NA         68.95754        72.65625       79.06433
```

Vemos la representación gráfica de las variables:

[1] Lo calculamos de la siguiente forma

$$\text{Porcentaje de Ocupación} = \frac{\text{Habitaciones ocupadas}}{\text{Habitaciones Disponibles}}$$



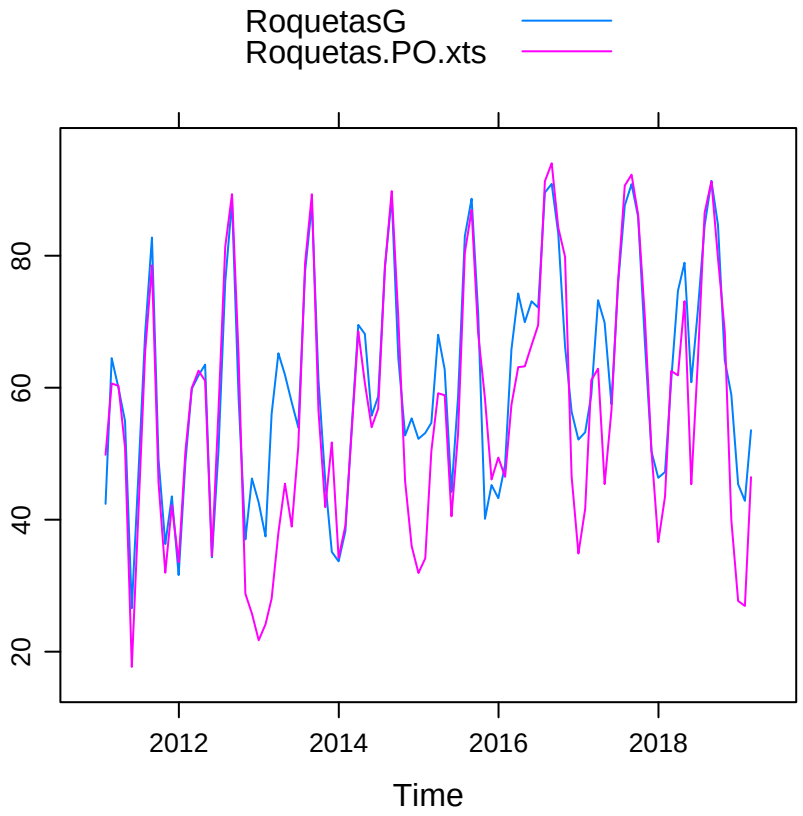
Ahora importamos los procedentes del *Instituto Nacional de Estadística* y los pasamos a formato xts, para poder trabajar simultaneamente con los datos generales y particulares. Obteniendo así el siguiente *data.frame*:

```
##            MojácarG RoquetasG CádizG  GranadaG MadridG BenalmádenaG
## 2011-01-31 "41.25"  "42.40"   "32.41" "39.56"  "54.26" "50.83"
## 2011-02-28 "46.28"  "64.48"   "58.12" "51.04"  "62.35" "71.53"
## 2011-03-31 "48.23"  "59.96"   "66.20" "56.50"  "68.01" "76.87"
## 2011-04-30 "55.26"  "54.96"   "60.25" "68.86"  "71.23" "76.32"
## 2011-05-31 "36.51"  "26.59"   "58.66" "67.85"  "73.12" "72.57"
## 2011-06-30 "51.67"  "46.74"   "65.40" "59.94"  "71.40" "83.33"
##            MarbellaG MurciaG SevillaG ValenciaG BarcelonaG
## 2011-01-31 "30.94"   "33.15" "41.11"  "40.91"   "45.92"
## 2011-02-28 "39.10"   "42.13" "49.44"  "55.36"   "61.39"
## 2011-03-31 "42.78"   "41.40" "63.78"  "64.04"   "67.81"
## 2011-04-30 "56.65"   "42.76" "73.68"  "61.57"   "79.54"
## 2011-05-31 "58.37"   "53.67" "75.62"  "60.15"   "81.96"
## 2011-06-30 "67.57"   "41.64" "64.30"  "66.93"   "80.74"
```

Una vez hecho esto juntamos ambos *data.frame*, obteniendo así el fichero de datos con el que al final trabajaremos, al que llamaremos PO.xts

Análisis de los datos

Tras el planteamineto del problema del que vamos a hacer el análisis e importados los datos vamos a dedicar este apartado para hacer un análisis detallado de una de las variables, este mismo proceso se podrá hacer con cualquiera de las variables, nosotros hemos seleccionado **Roquetas de Mar**. En primer lugar veremos una gŕafica donde aparezca simultaneamente el porcentaje de ocupación particular y general de dicha zona turística. Hemos denotado por Roquetas.PO.xts, al porcentaje de ocupación particular de esta zona, y al general RoquetasG.



Antes de estudiar el comportamimento de las series, que es lo que nos interesa realmente, vamos a hacer un estudio descriptivo de esta.

	Roquetas.PO.xts	RoquetasG
Mean	56.61	60.97
Std.Dev	19.10	16.07
Min	17.69	26.59
Q1	41.90	48.80
Median	56.56	59.39
Q3	68.54	72.15
Max	94.01	91.35
MAD	20.05	17.81
IQR	26.56	23.15
CV	0.34	0.26
Skewness	0.18	0.16
SE.Skewness	0.24	0.24
Kurtosis	-0.83	-0.80
N.Valid	98.00	98.00
Pct.Valid	100.00	100.00

Aquí podemos ver, ya alguna diferencia, puesto que el porcentaje de ocupación en media general de estos últimos años es mayor que el de la cadena, gráficamente vemos que esto se debe a que en invierno la cadena tiene bajadas muy bruscas, y en general se mantiende algo más estable.

Ahora bien, para estudiar simultaneamente el comportamiento de las series, es importante tener en cuenta su comportamiento individual, para ello observamos la tendencia de ambas y vemos si existe estacionariedad en media (esto nos permite ver si las series tienen una tendencia común, creciente, decreciente, o ausencia de la misma). Para ello aplicamos el **test ADF**.

```
adf.Roquetas.g<-adf.test(PO.xts[,c(2)])
adf.Roquetas.g$p.value

## [1] 0.01

adf.Roquetas.p<-adf.test(PO.xts[,c(19)])
adf.Roquetas.p$p.value

## [1] 0.01
```

En ambos casos vemos que el p-valor es más pequeño que 0.05, por tanto se da la estacionariedad en media en ambos, por lo que ninguna de las dos series presenta una tendencia clara. Por lo que debemos estimar un modelo **VAR**.

Para la estmación de este es necesario cargar en nuestra librería el paquete vars. El primer paso que damos, es ver cuantos retardos debemos incluir en nuestro modelo, para ello usamos la función VARselect.

```
library(vars)
Varselect<-VARselect(PO.xts[,c(2,19)], lag.max = 29)
Varselect$selection
```

```
## AIC(n)  HQ(n)  SC(n) FPE(n)
##     29     12      2     13
```

En el resultado se nos muestra los retardos más adecuados para el *Criterio de información de Akaike*(AIC), el de *Hannan y Quinn* (HQ), el *criterio de información bayesiana de Schwarz*(SC) y el *error de predicción final de Akaike*. Estos criterios nos indican información relativa perdida cuando los datos se ajustan. A veces, como se da en el caso, el número de retardos seleccionados por cada creterio es diferente. El SC y HQ proporcionan estimaciones consistentes del verdadero orden del retardo ,y FPE y AIC sobreestiman el orden del retardo con probabolidad positiva.

Nosotros nos con el número menor de retardos, puesto que es el más sencillo, por tanto, tomamos dos.

El siguiente paso para la construcción de nuestro modelo es hacer una estimación de los parámetros, esto lo hacemos con la función VAR, del paquete con el que hemos trabajado anteriormente.

```
(ModeloVar<- VAR (PO.xts[,c(2,19)],p=2))
```

El cual, si llamamos

$$Y_{1t} = \text{Roquetas.PO.xts}$$

$$Y_{2t} = \text{RoquetasG}$$

quedaría de la siguiente forma,

$$\begin{bmatrix} Y_{1t} \\ Y_{2t} \end{bmatrix} = \begin{bmatrix} 0{,}9164 & 0{,}0983 \\ 0{,}4190 & 0{,}5063 \end{bmatrix} \begin{bmatrix} Y_{1t-1} \\ Y_{2t-1} \end{bmatrix} + \begin{bmatrix} -0{,}4349 & -0{,}0793 \\ -0{,}4427 & -0{,}0520 \end{bmatrix} \begin{bmatrix} Y_{1t-2} \\ Y_{2t-2} \end{bmatrix} + \begin{bmatrix} 28{,}2360 \\ 34{,}75390 \end{bmatrix}$$

Vamos a ver si nuestro modelo es estable, para ello vamos a usar la función roots para ver si estas se encuentran dentro del disco unidad.

```
roots(ModeloVar)

## [1] 0.71607625 0.71607625 0.47206599 0.05172112
```

Vemos que estas son menores que uno, por lo que el modelo no es estable, pero esto no quiere decir que no sea válido, por lo que procedemos a su validación usando los contrastes de hipótesis ya definidos en secciones anteriores.

Para el *Contraste de Portmanteau*, se usa la función serial.test, para el *test Arch-LM*, arch.test, y, por último, la función normality.test para el *test de normalidad de Jarque-Bera*.

```
(Var.serial<-serial.test(ModeloVar))

##
##  Portmanteau Test (asymptotic)
##
## data:  Residuals of VAR object ModeloVar
## Chi-squared = 84.484, df = 56, p-value = 0.008274
```

```
(Var.arch <- arch.test(ModeloVar))

##
##  ARCH (multivariate)
##
## data:  Residuals of VAR object ModeloVar
## Chi-squared = 39.183, df = 45, p-value = 0.7159

(Var.normalidad <- normality.test(ModeloVar))

## $JB
##
##  JB-Test (multivariate)
##
## data:  Residuals of VAR object ModeloVar
## Chi-squared = 8.1867, df = 4, p-value = 0.08497
##
##
## $Skewness
##
##  Skewness only (multivariate)
##
## data:  Residuals of VAR object ModeloVar
## Chi-squared = 4.52, df = 2, p-value = 0.1044
##
##
## $Kurtosis
##
##  Kurtosis only (multivariate)
##
## data:  Residuals of VAR object ModeloVar
## Chi-squared = 3.6668, df = 2, p-value = 0.1599
```

En el caso del *test de Portmanteau*, el p-valor es menor que 0.05 por tanto rechazamos la hipótesis nula , por lo que los residuos están correlacionados. En los otros casos se acepta la hipótesis nula, es decir presencia de homocedasticidad y sí presencia de normalidad. Por lo que no podemos validar el modelo.

Teóricamente hemos visto que cuando el metodo no es estable es adecuado estudiar la cointegración de las variables, y en caso de que lo sean estimar un modelo **VECM**. Usaremos la metodología de Engle y Granger. Como nuestro β es desconocido aplicaremos el **contraste de Phillips-Oularts** (PO), puesto que, el parámetro β de cointegración es desconocido. Para ello usamos la función po.test, del paquete tseries:

```
(po <- po.test(PO.xts[,c(2,19)]))

##
```

```
##  Phillips-Ouliaris Cointegration Test
##
## data:  PO.xts[, c(2, 19)]
## Phillips-Ouliaris demeaned = -55.12, Truncation lag parameter = 0,
## p-value = 0.01
```

Claramente se rechaza la hipótesis nula, por lo que las series estan cointegradas. Asṕues, sería conveniente estimar un modelo **VECM**, no obstante vamos a recurrir a otra serie para mostrar una construcción de éste, podrámos aplicar este procedimiento a esta serie.

Para construir un modelo de de cointegración ,hacemos uso de la metodología de Johansen, comparando el ADR *Avarage Daily Rate* [2] general, con el particular de esta misma zona.

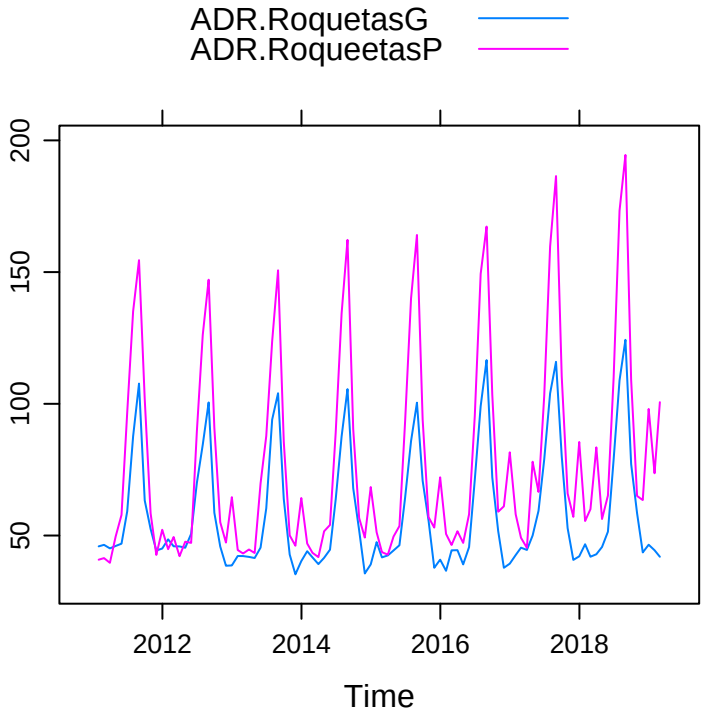
Haciendo uso de los *data.frame* ImportesG y Planning, realizamos el cálculo de la variable. Una vez hecho esto nos construimos un nuevo *data.frame* en el que se enfrenten el ADR de dicha zona general y particular, vemos dicho *data.frame* al que hemos llamado ADR.

```
##            ADR.RoquetasG ADR.RoqueetasP
## 2011-01-31         45.89       40.82701
## 2011-02-28         46.46       41.42812
## 2011-03-31         45.12       39.63984
## 2011-04-30         46.03       49.55910
## 2011-05-31         46.90       57.81411
## 2011-06-30         59.18       96.71623
```

Lo vemos representados gráficamente.

[2]Ingresos por habitación ocupada, se calcula de la siguiente forma

$$ADR = \frac{\text{Ingresos por habitación}}{\text{Habitaciones Ocupadas}}$$



Para la estimación del modelo necesario tener en cuenta el comportamiento independiente de las series para el uso de una buena técnica para determinar un modelo adecuado. Para ello estimamos un modelo ARIMA para cada una de las series, el parmetro d de esta nos dirá el orden de diferencia para que las series sean estacionarias.

```
auto.arima(ts(ADR[,c(1)],frequency = 12, start = 2011))

## Series: ts(ADR[, c(1)], frequency = 12, start = 2011)
## ARIMA(1,1,1)(0,1,1)[12]
##
## Coefficients:
##          ar1      ma1     sma1
##       0.5096  -0.9327  -0.5916
## s.e.  0.1198   0.0479   0.1146
##
## sigma^2 estimated as 17.94:  log likelihood=-245.19
## AIC=498.37   AICc=498.87   BIC=508.14

auto.arima(ts(ADR[,c(2)],frequency = 12, start = 2011))

## Series: ts(ADR[, c(2)], frequency = 12, start = 2011)
## ARIMA(0,1,1)(0,1,2)[12]
##
## Coefficients:
##          ma1     sma1    sma2
##       -0.8724  -0.5638  0.2707
## s.e.   0.0563   0.1445  0.1620
```

```
##
## sigma^2 estimated as 70.64:  log likelihood=-303.08
## AIC=614.16   AICc=614.66   BIC=623.93
```

Como vemos, en ambos casos, d=1, por tanto estas series serán estacionarias en su primera diferencia, por lo que es adecuado estimar un modelo **VECM** para estos datos.

Vamos a proceder a la construcción del modelo.

Procedemos a ello, como lo hemos hecho con el modelo **VAR** anterior, especificamos el número de retardos óptimos para este.

```
## AIC(n)  HQ(n)  SC(n) FPE(n)
##     29     29     12     29
```

Tomamos 12 retardos. Ahora vamos a determinar en número de relaciones, para ello usamos la función ca.jo, perteneciente al paquete con el que estamos trabajando esta sección, usando el procedimiento de Johansen.

```
 jo <-ca.jo(ADR[,c(1,2)],K=12)
 summary(jo)

##
## ######################
## # Johansen-Procedure #
## ######################
##
## Test type: maximal eigenvalue statistic (lambda max) , with linear trend
##
## Eigenvalues (lambda):
## [1] 0.2051817 0.0587944
##
## Values of teststatistic and critical values of test:
##
##           test 10pct  5pct  1pct
## r <= 1 |  5.21  6.50  8.18 11.65
## r = 0  | 19.75 12.91 14.90 19.19
##
## Eigenvectors, normalised to first column:
## (These are the cointegration relations)
##
##                    ADR.RoquetasG.112 ADR.RoqueetasP.112
## ADR.RoquetasG.112          1.0000000          1.0000000
## ADR.RoqueetasP.112        -0.5402333         -0.2183228
##
## Weights W:
## (This is the loading matrix)
##
##                   ADR.RoquetasG.112 ADR.RoqueetasP.112
```

```
## ADR.RoquetasG.d          -1.014212          -0.1859818
## ADR.RoqueetasP.d         -1.136921           0.8996781
```

Vemos que se rechaza la hipótesis de que $r = 0$, y con esto se rechaza la no cointegración. En cambio se acepta que $r = 1$, es decir que hay una ecuación de cointegración.

Pasamos a estimar los parámetros del modelo de cointegración **VECM**, para ello usamos la función cajorls.

```
Vecm<-cajorls(jo,1)
```

El resultado de este nos da los siguientes coeficientes del modelo ya definido.

$$\Delta \mathbf{Y}_t = \mathbf{CD}_t + \mathbf{\Phi} \mathbf{Y}_{t-1} + \mathbf{\Gamma}_1 \Delta \mathbf{Y}_{t-1} + \ldots + \mathbf{\Gamma}_{12-1} \Delta \mathbf{Y}_{t-12+1} + a_t,$$

Sea $\mathbf{Y}_t = (Y_{1t}, Y_{2t})$, con

$$Y_{1t} = \text{ADR.RoquetasG}$$
$$Y_{1t} = \text{ADR.RoquetasP}$$

entonces,

$$a_t = \begin{bmatrix} -1{,}0114 \\ -1{,}1369 \end{bmatrix} \quad \mathbf{CD}_t = \begin{bmatrix} 16{,}8313 \\ 20{,}6505 \end{bmatrix}$$

$$\Phi_1 = \begin{bmatrix} -0{,}6920 & -0{,}1492 \\ 0{,}1257 & -0{,}8581 \end{bmatrix} \quad \Phi_2 \begin{bmatrix} -0{,}6248 & -0{,}0963 \\ 0{,}1517 & -0{,}8246 \end{bmatrix} \quad \Phi_3 \begin{bmatrix} -0{,}8031 & -0{,}0730 \\ -0{,}4571 & -0{,}8191 \end{bmatrix}$$

$$\Phi_4 = \begin{bmatrix} -0{,}9686 & -0{,}0777 \\ -0{,}1517 & -0{,}6867 \end{bmatrix} \quad \Phi_5 = \begin{bmatrix} -0{,}8864 & 0{,}0223 \\ -0{,}5903 & -0{,}4682 \end{bmatrix} \quad \Phi_6 = \begin{bmatrix} -0{,}9969 & 0{,}0192 \\ -0{,}6192 & -0{,}509 \end{bmatrix}$$

$$\Phi_7 = \begin{bmatrix} -1{,}1197 & 0{,}0688 \\ -0{,}7696 & -0{,}4160 \end{bmatrix} \quad \Phi_8 = \begin{bmatrix} -0{,}8864 & 0{,}0223 \\ -0{,}5903 & -0{,}4682 \end{bmatrix} \quad \Phi_9 = \begin{bmatrix} -1{,}3224 & 0{,}17645 \\ -1{,}0999 & -0{,}3032 \end{bmatrix}$$

$$\Phi_{10} = \begin{bmatrix} -1{,}3224 & 0{,}1764 \\ -1{,}0357 & -0{,}3423 \end{bmatrix} \quad \Phi_{11} = \begin{bmatrix} -1{,}4282 & 0{,}3055 \\ -1{,}6062 & 0{,}1972 \end{bmatrix}$$

```
(vecvar<- vec2var(jo,1))
```

Siendo las relaciones mostradas:

$$\mathbf{\Phi}_1 = \mathbf{\Gamma}_1 + \mathbf{\Phi} + \mathbf{I}_4$$
$$\mathbf{\Phi}_k = \mathbf{\Gamma}_k - \mathbf{\Gamma}_{k-1}, \quad k = 2, \ldots, 12.$$

Se estiman los siguiente parámetros:

$$\Phi_1 = \begin{bmatrix} 0{,}3080 & -0{,}1492 \\ 0{,}1257 & 0{,}1419 \end{bmatrix} \quad \Phi_2 = \begin{bmatrix} 0{,}06725 & 0{,}0530 \\ 0{,}02946 & 0{,}0334 \end{bmatrix} \quad \Phi_3 = \begin{bmatrix} -0{,}1784 & 0{,}0232 \\ -0{,}6123 & 0{,}0055 \end{bmatrix}$$

$$\Phi_4 = \begin{bmatrix} -0{,}1654 & -0{,}0046 \\ 0{,}3054 & 0{,}1324 \end{bmatrix} \quad \Phi_5 = \begin{bmatrix} 0{,}0821 & 0{,}0999 \\ -0{,}4385 & 0{,}2184 \end{bmatrix} \quad \Phi_6 = \begin{bmatrix} -0{,}1105 & -0{,}0031 \\ -0{,}0290 & -0{,}0408 \end{bmatrix}$$

$$\Phi_7 = \begin{bmatrix} -0{,}1228 & 0{,}0497 \\ -0{,}1504 & 0{,}0929 \end{bmatrix} \quad \Phi_8 = \begin{bmatrix} -0{,}16692 & 0{,}14454 \\ -0{,}14947 & 0{,}19347 \end{bmatrix} \quad \Phi_9 = \begin{bmatrix} -0{,}03584 & -0{,}03696 \\ -0{,}18088 & -0{,}08068 \end{bmatrix}$$

$$\Phi_{10} = \begin{bmatrix} -0{,}12279 & 0{,}04967 \\ 0{,}06418 & -0{,}039124 \end{bmatrix} \quad \Phi_{11} = \begin{bmatrix} -0{,}08855 & 0{,}02922 \\ 0{,}06419 & -0{,}03912 \end{bmatrix} \quad \Phi_{12} = \begin{bmatrix} 0{,}41399 & 0{,}24240 \\ 0{,}46928 & 0{,}41693 \end{bmatrix}$$

5 Conclusiones

Durante el transcurso de este documento se ha trabajado con datos reales, no solo los proporcionados por la empresa, si no que también datos externos de interés, los cuales están en un contexto de actualidad. Estos nos abrían un amplio abanico de variables a analizar.

El estudio de estas no siempre ha sido exitoso, puesto que, en ocasiones, y tras un análisis de dichas variables se nos hace imposible estimar un buen modelo. Por ello, se han mostrado los resultados que han resultado más interesantes, haciendo un recorrido por las diversas formas en las que nos podemos encontrar una serie temporal, es decir, la presencia o ausencia de tendencia, variabilidad o estacionalidad, así como las pertinentes transformaciones que podemos hacer para encontrar una serie con la que nos resulte sencillo trabajar.

Para el estudio de este comportamiento ha tenido gran utilidad el uso de contrastes de hipótesis, los cuales vienen implementados en el paquete tseries.

Una vez analizado su comportamiento y en base a este se ha dado lugar a la estimación de diferentes modelos, los cuales no lo solo estudian el comportamiento pasado de la serie, sino que también nos han permitido hacer previsiones futuras.

Si hablamos del caso multivariante, a parte del estudio individual de cada serie, ha sido posible estudiar el comportamineto de series de forma simultánea (para el cual también ha sido fundamental el uso de contrastes de hipótesis), y establecer modelos partiendo de este comportamiento y haciendo uso de diferentes metodologías.

Para la construcción de los modelos univariantes la función auto.arima nos ha facilitado mucho el trabajo, puesto que esta busca el modelo que más se nos adecúa a nuestra serie. Por otro lado, en cuanto al multivariante, el paquete vars ha sido nuetra herramienta más fuerte, ayudandonos a aplicar las metodologías definidas correctamente. Otra utilidad de **R** a la que se le ha sacado gran partido, aunque no sea propia de series temporales, ha sido la representación de gráficas en **R**, puesto que de estas se han podido sacar diferentes conclusiones sobre el comportamiento de nuestras series.

Al inicio del trabajo se plantearon una serie de objetivos, los cuales han sido cumplidos de forma exitosa, viendo como es posible aplicar conceptos teóricos a un contexto de realidad con el uso de un software de actualidad, así como todas las herramientas, que en nuestro caso **R** nos proporciona para ello.

Bibliografía

[1] Gónzalez Casimiro, M.P. *Análisis de series temporales: Modelos ARIMA*, Universidad del País Vasco, 2009.

[2] Mogni, A.P. *Modelos de Series de Tiempo con aplicaciones en la industria aerocomercial* , Universidad de Buenos Aires, 2015.

[3] Paff, B. *Analysis of integrated and cointegrated time series whit R*, Sprinter, 2011.

[4] Reche Lorite, F. *Formato de los datos temporales en R*, Universidad de Almería, 2015.

[5] Zivot, E.; Wang, J. *Modelling financial time series with S-Plus*, Sprinter, 2015.

[6] Página de consulta de paquetes de R. http://www.tug.org.