Trabajo Fin de Máster

Optimización del reparto mensual del presupuesto

Irene Collazo Arnoso

Máster en Técnicas Estadísticas

Curso 2019-2020

Propuesta de Trabajo Fin de Máster

Título en galego: Optimización do reparto mensual do presuposto
Título en español: Optimización del reparto mensual del presupuesto
English title: Optimization of monthly budget distribution
Modalidad: Modalidad B
Autor/a: Irene Collazo Arnoso, Universidad de Santiago de Compostela
Director/a: Manuel Febrero Bande, Universidad de Santiago de Compostela
Tutor/a: Jorge López Muñiz, Estrella Galicia
Breve resumen del trabajo: Desarrollo de un modelo que reparta el presupuesto anual a nivel mes, por sku y cliente, optimizando la diferencia con los valores reales.

Don Manuel Febrero Bande, catedrático de Estadística e Investigación Operativa (Departamento de Estadística, Análisis Matemático y Optimización) de la Universidad de Santiago de Compostela y don Jorge López Muñiz, director del departamento tLab & Advanced Analytics de Estrella Galicia, informan que el Trabajo Fin de Máster titulado

Optimización del reparto mensual del presupuesto

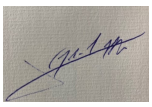
fue realizado bajo su dirección por doña Irene Collazo Arnoso para el Máster en Técnicas Estadísticas. Estimando que el trabajo está terminado, dan su conformidad para su presentación y defensa ante un tribunal.

En Santiago de Compostela, a 14 de julio de 2020.

El director:

Don Manuel Febrero Bande

El tutor:



Don Jorge López Muñiz

La autora:

Doña Irene Collazo Arnoso

Índice general

Resumen

Resumen en español

Las series de tiempo se analizan para entender el pasado y predecir el futuro, permitiendo así tomar decisiones informadas de forma apropiada. En este proyecto hemos trabajado con las bases de datos de la empresa Hijos de Rivera con el objetivo de obtener un modelo de mensualización del presupuesto anual por producto y ruta de distribución. Para ello nos basaremos en la predicción a través de las series de tiempo y que nos servirá para realizar un reparto apropiado. Para la construcción del modelo disponemos del historial de ventas mensuales por producto y ruta de distribución en la comunidad autónoma de Galicia. Comenzaremos con una breve introducción sobre series temporales. Después trataremos los modelos clásicos para el análisis de series de tiempo, los modelos ARIMA. Debido a que las series de tiempo pueden estar relacionadas con otras series, introduciremos los correspondientes modelos de regresión con múltiples variables regresoras. Por último, comentaremos brevemente la implementación propuesta y los resultados obtenidos.

English abstract

Time series analysis is usually carried out in order to understand the past and predict the future, allowing for decision making in an appropriate way. In this thesis we have worked with databases from Hijos de Rivera, focusing on the goal of modelling the monthly payment of the annual budget by product and distribution route. With that goal in mind, we will base our predictions in time series, which will allow us to perform a suitable distribution. Historical data on monthly sales by product and route of distribution in the autonomous community of Galicia are available for the construction of the model. A brief introduction to time series is given as a starting point. Afterwards, classical models for time series analysis, ARIMA models, are presented. Due to the fact that time series may be related to other series, regression models with multiple covariates will be introduced. To conclude, the implementation of the proposal and the obtained results will be discussed.

Introducción

Un presupuesto es un plan integral y formal que estima los gastos e ingresos para una organización durante un período específico, debido a que los presupuestos son herramientas valiosas a la hora de la planificación y del control de las finanzas. El presupuesto afecta a casi todo tipo de organizaciones, desde gobiernos y grandes corporaciones hasta pequeñas empresas o a familias e individuos. Además de asignar recursos, un presupuesto también permite establecer objetivos y medir resultados y riesgos. Habitualmente los presupuestos de las compañías son anuales, sin embargo, realizar también un reparto mensual puede resultar beneficioso. En función de la empresa podríamos encontrar múltiples repartos posibles.

Este trabajo ha sido desarrollado en la corporación Hijos de Rivera S.A.U., empresa centrada en la producción y venta de bebidas. El objetivo consistirá en la obtención de un modelo de reparto del presupuesto anual a nivel mensual, por producto y cliente, basándonos en la predicciones de las ventas. Para ello, utilizaremos las denominadas series de tiempo.

Para la construcción del modelo, disponemos del historial de ventas mensuales en litros por producto de Hijos de Rivera correspondiente a cada ruta de distribución en la comunidad autónoma de Galicia. Estos datos están almacenados dentro de la base de datos de la compañía en el programa SQL Server, que es un sistema de gestión de base de datos.

Los productos, las rutas de distribución y los datos con los que trabajaremos son reales, pero hemos decidido no mencionarlos por motivos de confidencialidad de la empresa. Por este mismo motivo, nos referiremos a los productos concretos mediante un alias y en todas las gráficas que puedan arrojar información sobre el volumen real de ventas de la compañía se elimnará el eje y.

Primeramente haremos una introducción a las series de tiempo, introduciendo los principales conceptos y presentando diversos ejemplos.

En segundo lugar estudiaremos cómo construir un modelo que represente la evolución de una serie temporal y generar predicciones futuras. Las predicciones obtenidas con estos modelos se basan en la hipótesis de que las condiciones futuras serán análogas a las pasadas y son especialmente útiles para la previsión a corto plazo.

En el tercer capítulo estudiaremos métodos para encontrar la relación de dependencia dinámica entre una serie de interés y un grupo de posibles variables explicativas. Las previsiones univariantes pueden mejorarse incorporando la información de la evolución de otras variables y construyendo modelos que tengan en cuenta esta dependencia. Estos modelos se conocen como modelos de regresión dinámica.

Por último, aplicaremos los modelos de regresión dinámica a los datos de ventas de Estrella Galicia, con variables regresoras como la temperatura media y el número de días con precipitaciones, para obtener una predicción de ventas en el siguiente año. Una vez obtenida la estimación y establecido el presupuesto planificado por la empresa crearemos un vector de pesos proporcional a la estimación de las ventas para así distribuir el presupuesto mensualmente.

En el desarrollo de este trabajo, las principales referencias utilizadas han sido Peña (2005), Cowpertwait y Metcalfe (2009), Cryer y Chan (2010), Brockwell y Davis (1991), Hyndman y Athanasopoulos (2018).

Capítulo 1

Procesos estocásticos y series de tiempo

Los datos obtenidos de observaciones recogidas secuencialmente en el tiempo son muy frecuentes. En economía, observamos el producto interior bruto anual, tasas de inflación, ventas mensuales, tasa de desempleo, etc. En meteorología, observamos las temperaturas máximas, medias o mínimas diarias, índices anuales de precipitaciones y sequía, etc. En ecología, registramos la concentración media mensual de nitratos en agua, emisiones anuales de CO_2, etc. La lista de áreas en las que se estudian las series de tiempo es prácticamente interminable. El objetivo del análisis de series de tiempo es doble, por una parte comprender o modelar el mecanismo estocástico que da lugar a una serie observada y por otra predecir los valores futuros de una serie en función del histórico de esa serie y, posiblemente, de otras series o factores relacionados.

En este capítulo se hará una introducción a las series de tiempo, presentando diversos ejemplos de series temporales de distintas áreas de aplicación. Además, se describirán los conceptos fundamentales en la teoría de los modelos de series de tiempo. En particular, introducimos los conceptos de proceso estocástico, función de medias y covarianzas, procesos estacionarios y funciones de autocorrelación.

1.1. Procesos estocásticos

Denotando por X_n el stock disponible en una n-ésima unidad de tiempo, podríamos representar su evolución con una familia de variables aleatorias $\{X_0, X_1, \dots\}$ indexadas por un parámetro de tiempo discreto $n \in \mathbb{Z}_+$. El número X_t de nacimientos en el intervalo de tiempo $[0, t]$ da lugar a una colección de variables aleatorias $\{X_t,\ t \geq 0\}$, indexadas por el parámetro de tiempo continuo t. La posición X_r de un pez dentro de un acuario proporciona una familia de variables aleatorias $\{X_r,\ r \in \mathbb{R}^3\}$ indexadas por un parámetro espacial multidimensional r. De forma general, tenemos la siguiente definición.

Definición 1.1.1. Dado un conjunto de índices I, un proceso estocástico indexado por I es una colección de variables aleatorias $\{X_\lambda,\ \lambda \in I\}$ definida en un espacio de probabilidad $(\Omega, \mathcal{F}, \mathbb{P})$ y que toma valores en el espacio de estados S.

El conjunto de índices I puede identificarse con los números naturales o con los números reales. Si $I \subset \mathbb{N}$, diremos que el proceso estocástico es de tiempo discreto, y en caso de que $I \subset \mathbb{R}$, diremos que es de tiempo continuo. Además, las variables aleatorias X_λ pueden ser discretas o continuas. Así, cuando $S \subset \mathbb{N}$ nos referiremos a procesos estocásticos de espacio de estados discreto y cuando $S \subset \mathbb{R}$ nos referiremos a procesos estocásticos de espacio de estados continuo.

Para simplificar la notación, el conjunto de índices I a menudo se suprime cuando el contexto lo deja claro. En particular, habitualmente se escribe $\{X_n\}$ en $\{X_n,\ n = 0, 1, 2, \dots\}$ y $\{X_t\}$ en vez de $\{X_t,\ t > 0\}$

Si denotamos por X_t el número de pacientes que esperan ser atendidos en el servicio de urgencias de un hospital, $\{X_t,\ t \geq 0\}$ es un proceso estocástico de tiempo continuo con espacio de estados discreto $S = \{0, 1, 2 \dots\}$. Siendo X_n el n-ésimo lanzamiento de una moneda, $\{X_n,\ n \geq 0\}$ es un proceso estocástico de tiempo discreto con espacio de estados finito $S = \{C, +\}$.

Recordando la definición de variable aleatoria notamos que para cada $\lambda \in I$ fijo, X_λ es de hecho una función $X_\lambda(\cdot)$ en el conjunto Ω. Por otra parte, para cada $\omega \in \Omega$ fijado, $X.(\omega)$ es una función en I.

Definición 1.1.2. Las funciones $\{X.(\omega), \omega \in \Omega\}$ se denominan realizaciones del proceso $\{X_\lambda\}_{\lambda \in I}$.

Una serie de tiempo es un proceso estocástico con espacio de estados continuo e I discreto y regularmente espaciado. Se trata de una colección de observaciones de una variable, X, tomadas secuencialmente a lo largo del tiempo. Dichas observaciones se toman en intervalos regulares de tiempo (cada hora, cada día, cada mes,...). Sin pérdida de generalidad supondremos que la variable X ha sido observada en los instantes $1, 2, \dots, T$. Por tanto, la serie de tiempo observada se representará por $x_1, x_2, \dots, x_T$. Con el objetivo de realizar inferencia estadística, supondremos que la serie de tiempo ha sido generada por un proceso estocástico.

Por tanto, la serie de tiempo $x_1, x_2, \dots, x_T$ es una realización o trayectoria parcial de un proceso estocástico. Utilizaremos el término serie de tiempo para referirnos tanto a los datos como al proceso del cual es una realización.

Por ejemplo, la Figura 1.1 presenta 11 realizaciones (años) de la serie de lluvia mensual en Santiago de Compostela (datos disponibles en el IGE)[1]. Tenemos por tanto 11 valores de 12 variables aleatorias, una para cada mes, y la trayectoria de los 12 valores en un año dado representa una realización del proceso estocástico. También podría considerarse que tenemos 11 realizaciones del mismo proceso estocástico o que este proceso estocástico se observa en indices 11×12.

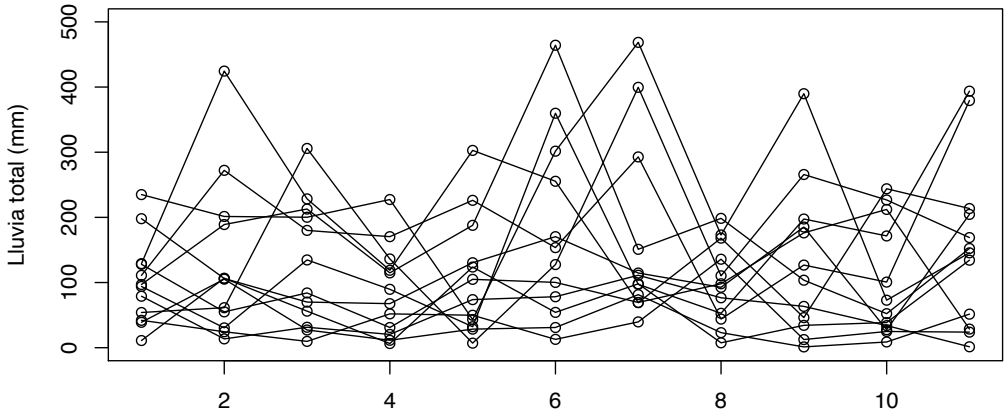


Figura 1.1: Series mensuales de la cantidad de lluvia en Santiago de Compostela entre los años 2008 y 2018.

Ejemplo 1.1.3 (Ruido blanco)**.** Se llama ruido blanco a una serie de tiempo $\{a_t\}$ en la que las variables aleatorias a_t son independientes e idénticamente distribuidas con media cero. Esto implica que todas

[1]Sitio web del IGE (*Instituto Galego de Estatística*): `https://www.ige.eu`.

las variables tienen la misma varianza, σ_a^2, y que $Corr(a_i, a_j) = 0$ para todo $i \neq j$. Si, además, las variables también siguen una distribución normal, es decir, $a_t \sim N(0, \sigma_a^2)$, la serie se denomina ruido blanco gaussiano.

En la Figura 1.2 podemos ver la representación de una simulación de una serie de ruido blanco gaussiano de tamaño 100.

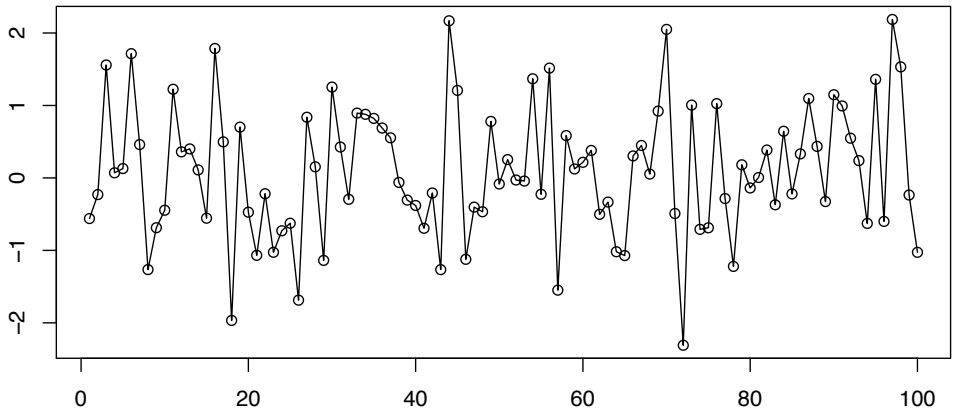


Figura 1.2: Gráfico secuencial de una serie simulada de ruido blanco gaussiano.

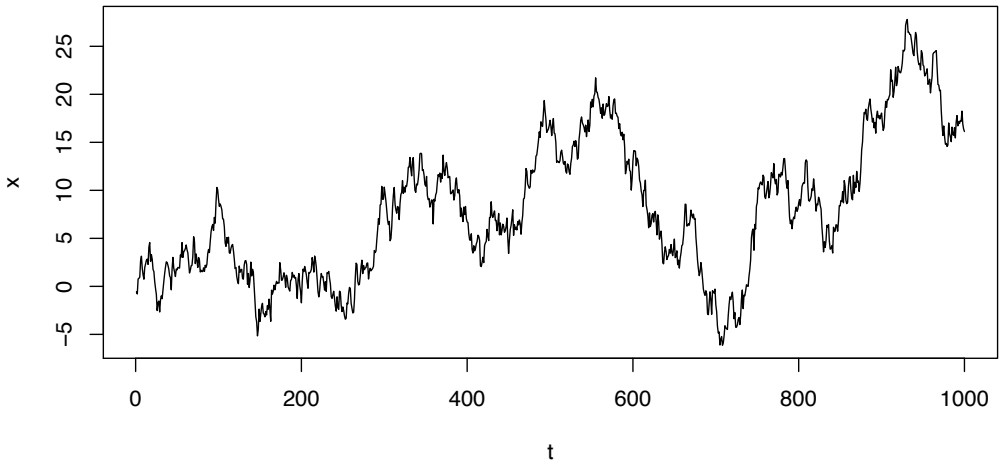


Figura 1.3: Gráfico secuencial de una serie simulada de un paseo aleatorio.

Ejemplo 1.1.4 (Paseo aleatorio)**.** Se llama paseo aleatorio a la serie de tiempo $\{X_t\}$ definida como

$$X_t = X_{t-1} + a_t,$$

donde $\{a_t\}$ es una serie de ruido blanco. Si sustituimos $X_{t-1} = X_{t-2} + a_{t-1}$ en la ecuación anterior y después sustituimos por X_{t-2}, seguido de X_{t-3} y así sucesivamente obtenemos

$$X_t = X_0 + a_t + a_{t-1} + a_{t-2} + \cdots$$

En la práctica, esta serie no será infinita sino que empezará en algún tiempo $t = 1$. Por lo tanto,

$$X_t = X_0 + \sum_{i=1}^{t} a_i.$$

En la Figura 1.3 podemos ver la representación de una simulación de una serie de un paseo aleatorio donde $\{a_t\}$ es gaussiano con $\sigma_a^2 = 1$.

Definición 1.1.5. Para un proceso estocástico $\{X_t\}_{t\in T}$, la función media se define como

$$\mu_t = \mathbb{E}(X_t).$$

Es decir, μ_t es el valor esperado del proceso en el tiempo t. En general, μ_t puede ser diferente en cada tiempo t.

Un caso particular importante aparece cuando todas las variables tienen la misma media y entonces la función de medias es una constante. Las realizaciones del proceso no mostrarán ninguna tendencia y diremos que el proceso es estable en la media. En caso de que las medias cambien con el tiempo, las observaciones de distintos momentos mostrarán dicho cambio.

Definición 1.1.6. La función de varianzas del proceso que proporciona las varianzas en cada instante temporal se define como

$$Var(X_t) = \sigma_t^2.$$

Diremos que el proceso es estable en la varianza si esta es constante en el tiempo. Un proceso puede ser estable en la media y no en la varianza y viceversa.

La falta de independencia entre dos observaciones X_t y X_{t+k} puede evaluarse numéricamente utilizando las nociones de covarianza y correlación. Suponiendo que la varianza de X_t es finita, tenemos la siguiente definición.

Definición 1.1.7. La función de autocovarianza, $\gamma(t, t+k)$, se define como

$$\gamma(t, t+k) = Cov(X_t, X_{t+k}) = \mathbb{E}\left[(X_t - \mu_t)(X_{t+k} - \mu_{t+k})\right] = \mathbb{E}(X_t X_s) - \mu_t \mu_s.$$

En particular,

$$\gamma(t, t) = Var(X_t) = \sigma_t^2.$$

La función de autocovarianzas describe las covarianzas entre dos variables del proceso en dos instantes cualesquiera. En un proceso estocástico la función de medias y la de autocovarianzas cumplen el mismo papel que la media y la varianza para una variable escalar.

Las autocovarianzas tienen dimensiones, por lo que no son convenientes para comparar series medidas en unidades distintas. Podemos obtener una medida adimensional de la dependencia lineal generalizando la idea del coeficiente de correlación lineal entre dos variables.

Definición 1.1.8. La función de autocorrelación simple (*fas*), $\rho(t, t+k)$, está dada por

$$\rho(t, t+k) = Corr(X_t, X_{t+k}) = \frac{Cov(X_t, X_{t+k})}{\sigma_t \sigma_{t+k}} = \frac{\gamma(t, t+k)}{\sqrt{\gamma(t,t)\gamma(t+k, t+k)}}.$$

La función de autocorrelaciones simples mide la correlación entre dos variables separadas por k períodos, es decir, la correlación entre dos variables del proceso en dos instantes cualesquiera. Para medir la correlación entre dos variables del proceso en dos instantes cualesquiera, una vez que se les ha sustraído el efecto lineal que ejercen sobre cada una de ellas las variables medidas en los instantes intermedios existentes entre ambas, utilizaremos la denominada función de autocorrelaciones parciales.

Definición 1.1.9. La función de autocorrelación parcial (*fap*), $P(t, t+k)$, está dada por

$$P(t,t+k) = \frac{Cov\left(X_t - \hat{X}_t^{(t,t+k)}, X_{t+k} - \hat{X}_{t+k}^{(t,t+k)}\right)}{\sqrt{Var\left(X_t - \hat{X}_t^{(t,t+k)}\right) Var\left(X_{t+k} - \hat{X}_{t+k}^{(t,t+k)}\right)}},$$

donde $\hat{X}_j^{(t,t+k)}$ denota al mejor predictor lineal de X_j construido a partir de las variables medidas en los instantes comprendidos entre t y $t+k$.

1.2. Series de tiempo estacionarias

Como vimos, una serie de tiempo es una colección de observaciones de una variable tomadas en intervalos de tiempo regulares. Las características principales de muchas series de tiempo son tendencias y variaciones estacionales, que pueden modelarse de manera determinista con funciones matemáticas del tiempo. Pero otra característica importante de la mayoría de las series de tiempo es que las observaciones próximas en el tiempo tienden a estar correlacionadas (dependientes en serie). Gran parte de la metodología en un análisis de series de tiempo tiene como objetivo explicar esta correlación y las características principales de los datos utilizando modelos estadísticos y métodos descriptivos apropiados. Una vez que se encuentra un buen modelo y se ajusta a los datos, el modelo puede usarse para predecir valores futuros o generar simulaciones para guiar decisiones de planificación. Los modelos ajustados también se utilizan como base para pruebas estadísticas. Finalmente, un modelo estadístico ajustado proporciona un resumen conciso de las características principales de una serie temporal, que a menudo puede ser esencial para la toma de decisiones.

Las series temporales pueden clasificarse en dos clases. Por una parte están las estacionarias, que toman valores estables en el tiempo alrededor de un nivel constante, sin mostrar una tendencia creciente o decreciente a corto plazo. Por otra parte están las series no estacionarias, que son aquellas que incumplen la estabilidad en media, varianza o covarianza. En la práctica la clasificación de una serie como estacionaria o no depende del periodo de observación, ya que una serie puede ser estable en un periodo corto y no estacionario en un periodo mayor.

Aunque las funciones teóricas que hemos descrito en la sección anterior son útiles para describir las propiedades de ciertos modelos hipotéticos, la mayoría de los análisis deben realizarse utilizando datos muestrales. En muchas ocasiones solo podemos observar una realización del proceso. Conceptualmente, el proceso estocástico existe pero no es posible obtener muestras sucesivas o realizaciones independientes del mismo. Para poder estimar las características del proceso (medias, varianzas, etc.) a partir de su evolución en el tiempo es necesario suponer que estas son estables a lo largo del tiempo. Esto conduce al concepto de estacionaridad.

Definición 1.2.1. Diremos que una serie de tiempo es estacionaria en sentido estricto si para todo t:

1. Las distribuciones marginales de todas las variables son idénticas.
2. Las distribuciones finito-dimensionales de cualquier conjunto de variables solo dependen de los retardos entre ellas.

La estacionaridad estricta es una condición muy fuerte, ya que para contrastarla sería necesario disponer de todas las distribuciones conjuntas para cualquier selección de variables del proceso. Una definición que es similar a la de estacionariedad estricta, pero más fácil de contrastar en la práctica,

es la estacionaridad en sentido débil, que conlleva estabilidad de la media, la varianza y la estructura de covarianzas a lo largo del tiempo.

Definición 1.2.2. Diremos que un proceso estocástico es estacionario en sentido débil si para todo t:

1. $\mu_t = \mu = cte.$
2. $\sigma_t^2 = \sigma^2 = cte.$
3. $\gamma(t, t+k) = \gamma_k$, para todo k.

Las dos primeras condiciones indican que la media y la varianza son constantes, y la tercera que la covarianza entre dos variables depende solo de su separación. En un proceso estacionario las autocovarianzas y autocorrelaciones solo dependen del retardo entre las observaciones y, en particular, la relación entre x_t y x_{t+k}, es siempre igual a la relación entre x_t y x_{t-k}. En consecuencia, en los procesos estacionarios,

$$Cov(x_t, x_{t+k}) = Cov(x_{t+j}, x_{t+j+k}) = \gamma_k, \qquad j = 0, \pm 1, \pm 2, \ldots$$

y también, para las autocorrelaciones,

$$\rho_k = \frac{Cov(x_t, x_{t+k})}{\sqrt{Var(x_t)Var(x_{t+k})}} = \frac{\gamma_k}{\gamma_0}.$$

En resumen, en los procesos estacionarios $\gamma_0 = \sigma^2$, $\gamma_k = \gamma_{-k}$ y $\rho_k = \rho_{-k}$.

Para simplificar, utilizaremos la expresión proceso estacionario cuando nos refiramos a un proceso estacionario en sentido débil.

Generalmente las series temporales evolucionan con cierta inercia, que se manifiesta en dependencia entre sus valores recientes y sus valores pasados. Con el objetivo de medir la dependencia lineal entre presente y pasado se utilizan las funciones de autocovarianzas y autocorrelaciones, las cuales generalizan la idea de covarianza y correlación entre dos variables.

Aunque las funciones teóricas de autocorrelación y correlación cruzada son útiles para describir las propiedades de ciertos modelos hipotéticos, la mayoría de los análisis deben realizarse utilizando datos muestrales. Esta limitación significa que solo están disponibles las observaciones $x_1, x_2, \ldots, x_T$ para estimar la media, la autocovarianza y las funciones de autocorrelación.

En consecuencia, si una serie de tiempo es estacionaria, como la función de medias $\mu_t = \mu$ es constante, ya que su media no depende del tiempo t, podemos estimarla por la media muestral,

$$\bar{x} = \frac{1}{T}\sum_{t=1}^{T} x_t.$$

En nuestro caso $\mathbb{E}(\bar{x}) = \mu$.

Asimismo, las funciones teóricas de autocovarianzas, autocorrelaciones simples y parciales, se estimarán a través de su contraparte muestral y se denotarán por $\hat{\gamma}_k$, $\hat{\rho}_k$ y $\hat{P}_k$, respectivamente.

Ejemplo 1.2.3 (Ruido blanco)**.** Recordemos que en un proceso de ruido blanco $\{a_t\}$ las variables aleatorias son incorreladas, con media cero y varianza finita σ_a^2. Entonces, para todo t:

1. $\mu_t = 0.$
2. $\sigma_t^2 = \sigma_a^2.$
3. $\gamma(t, t+k) = Cov(a_t, a_{t+k}) = \mathbb{E}(a_t a_{t+k}) = \begin{cases} \sigma_a^2, & \text{si } k = 0, \\ 0, & \text{si } k \neq 0. \end{cases}$

Por tanto, el ruido blanco es un proceso estacionario. Además, si el tamaño de la serie, T, es grande, se verifica que $\hat{\rho}_k \approx N\left(0, \frac{1}{\sqrt{T}}\right)$ y que $\hat{P}_k \approx N\left(0, \frac{1}{\sqrt{T}}\right)$, y por lo tanto debería cumplirse que $\hat{\rho}_k, \hat{P}_k \in \left(-\frac{1.96}{\sqrt{T}}, \frac{1.96}{\sqrt{T}}\right)$, para $k = 1, 2, \ldots, T$.

Ejemplo 1.2.4 (Paseo aleatorio)**.** Recordemos que un paseo aleatorio es una serie de tiempo $\{X_t\}$ definida como $X_t = X_0 + \sum_{i=1}^{t} a_i$, donde $\{a_t\}$ es una serie de ruido blanco. Si comienza en $t = 0$ y suponemos que $X_0 = 0$, para todo t se tiene que:

1. $\mu_t = 0$.
2. $\sigma_t^2 = Var\left(\sum_{i=1}^{t} a_i\right) = \sum_{i=1}^{t} Var(a_i) + 2\sum_{1\leq i<j\leq t} Cov(a_i, a_j) = t\sigma_a^2$.
3. $\gamma(t, t+k) = Cov(X_t, X_{t+k}) = Cov(X_t, X_t + \sum_{i=1}^{k} a_{t+i}) = Var(X_t) = t\sigma_a^2$.

Por tanto un paseo aleatorio no es un proceso estacionario. Este proceso no es estable en la varianza, de hecho esta es creciente, y tampoco en las autocovarianzas.

1.3. Ejemplos de series de tiempo

Las series temporales se analizan para entender el pasado y predecir el futuro, permitiendo así tomar decisiones informadas de forma adecuada. Un análisis de series de tiempo cuantifica las características principales de los datos y la variación aleatoria. Estos motivos, combinados con una potencia informática mejorada, hacen que los métodos de series temporales sean ampliamente aplicables en el gobierno, la industria y el comercio.

Uno de los pasos más importantes en un análisis preliminar de una serie de tiempo es representar su gráfico secuencial, es decir, representar cada observación frente al instante en el que se observa. Este gráfico permite observar cómo evoluciona la serie a lo largo del tiempo e identificar las características principales de una serie temporal: tendencia, estacionalidad y heterocedasticidad. Si una serie oscila alrededor de un nivel constante diremos que es estacionaria y en caso contrario no estacionaria. Cuando la serie presenta un comportamiento superpuesto que se repite a lo largo del tiempo, diremos que la serie es estacional. En el caso de que la variabilidad de la serie no sea constante diremos que es heterocedástica.

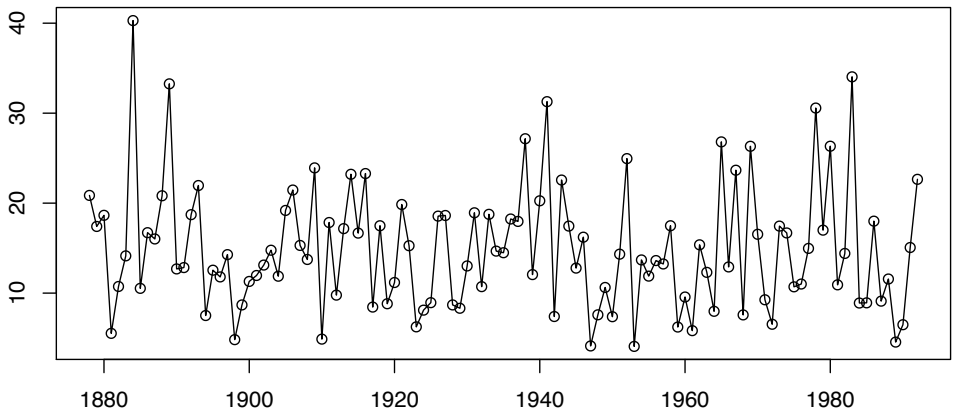


Figura 1.4: Precipitación anual en Los Ángeles en el periodo 1878 a 1992.

En la Figura 1.4 tenemos un ejemplo de serie estacionaria. Muestra el conjunto `larain`, disponible en el paquete `TSA` (Chan y Ripley, 2018), que corresponde a la precipitación anual (en pulgadas) en

Los Ángeles entre 1878 y 1992. Observamos que la serie es estable, con valores que oscilan alrededor de una precipitación promedio anual aproximada de 15 pulgadas, y además el gráfico no muestra ninguna clara tendencia creciente o decreciente.

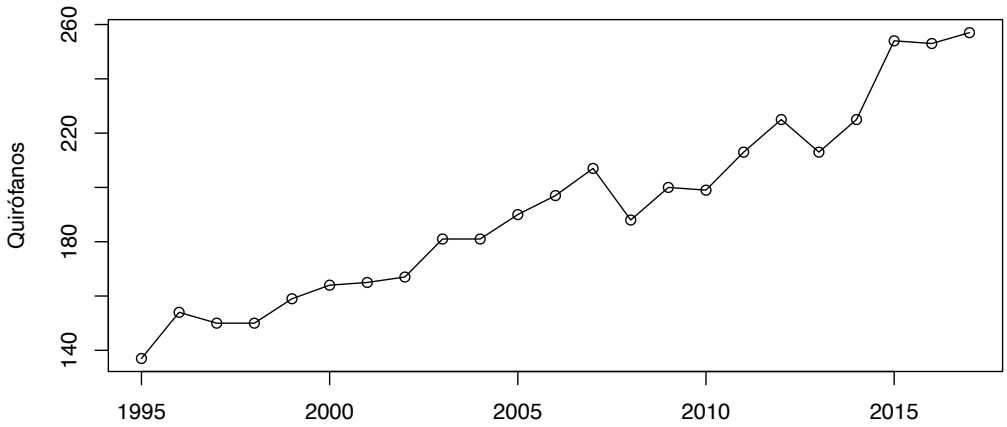


Figura 1.5: Quirófanos disponibles en Galicia entre 1995 y 2017.

La Figura 1.5 presenta el número de quirófanos disponibles en Galicia (datos disponibles en el IGE). El gráfico muestra que la serie anual no oscila alrededor de ningún nivel fijo, ya que su nivel aumenta con el tiempo. Por tanto, diremos que esta serie tiene una clara tendencia positiva. La tendencia de esta serie sería variable en el tiempo en lugar de constante. Esta propiedad es frecuente en las series reales, donde es raro encontrar una tendencia constante en periodos largos de observación.

La Figura 1.6 es otro ejemplo de serie estacionaria. Presenta la temperatura media mensual en Santiago de Compostela (datos disponibles en el IGE). La serie oscila alrededor de una temperatura media mensual de aproximadamente 10ºC. Al igual que la serie de precipitaciones en Los Ángeles, esta serie tiene un nivel fijo, sin tendencia clara a crecer o decrecer en el tiempo. Sin embargo, en este caso podemos observar que además de oscilar los valores de la serie alrededor de un valor central, en algunos meses las temperaturas son más altas que en otros. Por ejemplo, los meses de verano tienen temperaturas más altas que los meses de invierno. Este efecto puede verse mejor haciendo gráficos para cada uno de los meses del año. Como ilustración, la Figura 1.7 presenta la temperatura media de los meses de enero y agosto de distintos años. En ambos casos las series oscilan alrededor de un valor central, habiendo años en los que las temperaturas son más altas y otros en los que son más bajas. Lo más destacado del gráfico es que los valores medios de ambas series son muy distintos, siendo la temperatura media en agosto mucho mayor que en enero. Este fenómeno, en el que el valor medio de la variable observada depende del mes considerado se denomina estacionalidad.

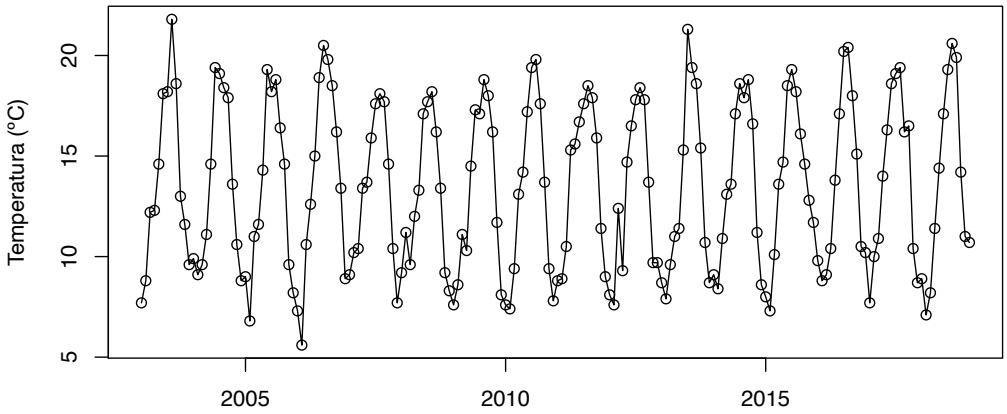


Figura 1.6: Temperatura media mensual en Santiago de Compostela entre 2003 y 2018.

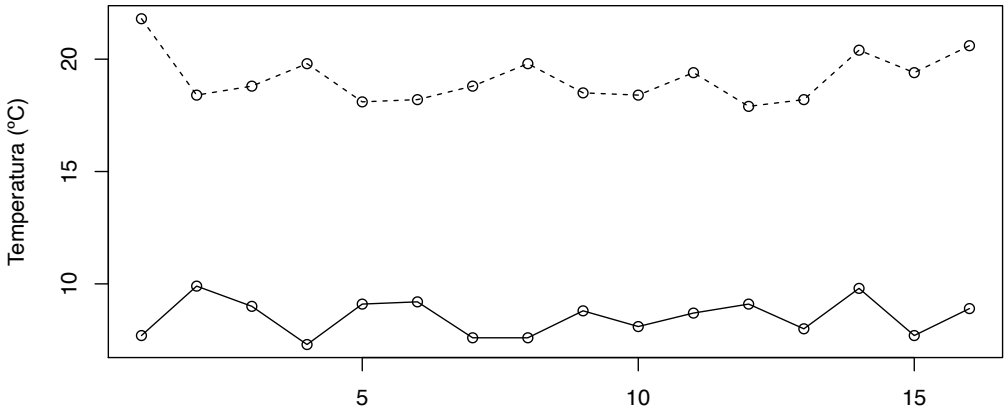


Figura 1.7: Temperatura media mensual en Santiago de Compostela en los meses de enero (línea continua) y agosto (línea discontinua).

Por último, consideremos el conjunto `AirPassengers` disponible en el paquete `datasets` (R Core Team, 2018), que son los totales mensuales de los pasajeros de aerolíneas internacionales entre 1949 y 1960, representado en la Figura 1.8. Es un ejemplo de serie con tendencia positiva y estacionalidad, pero además, en este caso, es heterocedástica, ya que la variabilidad aumenta con el nivel, siendo mayor en las observaciones más recientes.

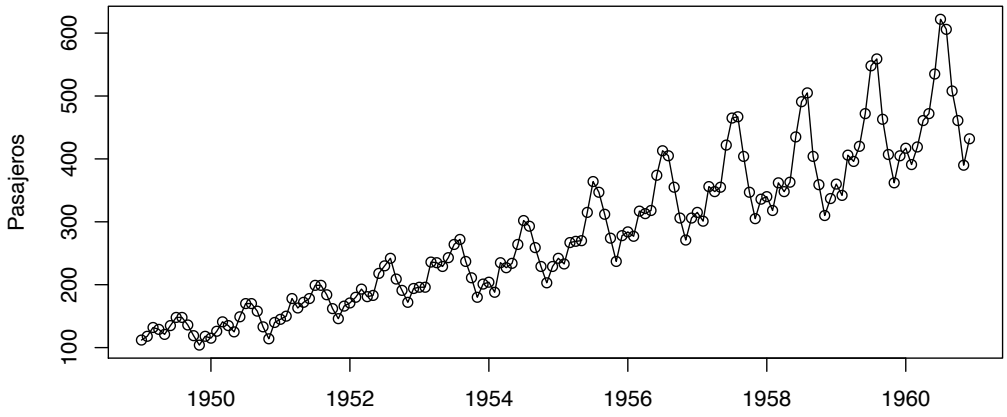


Figura 1.8: Pasajeros mensuales de aerolíneas internacionales entre 1949 y 1960.

Las series temporales pueden tener o no un nivel estable en el tiempo. En caso de no tenerlo puede presentar tendencias más o menos constantes. Cuando el nivel de la serie no es estable decimos que la serie no es estacionaria. Un caso particular de no estacionaridad es cuando el nivel varía manteniendo un período. Diremos entonces que la serie es estacional y esta faceta puede combinarse con tendencias más o menos acusadas en el nivel general. El gráfico de la serie es siempre una herramienta muy valiosa para entender su comportamiento.

Las series de tiempo se analizan para entender el pasado y predecir el futuro, permitiendo así una toma de decisiones informada. El principal objetivo del modelado de series temporales es comprender la estructura de dependencia temporal de las observaciones de una serie (análisis de series temporales univariantes) y determinar las principales relaciones entre varias series (análisis de series de tiempo multivariantes). En este trabajo nos centraremos en variables unidimensionales, por tanto trataremos con series de tiempo univariantes.

Es necesario buscar un modelo matemático que se ajuste adecuadamente los datos. Una vez elegido un modelo, pueden estimarse sus parámetros, verificar la adecuación de los datos y emplear el modelo ajustado para mejorar nuestra comprensión del mecanismo que genera los datos. Una vez que se ha desarrollado un modelo satisfactorio, puede utilizarse para predecir observaciones futuras. Para poder hacer inferencia estadística, supondremos que la serie de tiempo ha sido generada por un modelo estocástico.

Encontrar modelos apropiados para series de tiempo no es una tarea trivial. La estrategia de construcción de modelos tiene tres pasos principales:

1. Identificación del modelo.
2. Ajuste del modelo.
3. Diagnóstico del modelo.

En la identificación del modelo, se seleccionan las clases de modelos de series de tiempo que pueden ser apropiadas para una serie observada dada. En este paso, observamos el gráfico secuencial de la serie, calculamos distintos estadísticos de los datos y también aplicamos cualquier conocimiento del tema en el que surgen los datos, como biología, negocios o ecología. Cabe destacar que el modelo

elegido en este punto es provisional y está sujeto a revisión más adelante en el análisis. Al elegir un modelo, intentaremos que el modelo utilizado requiera el menor número de parámetros que representen adecuadamente la serie de tiempo.

Necesariamente el modelo involucrará uno o más parámetros cuyos valores deben estimarse a partir de las series observadas. El ajuste del modelo consiste en encontrar las mejores estimaciones posibles de estos parámetros desconocidos dentro de un modelo dado.

Finalmente, el diagnóstico del modelo se refiere a evaluar la calidad del modelo que hemos identificado y estimado. Si no se encuentran deficiencias, puede suponerse que el modelado está completo, y el modelo puede usarse, por ejemplo, para predecir valores futuros. En caso contrario, elegimos otro modelo, es decir, volvemos al paso de identificación del modelo. De esta manera, recorreremos los tres pasos hasta que, idealmente, encontremos un modelo aceptable.

1.4. Tranformaciones para conseguir estacionaridad

Hemos visto que un proceso es estacionario cuando tiene media y varianza constantes, y las autocovarianzas solo dependen del retardo. Cuando la falta de estacionaridad se debe a la inestabilidad en media y/o en varianza, a través de ciertas transformaciones, la serie de tiempo puede transformarse en una serie estacionaria.

1.4.1. Transformaciones para estabilizar la media

La mayoría de las series de tiempo reales no son estacionarias y su nivel medio varía con el tiempo. El método más habitual para eliminar la tendencia es la diferenciación.

Definición 1.4.1. Se define el operador diferencia regular como

$$\nabla = (1 - B),$$

donde B denota el operador retardo, definido por $Bx_t = x_{t-1}$. En general, se define el operador diferencia de orden d como

$$\nabla^d = (1 - B)^d.$$

Cuando la decisión de diferenciar no sea clara analizando el gráfico secuencial de la serie, estudiaremos su función de autocorrelación simple. En el caso de las series no estacionarias, mostrará autocorrelaciones positivas, con decrecimiento lento y lineal. Por tanto, si la *fas* no decrece rápidamente, en general será necesario diferenciar para obtener un proceso estacionario.

Ejemplo 1.4.2 (Paseo aleatorio)**.** El paseo aleatorio $X_t = X_{t-1} + a_t$, donde $\{a_t\}$ es una serie de ruido blanco, no es un proceso estacionario. Sin embargo, el paseo aleatorio diferenciado regularmente, $\nabla X_t = a_t$, sí es estacionario.

A veces será necesario diferenciar más de una vez una serie para obtener un proceso estacionario. Diremos que 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 $I(0)$. En la práctica la mayoría de las series no estacionarias que son integradas tienen un orden $h \leq 3$.

Hemos visto que una serie no estacionaria se puede convertir en una estacionaria tomando diferencias regulares, es decir, entre períodos consecutivos. Asimismo, una serie estacional puede convertirse en una serie no estacional aplicando una diferencia estacional.

Definición 1.4.3. Se define el operador diferencia estacional de período s como

$$\nabla_s = 1 - B^s$$

(nótese que $\nabla_s \neq \nabla^s = (1 - B)^s$).

Si aplicamos dicho operador a una serie, obtenemos una serie transformada que sustituye en cada instante, t, el valor de la serie por la diferencia entre el valor en t, y el valor en $t-s$. En efecto,

$$\nabla_s x_t = (1 - B^s)x_t = x_t - x_{t-s}.$$

Concluimos entonces que podemos convertir una serie no estacionaria con estacionalidad en estacionaria mediante la transformación

$$y_t = \nabla_s^D \nabla^d x_t,$$

donde D es el número de diferencias estacionales (si hay estacionalidad casi siempre $D = 1$, y si no la hay $D = 0$) y d el número de diferencias regulares ($d \leq 3$).

En la práctica, la decisión de aplicar diferencias regulares y estacionales puede basarse en el gráfico secuencial de la serie y en la función de autocorrelación muestral, pero también pueden realizarse contrastes formales como el de raíces unitarias o el de Dickey-Fuller (Peña et al., 2001).

1.4.2. Transformaciones para estabilizar la varianza

Frecuentemente nos encontramos con series de tiempo donde una mayor dispersión parece estar asociada con niveles más altos de la serie, es decir, cuanto mayor es el nivel de la serie, más variación hay alrededor de ese nivel y viceversa.

Si las observaciones originales son $X_1, X_2, \ldots, X_t$, la transformación Box-Cox f_λ las convierte en $f_\lambda(X_1), f_\lambda(X_2), \ldots, f_\lambda(X_t)$, donde

$$f_\lambda(x) = \begin{cases} \dfrac{x^\lambda - 1}{\lambda}, & \text{si } \lambda \neq 0, \\ \log(x), & \text{si } \lambda = 0. \end{cases}$$

Esta transformación se aplica solo a valores de datos positivos. Si algunos de los valores son negativos o cero, se puede agregar una constante positiva a todos los valores para hacerlos todos positivos antes de realizar la transformación.

Si la desviación típica es una función potencial de la media del tipo $\sigma_t = k\mu_t^{1-\lambda}$, la transformación Box-Cox con parámetro λ estabiliza la varianza (Peña, 2005).

Estas transformaciones son útiles cuando la variabilidad de los datos aumenta o disminuye con el nivel. Mediante una elección adecuada de λ, la variabilidad a menudo puede hacerse casi constante[2]. En particular, para datos positivos cuya desviación estándar aumenta linealmente con el nivel, la variabilidad puede estabilizarse eligiendo $\lambda = 0$.

Ejemplo 1.4.4. Consideremos de nuevo el conjunto `AirPassengers` representado en la Figura 1.8. Habíamos observado que la serie era heterocedástica. Sin embargo, tras aplicar una transformación Box-Cox apropiada, la serie transformada resultante, que puede observarse en la Figura 1.9, sí es homocedástica.

[2]Los métodos para la selección de λ pueden verse en Johnson y Wichern (2007, p. 193-194).

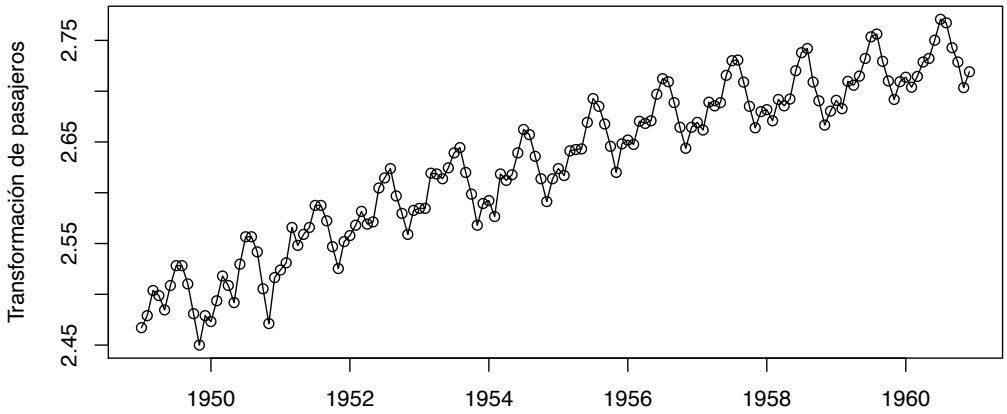


Figura 1.9: Transformación Box-Cox de la serie `AirPassengers` con $\hat{\lambda} = -0.31$.

Capítulo 2

Modelización de series temporales

Como hemos visto anteriormente, muchas series son no estacionarias debido a tendencias o varianzas no constantes. En particular, los paseos aleatorios, que caracterizan muchos tipos de series, no son estacionarios, pero pueden transformarse en series estacionarias mediante una diferenciación de primer orden. En este capítulo primeramente extenderemos el paseo aleatorio para incluir términos autorregresivos y de medias móviles. Como las series diferenciadas deben agregarse (o "integrarse") para recuperar la serie original, el proceso estocástico subyacente se denomina ARIMA.

El proceso ARIMA puede extenderse para incluir términos estacionales, dando un proceso ARIMA estacional no estacionario. Los modelos ARIMA estacionales son herramientas potentes en el análisis de series de tiempo, ya que son capaces de modelar una gama muy amplia de series. Gran parte de la metodología fue iniciada por Box y Jenkins en la década de 1970.

2.1. Modelos ARMA

Vamos a iniciar el estudio de modelos de procesos estacionarios que son útiles en la práctica para representar la dependencia de los valores de una serie temporal de su pasado. Los modelos más simples son los autorregresivos, propuestos por Yule (1927), que generalizan la idea de regresión para representar la dependencia lineal entre dos variables aleatorias.

Los modelos autorregresivos se basan en la idea de que el valor actual de la serie, x_t, puede explicarse como función de p valores pasados, $x_{t-1}, x_{t-2}, \ldots, x_{t-p}$, donde p determina el número necesario de pasos hacia el pasado para predecir el valor actual.

Definición 2.1.1. Un modelo autorregresivo de orden p, denotado como AR(p), tiene la forma

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p} + a_t, \tag{2.1}$$

donde x_t es estacionario, $\phi_1, \ldots, \phi_p$ ($\phi_p \neq 0$) son parámetros y a_t es un proceso de ruido blanco con varianza σ_a^2. La media de x_t en (2.1) es cero. Si la media, μ, de x_t no es nula, fijamos $c = \mu(1 - \phi_1 - \cdots - \phi_p)$ y escribimos el modelo como

$$x_t = c + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p} + a_t.$$

El proceso AR(p) puede escribirse utilizando la notación del operador de retardo, B, definido por $Bx_t = x_{t-1}$, como

$$(1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p) x_t = a_t,$$

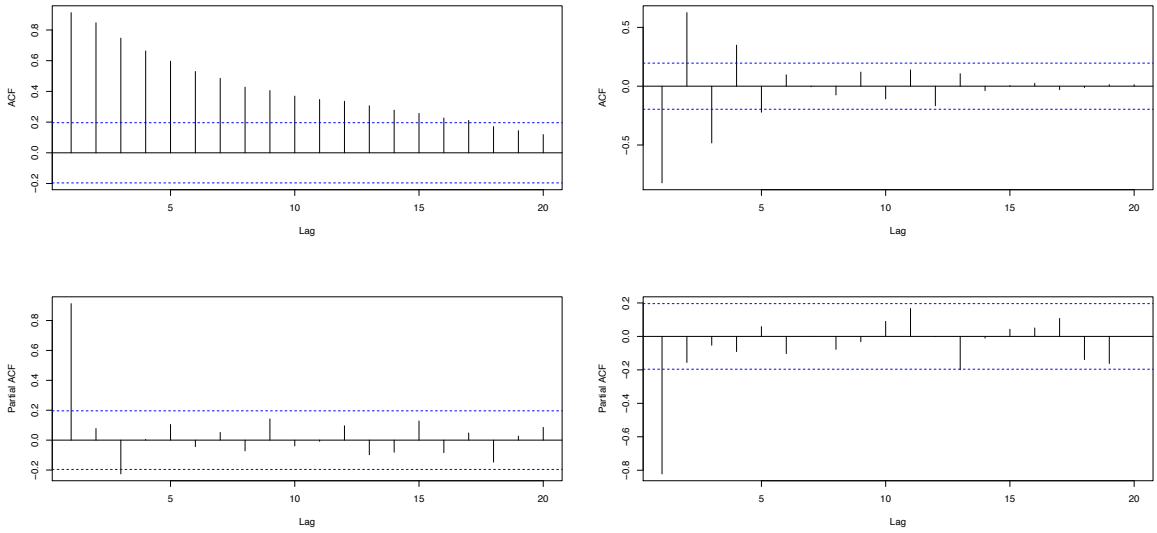
y llamando $\phi_p(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$ al polinomio de grado p en el operador de retardo, cuyo primer término es la unidad, tenemos:

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

que es la expresión general de un proceso autorregresivo.

Cuando las raíces de la ecuación característica $\phi_p(B) = 0$ estén fuera del círculo unidad el proceso AR(p) será estacionario (Brockwell y Davis, 1991).

Un proceso AR(p) tendrá los p primeros coeficientes de autocorrelación parcial distintos de cero, y, por tanto, en la *fap* el número de coeficientes distintos de cero indica el orden del proceso AR (Peña, 2005). Esta propiedad va a ser clave para identificar el orden de un proceso autorregresivo. La Figura 2.1 resume la *fas* y la *fap* de distintos procesos AR. Observamos que, en ambos casos, la última $\hat{P}_k$ en salir del intervalo $\left(-\frac{1.96}{\sqrt{T}}, \frac{1.96}{\sqrt{T}}\right)$ es $\hat{P}_1$.



(a) AR(1) con $\phi_1 > 0$.

(b) AR(1) con $\phi_1 < 0$.

Figura 2.1: Funciones de autocorrelación simple y parcial para procesos AR.

Como alternativa a la representación autorregresiva, el modelo de medias móviles de orden q, abreviado como MA(q), asume que el ruido blanco a_t se combina linealmente para formar los datos observados. Estos modelos fueron considerados por primera vez por Slutsky (1937) y Wold (1938).

Definición 2.1.2. Un modelo de medias móviles de orden q, o modelo MA(q), tiene la forma

$$x_t = a_t + \theta_1 a_{t-1} + \theta_2 a_{t-2} + \cdots + \theta_q a_{t-q},$$

donde $\theta_1, \theta_2, \ldots, \theta_q$ ($\theta_q \neq 0$) son parámetros y a_t es un proceso de ruido blanco con varianza σ_a^2.

Introduciendo la notación de operadores, el proceso MA(q) puede escribirse como

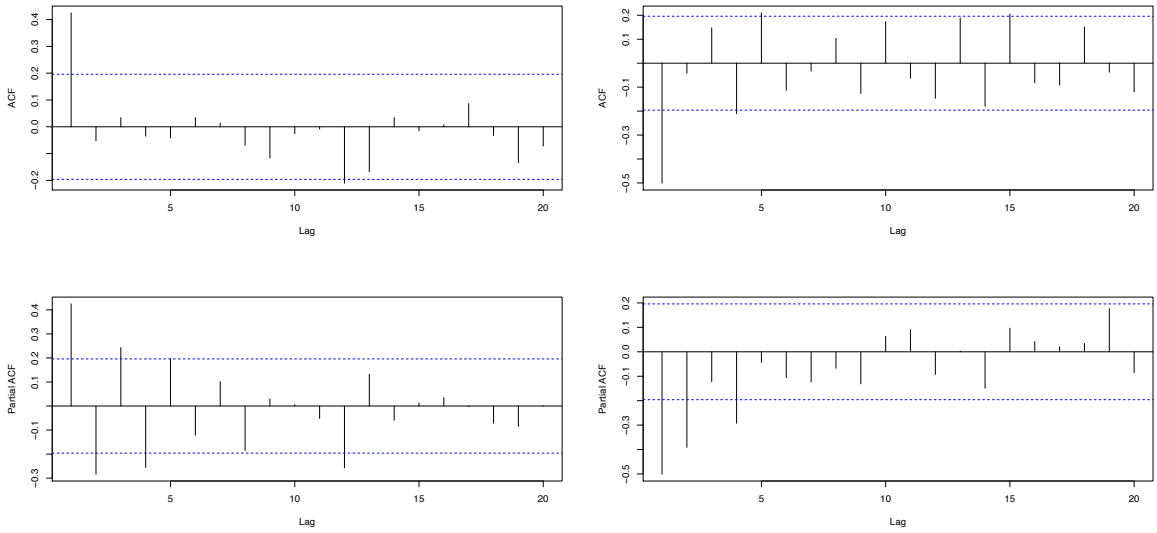
$$x_t = (1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q) a_t$$

y a su vez, siendo $\theta_q(B) = 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q$ el polinomio de grado q en el operador de retardo, puede escribirse de forma más compacta como

$$x_t = \theta_q(B) a_t.$$

Debido a que los procesos de medias móviles consisten en una suma finita de términos de ruido blanco estacionario, son estacionarios y, por lo tanto, tienen una media y una autocovarianza invariantes en el tiempo.

En el caso de los procesos MA(q), los q primeros coeficientes de autocorrelación simple serán distintos de cero, y, por tanto, en la *fas* el número de coeficientes distintos de cero indicará el orden del proceso MA. La Figura 2.2 presenta las funciones *fas* y *fap* de distintos procesos MA. Observamos que, en ambos casos, la última $\hat{\rho}_k$ en salir del intervalo $\left(-\frac{1.96}{\sqrt{T}}, \frac{1.96}{\sqrt{T}}\right)$ es $\hat{\rho}_1$.



(a) MA(1) con $\theta_1 > 0$.

(b) MA(1) con $\theta_1 < 0$.

Figura 2.2: Funciones de autocorrelación simple y parcial para procesos MA.

Los procesos autorregresivos y de medias móviles son casos particulares de una representación general de procesos estacionarios obtenida por Wold (1938). Demostró que todo proceso estocástico x_t estacionario de media finita μ, que no contenga componentes deterministas, puede escribirse como un MA(∞), es decir, como una función lineal de variables aleatorias incorreladas a_t de la forma

$$x_t = \mu + \sum_{i=0}^{\infty} \psi_i a_{t-i}, \tag{2.2}$$

donde $\psi_0 = 1$, $\mathbb{E}(x_t) = \mu$, $\mathbb{E}(a_t) = 0$, $Var(a_t) = \sigma_a^2$ y $\mathbb{E}(a_t a_{t-k}) = 0$ si $k > 1$. Llamando $\tilde{x}_t = x_t - \mu$, y utilizando el operador de retardo, podemos escribir

$$\tilde{x} = \psi(B) a_t,$$

donde $\psi(B) = 1 + \psi_1 B + \psi_2 B^2 + \cdots$.

La varianza de x_t en (2.2) será

$$Var(x_t) = \sigma_a^2 \sum_{i=0}^{\infty} \psi_i^2,$$

y para que el proceso tenga varianza finita, la serie $\{\psi_i^2\}$ debe ser convergente. Observamos que si los coeficientes ψ_i se anulasen a partir de un retardo q, el modelo general se reduce a un MA(q).

Una consecuencia de (2.2) es que todo proceso estacionario también admite una representación autorregresiva, que puede ser de un orden infinito. Esta representación AR(∞) es la inversa de la de Wold, y escribiremos

$$\tilde{x}_t = a_t + \pi_1 \tilde{x}_{t-1} + \pi_2 \tilde{x}_{t-2} + \cdots,$$

que en notación de operadores se reduce a

$$\pi(B)\tilde{x}_t = a_t.$$

Si un proceso admite una representación del tipo MA(∞) diremos que es causal y si admite una representación del tipo AR(∞) diremos que es invertible.

Por una parte, es evidente que los procesos AR son invertibles y que los procesos MA son causales. Por la otra, Shumway y Stoffer (2017) demuestran tanto que un proceso AR es causal cuando las raíces del polinomio $\phi_p(B)$ están fuera del círculo unidad, como que un proceso MA es invertible cuando las raíces del polinomio $\theta_q(B)$ están fuera del círculo unidad.

Ahora procedemos con el desarrollo general de modelos autorregresivos, medias móviles y autorregresivo de medias móviles mixto (ARMA) para series de tiempo estacionarias.

Definición 2.1.3. Una serie de tiempo $\{x_t\}$ es ARMA(p,q) si es estacionaria y

$$x_t = \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + a_t + \theta_1 a_{t-1} + \cdots + \theta_q a_{t-q},$$

con $\phi_p \neq 0$ y $\theta_q \neq 0$. Los parámetros p y q se denominan orden autorregresivo y de media móvil, respectivamente. Si x_t tiene una media μ no nula, fijamos $c = \mu(1 - \phi_1 - \cdots - \phi_p)$ y escribimos el modelo como

$$x_t = c + \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + a_t + \theta_1 a_{t-1} + \cdots + \theta_q a_{t-q}. \tag{2.3}$$

Observamos que, cuando $q = 0$, el modelo resultante es el modelo autorregresivo de orden p, AR(p), y cuando $p = 0$, el modelo resultante es el modelo de medias móviles de orden q, MA(q). En particular, el modelo ARMA(p,q) en (2.3) se puede escribir en forma compacta como

$$\phi_p(B)x_t = c + \theta_q(B)a_t,$$

donde $\phi_p(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$, $\theta_q(B) = 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q$ y B denota el operador retardo, definido como $Bx_t = x_{t-1}$.

Un proceso ARMA será estacionario si las raíces de $\phi_p(B)$ están fuera del círculo unidad, e invertible si lo están las de $\theta_q(B)$.

La *fas* y la *fap* de los procesos ARMA son el resultado de la superposición de sus propiedades AR y MA: en la *fas* ciertos coeficientes iniciales que dependen del orden de la parte MA y después un decrecimiento dictado por la parte AR. En la *fap* valores iniciales dependientes del orden del AR seguidos de un decrecimiento debido a la parte MA. Esta estructura compleja hace que el orden de un proceso ARMA sea difícil de identificar en la práctica. El cuadro 2.1 resume estas características. La identificación de los órdenes p y q del modelo se realiza comparando las funciones estimadas de autocorrelación simple y parcial con las funciones teóricas de procesos ARMA.

	AR(p)	MA(q)	ARMA(p,q)
fas	Decrecimiento rápido en los valores	Último retardo no nulo en q	Primeros q coeficientes no nulos y luego decrecimiento rápido
fap	Último retardo no nulo en p	Decrecimiento rápido en los valores	Primeros p coeficientes no nulos y luego decrecimiento rápido

Cuadro 2.1: Comportamiento de las funciones de autocorrelación simple y parcial en modelos ARMA.

Los procesos ARMA son una familia muy flexible de procesos estacionarios. Sin embargo, no abundan series reales generadas por procesos estacionarios, ya que las series reales suelen presentar tendencia y/o patrones repetitivos. Por tanto es necesario ampliar la clase de procesos ARMA, de modo que la nueva clase permita incorporar estas características.

2.2. Modelos integrados

Los procesos no estacionarios más importantes son los procesos integrados, que como vimos, tienen la propiedad fundamental de que al diferenciarlos se obtienen procesos estacionarios. Una propiedad importante que distingue a los procesos integrados de los estacionarios es la forma en que desaparece la dependencia con el tiempo. En los procesos estacionarios ARMA las autocorrelaciones disminuyen geométricamente, y se hacen prácticamente cero a los pocos retardos. En los procesos integrados las autocorrelaciones disminuyen linealmente con el tiempo y es posible encontrar coeficientes de autocorrelación distintos de cero hasta retardos muy altos.

Ya hemos visto la importancia de la clase de modelos ARMA para representar series estacionarias. Los procesos ARIMA proporcionan una generalización de esta clase, que incorpora una amplia gama de series no estacionarias, es decir, procesos que, después de diferenciar muchas veces, se reducen a procesos ARMA.

Definición 2.2.1. Un proceso x_t se dice que es un ARIMA(p,d,q) si

$$\nabla^d x_t = (1-B)^d x_t$$

es un ARMA(p,q). En general, el modelo se escribe como

$$\phi_p(B)\nabla^d x_t = c + \theta_q(B)a_t, \tag{2.4}$$

donde $c = \mu(1-\phi_1-\cdots-\phi_p)$, $\phi_p(B) = 1-\phi_1 B-\phi_2 B^2-\cdots-\phi_p B^p$ y $\theta_q(B) = 1+\theta_1 B+\theta_2 B^2+\cdots+\theta_q B^q$.

Ejemplo 2.2.2. Consideremos el proceso $x_t = x_{t-1}+a_t+\theta a_{t-1}$, donde θ es un parámetro y a_t es ruido blanco. Operando y expresando en términos del operador de retroceso obtenemos que $(1-B)x_t = (1+\theta B)a_t$. Comparando esta ecuación con la ecuación (2.4), podemos ver que $\{x_t\}$ es un proceso ARIMA(0,1,1), que también se denomina IMA(1,1). En general, ARIMA(0,d,q)≡IMA(d,q).

Ejemplo 2.2.3. Consideremos el proceso $x_t = (1+\phi)x_{t-1} - \phi x_{t-2} + a_t$, donde $|\phi| < 1$ y a_t es ruido blanco. Operando y expresando en términos del operador de retroceso obtenemos que $(1-\phi B)(1-B)x_t = a_t$, y por tanto es un proceso ARIMA(1,1,0), que se denomina como ARI(1,1). En general, ARIMA(p,d,0)≡ARI(p,d).

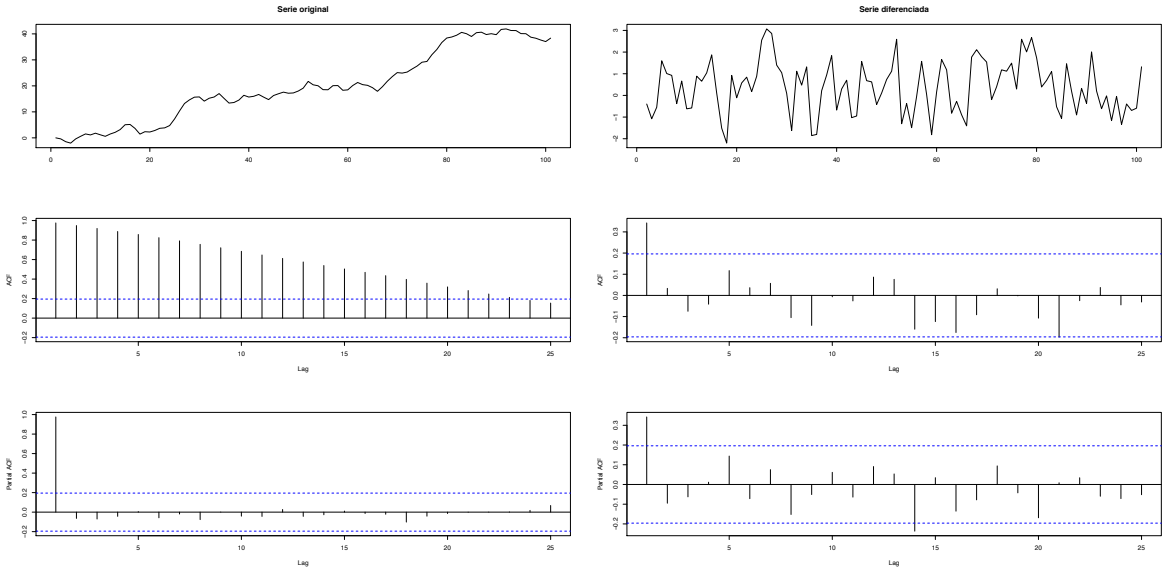


Figura 2.3: En la primera columna simulación de un proceso ARIMA(1,1,0) y sus funciones *fas* y *fap*. En la segunda columna, mismos gráficos sobre el proceso diferenciado.

En la práctica, se propone un modelo ARIMA como posible generador cuando el proceso muestre no estacionariedad provocada exclusivamente por la presencia de tendencia. Para detectar la tendencia observaremos el gráfico secuencial de la serie y la *fas* muestral, que tendrá coeficientes positivos que se amortiguarán aproximadamente de forma lineal a medida que crece el retardo. La Figura 2.3 muestra una serie de tiempo generada a partir de un proceso ARIMA(1,1,0). En el gráfico secuencial de la serie observamos una fuerte tendencia y, además, su correspondiente *fas* decae lentamente a cero. Sin embargo, en el caso de la serie diferenciada la tendencia del gráfico secuencial ha desaparecido y, además, la *fas* y la *fap* permiten identificar un proceso ARMA(1,0).

2.3. Modelos ARIMA estacionales

A continuación, presentaremos modificaciones relizadas al modelo ARIMA para tener en cuenta el comportamiento estacional y no estacionario. A menudo, la dependencia del pasado tiende a ocurrir con mayor fuerza en múltiplos de algún retardo estacional s. Por ejemplo, con los datos económicos mensuales, hay un fuerte componente anual que ocurre en retardos que son múltiplos de $s = 12$, debido a las fuertes conexiones de todas las actividades con el calendario anual. Los datos tomados trimestralmente tendrán el periodo repetitivo anual en $s = 4$. Los fenómenos naturales como la temperatura también tienen componentes fuertes que corresponden a las estaciones. Por lo tanto, la variabilidad natural de muchos procesos físicos, biológicos y económicos tiende a coincidir con las fluctuaciones estacionales.

Debido a esto, es apropiado introducir polinomios de autorregresión y media móvil que se identifiquen con los retardos estacionales. Modelando de forma separada la dependencia regular y la estacional, puede construirse un modelo que las incorpore a ambas de forma multiplicativa.

Definición 2.3.1. El modelo ARIMA estacional multiplicativo o modelo SARIMA viene dado por

$$\Phi_P(B^s)\phi_p(B)\nabla_s^D\nabla^d x_t = c + \Theta_Q(B^s)\theta_q(B)a_t,$$

donde a_t es el proceso habitual de ruido blanco gaussiano y $\nabla^d = (1-B)^d$ y $\nabla_s^D = (1-B^s)^D$ son los operadores diferencia regular y estacional. El modelo general se denota por ARIMA$(p,d,q)\times(P,D,Q)_s$. Las componentes de la media móvil y autorregresiva ordinaria están representadas por los polinomios $\phi_p(B)$ y $\theta_q(B)$ de órdenes p y q, respectivamente, y las componentes estacionales autorregresivas y de medias móviles por $\Phi_P(B^s)$ y $\Theta_Q(B^s)$ de órdenes P y Q, respectivamente.

Ejemplo 2.3.2. Un modelo AR simple con un período estacional de 12 unidades, denotado como ARIMA$(0,0,0)\times(1,0,0)_{12}$, es $x_t = c + x_{t-12} + a_t$. Dicho modelo sería apropiado para datos mensuales cuando solo el valor en el mismo mes del año anterior influye en el valor mensual actual.

Ejemplo 2.3.3. Un modelo trimestral simple de media móvil estacional es $x_t = a_t + \beta a_{t-4}$. Este proceso es estacionario y solo es adecuado para datos sin tendencia. Si los datos también contienen una tendencia, el modelo podría extenderse para incluir diferencias de primer orden, $x_t = x_{t-1} + a_t + \beta a_{t-4}$, que es un proceso ARIMA$(0,1,0)\times(0,0,1)_4$. Alternativamente, si los términos estacionales contienen una tendencia, puede aplicarse la diferencia en el período estacional para dar $x_t = x_{t-4} + a_t + \beta a_{t-4}$, que es un ARIMA$(0,0,0)\times(0,1,1)_4$.

El cuadro 2.2 presenta un resumen de las principales propiedades de la *fas* y *fap* para procesos ARIMA estacionales multiplicativos. En la práctica, una vez diferenciada la serie, para identificar los valores de p, q, P y Q, lo recomendable es fijarse en los retardos iniciales $1, 2, 3, 4, \ldots$ para identificar la estructura regular y en los retardos estacionales, $s, 2s, 3s, \ldots$, para identificar la estructura estacional.

fas	*fap*
1. En los retardos bajos ($j = 1, \ldots, \lfloor s/2 \rfloor$) se observará únicamente la parte regular y en los retardos estacionales se observará la parte estacional. 2. Alrededor de los retardos estacionales observaremos la interacción entre la parte regular y estacional, que se manifestará en la repetición a ambos lados de cada retardo estacional de la *fas* de la parte regular.	1. En los retardos bajos aparecerá la *fap* de la parte regular y en los retardos estacionales se observará la parte estacional. 2. A la derecha de cada retardo estacional aparecerá la *fap* de la parte regular (invertida, en caso de que la *fap* en el retardo estacional es positiva). A la izquierda de los retardos estacionales, observaremos la función de autocorrelación simple de la parte regular.

Cuadro 2.2: Comportamiento de las funciones de autocorrelación simple y parcial en procesos ARIMA estacionales multiplicativos.

Esta clase de modelos, que fueron introducidos por Box y Jenkins (1976), representan bien muchas series estacionales que se encuentran en la práctica. Los procesos ARIMA estacionales multiplicativos generalizan todos los procesos estudiados en este capítulo. Dichos procesos consiguen capturar la falta de estacionaridad provocada por la presencia de tendencia y componente estacional, además de modelar la dependencia regular y estacional.

2.4. Estimación y selección de modelos ARIMA

Supongamos que disponemos de una serie estacionaria con T observaciones $\{X_t\}_{t\in T}$. Por simplicidad, dejaremos que la serie $\{X_t\}_{t\in T}$ denote un proceso estacionario observado a pesar de que puede ser una diferencia apropiada de la serie original. Una vez especificado un modelo ARMA, queremos estimar sus parámetros por máxima verosimilitud. Para ello debemos escribir la función de densidad conjunta y maximizarla respecto a los parámetros, considerando a los datos como fijos.

Por tanto, los valores $\hat{c}, \hat{\phi}_1, \ldots, \hat{\phi}_p, \hat{\theta}_1, \ldots, \hat{\theta}_q$ y $\hat{\sigma}_a^2$ que maximizan la función de verosimilitud anterior son las estimaciones de los parámetros $c, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q$ y σ_a^2 por el método de máxima verosimilitud.

Supongamos ahora que disponemos de varios modelos ARIMA estimados para una serie de tiempo y que queremos seleccionar el que mejor explica la serie observada. Seleccionar el modelo por su ajuste a una muestra dada no resulta adecuado, ya que el modelo con más parámetros siempre conducirá a una mayor verosimilitud y a una menor suma de cuadrados de los errores dentro de la muestra. Por tanto debemos recurrir a otros criterios.

Uno de los criterios más utilizados es el Criterio de Información de Akaike (AIC), que consiste en seleccionar el modelo que minimice

$$AIC = T \ln \hat{\sigma}_a^2 + 2k,$$

donde T es el tamaño muestral utilizado para estimar el modelo, $\hat{\sigma}_a^2$ es el estimador máximo verosímil de la varianza de las innovaciones y k el número de parámetros estimados. La suma del término $2k$ funciona como una "función de penalización" para ayudar a garantizar la selección de modelos lo más sencillo posible y evitar elegir modelos con demasiados parámetros.

El problema con el AIC es que tiende a sobreestimar el número de parámetros en el modelo. Una alternativa que corrige dicha sobreestimación es el criterio AIC corregido, AICc, que consiste en

minimizar

$$AICc = T \ln \hat{\sigma}_a^2 + T\frac{(1 + k/T)}{1 - (k + 2)/T}.$$

Por último, otro enfoque es seleccionar un modelo que minimice el Criterio de Información Bayesiano de Schwarz (BIC) definido como

$$BIC = T \ln \hat{\sigma}_a^2 + k \ln T.$$

Si comparamos esta expresión con la del AIC, el BIC penaliza por introducir nuevos parámetros más que el AIC, con lo que tiende a elegir modelos más parsimoniosos. La diferencia entre ellos puede ser grande si T es grande. Un modelo que minimice alguna de las funciones anteriores consigue un buen ajuste sin demasiados parámetros.

2.5. Diagnosis del modelo

Una vez que se ha ajustado un modelo a los datos, debemos realizar una serie de verificaciones de diagnóstico. Si el modelo se ajusta bien, los residuos deberían comportarse esencialmente como ruido blanco. En otras palabras, los residuos tienen media cero, son incorrelados y su varianza es constante. Además, en el modelado, frecuentemente suponemos que los errores se distribuyen normalmente, por lo tanto, esperamos que los residuos también se distribuyan más o menos normalmente.

La diagnosis está relacionada con la selección de modelos que hemos visto. Podría suceder que el mejor modelo seleccionado lleve a residuos que no verifican las hipótesis anteriores y entonces tendremos que buscar un nuevo modelo. También es posible que dispongamos de varios modelos cuyos residuos verifican las hipótesis anteriores, y entonces podemos seleccionar el mejor entre ellos con un criterio de selección. Por tanto, la diagnosis del modelo es una etapa que complementa a la selección del mejor modelo.

Para verificar si los residuos estimados están incorrelados, normalmente se calculan la *fas* y la *fap*, dibujando dos líneas paralelas a distancia $1.96/\sqrt{T}$ del origen, y se comprueba si todos los coeficientes $\hat{\rho}_k$ están dentro de estos límites de confianza. Como estos intervalos son aproximadamente del 95 %, en promedio uno de cada veinte coeficientes de correlación estimados saldrá fuera, por lo que la aparición de un valor significativo en un retardo elevado es esperable. Sin embargo, un valor próximo a los límites de confianza $\pm 1.96/\sqrt{T}$ en los retardos iniciales debe considerarse un indicio claro de que el modelo es inadecuado. Otra forma de verificar que las autocorrelaciones estimadas $\hat{\rho}_k$, $k = 1, 2, \ldots, K$, en su conjunto, no indican una insuficiencia del modelo, es calculando el estadístico de Ljung-Box-Pierce modificado

$$Q = T(T + 2)\sum_{i=1}^{K} \frac{\hat{\rho}_i^2}{T - i}.$$

Este estadístico se distribuye aproximadamente como una χ^2 con $(K - p - q)$ grados de libertad. Por tanto, concluiremos que el modelo no es apropiado si el valor obtenido de Q es mayor que el percentil 0.95 de la distribución χ^2 con $(K - p - q)$ grados de libertad.

Para contrastar la hipótesis de que los residuos tienen media cero, suponiendo T residuos y $p + q$ parámetros, calcularemos su media y varianza,

$$\bar{a} = \frac{1}{T}\sum_{t=1}^{T} \hat{a}_t \qquad \text{y} \qquad \hat{\sigma}_a^2 = \frac{1}{T}\sum_{t=1}^{T} (\hat{a}_t - \bar{a})^2 ,$$

y concluiremos que $\mathbb{E}(\hat{a}_t) \neq 0$, si

$$\frac{\bar{a}}{\hat{\sigma}_a/\sqrt{T}}$$

es significativamente grande con relación a la distribución $N(0, 1)$. Este contraste debe aplicarse después de comprobar que los residuos están incorrelados, para asegurar que $\hat{\sigma}_a^2$ es un estimador razonable de la varianza.

Por último, la hipótesis de normalidad de los residuos se comprueba con cualquiera de los contrastes habituales, como el test de Shapiro-Wilks o el de Jarque-Bera. También conviene estudiar el gráfico de los residuos estimados $\hat{a}_t$ a lo largo del tiempo, ya que nos permitirá detectar posibles valores atípicos.

2.6. Predicción con modelos ARIMA

Supongamos que disponemos de una realización de tamaño T, $X_T = (x_1, \ldots, x_T)$, de un proceso ARIMA(p,d,q). Siendo conocidos los parámetros podemos obtener todas las innovaciones a_t fijando unos valores iniciales. Por ejemplo, en el caso de un proceso ARMA(1,1), las innovaciones $a_2, \ldots, a_T$ se calculan de forma recursiva mediante la ecuación

$$a_t = x_t - c - \phi x_{t-1} + \theta a_{t-1}, \qquad t = 2, \ldots, T.$$

En el caso $t = 1$,

$$a_1 = x_1 - c - \phi x_0 + \theta a_0,$$

pero ni x_0 ni a_0 son conocidas, por lo que no podemos calcular a_1. Podemos sustituirla por su esperanza, $\mathbb{E}(a_1) = 0$, y calcular las demás innovaciones con esta condición inicial. En consecuencia, en adelante supondremos que tanto las observaciones como las innovaciones son conocidas o estimadas hasta el instante T.

La predicción que minimiza el error cuadrático medio de x_{T+k}, a la que nos referiremos como predicción óptima de x_{T+k}, es la esperanza de la variable condicionada a los valores observados (Peña et al., 2001). Definimos

$$\hat{x}_T(k) = \begin{cases} \mathbb{E}(x_{T+k}|X_T), & \text{si } k = 1, 2 \ldots \\ x_{T+k}, & \text{si } k = -1, -2, \ldots \end{cases} \quad \text{y} \quad \hat{a}_T(k) = \begin{cases} \mathbb{E}(a_{T+k}|X_T), & \text{si } k = 1, 2 \ldots \\ a_{T+k}, & \text{si } k = -1, -2, \ldots \end{cases}$$

donde T representa el origen de la predicción, que suponemos fijo, y k el horizonte de la misma, que irá cambiando para generar predicciones de distintas variables futuras desde el origen T. Denotando por $\varphi_h(B) = \phi_p(B)\nabla^d$ al operador de orden $h = p + d$, tenemos que

$$x_{T+k} = c + \varphi_1 x_{T+k-1} + \cdots + \varphi_h x_{T+k-h} + a_{T+k} - \theta_1 a_{T+k-1} - \cdots - \theta_q a_{T+k-q}. \tag{2.5}$$

Tomando esperanzas condicionadas a X_T en todos los términos de esta expresión, obtenemos

$$\hat{x}_T(k) = c + \hat{\varphi}_1 \hat{x}_T(k-1) + \cdots + \hat{\varphi}_h \hat{x}_T(k-h) - \hat{\theta}_1 \hat{a}_T(k-1) - \cdots - \hat{\theta}_q \hat{a}_T(k-q). \tag{2.6}$$

Esta expresión tiene dos partes. La primera, que depende de los coeficientes autorregresivos, determinará la forma de la predicción a largo plazo. La segunda, que depende de los coeficientes de la media móvil, desaparecerá para $k > q$. Cuando $k > 0$, $\hat{x}_T(k)$ es la esperanza condicionada de la variable x_{T+k} que aun no ha sido observada. En cambio, cuando $k \leq 0$, $\hat{x}_T(k)$ es la esperanza condicionada de la variable $x_{T-|k|}$, que ya se ha observado y es conocida, por lo que esta esperanza coincidirá con la observación y $\hat{x}_T(-|k|) = x_{T-|k|}$. En cuando a las innovaciones, cuando $j > 0$, las $\hat{a}_T(k)$ serán cero y, cuando $k \leq 0$, $\hat{a}_T(-|k|) = a_{T-|k|}$. Así, podemos calcular recursivamente las predicciones empezando con $k = 1$ y continuando con $k = 2, 3, \ldots$, hasta el horizonte deseado.

En el caso $k = 1$, restando (2.5) de (2.6) se obtiene que

$$a_{T+1} = x_{T+1} - \hat{x}_T(1),$$

y las innovaciones se pueden interpretar como los errores de predicción a un período por delante cuando los parámetros del modelo son conocidos.

Ahora bien, las predicciones tienen poca utilidad sin una medida de su precisión, por lo que calcularemos intervalos de confianza. Para ello, utilizaremos la distribución muestral de los errores de predicción

$$\hat{e}_T(k) = x_{T+k} - \hat{x}_T(k).$$

Sea $x_t = \psi(B)a_t$ la representación de Wold del proceso. Entonces

$$x_{T+k} = \sum_{i=0}^{\infty} \psi_i a_{T+k-i} \qquad (\psi_0 = 1), \tag{2.7}$$

y la predicción óptima será,

$$\hat{x}_T(k) = \sum_{i=0}^{\infty} \psi_{k+i} a_{T-i}. \tag{2.8}$$

Restando (2.8) de (2.7), se obtiene el error de predicción,

$$\hat{e}_T(k) = x_{T+k} - \hat{x}_T(k) = a_{T+k} + \psi_1 a_{T+k-1} + \cdots + \psi_{k-1} a_{T+1},$$

cuya varianza será

$$Var(\hat{e}_T(k)) = \sigma_a^2(1 + \psi_1^2 + \cdots + \psi_{k-1}^2).$$

Si suponemos que la distribución de las innovaciones es normal, los resultados anteriores nos permiten calcular intervalos de confianza para la predicción. Entonces x_{T+k} será una variable normal de media $\hat{x}_T(k)$ y varianza $\sigma_a^2(1+\psi_1^2+\cdots+\psi_{k-1}^2)$, siendo ψ_i los coeficientes de la representación de Wold. Por tanto, podemos escribir

$$x_{T+k} \in \left(\hat{x}_T(k) \pm z_{1-\alpha}\sqrt{\sigma_a^2(1 + \psi_1^2 + \cdots + \psi_{k-1}^2)}\right)$$

donde $z_{1-\alpha}$ son los percentiles de la distribución normal estándar.

2.7. Aplicación de los modelos ARIMA a los datos de Estrella Galicia

Vamos a aplicar la teoría desarrollada en las secciones anteriores a una serie de tiempo cuyas observaciones son las ventas mensuales en litros de la cerveza A en una ruta de distribución concreta. Estudiaremos las características de la serie, ajustaremos un modelo ARIMA adecuado a la misma y realizaremos predicciones para el siguiente año.

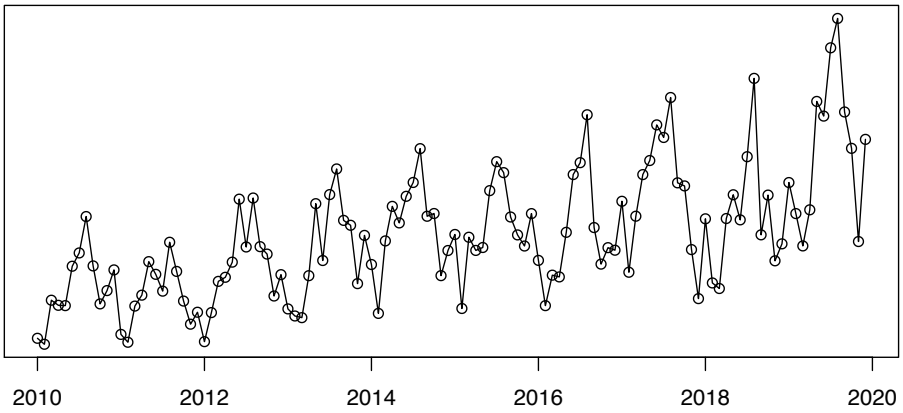


Figura 2.4: Gráfico secuencial de la serie de ventas mensuales.

En la Figura 2.4 podemos ver el gráfico secuencial de la serie. Parece hay una estacionalidad clara de periodo $s = 12$, sin embargo la tendencia no parece tan evidente. Además también podemos detectar heterocedasticidad en la serie, ya que la variabilidad aumenta con el nivel. Por tanto, en primer lugar transformaremos la serie para conseguir que la variabilidad sea constante.

Ya que los valores de la serie son positivos, realizaremos una transformación Box-Cox a los datos. Posiblemente la transformación logarítmica (transformacion de Box-Cox con $\lambda = 0$) sea apropiada para estabilizar la varianza. Sin embargo, calculando el óptimo para dicha transformación obtenemos $\hat{\lambda} = -0.38$. En la Figura 2.5 podemos ver que ambas transformaciones son muy similares, por lo que utilizaremos la transformacion logarítmica ($\lambda = 0$), ya que funciona bien y es más sencilla de interpretar.

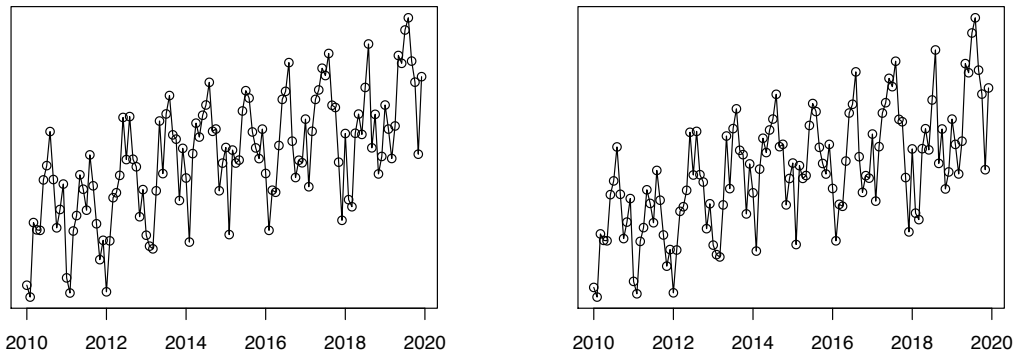


Figura 2.5: Gráfico secuencial de la serie transformada por Box-Cox (izquierda) y por logaritmos (derecha).

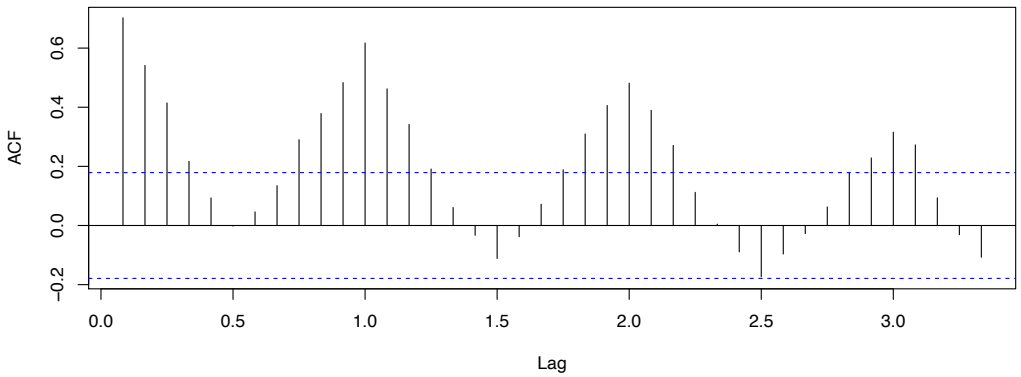


Figura 2.6: Función de autocorrelaciones simples.

Ahora propondremos y ajustaremos un modelo ARIMA para la serie transformada. Tanto en su

gráfico secuencial como en el gráfico de la *fas*, que se muestra en la Figura 2.6, podemos ver que la serie no es estacionaria debido a la presencia de una componente estacional de periodo $s = 12$, ya que podemos ver que la *fas* en los retardos estacionales tarda muchos retardos en anularse. Por tanto es necesaria una diferenciación estacional.

Tras aplicar una diferencia estacional de periodo $s = 12$, en la Figura 2.7 observamos que la componente estacional ha desaparecido. Además, observamos que no hay ninguna tendencia clara, por tanto podemos concluir que la serie diferenciada es estacionaria. Por tanto, nuestro modelo será un ARIMA$(p,0,q)\times(P,1,Q)_{12}$. Para idenfiticar los órdenes p, q, P y Q observamos en la Figura 2.8 la *fas* y la *fap* de la serie diferenciada.

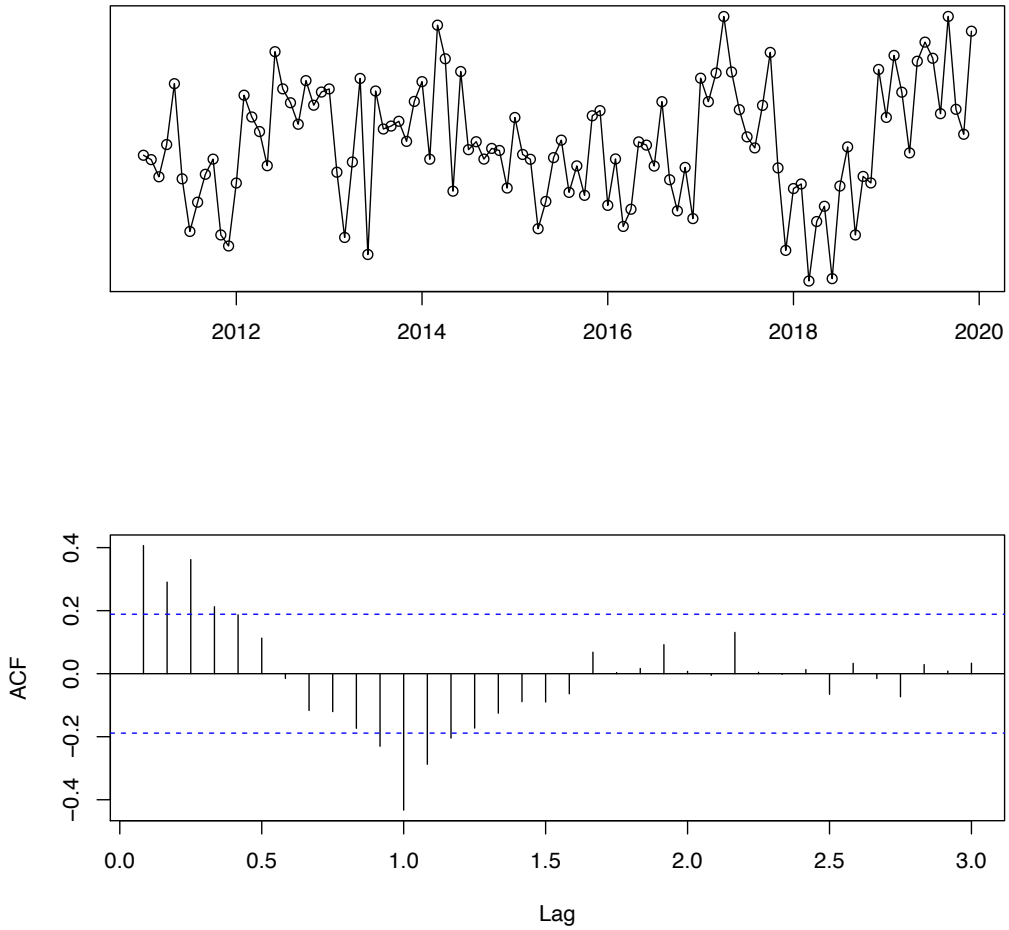


Figura 2.7: Gráfico secuencial y *fas* de la serie diferenciada estacionalmente.

Para la identificación de la estructura regular nos fijaremos en los retardos iniciales. Si analizamos la gráfica de la *fas*, el último retardo en salirse de las bandas es el cuarto, por tanto sugeriríamos un modelo MA(4) para la parte regular. Analizando la gráfica de la *fap*, el útlimo retardo en salirse de las bandas es el tercero, por lo que sugeriríamos un modelo AR(3) para la parte regular. Ya que pretendemos estimar el modelo más sencillo posible, prondremos un modelo AR(3) para la parte regular.

Para identificar la estructura estacional nos fijaremos en los retardos estacionales, $s, 2s, 3s, \ldots$ siendo $s = 12$. En la *fas* solo el primer retardo estacional se sale de la banda, mientras que en la *fap*, el último retardo estacional en salirse de la banda es el segundo. Por tanto, propondremos un modelo MA(1) para la parte estacional ya que parece más claro.

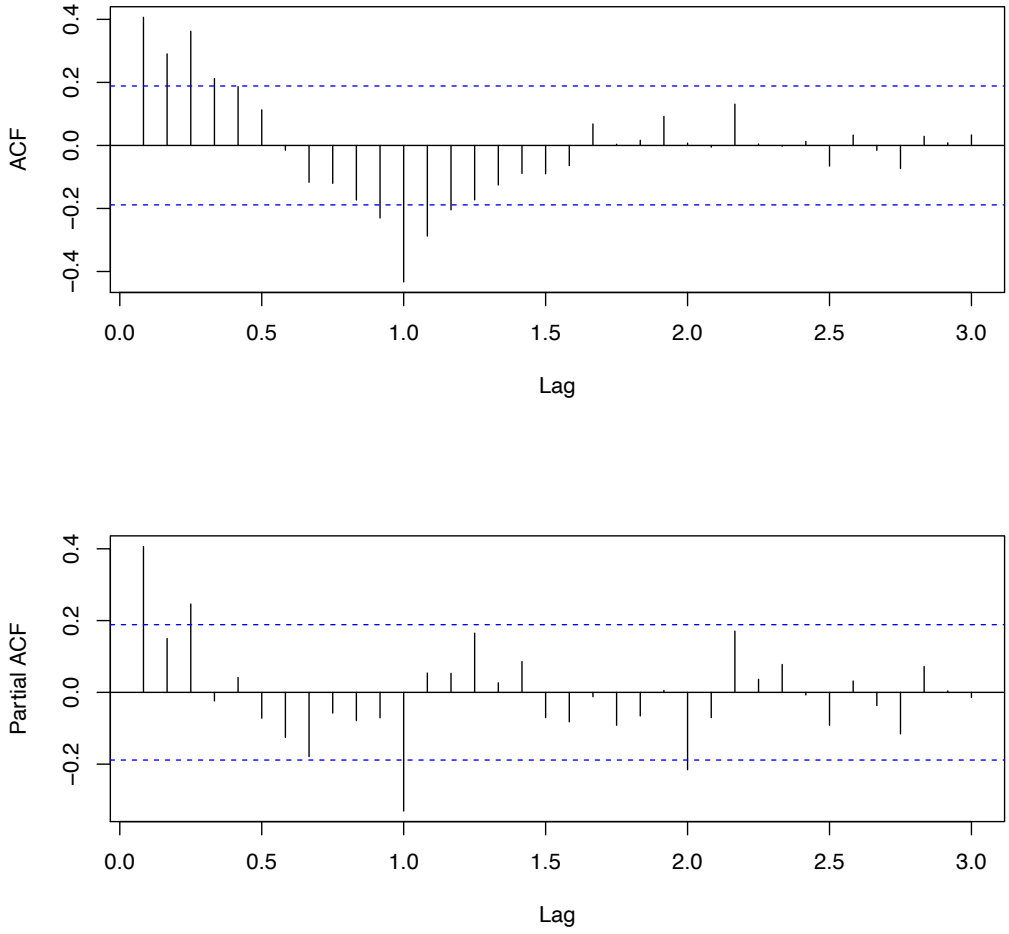


Figura 2.8: Funciones de autocorrelaciones simples y parciales de la serie diferenciada.

Así, el modelo tentativo es un ARIMA(3,0,0)×(0,1,1)$_{12}$ para el logaritmo de la serie de ventas, es decir,

$$\phi_3(B)\nabla_{12} y_t = c + \Theta_1(B^{12})a_t.$$

Ajustando el modelo obtenemos las siguientes estimaciones por máxima verosimilitud de los parámetros

$$\phi_1 = 0.2062, \quad \phi_2 = 0.0145, \quad \phi_3 = 0.3362, \quad c = 0.0554 \quad \text{y} \quad \Theta_1 = -0.7162.$$

Todos los parámetros resultan ser significativos excepto el ϕ_2, por lo que fijaremos ϕ_2 a cero. Por tanto,

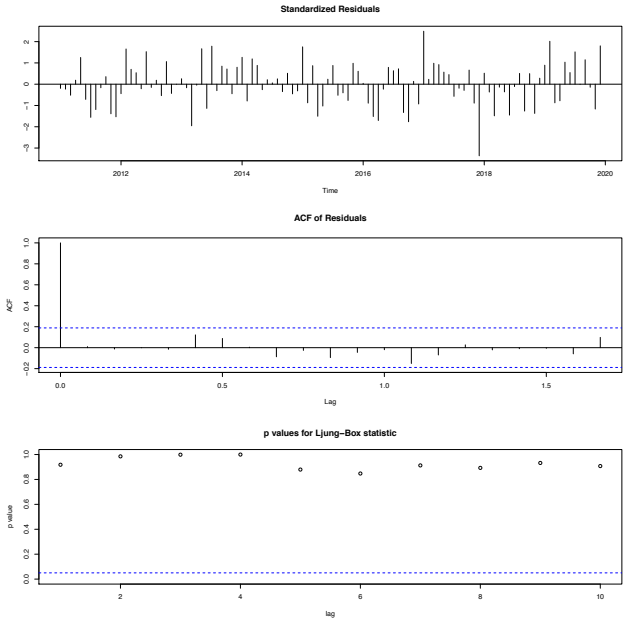
$$Y_t = c + \phi_1 Y_{t-1} + \phi_3 Y_{t-3} + Y_{t-12} - \phi_1 Y_{t-13} - \phi_3 Y_{t-15} + a_t + \Theta_1 a_{t-12}.$$

Sustituyendo los parámetros por las estimaciones obtenidas en el ajuste, tenemos que

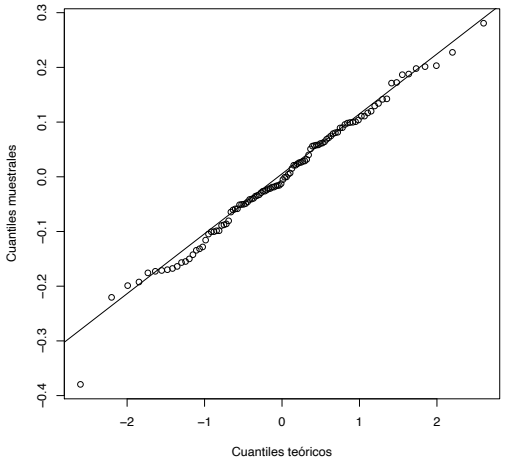
$$Y_t = 0.0554 + 0.2062\, Y_{t-1} + 0.3362\, Y_{t-3} + Y_{t-12} - 0.2062\, Y_{t-13} - 0.3362\, Y_{t-15} + a_t - 0.7162\, a_{t-12}.$$

Llegados a este punto es necesario realizar un diagnóstico del modelo, verificando si sus residuos son o no ruido blanco con distribución gaussiana. Recordemos que estos contrastes deberían realizarse sobre las innovaciones a_t, pero ya que estas no son observables, los aplicaremos sobre los residuos.

Como podemos ver en la Figura 2.9a, el contraste de Ljung-Box mantiene la incorrelación. Una vez comprobado que los residuos están incorrelados, el contraste de media cero obtiene un p-valor igual a 0.9107 por lo que no se rechaza la hipótesis nula. Finalmente, en la Figura 2.9b observamos que el gráfico Q-Q normal parece ser bastante lineal y, además, los contrastes de normalidad de Jarque-Bera y Shapiro-Wilks obtienen p-valores iguales a 0.6902 y 0.6738, respectivamente, por lo que la normalidad tampoco se rechaza. Por tanto, podemos aceptar que las innovaciones son ruido blanco gaussiano y, en consecuencia, el modelo que hemos estimado es apropiado como generador de la serie analizada.



(a) Contrastes de independencia.



(b) Gráfico Q-Q normal de los residuos.

Figura 2.9: Diagnosis de los residuos.

A continuación, utilizaremos el ARIMA(3,0,0)×(0,1,1)$_{12}$ que hemos seleccionado, estimado y chequeado para realizar predicciones a horizontes de predicción $k = 1, \ldots, 12$ de la serie de ventas. Estas se muestran en la Figura 2.10 (azul), junto con la serie histórica (negro). Además, como tenemos gaussianidad, también representamos los intervalos de predicción con nivel de confianza del 80 % y 95 %.

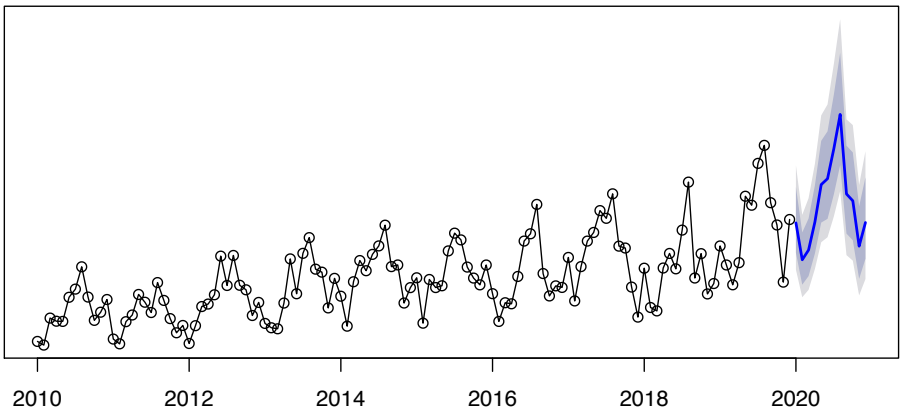


Figura 2.10: Grafico secuencial de la serie de ventas mensuales con las predicciones para el siguiente año.

Capítulo 3

Regresión con series de tiempo

El principal objetivo de la construcción de modelos de series de tiempo es la predicción de valores futuros. Frecuentemente, las series de tiempo a estudiar pueden estar relacionadas o dirigidas por otras series de tiempo. Por ejemplo, Fisher (1925) estudió la dependencia de la producción de trigo y la lluvia en la estación agraria de Rothamsted, en Inglaterra. En estos casos, mediante la incorporación de covariables relevantes en el modelo puede alcanzarse una mejor comprensión del proceso subyacente y predicciones más precisas.

Los modelos de series de tiempo que hemos visto en el el Capítulo 2 permiten incluir información de observaciones pasadas de una serie, pero no otra información que también pueda ser relevante como variables externas que puedan explicar algunas de las variaciones históricas y que puedan conducir a pronósticos más precisos. En este capítulo consideramos cómo extender los modelos ARIMA para permitir que se incluya otra información en los modelos.

Los modelos de regresión dinámica sirven para modelizar la dependencia existente entre dos o más series temporales. Para explicar el comportamiento de y_t a partir de x_t, podemos utilizar el modelo de regresión lineal entre series de tiempo

$$\Phi_P(B^s)\phi_p(B)\nabla_s^D\nabla^d y_t = c + \delta(B)x_t + \Theta_Q(B^s)\theta_q(B)a_t, \tag{3.1}$$

donde $\delta(B) = \delta_0 + \delta_1 B + \cdots + \delta_m(B^m)$ y las variables a_t son las innovaciones del modelo que también siguen un proceso de ruido blanco. En este modelo, y_t depende de su pasado, depende de x_t y del pasado de x_t.

Estos modelos son conocidos como modelos dinámicos y describen cómo se transmiten los efectos desde una variable x_t a otra y_t cuando no existe realimentación o causalidad bidireccional y son un instrumento muy utilizado para evaluar respuestas dinámicas.

Si tenemos solo dos series de tiempo, el modelo más genérico es el modelo de regresión dinámica (3.1). Sin embargo, si tratamos con varias variables regresoras, es complicado estimar el polinomio $\delta(B)$ y entonces la única posibilidad es tratar con el modelo instantáneo concurrente, que es aquel en el que y_t depende de los valores del mismo índice de las otras series, por tanto nunca hay retardos.

Debido a que el objetivo del trabajo es automatizar un modelo de regresión para series de ventas con varias variables explicativas, y que para ello utilizaremos función `auto.arima` disponible en el paquete `forecast` (Hyndman y Khandakar, 2008), el modelo con el que realmente trataremos será el instantáneo concurrente, es decir,

$$y_t = \beta_0 + \beta_1 x_{1,t} + \cdots + \beta_k x_{k,t} + \varepsilon_t, \tag{3.2}$$
$$\Phi_P(B^s)\phi_p(B)\nabla_s^D\nabla^d \varepsilon_t = \Theta_Q(B^s)\theta_q(B)a_t,$$

donde a_t es ruido blanco.

El modelo (3.2) supone que la relación entre y_t y los predictores es instantánea, sin embargo, a veces el impacto de un predictor que se incluye en un modelo de regresión no lo es. Por ejemplo, una

campaña publicitaria puede afectar las ventas durante un tiempo más allá del final de la campaña, y las ventas en un mes dependerán del gasto publicitario en cada uno de los últimos meses.

En estas situaciones, debemos permitir los efectos retardados del predictor. Supongamos que tenemos un solo predictor en nuestro modelo. Hyndman y Athanasopoulos (2018) proponen un modelo que permitiría efectos rezagados y que puede escribirse como

$$y_t = \beta_0 + \gamma_0 x_t + \gamma_1 x_{t-1} + \cdots + \gamma_k x_{t-k} + \varepsilon_t,$$

donde ε_t es un proceso ARIMA. El valor de k puede seleccionarse utilizando el AICc, junto con los valores de p y q para el error ARIMA.

3.1. Relaciones entre series estacionarias

Comenzamos con varias definiciones que son de utilidad para asesorarnos acerca de la existencia, o no, de relación lineal entre dos series de tiempo.

Definición 3.1.1. Dados dos procesos estacionarios x_t e y_t, se define la función de covarianzas cruzadas, $\gamma_{x,y}$, como

$$\gamma_{xy}(t, t+k) = Cov(x_t, y_{t+k}).$$

La función anterior mide el grado de dependencia lineal existente entre las dos variables x_t e y_{t+k}. El valor de $\gamma_{xy}(t, t+k)$ puede ser positivo o negativo y además $\gamma_{xy}(t, t+k) = \gamma_{yx}(t+k, t)$.

Definición 3.1.2. Se dice que dos procesos estacionarios x_t e y_t son conjuntamente estacionarios si cada uno de ellos es estacionario y las covarianzas cruzadas solo dependen del retardo entre las variables, k, y no del instante inicial considerado, es decir,

$$\gamma_{xy}(t, t+k) = \gamma_{xy}(k), \qquad k \in \mathbb{Z}.$$

Para simplificar, diremos que dos variables son estacionarias para indicar que son conjuntamente estacionarias. La función de covarianzas cruzadas para procesos estacionarios no varía si intercambiamos las variables y el signo del retardo simultáneamente, es decir, $\gamma_{xy}(k) = \gamma_{yx}(-k)$. Por tanto, la función $\gamma_{xy}(k)$ resume toda la dependencia lineal entre ambas variables y no es necesario calcular la γ_{yx}. Además, si $\gamma_{xy}(0) \neq 0$, diremos que existe relación instantánea entre las dos variables.

De la misma forma que las autocovarianzas permiten identificar el orden del modelo ARMA, podría pensarse en utilizar la función de covarianzas cruzadas para identificar si la relación es en una dirección o bidireccional y también obtener el número de retardos distintos de cero en la relación entre dos variables. Sin embargo no es así, ya que esta función tiene muchas limitaciones que detalla Peña (2005).

La estandarización de la función de covarianzas cruzadas es la función de correlación cruzada.

Definición 3.1.3. Se define la función de correlación cruzada, $\rho_{xy}(k)$, de dos procesos estocásticos conjuntamente estacionarios, x_t e y_t, como

$$\rho_{xy} = \frac{\gamma_{xy}(k)}{\sigma_x \sigma_y} \qquad k \in \mathbb{Z}.$$

En el primer capítulo vimos que las autocovarianzas teóricas de un proceso x_t se estimaban mediante las muestrales. Del mismo modo, dadas dos series temporales x_t e y_t, estimaremos sus covarianzas cruzadas mediante las muestrales

$$\hat{\gamma}_{xy}(k) = \frac{1}{T} \sum_{t=1}^{T-k} (x_t - \bar{x})(y_{t+k} - \bar{y}),$$

y las correlaciones cruzadas muestrales por

$$\hat{\rho}_{xy}(k) = \frac{\hat{\gamma}_{xy}(k)}{s_x s_y},$$

donde $s_x = \hat{\gamma}_x^{1/2}(0)$ y $s_y = \hat{\gamma}_y^{1/2}(0)$.

En el caso particular de que uno de los procesos sea ruido blanco, si ambos procesos están incorrelados, podemos aproximar las desviaciones típicas de las estimaciones de los coeficientes de correlación cruzada de manera idéntica a las de las autocorrelaciones mediante $1/\sqrt{T}$. Por tanto, si x_t es ruido blanco podemos construir bandas $1.96/\sqrt{T}$ sobre las autocorrelaciones y si $\hat{\rho}_{xy}(k) = 0$ para $k < 0$ concluimos que no hay relación desde y_t hacia x_{t+k}. La parte de esta función para $k > 0$ permitirá estudiar si existe o no relación desde x_{t-k} hacia y_t. Si la variable x_t no es ruido blanco, podría transformarse a través del método de preblanqueado de series en un proceso de ruido blanco (ver Cryer y Chan, 2010). El problema de este método es que no puede extenderse fácilmente cuando existe más de un regresor.

3.2. Estimación

Para estimar los parámetros del modelo necesitamos minimizar la suma de los valores a_t al cuadrado. Sin embargo, si minimizamos la suma de los valores de ε_t al cuadrado surgen varios problemas.

1. Los coeficientes estimados $\hat{\beta}_0, \ldots, \hat{\beta}_k$ dejan de ser los mejores estimadores, ya que se ha omitido información para su cálculo.

2. Cualquier test estadístico asociado al modelo, como por ejemplo el test T en los coeficientes, será incorrecto.

3. Los valores de AICc de los modelos ajustados ya no son una buena referencia sobre cuál es el mejor modelo para realizar predicciones.

4. En la mayoría de los casos, los p-valores asociados a los coeficientes serán muy pequeños. Entonces, algunas variables predictoras parecerán importantes cuando realmente no lo son. Esto se denomina regresión espuria.

Minimizar la suma de los valores a_t al cuadrado evita estos problemas. Alternativamente, puede utilizarse la estimación de máxima verosimilitud, dando lugar a estimaciones similares de los coeficientes.

Una consideración importante a la hora de estimar una regresión con errores ARMA es que primeramente todas las variables en el modelo deben ser estacionarias. Por lo tanto, primero tenemos que verificar que y_t y todos los predictores $x_{1,t}, \ldots, x_{k,t}$ parezcan estacionarios. Si estimamos el modelo cuando alguno de estos no es estacionario, los coeficientes estimados no serán estimaciones consistentes (y, por lo tanto, pueden no ser significativos). Una excepción a esto es el caso donde las variables no estacionarias están cointegradas, es decir, si existe una combinación lineal del proceso y_t no estacionario y los predictores que son estacionarios, entonces los coeficientes estimados serán consistentes[1].

Por lo tanto, primero diferenciaremos las variables no estacionarias en el modelo. A menudo es deseable mantener la forma de la relación entre y_t y los predictores, y en consecuencia comúnmente se diferencian todas las variables si alguna de ellas necesita diferenciarse. El modelo resultante se denomina modelo en diferencias.

Si todas las variables del modelo son estacionarias, entonces solo es necesario considerar errores ARMA para los residuos. Se puede ver fácilmente que un modelo de regresión con errores ARIMA es

[1] Para el estudio de los modelos cointegrados consultar Peña (2005).

equivalente a un modelo de regresión en diferencias con los errores ARMA. Por ejemplo, si el modelo de regresión anterior con errores ARIMA(1,1,0) se diferencia, obtenemos el modelo

$$y'_t = \beta_1 x'_{1,t} + \cdots + \beta_k x'_{k,t} + \varepsilon'_t,$$

$$(1 - \phi_1 B)\varepsilon'_t = a_t,$$

donde $y'_t = \nabla y_t$, $x'_{t,i} = \nabla x_{t,i}$ y $\varepsilon'_t = \nabla \varepsilon_t$, que es un modelo de regresión en diferencias con los errores ARMA.

Ejemplo 3.2.1 (Estrella Galicia)**.** Consideremos la serie de ventas mensuales en litros de la cerveza A en una ruta de distribución concreta y la temperatura media mensual en la correspondiente provincia (datos disponibles en el IGE). La Figura 3.1 muestra el logaritmo de las ventas mensuales en litros y la temperatura media desde enero de 2010 hasta diciembre de 2019.

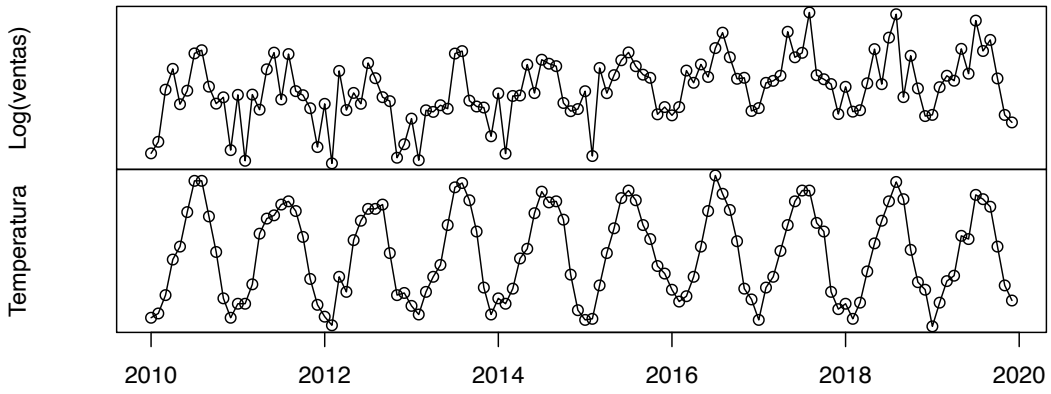


Figura 3.1: Gráfico secuencial del logaritmo de la serie de ventas en litros y la temperatura media.

Vamos a considerar incluir la temperatura media hasta cuatro meses, es decir, el modelo podrá incluir la temperatura en el mes actual y los tres meses anteriores. Por tanto, estimaremos los siguientes modelos

$$y_t = \beta_0 + \gamma_0 x_t + \varepsilon_t, \tag{3.3}$$
$$y_t = \beta_0 + \gamma_0 x_t + \gamma_1 x_{t-1} + \varepsilon_t, \tag{3.4}$$
$$y_t = \beta_0 + \gamma_0 x_t + \gamma_1 x_{t-1} + \gamma_2 x_{t-2} + \varepsilon_t, \tag{3.5}$$
$$y_t = \beta_0 + \gamma_0 x_t + \gamma_1 x_{t-1} + \gamma_2 x_{t-2} + \gamma_3 x_{t-3} + \varepsilon_t, \tag{3.6}$$

donde y_t son las ventas en litros en el mes t y x_t la temperatura media en el mes t.

Una vez estimados los modelos anteriores a través de la función `auto.arima`, elegiremos el retardo óptimo para la temperatura basándonos en el criterio AICc. Los valores del AICc correspondientes a los modelos (3.3), (3.4), (3.5) y (3.6) son -102.10, -114.94, -113.08, -111.17, respectivamente. Por

tanto, el mejor modelo (con el menor valor AICc) tiene dos predictores retardados, es decir, incluye la temperatura en el mes actual y en el mes anterior. Así que el modelo ajustado es

$$y_t = 9.6092 + 0.0564x_t - 0.0156x_{t-1} + \varepsilon_t,$$
$$\varepsilon_t = 0.9747\varepsilon_{t-1} + a_t - 1.3384a_{t-1} + 0.4954a_{t-2},$$

donde a_t es ruido blanco. Además, todos los parámetros resultan ser significativos.

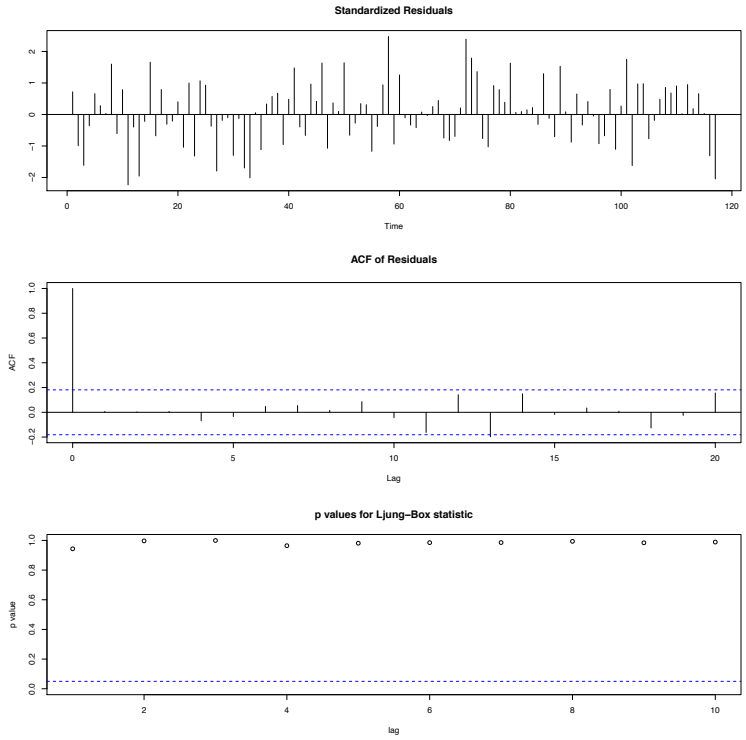


Figura 3.2: Contrastes de independencia.

Comprobemos que, efectivamente, los residuos son ruido blanco con distribución gaussiana. En la Figura 3.2 observamos que el contraste de Ljung-Box mantiene la incorrelación. En cuanto al contraste de media cero, obtenemos un p-valor igual a 0.6311 por lo que no se rechaza la hipótesis nula. Finalmente, los contrastes de normalidad de Jarque-Bera y Shapiro-Wilks obtienen p-valores iguales a 0.8273 y 0.8542, respectivamente, por lo que la normalidad tampoco se rechaza. Por tanto, podemos aceptar que los residuos, es decir, los errores ARIMA, no son significativamente distintos de ruido blanco. Por tanto, concluimos que el modelo que hemos estimado es apropiado como generador de la serie analizada.

3.3. Predicción

Para realizar predicciones utilizando un modelo de regresión con errores ARIMA, necesitamos predecir la parte de regresión y la parte ARIMA del modelo, y combinar los resultados. Al igual que con los modelos de regresión ordinarios, para obtener predicciones primero necesitamos predecir los predictores. Cuando los predictores se conocen en el futuro (por ejemplo, variables relacionadas con el calendario, como la hora, el día de la semana, etc.), esto es sencillo. Pero cuando los predictores son desconocidos, debemos modelarlos por separado o usar valores futuros supuestos para cada predictor.

Ejemplo 3.3.1 (Estrella Galicia)**.** Utilizando el modelo estimado en la sección anterior, calcularemos las previsiones para el próximo año suponiendo que las temperaturas medias futuras serán iguales a las temperaturas medias mensuales de los últimos 10 años.

Las predicciones se muestran en la Figura 3.3 (azul), junto con la serie histórica (negro). Además, como hemos comprobado que tenemos gaussianidad, también se presentan los intervalos de confianza con nivel de confianza del 80 % y 95 %.

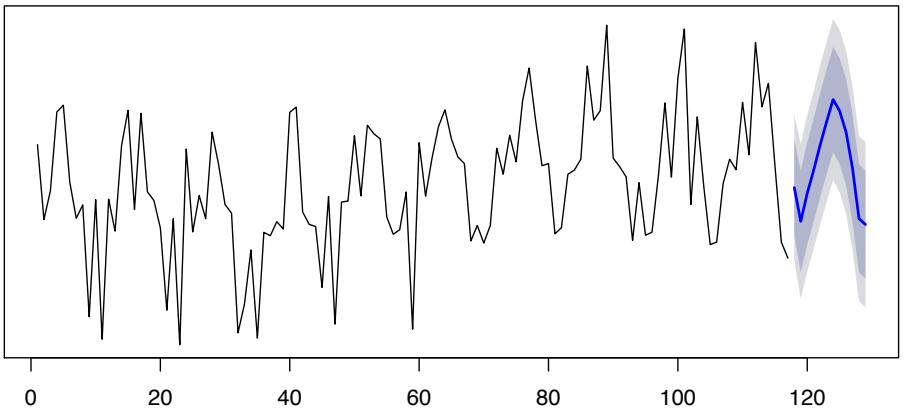


Figura 3.3: Gráfico secuencial de la serie del logaritmo de ventas mensuales con las predicciones obtenidas mediante la regresión para el siguiente año.

Es importante darse cuenta de que los intervalos de predicción de los modelos de regresión (con o sin errores ARIMA) no tienen en cuenta la incertidumbre en los pronósticos de los predictores. Por tanto, deben interpretarse que están condicionados a los valores futuros supuestos (o estimados) de las variables predictoras.

Capítulo 4

Aplicación a los datos de Estrella Galicia

Por la naturaleza de los datos, la series mensuales muestran una volatilidad alta y aunque las predicciones puntuales obtenidas mediante los modelos que hemos estudiado en el Capítulo 2 pueden ser aceptables, los correspondientes intervalos de confianza pueden ser demasiado amplios. Como solución, planteamos la posibilidad de incorporar covariables externas continuas, como la temperatura media, y discretas, como la estación, realizanzo un análisis de la serie temporal de ventas con variables regresoras. Una vez obtenida la estimación y establecido el presupuesto planificado por la empresa crearemos un vector de pesos proporcional a la estimación de las ventas para así distribuir el presupuesto mensualmente.

Las variables extraidas de la base de datos de Estrella Galicia son las siguientes:

- Ruta ID: asigna a cada ruta de cada camión una identidad numérica.
- Marca desglose: desglose de los productos dentro de la marca. Por ejemplo, la marca 1906 se desglosa en 1906, Red Vintage y Black Coupage.
- Mes: proporciona el mes y año de la venta. Contiene seis dígitos, correspondiendo los cuatro primeros al año y los dos últimos al mes.
- Provincia: indica la provincia en la que se realiza la venta.
- Ventasacummeslitros: ventas mensuales en litros de un producto determinado.

Por tanto, por cada ruta tendremos tantas series mensuales de ventas como productos haya en marca desglose.

Por la heterocedasticidad de las series de ventas, se ha decidido trabajar con la transformación logarítmica de las mismas. Recordemos que el logaritmo se aplica a número positivos, ya que contrae los números grandes (mayores que 1) y dilata los números pequeños (entre 0 y 1), produciendo así un efecto de simetrización en variables positivas, como es el caso de las ventas. Así, por ejemplo, las ventas de un producto pueden desviarse hacia valores inferiores a la media, solo hasta llegar a cero, mientras que por encima de la media no hay cota, y de hecho se podran observar ventas muy elevadas. Aplicando un logaritmo se contraen las ventas elevadas, y se separan las ventas bajas, consiguiendo así un comportamiento simétrico.

Al trabajar con el logaritmo de la serie, hemos decidido trabajar solo con aquellas series en las que las ventas son superiores a uno y que no tienen valores faltantes. Además, nos restringiremos a aquellas series que tengan el registro completo de ventas desde enero de 2010 hasta diciembre de 2019.

Después de sopesar algunas posibilidades, hemos decidido que las variables explicativas que tienen más sentido estudiar son la temperatura media, los días de lluvia, el número de días laborables, un

efecto de la Semana Santa, que explicaremos más adelante, y la estación. Los datos meteorológicos se han extraido del IGE[1] y los días laborables de la propia base de datos de Hijos de Rivera.

Cuando desarrollamos los modelos de regresión dinámicos, asumimos que cada predictor tomaba valores numéricos. Pero en nuestro caso, estamos tratando con datos mensuales y queremos tener en cuenta la estación del año como predictor. Esta situación puede manejarse dentro del marco de modelos de regresión múltiple creando variables *dummy* o indicadoras $E_{1,t}$, $E_{2,t}$ y $E_{3,t}$ para invierno, primavera y verano, respectivamente, que valen cero o uno dependiendo de si el mes correspondiente está en la estación o no como podemos ver en el cuadro 4.1.

	$E_{1,t}$	$E_{2,t}$	$E_{3,t}$		$E_{1,t}$	$E_{2,t}$	$E_{3,t}$
Enero	1	0	0	Junio	0	0	1
Febrero	1	0	0	Julio	0	0	1
Marzo	0	1	0	Agosto	0	0	1
Abril	0	1	0	Septiembre	0	0	0
Mayo	0	1	0	Octubre	0	0	0
Junio	0	0	1	Noviembre	0	0	0
Julio	0	0	1	Diciembre	1	0	0

Cuadro 4.1: Variables indicadoras para las estaciones.

Tengamos en cuenta que solo se necesitan tres variables indicadoras para codificar cuatro categorías. Esto se debe a que la cuarta categoría, en este caso otoño, es capturada por el intercepto, y se especifica cuando todas las variables indicadoras se fijan a cero.

La interpretación de cada uno de los coeficientes asociados a las variables indicadoras es que es una medida del efecto de esa categoría en relación con la categoría omitida. En el caso de las estaciones, el coeficiente de $E_{1,t}$ asociado con el invierno medirá el efecto del invierno en la variable de pronóstico en comparación con el efecto del otoño.

Una variable indicadora también nos serviría para tener en cuenta el efecto que pueda tener la Semana Santa. Esta festividad difiere de la mayoría de los días festivos porque no se celebra en la misma fecha cada año, y podría ser de interés tener en cuenta su efecto en las ventas. Dado que tratamos con datos mensuales, si la Semana Santa cae en marzo, la variable indicadora tomará el valor uno en marzo, y si cae en abril, tomará el valor uno en abril. Cuando empieza en marzo y termina en abril, la variable indicadora se repartirá proporcionalmente entre los dos meses.

Por tanto, el modelo máximo que estimaremos será

$$y_t = \beta_0 + \sum_{i=1}^{4} \beta_i x_{i,t} + \sum_{j=0} \gamma_j E_{j,t} + \varepsilon_t,$$

donde las $x_{i,t}$ son las variables temperatura media, días de lluvia, número de días laborables, efecto de la Semana Santa, $E_{j,t}$ son las variables indicadoras de la estación y ε_t es un proceso ARIMA.

Una vez definidas las variables a tener en cuenta, vamos a comprobar que realmente hay motivos para sospechar de la relación entre las ventas y las distintas variables.

[1] Los datos faltantes de cada variable se han sustituido por el valor medio en los demás años de la variable en el mismo mes.

Tras distintos experimentos, hemos comprobado que existe relación positiva con la temperatura media y relación negativa con los días de lluvia. Como ejemplo, en la Figura 4.1 podemos ver los diagramas de dispersión de tres variables para una ruta de distribución concreta. La primera columna muestra las relaciones entre la variable de pronóstico (ventas) y cada uno de los predictores. Observamos que, efectivamente, los diagramas de dispersión muestran relación positiva con la temperatura media y relación negativa con los días de lluvia. La fuerza de estas relaciones se muestra mediante los coeficientes de correlación en la primera fila. Los diagramas de dispersión restantes y los coeficientes de correlación muestran las relaciones entre los predictores.

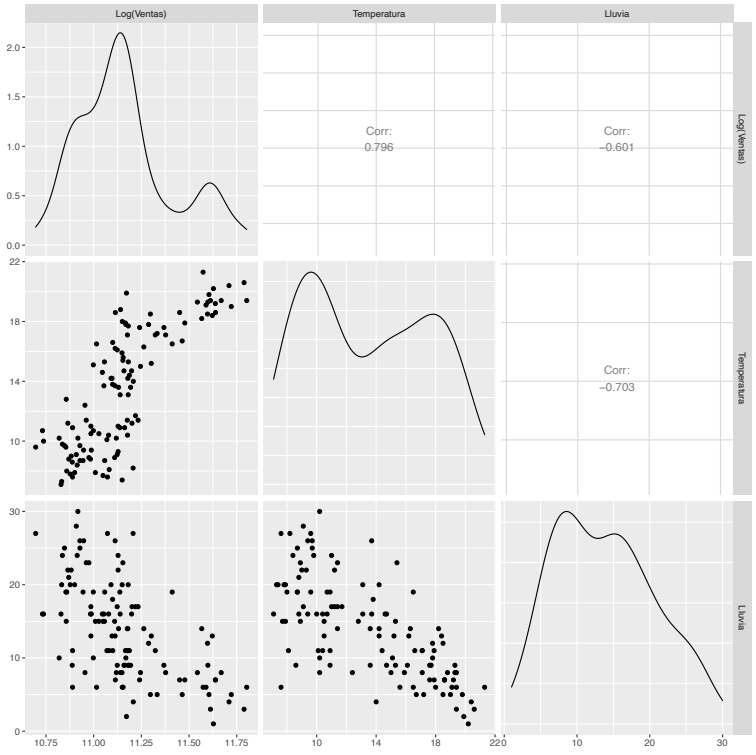


Figura 4.1: Diagramas de dispersión del logaritmo de ventas de la cerveza A y dos predictores.

Ajustando el modelo de regresión con las variables temperatura media y número de días de lluvia, observamos que en casi el 40 % de los modelos ajustados, ambas variables resultan significativas. Por tanto, incluiremos ambas variables en el modelo, que aplicaremos de forma automática, por lo que en algunos casos alguna de estas variables podría resultar no significativa.

En cuando a los días laborables, si tratásemos con datos semanales, esta variable podría ser una buena candidata a ser predictora ya que en las semanas vacacionales como la Semana Santa o con puentes se observarían picos en las series de ventas. Además, ya que algunas de estas festividades varían de semana de año en año, resultaría difícil recoger estos eventos a través de la estacionaldidad de la serie. Es evidente que este problema es propio de las series con frecuencia semanal y diaria, no de las mensuales, ya que en tal caso la estacionalidad de la serie sí capturaría su efecto. A excepción de la Semana Santa, los demás festivos se mantienen siempre dentro del mismo mes y por tanto, en la serie mensual, la estacionalidad ya los tiene en cuenta.

Concluimos entonces, que en nuestro caso, al tratar con series mensuales, carece de sentido incluir esta variable como predictora. Sin embargo, como ya comentamos, la Semana Santa oscila, según el año, entre marzo y abril, y por tanto podría existir un efecto que sería interesante tener en cuenta.

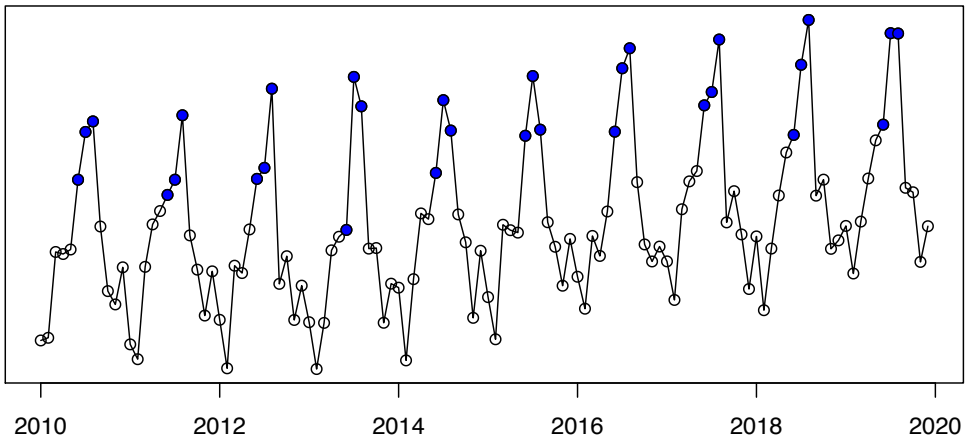


Figura 4.2: Gráfico secuencial de la serie de tiempo del logaritmo de ventas de la cerveza A en Galicia. En azul, los meses de junio, julio y agosto.

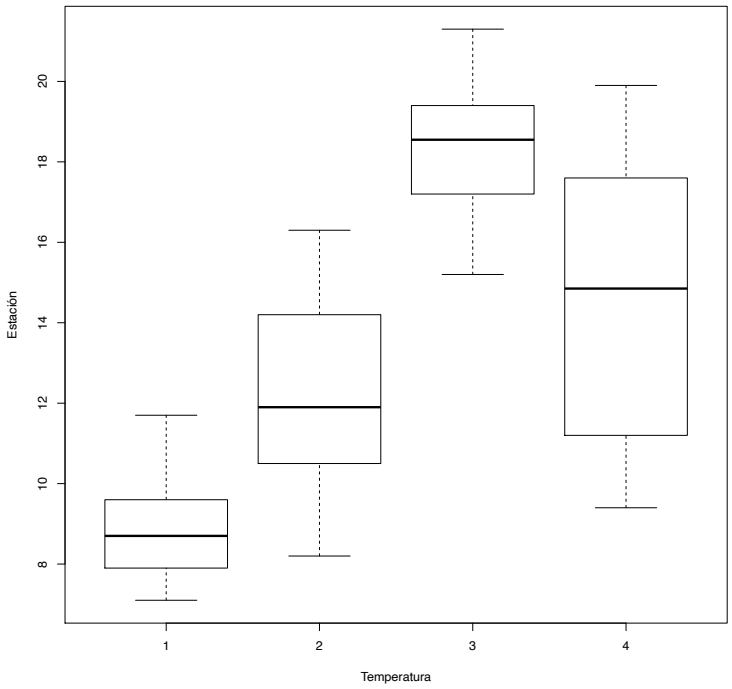


Figura 4.3: Boxplot de la temperatura media frente a la estación correspondiente.

En cuanto a la estación, como cabría esperar, las ventas durante el verano son más altas que entre las demás estaciones. En la Figura 4.2 observamos la serie del logaritmo de las ventas de Estrella Galicia con los meses de verano en color azul y podemos ver que, efectivamente en los meses del verano se producen ventas más elevadas. Esta estructura también se reproduce cuando tratamos con zonas geográficas más pequeñas, por lo que parece que la estación, como variable explicativa, podría estar justificada.

Sin embargo, la estación está fuertemente relacionada con la temperatura. En la Figura 4.3 observamos el diagrama de cajas de la temperatura con la estación. Claramente existen diferencias en las temperaturas medias dependiendo de la estación. Además, la correlación entre ambas variables es alta. En consecuencia, hemos decidido prescindir de esta variable.

A continuación resumimos las variables con las que trataremos finalmente, la nomenclatura utilizada y la unidad en las que se miden.

- $\{Y_t\}$: logaritmo de ventas mensuales en litros de un producto determinado en una ruta concreta.
- $\{T_t\}$: temperatura media mensual en grados Celsius en la provincia.
- $\{L_t\}$: número de días de lluvia mensuales en la provincia.
- $\{S_t\}$: indicadora de la Semana Santa.

Por tanto, el modelo de regresión múltiple que trataremos de estimar para cada ruta y producto es

$$y_t = \beta_0 + \beta_1 T_t + \beta_2 L_t + \beta_3 S_t + \varepsilon_t,$$

donde ε_t es un proceso ARIMA.

4.1. Implementación en R

Para realizar automáticamente la estimación y predicción del modelo de regresión múltiple en `R` se seguirá el siguiente esquema:

1. La información de entrada consistirá en un archivo `.csv`, generado con SQL Server en conexión con la base de datos de Estrella Galicia, que recogerá el registro en de todas las ventas de cerveza en litros por ruta marca desglose.
2. El archivo anterior será tratado en el script `trat_datos.R`, transformándolo adecuadamente para adaptarse al modelo que luego queremos aplicar. Incluirá las variables meteorológicas (temperatura media y días de lluvia) por provincias, completando además en el caso de datos faltantes de cada variable con el valor medio de dicha variable en el mes correspondiente. También eliminará las ventas con valores `NA` y las inferiores a 1, ya que trabajaremos con la transformación logarítmica de las series.
3. El script `mensualizacion.R` calculará mediante la función `pred` para cada marca desglose las predicciones de ventas por ruta, tomando como valores futuros de las variables regresoras las predicciones de modelos ARIMA ajustados a cada variable, creando el archivo `predicciones.csv`. Finalmente, introduciendo el presupuesto, la función `mens` devolverá la mensualización propuesta, creando el archivo `mensualizacion.csv`.[2]

Como ejemplo del procedimiento anterior, vamos a ver en detalle como actuarían las funciones anteriores en `R` para el caso de la cerveza A.

En primer lugar, el script `mensualizacion.R` se encarga de seleccionar aquellas rutas en las que la cerveza A tiene el registro completo de las ventas en litros entre enero de 2010 y diciembre de 2019, resultando para este producto que 66 rutas son adecuadas para continuar el proceso.

[2]Pueden consultarse el código de ambas funciones en el Apéndice A.

A continuación, para cada una de las rutas factibles, con la función `auto.arima` ajustamos el modelo de regresión múltiple

$$y_t = \beta_0 + \beta_1 T_t + \beta_2 L_t + \beta_3 S_t + \varepsilon_t \tag{4.1}$$

donde ε_t es un proceso ARIMA. Este modelo será el que utilizaremos para calcular la predicción para el siguiente año. Sin embargo, para ello primeramente debemos "determinar" los valores futuros de las series de temperatura, días de lluvia y el efecto de la Semana Santa.

En el caso de las dos primeras variables, lo que haremos será estimar automáticamente un modelo ARIMA apropiado para cada una de las series y predecir los valores para el próximo año. En el caso de la variable indicadora de la Semana Santa, a través de la función `easter` (disponible en el paquete `forecast`) obtendremos un vector de ceros y unos o un reparto proporcional si la Semana Santa abarca marzo y abril en el período de tiempo observado. Estos resultados son los que introduciremos en la función `forecast` para realizar la predicción.

Ya que hemos trabajado con la serie de logaritmos, se realiza una transformación exponencial a las predicciones obtenidas para obtener los resultados en función de los litros vendidos.

Finalmente, introduciendo un presupuesto, la función `mens` calculará, en función de las predicciones obtenidas, la mensualización del presupuesto.

A continuación seleccionamos una de las 66 rutas aplicables a la cerveza A para estudiar los resultados obtenidos. En la Figura 4.4 observamos la serie logaritmo de ventas y las series de temperatura y días de lluvia correspondiente a la ruta escogida.

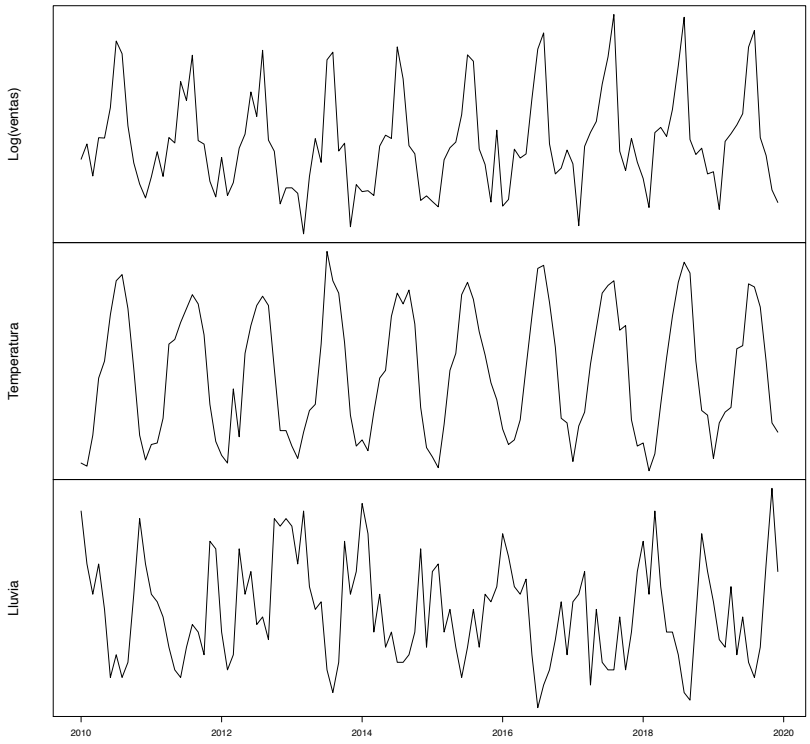


Figura 4.4: Gráficos secuenciales de las series logaritmo de ventas, temperatura media y número de días de lluvia.

Ajustando el modelo de regresión múltiple (4.1), tenemos que

$$y_t = 0.0151T_t - 0.0004L_t + 0.0380S_t + \varepsilon_t$$

$$\varepsilon_t = 0.9552\varepsilon_{t-1} + \varepsilon_{t-12} - 0.9552\varepsilon_{t-13} + a_t + 0.3326a_{t-12} - 0.3088a_{t-24}.$$

Observamos que los coeficientes relativos a las variables de temperatura y efecto de la Semana Santa son positivos. Esto implica que el aumento de la temperatura y la presencia de la Semana Santa se traducirán en ventas más altas. En cambio, en el caso de la lluvia, el coeficiente es negativo, por lo que cuantos más días de lluvia, menores serán las ventas.

Tras ajustar y predecir modelos ARIMA para las series de tiempo de temperaturas y lluvia, hemos realizado la predicción para el próximo año de nuestra serie y_t. En la Figura 4.5 podemos observar las predicciones (azul) junto con la serie histórica (negro).

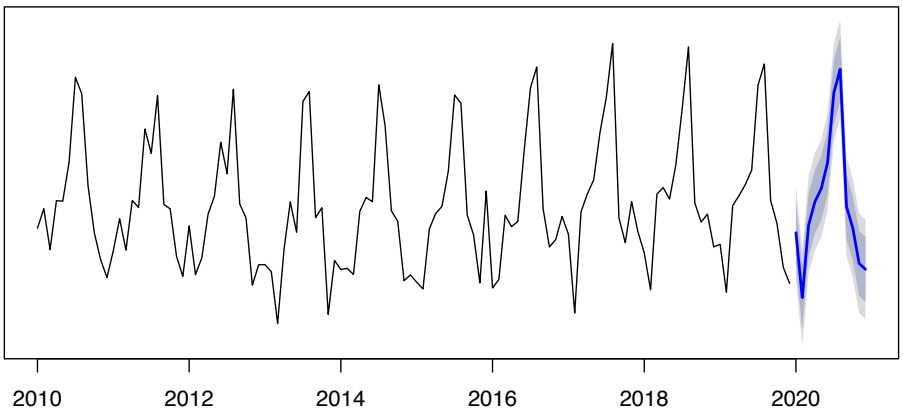


Figura 4.5: Gráficos secuencial de la serie del logaritmo de ventas con las predicciones obtenidas mediante la regresión.

Para que las predicciones sean útiles en la toma de decisiones de la empresa, realizamos una transformación exponencial sobre las predicciones, para así tener la predicción de las ventas en litros.

Por último, definiendo el presupuesto para el próximo año para la cerveza A en esta ruta, se calcularía la correspondiente mensualización.

4.2. Resultados en la cerveza A

A continuación comentaremos brevemente algunos de los resultados obtenidos para el caso de la cerveza A.

En el caso de esta cerveza, son 66 las rutas que verifican las condiciones de registro completo y ventas mensuales en litros superiores a 1. Los cuadros B.1, B.2, B.3 y B.4 del Apéndice B muestran un resumen de los modelos obtenidos para la serie del logaritmo de las ventas en litros de la cerveza A en cada una de las 66 rutas.

Los modelos obtenidos para los residuos de la regresión, ε_t, han sido muy diversos, alcanzando un total de 46 modelos distintos. El modelo más frecuente ha sido el ARIMA$(0,1,1)\times(1,0,0)_{12}$, que ha

aparecido hasta en ocho ocasiones. Salvo en un caso, todos los modelos han sido ARIMAs estacionales multiplicativos con periodo estacional $s = 12$.

Si nos fijamos en las estimaciones que hemos obtenido del coeficiente β_1, es decir, el coeficiente relativo a la variable regresora temperatura media, podemos observar que estas son siempre positivas, salvo en una de las rutas. Esto quiere decir, que el aumento en la temperatura provoca incrementos en las ventas. Además, en el 97 % de los casos esta variable resulta ser significativa.

En cambio, la estimación del coeficiente β_2, relativo a la variable regresora del número de días de lluvia, es mayoritariamente negativo. Por tanto, cuantos más días de lluvia hay en un mes, menores serán las ventas. También observamos que en casi en el 30 % de las rutas, este coeficiente resulta no ser significativo, por lo que en esos casos podría fijarse a cero.

En relación a la estimación del coeficiente relativo a la variable efecto de la Semana Santa, β_3, su signo no tiene un patrón tan claro como en los casos anteriores, siendo en el 46 % de las rutas negativo y en el 53 % positivo. Entonces, habrá casos en los que la presencia de la Semana Santa esté asociada a un aumento de las ventas en el mes correspondiente y otros en los que estará asociada a una disminución. Además, en más del 60 % de los casos, la variable relativa al efecto de la Semana Santa resulta no ser significativa, por lo que podría fijarse β_3 a cero en esos casos.

En líneas generales, podemos concluir que la covariable temperatura media es muy buena predictora para este producto, ya que ha resultado ser significativa casi en la totalidad de las rutas. Además, unas temperaturas altas están asociadas a valores mayores de ventas. En cuanto al número de días de lluvia observamos que mayoritariamente tiene un impacto negativo sobre las ventas.

Sin realizar un análisis previo, la intuición más razonable es que existe una relación entre la climatología y el consumo de cerveza. Una vez realizado el análisis desarrollado en este trabajo, encontramos que esta intuición se ve confirmada con los resultados obtenidos.

Apéndice A

Código de R

A.1. Predicción de ventas

```
pred<-function(MARCA_DESGLOSE_ID){
ii<-numeric()
aux<-levels(datos$RUTA_ID)
for (i in 1:length(aux)){
if (length(datos$MES[datos$RUTA_ID==aux[i] &
datos$MARCA_DESGLOSE_ID==MARCA_DESGLOSE_ID])==120){ii[i]<-i}
else {ii[i]<-0}}
ruta<-aux[ii]
sal<-matrix(nrow=length(ruta),ncol=14)
colnames(sal)<-c("RUTA_ID","MARCA_DESGLOSE_ID",month.name)
for (i in 1:length(ruta)){
dat<-datos[datos$RUTA_ID==ruta[i] & datos$MARCA_DESGLOSE_ID==MARCA_DESGLOSE_ID,]
serie<-ts(cbind(log(dat$VENTASACUMMESLITROS),dat$tm,dat$llu),
start=c(2010,1),frequency=12)
mreg<-matrix(c(dat$tm,dat$llu,easter(serie)),nrow=nrow(dat),byrow=FALSE)
colnames(mreg)<-c("tm","llu","ss")
fit<-auto.arima(serie[,1],xreg=mreg)
mregf<-ts(cbind(rep(0,12),rep(0,12),rep(0,12)),start=c(2020,1),frequency=12)
colnames(mregf)<-c("tm","llu","ss")
mregf[,3]<-easter(mregf[,3])
for (j in 1:2){
aux2<-as.data.frame(forecast(auto.arima(serie[,j+1]),h=12))
mregf[,j]<-aux2[,1]}
fcast<-forecast(fit, xreg=mregf)
pred<-exp(as.data.frame(fcast)[,1])
sal[i,]<-c(ruta[i],MARCA_DESGLOSE_ID,pred)}
write.csv2(sal,file="predicciones.csv")}
```

A.2. Mensualización del presupuesto

```
mens<-function(RUTA_ID,MARCA_DESGLOSE_ID,presupuesto){
v<-numeric()
pred<-numeric()
for (i in 1:12){
```

```
pred[i]<-sal[(sal$RUTA_ID==RUTA_ID &
sal$MARCA_DESGLOSE_ID==MARCA_DESGLOSE_ID),i+2]}
prop.pred<-pred*100/sum(pred)
mp<-prop.pred*presupuesto/100
v<-t(c(RUTA_ID,MARCA_DESGLOSE_ID,mp))
colnames(v)<-c("RUTA_ID","MARCA_DESGLOSE_ID",month.name)
write.csv2(v,file="mensualizacion.csv")}
```

Apéndice B

Tablas de resultados de la cerveza A

Ruta	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	Modelo ARIMA de ε_t
1	0.0151	-0.0004*	0.0380	ARIMA(1,0,2)×$(0,1,2)_{12}$
2	0.0347	0.0032	0.0094*	ARIMA(0,1,1)×$(1,0,1)_{12}$
3	0.0424	-0.0004*	0.0329	ARIMA(0,1,2)×$(2,0,0)_{12}$
4	0.0320	-0.0085	0.1806	ARIMA(1,1,1)×$(1,0,0)_{12}$
5	0.0395	0.0004*	-0.0312	ARIMA(1,1,1)×$(1,0,0)_{12}$
6	0.0178	-0.0006*	-0.0792	ARIMA(0,1,1)×$(1,0,0)_{12}$
7	0.0256	0.0008	0.0000*	ARIMA(0,1,1)×$(1,0,1)_{12}$
8	0.0397	0.0019	-0.0049*	ARIMA(0,0,0)×$(0,0,2)_{12}$
9	0.0349	0.0014*	0.0623	ARIMA(0,1,1)×$(1,0,0)_{12}$
10	0.0116*	-0.0065	0.0937	ARIMA(0,1,1)×$(0,1,1)_{12}$
11	0.0219	-0.0059	0.0369*	ARIMA(3,0,0)×$(0,1,2)_{12}$
12	0.0546	0.0016	-0.0813*	ARIMA(0,1,4)×$(2,0,0)_{12}$
13	0.0156	-0.0030	-0.0227*	ARIMA(0,0,0)×$(0,1,1)_{12}$
14	0.0150	-0.0019*	-0.0789	ARIMA(1,0,1)×$(1,0,0)_{12}$

Cuadro B.1: Coeficientes del modelo de regresión de la cerveza A.

Ruta	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	Modelo ARIMA de ε_t
15	0.0405	-0.0016*	-0.0600	ARIMA(0,0,0)×(2,0,0)$_{12}$
16	0.0522	-0.0051	-0.0041*	ARIMA(2,1,1)×(1,0,0)$_{12}$
17	0.0365	-0.0088	-0.0451*	ARIMA(1,1,1)×(2,0,0)$_{12}$
18	0.0145	-0.0009	-0.0142*	ARIMA(3,0,0)×(2,1,1)$_{12}$
19	0.0229	-0.0035	-0.0112*	ARIMA(0,1,2)×(0,1,1)$_{12}$
20	0.0239	-0.0027	0.0328*	ARIMA(0,1,1)×(2,0,0)$_{12}$
21	0.0471	-0.0030*	0.1505	ARIMA(0,1,2)×(0,0,2)$_{12}$
22	0.0276	-0.0052	0.0588	ARIMA(1,0,0)×(2,0,0)$_{12}$
23	0.0219	-0.0041	-0.0555	ARIMA(3,1,1)×(1,0,0)$_{12}$
24	0.0326	-0.0027	0.0118*	ARIMA(0,1,1)×(1,0,0)$_{12}$
25	0.0371	0.0014	-0.0101*	ARIMA(0,0,0)×(0,0,2)$_{12}$
26	0.0217	0.0001	-0.0226*	ARIMA(2,1,1)×(0,0,2)$_{12}$
27	0.0285	-0.0039	-0.0072*	ARIMA(0,1,2)×(1,0,0)$_{12}$
28	0.0298	-0.0059	0.0246*	ARIMA(3,0,0)×(1,0,0)$_{12}$
29	0.0115	-0.0041*	-0.0502	ARIMA(1,1,1)×(1,1,2)$_{12}$
30	0.0369	-0.0046	0.0262*	ARIMA(0,0,0)×(1,0,0)$_{12}$
31	0.0139	-0.0023*	0.0487	ARIMA(0,1,1)×(0,0,1)$_{12}$
32	0.0325	-0.0011	-0.0071*	ARIMA(0,1,2)×(0,0,2)$_{12}$
33	0.0238	-0.0036	0.0593*	ARIMA(0,1,2)×(0,0,2)$_{12}$
34	0.0245	-0.0027	-0.0058*	ARIMA(0,1,1)×(0,0,1)$_{12}$
35	0.0175	-0.0027	0.0144*	ARIMA(0,0,3)×(1,0,0)$_{12}$
36	0.0247	-0.0058	0.0035*	ARIMA(1,0,2)×(0,1,1)$_{12}$
37	0.0241	-0.0040	-0.0377*	ARIMA(2,1,1)×(0,0,2)$_{12}$

Cuadro B.2: Coeficientes del modelo de regresión de la cerveza A.

Ruta	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	Modelo ARIMA de ε_t
38	0.0275	0.0009	-0.0064*	ARIMA(2,1,0)×$(1,0,0)_{12}$
39	0.0269	-0.0032	-0.0285*	ARIMA(0,1,1)×$(1,0,1)_{12}$
40	0.0337	-0.0024*	-0.0957	ARIMA(0,0,1)×$(2,0,0)_{12}$
41	0.0226	-0.0040	0.0528*	ARIMA(0,1,1)×$(1,0,0)_{12}$
42	0.0374	-0.0047	-0.0196*	ARIMA(0,1,1)×$(1,0,0)_{12}$
43	0.0186	-0.0005*	-0.0358	ARIMA(4,0,0)×$(1,0,0)_{12}$
44	0.0436	0.0001*	-0.0567	ARIMA(0,1,1)×$(1,0,0)_{12}$
45	0.0357	-0.0051	0.0769*	ARIMA(4,1,1)×$(1,0,0)_{12}$
46	0.0557	-0.0071	0.1190*	ARIMA(0,0,1)×$(1,0,0)_{12}$
47	0.0446	-0.0073	0.1226*	ARIMA(0,0,1)×$(1,0,1)_{12}$
48	0.0120	-0.0032	-0.0100*	ARIMA(5,1,1)×$(0,0,2)_{12}$
49	0.0239	-0.0067	0.1085	ARIMA(1,0,2)×$(0,1,1)_{12}$
50	0.0235	-0.0048	0.0123*	ARIMA(3,1,1)×$(1,0,0)_{12}$
51	0.0237	-0.0036*	0.0993	ARIMA(3,0,0)×$(0,1,1)_{12}$
52	0.0270	-0.0038	0.0236*	ARIMA(1,0,2)×$(0,0,1)_{12}$
53	0.0187	0.0006*	-0.0204	ARIMA(0,1,1)×$(1,0,0)_{12}$
54	-0.0074	0.0043*	-0.1225	ARIMA(0,1,3)×$(1,0,0)_{12}$
55	0.0351	-0.0085	0.1112	ARIMA(0,1,2)×$(2,0,0)_{12}$
56	0.0396	0.0022	-0.0233*	ARIMA(1,1,1)×$(0,0,2)_{12}$
57	0.0157	-0.0048	-0.0150*	ARIMA(2,1,1)×$(1,0,1)_{12}$
58	0.0296	-0.0053	0.0663*	ARIMA(0,1,2)×$(0,0,2)_{12}$
59	0.0158	0.0002*	-0.0443	ARIMA(0,1,1)×$(1,0,0)_{12}$
60	0.0273	-0.0043	0.0078*	ARIMA(1,1,1)×$(2,0,0)_{12}$

Cuadro B.3: Coeficientes del modelo de regresión de la cerveza A.

Ruta	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	Modelo ARIMA de ε_t
61	0.0367*	0.0000	-0.0400	ARIMA(0,0,1)×(2,0,0)$_{12}$
62	0.0143	0.0055	0.0672*	ARIMA(0,1,1)
63	0.0363	-0.0058	0.0474*	ARIMA(0,1,2)×(1,0,0)$_{12}$
64	0.0496	-0.0050	0.0429*	ARIMA(0,1,1)×(0,0,1)$_{12}$
65	0.0757	0.0038*	0.3826	ARIMA(0,0,0)×(1,0,0)$_{12}$
66	0.0441	-0.0023	-0.0286*	ARIMA(3,1,1)×(2,0,0)$_{12}$

Cuadro B.4: Coeficientes del modelo de regresión de la cerveza A.

Bibliografía

[1] Box GEP, Jenkins GM (1976) Time Series Analysis: Forecasting and Control. Holden Day.

[2] Brockwell PJ, Davis RA (1991) Time Series: Theory and Methods. Springer, New York.

[3] Chan KS, Ripley B (2018) TSA: Time Series Analysis. R package version 1.2. http://www.cran.r-project.org/package=TSA. Accedido 29 de abril de 2020.

[4] Cowpertwait PS, Metcalfe AV (2009) Introductory Time Series with R. Springer, New York.

[5] Cryer JD, Chan KS (2010) Time Series Analysis with Applications in R. Springer, New York.

[6] Fisher R (1925) The Influence of Rainfall on the Yield of Wheat at Rothamsted. Philosophical Transactions of the Royal Society of London. Series B, Containing Papers of a Biological Character 213:89-142.

[7] Hyndman RJ (2018) Data for "Forecasting: Principles and Practic" (2nd Edition). R package version 2.3. https://cran.r-project.org/web/packages/fpp2/fpp2.pdf. Accedido 28 de junio de 2020.

[8] Hyndman RJ, Athanasopoulos G (2018) Forecasting: principles and practic (2nd ed.). OTexts, Melburne.

[9] Hyndman RJ, Khandakar Y (2008) Forecasting Functions for Time Series and Linear Models. R package version 8.12. https://cran.r-project.org/web/packages/forecast/forecast.pdf. Accedido 28 de junio de 2020.

[10] Johnson RA, Wichern DW (2007) Applied Multivariate Statistical Analysis. Pearson Prentice Hall.

[11] Peña D (2005) Análisis de series temporales. Alianza Editorial, Madrid.

[12] Peña D, Tiao GC, Tsay RS (2001) A Course in Time Series Analysis. Wiley, New York.

[13] R Core Team (2018) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

[14] Shumway RH, Stoffer DS (2017) Time Series Analysis and Its Applications: With R Examples. Springer, Basingstoke.

[15] Slutsky E (1937) The summation of random causes as the source of cyclic processes. Econometrica 5(2):105-146.

[16] Wold HOA (1938) A Study of the Analysis of Stationary Time Series. Almqvist and Wiksells, Uppsala.

[17] Yule GU (1927) On a method of investigating periodicities disturbed series, with special reference to Wolfer's sunspot numbers. Philosophical Transactions of the Royal Society of London 226:267-298.