



UNIVERSIDAD DE BUENOS AIRES
Facultad de Ciencias Exactas y Naturales
Departamento de Matemática

Tesis de Licenciatura

Modelos de Series de Tiempo con aplicaciones en la industria
aerocomercial

Andrés Pablo Mogni

Director: Dra. Daniela Rodriguez

12 de Marzo de 2013

Índice general

1. Introducción	2
2. Modelos de Series de Tiempo	5
2.1. Modelos de Series de Tiempo Estacionarias	5
2.1.1. Definiciones	5
2.1.2. Representaciones Autoregresivas y de Promedio Movil	8
2.2. Modelos de Series de Tiempo No Estacionarias	18
2.2.1. No estacionariedad en media	18
2.2.2. No estacionariedad en varianza y autocovarianza	19
2.3. Series de Tiempo estacionales	21
3. Identificación, Estimación y Pronóstico	23
3.1. Identificación del Modelo	23
3.2. Estimación de Parámetros	25
3.3. Diagnóstico y Selección del Modelo	27
3.3.1. Algunas herramientas de validación	28
3.4. Forecasting	30
3.5. Modelos RegARIMA	32
4. Analisis de datos reales	34
4.1. AEP-CNQ-AEP	35
4.1.1. Caso $\lambda = -0,5$ y $d = D = 1$	37
4.1.2. Caso $\lambda = -0,5$ y $d = D = 0$	41
4.1.3. Caso $\lambda = 0$ y $d = D = 1$	45
4.1.4. Caso $\lambda = 0$ y $d = D = 0$	49

4.1.5. Modelos regARIMA	51
4.1.6. Pronósticos y conclusiones	55
4.2. AEP-FMA-AEP	57
4.2.1. ARIMA	57
4.2.2. RegARIMA	62
4.2.3. Pronosticos y conclusiones.	65

A mis padres, que desde chico me mostraron el valor del estudio y me acompañaron en cada etapa de la carrera, sosteniéndome y apoyándome en los malos momentos y disfrutando conjuntamente los buenos.

A Bel, mi compañera de vida que conocí en Análisis I y con quien disfruto cada segundo de mi vida. Sin ella hubiese sido muy difícil superar los momentos difíciles y gracias a ella sigo creciendo como persona.

A Dani, a quien tuve la suerte de tener como profesora en dos oportunidades y a quien considero uno de los grandes valores que tiene la facultad. Como directora fue una importante guía a la hora de aprender desde cero un tema y terminó saliendo un trabajo muy lindo. Gracias Dani!

A Caro Mosquera, a quien tuve como profesora en tres momentos clave de la carrera y fue quien me enseñó a pensar como matemático. Sin ella hubiese sido difícil haber madurado matemáticamente y haber logrado todo esto.

A mis compañeros, principalmente a Caro y Anto, con quienes transité este camino y formé una muy linda relación.

Capítulo 1

Introducción

En diferentes industrias surge la necesidad de pronosticar eventos con el objetivo de poder planear de un modo más eficiente las decisiones a tomar. En particular, el sector aerocomercial ve esa necesidad con mayor avidez ya que presenta muchos actores, escasos márgenes de ganancias y altos costos, fundamentalmente los producidos por el combustible. Por tal motivo, poder pronosticar la demanda de pasajeros consiste en un trabajo fundamental. Este tipo de información permite tomar acertadamente decisiones tales como aumento y/o disminución de frecuencias, cambios de equipos, inversión en equipos nuevos, aumento y/o disminución de tarifas, entre otros.

Es esta tesis abordaremos un tipo de modelo de pronóstico elaborado por Box y Jenkins denominado ARMA/ARIMA. Presentaremos la teoría correspondiente y los pasos a seguir para identificar y pronosticar los modelos. Finalmente aplicaremos la teoría para dos ejemplos reales, trataremos los casos de las rutas Aeroparque-Corrientes-Aeroparque y Aeroparque-Formosa-Aeroparque, volados por la principal línea aérea del país.

Los principales materiales de referencia de donde se obtuvo la idea central de la teoría fueron los libros “Times Series Analysis” de Wei (2006) y “Time Series Analysis: Forecasting and Control” de Box-Jenkins (1984). Otra bibliografía consultada e importante a la hora de profundizar en algunos conceptos corresponden a los libros de Cryer y Chang (2008), Greene (2008) y Brockwell y Davis (1996) entre otros.

En el capítulo 2 definiremos conceptos básicos necesarios para modelar, introduciremos las series de tiempo estacionarias y no estacionarias y observaremos que ocurre cuando hay estacionalidad. En el capítulo 3 presentaremos todos los pasos para modelar una serie de datos. En particular veremos como identificar los parámetros del modelo, como estimar los coeficientes de este, que modelo elegir y como pronosticar. Estudiaremos también los modelos RegARIMA, que combinan los modelos de regresión lineal modelando los residuos como un proceso ARIMA. El último capítulo

contiene el análisis y los resultados correspondientes para el modelado de las dos rutas aéreas, compararemos los modelos obtenidos y contrastaremos con los datos reales.

Capítulo 2

Modelos de Series de Tiempo

2.1. Modelos de Series de Tiempo Estacionarias

2.1.1. Definiciones

En este Capítulo presentaremos ciertas definiciones que nos servirán para el posterior desarrollo de los temas. Sea $Z(w, t)$ un proceso estocástico con w en algún espacio muestral, en este trabajo consideraremos $t \in \mathbb{Z}$. Sean $t_1, \dots, t_n \in \mathbb{Z}$ y llamemos $Z_{t_j} = Z(w, t_j)$ con $1 \leq j \leq n$ variables aleatorias provenientes de un proceso estocástico $Z(w, t)$. Definimos su función de distribución como:

$$F_{Z_{t_1}, \dots, Z_{t_n}}(x_1, \dots, x_n) = P(w : Z_{t_1} \leq x_1, \dots, Z_{t_n} \leq x_n).$$

Un proceso se dice de *primer orden estacionario en distribución* si:

$$F_{Z_{t_1}}(x_1) = F_{Z_{t_1+k}}(x_1) \quad \forall k.$$

En general, se dice que es de *orden n estacionario en distribución* si:

$$F_{Z_{t_1}, \dots, Z_{t_n}}(x_1, \dots, x_n) = F_{Z_{t_1+k}, \dots, Z_{t_n+k}}(x_1, \dots, x_n).$$

A partir de ahora denotaremos $Z(w, t)$ indistintamente como $Z(t)$ o Z_t e introduciremos también los siguientes conceptos:

- i) *Esperanza*: $\mu_t = E(Z_t)$.
- ii) *Varianza*: $\sigma_t^2 = E(Z_t - \mu_t)^2$.
- iii) *Covarianza*: $\gamma(t_1, t_2) = E(Z_{t_1} - \mu_{t_1})(Z_{t_2} - \mu_{t_2})$.
- iv) *Función de Correlación*: $\rho(t_1, t_2) = \frac{\gamma(t_1, t_2)}{\sqrt{\sigma_{t_1}^2 \sigma_{t_2}^2}}$.

Para procesos estacionarios tendremos que $\mu_t = \mu$ y $\sigma_t^2 = \sigma^2$ y dados t y $t+k$ tenemos que $\gamma(t, t+k)$ y $\rho(t, t+k)$ dependerán solo de k . Un proceso se definirá como de *Segundo Orden Débil* o *Covarianza Estacionaria* si sus momentos de orden 1 y 2 (su esperanza y varianza) no dependen de t . En muchas ocasiones se usa el término estacionario para procesos que son de covarianza estacionaria. En el caso particular de un proceso gaussiano, como su distribución queda determinada por μ y σ^2 se tiene que orden fuerte es lo mismo que orden débil. Pero en general un proceso puede ser fuertemente estacionario y no ser débilmente estacionario (como ejemplo podemos pensar en la distribución Cauchy).

Funciones de Autocorrelación, Autocovarianza y de Autocorrelación Parcial

Definición 2.1.1. Dado un proceso estacionario Z_t con $E(Z_t) = \mu$ y $V(Z_t) = \sigma^2$ constantes, se definen:

- Función de Autocovarianza: $\gamma_k = Cov(Z_t, Z_{t+k}) = E(Z_t - \mu)(Z_{t+k} - \mu)$.
- Función de Autocorrelación (ACF): $\rho_k = \frac{Cov(Z_t, Z_{t+k})}{\sqrt{V(Z_t)}\sqrt{V(Z_{t+k})}} = \frac{\gamma_k}{\gamma_0}$.

Para todo proceso estacionario se satisfacen las siguientes propiedades:

- i) Para todo t , $\gamma_0 = V(Z_t)$.
- ii) Del hecho que $|\rho_k| \leq 1$ tenemos que $|\gamma_k| \leq \gamma_0$ para todo k .
- iii) γ_k y ρ_k son simétricas con respecto a k , i.e. $\gamma_k = \gamma_{-k}$ y $\rho_k = \rho_{-k}$ para todo k .
- iv) γ_k y ρ_k son semidefinidas positivas. Es decir, valen las siguientes desigualdades

$$\sum_{i=1}^n \sum_{j=1}^n a_i a_j \gamma_{|t_i - t_j|} \geq 0,$$

y

$$\sum_{i=1}^n \sum_{j=1}^n a_i a_j \rho_{|t_i - t_j|} \geq 0,$$

para todos t_1, \dots, t_n y a_1, \dots, a_n .

Definiremos también la *Función de Autocorrelación Parcial* (PACF) como la correlación condicional

$$P_k = \text{Corr}(Z_t, Z_{t+k} | Z_{t+1}, \dots, Z_{t+k-1}).$$

Consideremos ahora la regresión con variable dependiente Z_{t+k} y covariables Z_{t+k-1}, \dots, Z_t . Es decir

$$Z_{t+k} = \phi_{k1}Z_{t+k-1} + \dots + \phi_{kk}Z_t + e_{t+k}$$

con ϕ_{ki} los parámetros de la regresión y e_{t+k} errores de media 0 y no correlacionados con $Z_{t+k-j} \forall j$. Supongamos sin pérdida de generalidad que $E(Z_t) = 0$, luego multiplicando a ambos lados de la regresión por Z_{t+k-j} y tomando esperanza obtenemos:

$$\gamma_j = \phi_{k1}\gamma_{j-1} + \dots + \phi_{kk}\gamma_{j-k},$$

$$\rho_j = \phi_{k1}\rho_{j-1} + \dots + \phi_{kk}\rho_{j-k}.$$

Usando la Regla de Cramer, podemos despejar los coeficientes ϕ_{ki} en particular se obtiene fácilmente que $\phi_{kk} = P_k$, la autocorrelación parcial entre Z_t y Z_{t+k} . Por otro lado podemos determinar que la autocorrelación parcial entre Z_t y Z_{t+k} se puede obtener como los coeficientes de la regresión asociada a Z_t de k pasos.

Un proceso estacionario particular es el denominado Ruido Blanco. Diremos que un proceso a_t es de *Ruido Blanco* si es una sucesión de variables aleatorias donde $E(a_t) = 0$, $V(a_t) = \sigma^2$ y $\gamma_k = Cov(a_t, a_{t+k}) = 0$ si $k \neq 0$.

Estimación de μ , σ^2 , γ_k y ρ_k

Las definiciones introducidas anteriormente son referidas a momentos poblacionales que no son observables. A continuación, resumiremos algunos posibles estimadores de los parámetros definidos anteriormente, obtenidos a partir de una realización del proceso en n instantes de tiempo.

Resulta natural considerar $\bar{Z} = \sum_{t=1}^n \frac{Z_t}{n}$, como un estimador de la media poblacional μ . Se puede probar fácilmente que el estimador resulta insesgado. Además podemos ver que si $\rho_k \rightarrow 0$, $Var(\bar{Z}) \rightarrow 0$ y por lo tanto el estimador resulta débilmente consistente.

Para estimar γ_k podemos considerar las siguientes alternativas

$$\hat{\gamma}_k = \sum_{t=1}^{n-k} \frac{(Z_t - \bar{Z})(Z_{t+k} - \bar{Z})}{n} \quad \text{o} \quad \hat{\gamma}_k = \frac{n\hat{\gamma}_k}{n-k}.$$

Desarrollando $\hat{\gamma}_k$ podemos aproximar

$$n\hat{\gamma}_k \sim \sum_{t=1}^{n-k} (Z_t - \mu)(Z_{t+k} - \mu) - (n-k)(\bar{Z} - \mu)^2.$$

En ese caso, $E(\hat{\gamma}_k) \sim \gamma_k - \frac{k}{n}\gamma_k - (\frac{n-k}{n})V(\bar{Z})$ y $E(\hat{\gamma}_k) \sim \gamma_k - V(\bar{Z})$.

Donde si se desprecia el termino $V(\bar{Z})$, que representa el efecto de estimar la varianza, $\hat{\gamma}_k$ resulta un estimador con menor sesgo que $\hat{\gamma}_k$. Pero $\hat{\gamma}_k$ al igual que γ_k es semidefinida positiva, sin embargo $\hat{\gamma}_k$ no necesariamente lo es. Es por ello que en general se considera $\hat{\gamma}_k$ para estimar γ_k .

Como estimador de ρ_k usaremos

$$\hat{\rho}_k = \frac{\hat{\gamma}_k}{\hat{\gamma}_0} = \frac{\sum_{t=1}^{n-k} (Z_t - \bar{Z})(Z_{t+k} - \bar{Z})}{\sum_{t=1}^n (Z_t - \bar{Z})^2}.$$

A partir de los datos de una series temporal se puede estimar un número finito de autocorrelaciones. El gráfico de las autocorrelaciones muestrales recibe el nombre de correlograma.

Para estimar la función de autocorrelación parcial ϕ_{kk} usaremos la siguiente recursión definida por Durbin (1960),

$$\begin{aligned} \hat{\phi}_{11} &= \hat{\rho}_1, \\ \hat{\phi}_{k+1 \ k+1} &= \frac{\hat{\rho}_{k+1} - \sum_{j=1}^k \hat{\phi}_{kj} \hat{\rho}_{k+1-j}}{1 - \sum_{j=1}^k \hat{\phi}_{kj} \hat{\rho}_j}, \\ \hat{\phi}_{k+1 \ j} &= \hat{\phi}_{kj} - \hat{\phi}_{k+1 \ k+1} \hat{\phi}_{k \ k+1-j}. \end{aligned}$$

2.1.2. Representaciones Autoregresivas y de Promedio Movil

En esta Sección analizaremos los modelos estacionarios que pueden clasificarse como autoregresivos (AR), de medias o promedios móviles (MA) o procesos mixtos que se denominan ARMA. En el análisis de datos reales, se encuentran muy raramente series generadas a partir de procesos estacionarios ya que generalmente las series presentan algún tipo de tendencia que no se corresponde con procesos de un modelo estacionario. Sin embargo esto no debe llevar a pensar que estos procesos no son interesante de estudiar. Como veremos más adelante, una gran cantidad de procesos no estacionarios pueden ser fácilmente transformados en estacionarios y a partir de esta transformación les son aplicables los métodos de identificación y estimación de los modelos estacionarios.

Antes de continuar, definiremos los operadores de retardo y diferenciación de una serie. Más precisamente, el operador de retardo B se define como

$$BY_t = Y_{t-1},$$

es decir el resultado de aplicar el operador B resulta la observación del período anterior, por lo tanto si aplicamos d veces el operador obtenemos $B^d Y_t = Y_{t-d}$.

El operador diferencia de una serie se define como

$$\Delta Y_t = Y_t - Y_{t-1} = (1 - B)Y_t,$$

en general $\Delta^d Y_t = (1 - B)^d Y_t$.

Consideremos ahora, un proceso de la forma

$$Z_t = \mu + a_t + \psi_1 a_{t-1} + \psi_2 a_{t-2} + \dots = \mu + \sum_{j=0}^{\infty} \psi_j a_{t-j},$$

y su respectivo proceso de media cero $\dot{Z}_t = Z_t - \mu$, donde las a_t son variables aleatorias no correlacionadas de media 0, notemos que

$$\begin{aligned} E(Z_t) &= \mu, \\ \text{Var}(Z_t) &= \sigma_a^2 \sum_{j=0}^{\infty} \psi_j^2, \\ \gamma_k &= E(\dot{Z}_t \dot{Z}_{t+k}) = E\left(\sum_{j=0}^{\infty} \sum_{i=0}^{\infty} \psi_i \psi_j a_{t-i} a_{t+k-j}\right) = \sigma_a^2 \sum_{i=0}^{\infty} \psi_i \psi_{i+k}, \\ \rho_k &= \frac{\sum_{i=0}^{\infty} \psi_i \psi_{i+k}}{\sum_{i=0}^{\infty} \psi_i^2}. \end{aligned}$$

Si observamos las funciones de autocorrelación y autocovarianza, al solo depender estas de k (y no de t) y ser sumas infinitas, para poder afirmar que el proceso es realmente estacionario, es decir de covarianza estacionario, basta ver que γ_k es finito para cada k . Por lo tanto, como

$$|\gamma_k| = |E(\dot{Z}_t \dot{Z}_{t+k})| \leq (\text{Var}(Z_t) \text{Var}(Z_{t+k}))^{1/2} = \sigma_a^2 \sum_{j=0}^{\infty} \psi_j^2. \quad (2.1)$$

si $\sum_{j=0}^{\infty} \psi_j^2 < \infty$ (y en consecuencia $\psi(B)$ converge en el círculo unitario) es una condición suficiente para que el proceso sea estacionario.

Otra forma útil de escribir un proceso es a través de su representación autoregresiva es decir

$$\dot{Z}_t = \pi_1 \dot{Z}_{t-1} + \pi_2 \dot{Z}_{t-2} + \dots + a_t,$$

o de manera equivalente

$$\pi(B) \dot{Z}_t = a_t,$$

donde $\pi(B) = 1 - \sum_{j=1}^{\infty} \pi_j B^j$ y $1 + \sum_{j=1}^{\infty} |\pi_j| < \infty$.

Un proceso se dirá *invertible* si puede ser representado de esta manera. Notemos que dado un proceso de la forma de $\dot{Z}_t = \psi(B) \dot{a}_t$, será invertible si las raíces de $\psi(B) = 0$ se encuentran fuera del círculo unitario.

Para comprender esto consideremos un ejemplo particular, sea un proceso de la forma $\dot{Z}_t = (1 - \theta B)a_t$. Si expresamos a a_t en función de los \dot{Z}_t , obtenemos que

$$a_t = (1 - \theta B)^{-1} \dot{Z}_t = (1 + \theta B + \theta B^2 + \dots + \theta B^k)(1 - \theta^{k+1} B^{k+1})^{-1} \dot{Z}_t,$$

es decir

$$\dot{Z}_t = -\theta \dot{Z}_{t-1} - \theta^2 \dot{Z}_{t-2} - \dots - \theta^k \dot{Z}_{t-k} + a_t - \theta^{k+1} a_{t-k-1}. \quad (2.2)$$

Luego, si $|\theta| < 1$, tendiendo k a infinito, obtenemos que

$$\dot{Z}_t = -\theta \dot{Z}_{t-1} - \theta^2 \dot{Z}_{t-2} - \dots + a_t \quad (2.3)$$

que define un proceso estacionario. Además, $\theta^j = \pi_j$. Luego, si $|\theta| \geq 1$, tenemos que (2.2) se agranda a medida que k se hace grande. Luego se pide $|\theta| < 1$ para que la serie sea invertible y se satisfará si

$$\pi(B) = (1 - \theta B)^{-1} = \sum_{j=0}^{\infty} \theta^j B^j$$

converge para todo $|B| \leq 1$, es decir dentro del círculo unitario.

En general tendremos que un proceso será estacionario si $\psi(B)$ converge dentro del círculo unitario y será invertible si $\pi(B)$ converge dentro del círculo unitario.

Finalmente presentaremos el concepto de la función generatriz de autocovarianza que denominaremos γ . En efecto, dada γ_k definiremos γ de la forma

$$\gamma(B) = \sum_{k=-\infty}^{\infty} \gamma_k B^k. \quad (2.4)$$

Utilizando (2.1) y la condición de estacionariedad, reescribimos la función generatriz de autocovarianza del siguiente modo

$$\begin{aligned} \gamma(B) &= \sigma_a^2 \sum_{k=-\infty}^{\infty} \sum_{i=0}^{\infty} \psi_i \psi_{i+k} B^k = \sigma_a^2 \sum_{j=0}^{\infty} \sum_{i=0}^{\infty} \psi_i \psi_j B^{j-i} \\ &= \sigma_a^2 \sum_{j=0}^{\infty} \psi_j B^j \sum_{i=0}^{\infty} \psi_i B^{-i} = \sigma_a^2 \psi(B) \psi(B^{-1}). \end{aligned}$$

Este método ayuda a calcular de forma conveniente la autocovarianza de un proceso. La correspondiente función generatriz de autocorrelación se define como

$$\rho(B) = \sum_{k=-\infty}^{\infty} \frac{\gamma(B)}{\gamma_0}. \quad (2.5)$$

Modelos MA

Los procesos de orden q de medias móviles, $MA(q)$, se definen de la siguiente forma

$$Z_t = \mu + a_t + \psi_1 a_{t-1} + \psi_2 a_{t-2} + \dots + \psi_q a_{t-q} = \mu + \sum_{j \geq 0}^q \psi_j a_{t-j} \quad (2.6)$$

donde a_t es un proceso de ruido blanco con las propiedades ya definidas.

Observemos que el proceso de medias móviles corresponde a una combinación lineal de variables ruido blanco, siendo los coeficientes ψ los ponderadores de la combinación lineal. Su nombre proviene del hecho que las variables forman parte de este promedio, aunque los coeficientes no sumen uno, que varia a lo largo del tiempo. Los modelos de infinitos términos los denotaremos con $MA(\infty)$.

Como vimos anteriormente sus momentos pueden ser calculados fácilmente, resultando:

$$E(Z_t) = \mu \quad V(Z_t) = \sigma_a^2 \sum_{j \geq 0} \psi_j^2$$

$$\gamma_k = \sigma_a^2 \sum_{i \geq 0} \psi_i \psi_{i+k} \quad \rho_k = \frac{\sum_{i \geq 0} \psi_i \psi_{i+k}}{\sum_{i \geq 0} \psi_i^2}.$$

Como su esperanza y su varianza son invariantes en el tiempo, este proceso será estacionario si $\sum \psi_j^2 < \infty$.

Por una cuestión de mera conveniencia el proceso se escribirá con los coeficientes precedidos por el signo negativo. Además si usamos el operador de retardo un proceso $MA(q)$ puede definirse de la siguiente manera.

$$\dot{Z}_t = \theta_q(B) a_t,$$

donde $\theta_q(B) = (1 - \theta_1 B - \theta_2 B^2 - \dots - \theta_q B^q)$ y $\dot{Z}_t = Z_t - \mu$. El polinomio $\theta_q(B)$ lo llamaremos polinomio de medias móviles.

Luego tenemos,

$$\gamma_k = \begin{cases} \sigma_a^2(-\theta_k + \theta_1 \theta_{k+1} + \dots + \theta_{q-k} \theta_q) & \text{si } k = 1, 2, \dots, q \\ 0 & \text{si } k > q \end{cases} \quad (2.7)$$

y

$$\rho_k = \begin{cases} \frac{(\theta_1 \theta_{k+1} + \dots + \theta_{q-k} \theta_q)}{(1 + \theta_1^2 + \dots + \theta_q^2)} & \text{si } k = 1, 2, \dots, q \\ 0 & \text{si } k > q \end{cases}$$

Observando (2.7) notemos que la función γ_k es 0 después del paso q . Es decir, en general, un proceso $MA(q)$ tendrá una función de autocorrelación que se corta después del paso q .

Analicemos en profundidad un proceso MA(1). En efecto, sea el proceso de la forma

$$\begin{aligned}\dot{Z}_t &= a_t - \theta_1 a_{t-1} \\ &= (1 - \theta_1 B)a_t,\end{aligned}$$

Usando la función generatriz de autocovarianza, obtenemos para este caso la siguiente expresión

$$\gamma(B) = \sigma_a^2(1 - \theta_1 B)(1 - \theta_1 B^{-1}) = \sigma_a^2(-\theta_1 B^{-1} + (1 + \theta_1^2) - \theta_1 B).$$

Luego, la función de autocovarianza es la siguiente

$$\gamma_k = \begin{cases} (1 + \theta_1^2)\sigma_a^2 & \text{si } k = 0 \\ -\theta_1\sigma_a^2 & \text{si } k = 1 \\ 0 & \text{si } k > 1. \end{cases}$$

Una vez obtenida la función de autocovarianza podemos calcular la función de autocorrelación. En este caso es

$$\rho_k = \begin{cases} \frac{-\theta_1}{1+\theta_1^2} & \text{si } k = 1 \\ 0 & \text{si } k > 1, \end{cases} \quad (2.8)$$

que se corta después de $k=1$. Notemos que el proceso es estacionario pues $1 + \theta_1^2 < \infty$. Para que sea invertible necesitamos que las raíces de $1 - \theta_1 B$ estén fuera del círculo unitario. Para ello debe valer que $|\theta_1| < 1$.

Calculemos ahora la función de autocorrelación parcial. Para ello, utilizaremos (2.8) y su definición, luego la función es de la forma

$$\begin{aligned}\phi_{11} &= \rho_1 = \frac{-\theta_1}{1 + \theta_1^2} = \frac{-\theta_1(1 - \theta_1^2)}{1 + \theta_1^4} \\ \phi_{22} &= \frac{-\rho_1^2}{1 - \rho_1^2} = \frac{-\theta_1^2}{1 + \theta_1^2 + \theta_1^4} \\ \phi_{33} &= \frac{\rho_1^3}{1 - 2\rho_1^2} = \frac{-\theta_1^3}{1 + \theta_1^2 + \theta_1^4 + \theta_1^6}.\end{aligned}$$

En general,

$$\phi_{kk} = \frac{-\theta_1^k(1 - \theta_1^2)}{1 - \theta_1^{2(k+1)}}. \quad (2.9)$$

Notemos así que el PACF decae exponencialmente, la forma en la que decae dependerá del signo de θ_1 . Si el signo alterna, empieza con valores positivos, si no, decae en los valores negativos.

Ejemplo:

Sea el modelo MA(1) de la forma $Z_t = (1 - 0,5B)a_t$. Según lo que vimos anteriormente, el ACF se debe cortar después del lag 1 y el PACF debe decaer exponencialmente. Generamos una muestra de este proceso y obtuvimos las siguiente estimaciones.

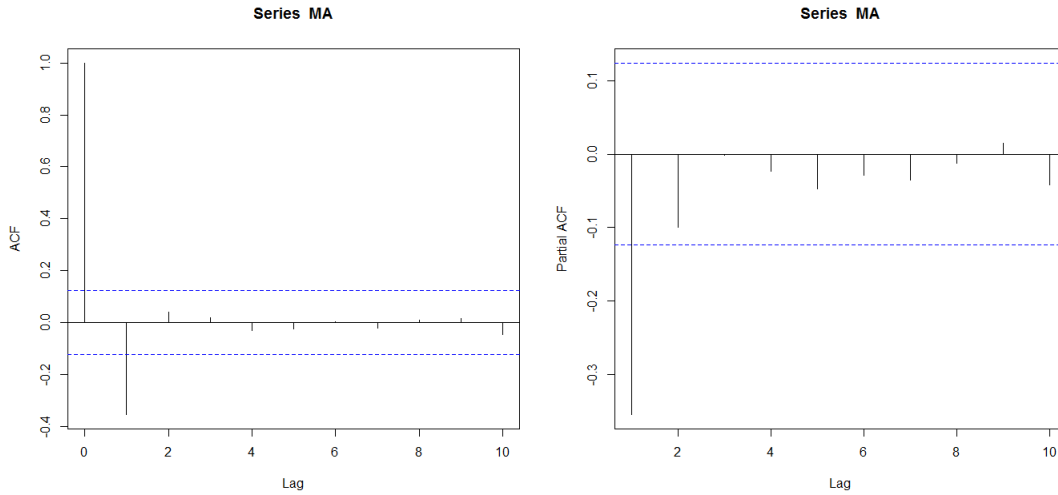


Figura 1: Proceso MA(1). Gráficos de ACF y PACF

Wold(1938) probó que cualquier proceso estacionario puede representarse univocamente como la suma de dos procesos mutuamente no correlacionados

$$Z_t = D_t + X_t$$

donde D_t es linealmente determinista y X_t es un proceso MA(∞). La parte D_t puede ser una función exacta del tiempo, en el caso más simple podría ser una constante.

Modelos AR

Un proceso se dirá autoregresivo de orden p o AR(p) si podemos representar a la serie de tiempo como:

$$\dot{Z}_t = \phi_1 \dot{Z}_{t-1} + \dots + \phi_p \dot{Z}_{t-p} + a_t. \quad (2.10)$$

Es decir, si el valor de la variable Z_t depende de sus valores pasados y de la innovación a tiempo t . Estos procesos sirven para describir fenómenos donde los eventos producen un efecto inmediato que solo duran un período corto de tiempo.

Si definimos el operador $\phi_p(B) = (1 - \phi_1 B - \dots - \phi_p B^p)$ podemos escribir el modelo AR(p) como:

$$\phi_p(B) \dot{Z}_t = a_t.$$

Es fácil ver que

$$\begin{aligned}\gamma_k &= \phi_1 \gamma_{k-1} + \dots + \phi_p \gamma_{k-p} \\ \rho_k &= \phi_1 \rho_{k-1} + \dots + \phi_p \rho_{k-p} \\ \psi_{kk} &= 0 \quad \forall j > p.\end{aligned}\tag{2.11}$$

Analizaremos en detalle el proceso AR(1), es decir,

$$\dot{Z}_t = \phi_1 \dot{Z}_{t-1} + a_t.$$

La función de autocovarianza se puede calcular de la siguiente forma:

$$\begin{aligned}E(\dot{Z}_{t-k} \dot{Z}_t) &= E(\phi_1 \dot{Z}_{t-k} \dot{Z}_{t-1}) + E(\dot{Z}_{t-k} a_t), \\ \gamma_k &= \phi_1 \gamma_{k-1}.\end{aligned}$$

y iterando y usando que $\rho_0 = 1$, obtenemos la función de autocorrelación

$$\rho_k = \phi_1 \rho_{k-1} = \phi_1^k \quad k \geq 1.\tag{2.12}$$

Observemos que si $|\phi_1| < 1$ y el proceso es estacionario, la ACF decae exponencialmente. Además, notemos que el PACF es de la forma:

$$\phi_{kk} = \begin{cases} \rho_1 = \phi_1 & \text{si } k = 1 \\ 0 & \text{si } k > 1 \end{cases}$$

Luego, el PACF de un proceso AR(1) presenta un pico con lag 1 y se corta posteriormente.

Ejemplo

Generamos el modelo AR(1) de la forma $(1 - 0,5B)Z_t = a_t$ y obtuvimos las siguientes funciones estimadas.

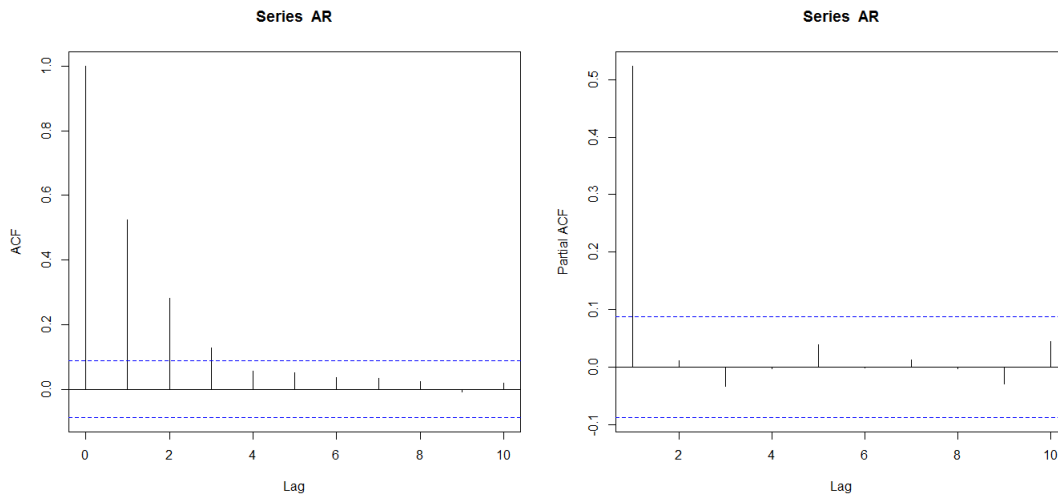


Figura 2: Proceso AR(1). Gráficos de ACF y PACF

Volviendo al modelo AR(p), notemos que podemos escribir a (2.11) como

$$\phi_p(B)\rho_k = 0$$

donde $\phi_p(B) = 1 - \phi_1 B - \dots - \phi_p B^p$ y B está operando sobre la autocorrelación. Si reescribimos la ecuación en la forma

$$\phi_p(B) = \prod_{i=1}^p (1 - G_i B)$$

obtenemos que para ciertos A_j

$$\rho_k = A_1 G_1^k + A_2 G_2^k + \dots + A_p G_p^k$$

Dado que el modelo es estacionario resulta $|G_i| < 1$. Luego si la raíz fuese real, el término $A_i G_i^k$ decaerá a 0 a medida que k crece. Es decir, la función de autocorrelación es una exponencial amortiguada. En el caso de que la raíz fuese compleja, tendríamos el término $d^k \sin(2\pi k f + F)$ en la función de autocorrelación, lo que lleva a un seno amortiguado. En general, la función de autocorrelación de un proceso autoregresivo será una mezcla entre un seno amortiguado y una exponencial amortiguada.

Por otro lado, observemos que existe una relación entre los procesos AR(p) y MA(q). En efecto, dada un proceso estacionario AR(p)

$$\phi_p(B)\dot{Z}_t = a_t,$$

podemos escribirlo de la siguiente forma,

$$\dot{Z}_t = \frac{1}{\phi_p(B)} a_t = \psi(B) a_t.$$

luego $\psi(B)\phi_p(B) = 1$. Esto implica que un proceso estacionario AR(p) finito lo podemos entender como un proceso MA(∞) y de manera equivalente, dado un proceso invertible MA(q),

$$\dot{Z}_t = \theta_q(B) a_t,$$

podemos escribirlo como

$$\pi(B)\dot{Z}_t = \frac{1}{\theta_q(B)} \dot{Z}_t = a_t,$$

así, un proceso invertible MA(q) puede ser explicado como un proceso AR(∞). Esta analogía se ve también en las funciones de autocorrelación y de autocorrelación parcial. Las autocorrelaciones de un AR(p) disminuyen y las autocorrelaciones parciales se cortan, mientras que las autocorrelaciones de un proceso MA(q) se cortan y las autocorrelaciones parciales disminuyen.

Modelos ARMA

La combinación de los procesos $AR(p)$ y $MA(q)$ da lugar a los procesos mixtos $ARMA(p,q)$ que se definen de la siguiente manera.

Definición 2.1.2. Diremos que Z_t sigue un modelo $ARMA(p,q)$ si

$$\phi_p(B)\dot{Z}_t = \theta_q(B)a_t,$$

donde $\phi_p(B) = 1 - \phi_1 B - \dots - \phi_p B^p$ y $\theta_q(B) = 1 - \theta_1 B - \dots - \theta_q B^q$.

Como discutimos anteriormente para los modelos AR y MA podemos ver que el proceso será invertible si las raíces de $\theta_q(B) = 0$ se encuentran fuera del círculo unitario y el proceso será estacionario si las raíces de $\phi_p(B) = 0$ toman valores también fuera del círculo unitario.

Asimismo notemos que un modelo ARMA puede escribirse como un modelo puramente AR o puramente MA, pues:

$$\dot{Z}_t = \frac{\theta_q(B)}{\phi_p(B)} a_t \quad \text{ó} \quad a_t = \frac{\phi_p(B)}{\theta_q(B)} \dot{Z}_t.$$

Por otro lado, a partir de cálculos sencillos podemos obtener que:

$$\begin{aligned} \gamma_k &= \phi_1 \gamma_{k-1} + \dots + \phi_p \gamma_{k-p}, \\ \rho_k &= \phi_1 \rho_{k-1} + \dots + \phi_p \rho_{k-p}. \end{aligned}$$

Notemos que la función de autocorrelación es similar a la del proceso $AR(p)$. Entonces, la función de autocorrelación de un proceso $ARMA(p,q)$ decae después del lag q como la de un proceso $AR(p)$. De manera análoga, como el ARMA contiene al proceso MA como caso particular, la función de autocorrelación parcial será una combinación de un seno amortiguado con una exponencial amortiguada.

A modo de ejemplo, trabajaremos con el proceso $ARMA(1,1)$

$$\dot{Z}_t = \phi_1 \dot{Z}_{t-1} + a_t - \theta_1 a_{t-1} \tag{2.13}$$

donde por las condiciones de invertibilidad y estacionaridad $|\phi_1| < 1$ y $|\theta_1| < 1$. Para obtener la función de autocovarianza, multiplicamos (2.13) por \dot{Z}_{t-k} a ambos lados y tomamos esperanza, por lo que obtenemos

$$\gamma_k = \phi_1 \gamma_{k-1} + E(\dot{Z}_{t-k} a_t) - \theta_1 E(\dot{Z}_{t-k} a_{t-1}). \tag{2.14}$$

En particular, si $k = 0$, por la simetría de γ_k obtenemos

$$\gamma_0 = \phi_1 \gamma_1 + E(\dot{Z}_t a_t) - \theta_1 E(\dot{Z}_t a_{t-1}),$$

luego usando que $E(\dot{Z}_t a_t) = \sigma_a^2$, multiplicando (2.13) por a_{t-1} y tomando esperanza, obtenemos que

$$\begin{aligned} E(\dot{Z}_t a_{t-1}) &= \phi_1 E(\dot{Z}_{t-1} a_{t-1}) + E(a_t a_{t-1}) - \theta_1 E(a_{t-1}^2) \\ &= (\phi_1 - \theta_1) \sigma_a^2. \end{aligned}$$

Por lo tanto,

$$\gamma_0 = \phi_1 \gamma_1 + \sigma_a^2 - \theta_1 (\phi_1 - \theta_1) \sigma_a^2. \quad (2.15)$$

Si ahora reemplazamos $k = 1$, tenemos que

$$\gamma_1 = \phi_1 \gamma_0 - \theta_1 \sigma_a^2. \quad (2.16)$$

Sustituyendo (2.16) en (2.15), obtenemos que

$$\begin{aligned} \gamma_0 &= \phi_1^2 \gamma_0 - \phi_1 \theta_1 \sigma_a^2 + \sigma_a^2 - \phi_1 \theta_1 \sigma_a^2 + \theta_1^2 \sigma_a^2, \\ \gamma_0 &= \frac{(1 + \theta_1^2 - 2\phi_1 \theta_1)}{(1 - \phi_1^2)} \sigma_a^2. \end{aligned}$$

Y por lo tanto γ_1 quedará

$$\begin{aligned} \gamma_1 &= \phi_1 \gamma_0 - \theta_1 \sigma_a^2, \\ &= \phi_1 \frac{(1 + \theta_1^2 - 2\phi_1 \theta_1)}{(1 - \phi_1^2)} \sigma_a^2 - \theta_1 \sigma_a^2, \\ &= \frac{(\phi_1 - \theta_1)(1 - \phi_1 \theta_1)}{(1 - \phi_1^2)}. \end{aligned}$$

Si $k \geq 2$, tenemos por (2.14) que

$$\gamma_k = \phi_1 \gamma_{k-1}.$$

Y así obtenemos finalmente el ACF de un ARMA(1,1)

$$\rho_k = \begin{cases} 1 & \text{si } k = 0 \\ \frac{(\phi_1 - \theta_1)(1 - \phi_1 \theta_1)}{(1 + \theta_1^2 - 2\phi_1 \theta_1)} & \text{si } k = 1 \\ \phi_1 \rho_{k-1} & \text{si } k \geq 2. \end{cases}$$

Para determinar el PACF de un proceso ARMA(1,1), basta notar que un proceso MA(1) es un caso particular. Entonces, el PACF del ARMA(1,1) también decaerá exponencialmente pero la forma en que lo hará esta directamente relacionada con los valores de ϕ_1 y θ_1 .

Ejemplo:

Si generamos un proceso ARMA(1,1) de la forma notada como en (2.13) con $\phi_1 = 0,5$ y $\theta_1 = 0,5$, obtenemos las siguientes funciones que reflejan lo estudiado anteriormente.

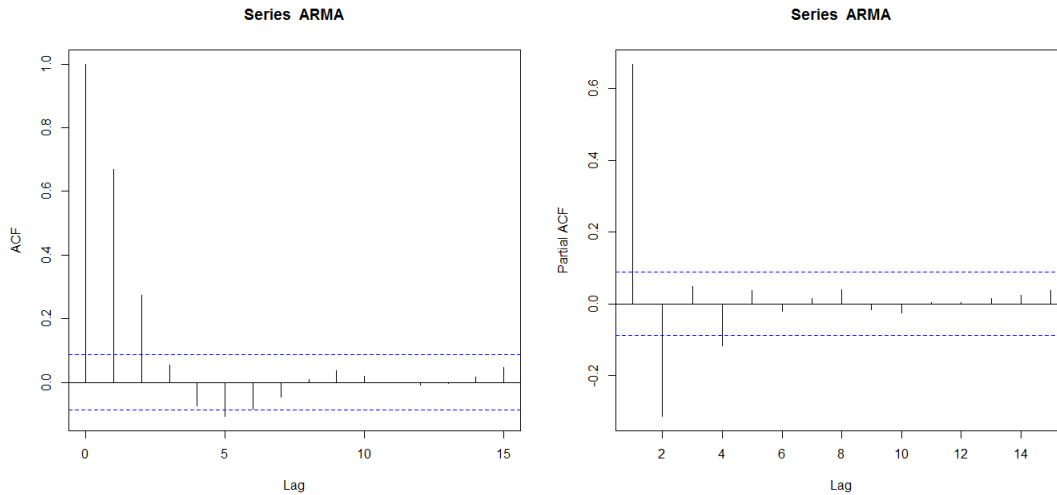


Figura 3: Proceso ARMA(1,1). Gráficos de ACF y PACF

2.2. Modelos de Series de Tiempo No Estacionarias

Los modelos de series de tiempo presentados en el capítulo anterior, suponen la existencia de la propiedad de estacionariedad; sin embargo, muchas de las series que observamos no cumplen con esta característica presentando una tendencia creciente a lo largo del tiempo o una clara inclinación a permanecer durante largos periodos por encima o por debajo de su media. Como veremos en el presente capítulo, la no estacionariedad en las series de tiempo se puede deber a la presencia de una tendencia determinística, al problema de las raíces unitarias o a la presencia de una varianza heterocedástica. Dependiendo del caso en que nos encontremos, la no estacionariedad tendrá determinadas implicancias teóricas y estadísticas, así como determinados procedimientos para transformar las series en estacionarias.

2.2.1. No estacionariedad en media

Como mencionamos anteriormente las series pueden presentar una tendencia a través del tiempo. Estas series con tendencia, claramente no son estacionarias, porque su valor medio cambia con del tiempo. Esta tendencia puede ser determinística. Es decir, la media del proceso no estacionario puede ser determinada mediante un función que depende del tiempo. Por ejemplo, la función μ_t puede seguir una tendencia lineal en cuyo caso se podría modelar a partir del siguiente modelo $\mu_t = \alpha_0 + \alpha_1 t$ y luego

utilizar un modelo para la serie temporal del tipo:

$$Z_t = \alpha_0 + \alpha_1 t + a_t \quad (2.17)$$

donde a_t es ruido blanco. En general se puede proponer un polinomio de grado k para ajustar la media del proceso o algún tipo de tendencia representada por ejemplo por senos y cosenos.

Otro tipo de tendencia en media es la tendencia estocástica. Aunque una serie sea estacionaria, algunas partes de esta se comportan de igual manera salvo una diferencia a nivel local. Tomemos por ejemplo un paseo aleatorio, es decir, un proceso estocástico de la forma

$$Z_t = \mu + Z_{t-1} + a_t,$$

donde μ es un parámetro y $a_t \sim N(0, \sigma_a^2)$ i.i.d.. Aplicando diferencias obtenemos que

$$\Delta Z_t = Z_t - Z_{t-1} = \mu + a_t.$$

Esto nos muestra que ΔZ_t sigue un modelo ARMA(0,0) y que es estacionario.

Esta forma de convertir una serie no estacionaria en estacionaria lleva a la siguiente definición:

Definición 2.2.1. Dada una diferenciación de nivel d , se denomina al siguiente modelo *ARIMA*(p, d, q):

$$\phi_p(B)(1 - B)^d Z_t = \theta_0 + \theta_q(B)a_t \quad (2.18)$$

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

2.2.2. No estacionaridad en varianza y autocovarianza

Un proceso que no es estacionario en esperanza, tampoco lo será en varianza y autocovarianza. Un ejemplo sencillo de esto son los modelos ARIMA introducidos anteriormente, no estacionario en varianza y autocovarianza. Ilustremos esto en el caso particular del modelo ARIMA(0,1,1), es decir

$$Z_t = Z_{t-1} + a_t - \theta a_{t-1}$$

o

$$(1 - B)Z_t = (1 - \theta B)a_t \quad (2.19)$$

Si fijamos n_0 , dado $t > n_0$, podemos hacer una recursión en (2.19) y obtenemos:

$$Z_t = Z_{n_0} + a_t + (1 - \theta)a_{t-1} + \dots + (1 - \theta)a_{n_0+1} - \theta a_{n_0}$$

Haciendo lo mismo para $t - k > n_0$ tenemos,

$$Z_{t-k} = Z_{n_0} + a_{t-k} + (1 - \theta)a_{t-k-1} + \dots + (1 - \theta)a_{n_0+1} - \theta a_{n_0}$$

Luego, calculamos su varianza, covarianza y correlación:

$$\begin{aligned} V(Z_t) &= \sigma_a^2(1 + (t - n_0 - 1)(1 - \theta)^2), \\ V(Z_{t-k}) &= \sigma_a^2(1 + (t - k - n_0 - 1)(1 - \theta)^2), \\ \text{cov}(Z_{t-k}, Z_t) &= \sigma_a^2((1 - \theta) + (t - k - n_0 - 1)(1 - \theta)^2), \\ \text{corr}(Z_{t-k}, Z_t) &= \frac{(1 - \theta) + (t - k - n_0 - 1)(1 - \theta)^2}{\sqrt{(1 + (t - n_0 - 1)(1 - \theta)^2)(1 + (t - k - n_0 - 1)(1 - \theta)^2)}}. \end{aligned}$$

De estas ecuaciones obtenemos lo siguiente:

- $V(Z_t)$ depende de t .
- La varianza no está acotada cuando $t \rightarrow \infty$.
- Si $t > n_0$, $\text{corr}(Z_{t-k}, Z_t) \sim 1$.

Si bien hay series no estacionarias en esperanza y por lo tanto en varianza, existen series que son estacionarias en la esperanza pero no estacionarias en la varianza. Este tipo de estacionaridad puede ser tratada con transformaciones de estabilización de varianza. Más precisamente, aplicaremos un transformación sobre la serie con el objetivo de que su varianza sea constante. Es decir, buscaremos T tal que $V(T(Z_t))$ sea constante.

En particular, vamos a considerar el caso en que la varianza cambie en función a cambios en la esperanza, es decir: $V(Z_t) = cf(\mu_t)$.

Para ello aproximamos T por su polinomio de Taylor centrado en μ_t . Entonces,

$$T(Z_t) \simeq T(\mu_t) + T'(\mu_t)(Z_t - \mu_t),$$

y por lo tanto

$$V(T(Z_t)) = [T'(\mu_t)]^2 V(Z_t) = c[T'(\mu_t)]^2 f(\mu_t).$$

De esta forma, elijeremos T tal que $T'(\mu_t) = \frac{1}{\sqrt{f(\mu_t)}}$, es decir:

$$T(\mu_t) = \int^{\mu_t} \frac{1}{\sqrt{f(s)}} ds.$$

Box y Cox (1964), propusieron usar transformaciones de potencias del tipo

$$T(Z_t) = \frac{Z_t^\lambda - 1}{\lambda},$$

donde el parámetro λ puede ser agregado al modelo y estimado por el método de cuadrados mínimos. La siguiente tabla muestra valores de λ comunes y sus respectivas transformaciones.

Valores	Transformación
-1	$\frac{1}{Z_t}$
-0.5	$\frac{1}{\sqrt{Z_t}}$
0	$\log(Z_t)$
0.5	$\sqrt{Z_t}$

Es importante aclarar que estas transformaciones se aplican en series positivas y que deben utilizarse antes del método de las diferencias. Esto no introduce restricción alguna ya que puede sumarse una constante a la serie en orden de hacerla positiva.

2.3. Series de Tiempo estacionales

Otro factor importante que pueden presentar las series de tiempo es la estacionalidad. Es decir, ciclos u oscilaciones estrictamente periódicas debido principalmente a las fluctuaciones en diferentes períodos del año. La presencia de esta componente se explica por la existencia de estaciones y su impacto sobre la actividad económica ya sea en la producción agropecuaria, el turismo, las costumbres como por ejemplo el fin de año que es estrictamente cultural o simplemente por procesos físicos tal es el caso de la temperatura o las lluvias.

Diremos que una serie exhibe un comportamiento estacional de período s si existen similitudes en la serie cada s intervalos. Así, el operador de retardo B^s jugará un papel importante en la explicación de la serie y la identificación del modelo.

Conociendo la periodicidad de los datos, será conveniente presentar la información utilizando la tabla de Buys-Ballot. Esta consistirá en una tabla de doble entrada con s columnas (o sea, tantas como sea la periodicidad). De esta manera uno podrá analizar los datos por columnas (corresponden a los datos $Z_t, Z_{t+s}, Z_{t+2s}, \dots$) o por filas (corresponden a los datos $Z_t, Z_{t+1}, \dots, Z_{t+s-1}$ para la primera fila).

Si tomáramos por ejemplo datos mensuales a través de los años, tendríamos los datos por años en las filas y por meses en las columnas. En este caso, si tuviésemos el dato de Agosto, este dependería de las observaciones de los meses anteriores como también de los pasados Agostos.

Luego, para relacionar observaciones de años distintos (los Agostos) podemos proponer un modelo de la forma

$$\Phi_P(B^s)(1 - B^s)^D Z_t = \Theta_Q(B^s)\alpha_t \quad (2.20)$$

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

$$\phi(B)\alpha_t = \theta(B)a_t \quad (2.21)$$

donde los a_t son ruido blanco. Finalmente, si sustituimos (2.21) en (2.20), utilizando las nociones de invertibilidad de los procesos obtenemos la fórmula general del modelo multiplicativo $\text{ARIMA}(p,d,q) \times \text{ARIMA}(P,D,Q)_s$

$$\Phi_P(B^s)\phi_p(B)(1-B)^d(1-B^s)^D\dot{Z}_t = \Theta_Q(B^s)\theta_q(B)a_t \quad (2.22)$$

Este análisis puede ser aplicado también para series con tres o más componentes periódicas.

Para identificar el modelo, será necesario analizar la función de autocorrelación de la serie completa, la serie correspondiente a $B^s(Z_t)$ y la correspondiente a $B(Z_t)$.

Capítulo 3

Identificación, Estimación y Pronóstico

Los modelos estadísticos en general requieren de tres etapas para su construcción: Identificación, Estimación y Validación. Si bien las diferentes etapas son sucesivas, de acuerdo al resultado de cada etapa puede surgir una retro alimentación. Donde por ejemplo luego de la estimación puede darse la necesidad de reespecificar el modelo, o en el proceso de validación puede dar como resultado un modelo insatisfactorio.

3.1. Identificación del Modelo

En la etapa de identificación el objetivo es definir si es necesario aplicar algún tipo de transformación, si el modelo debe incluir una media no nula y finalmente decidir cual de los modelos es el indicado y con que parámetros.

Comencemos identificando un modelo ARIMA(p,d,q). Es decir:

$$(1 - \phi_1 B - \dots - \phi_p B^p)(1 - B)^d Z_t = (1 - \theta_1 B - \dots - \theta_q B^q) a_t$$

Hemos visto que para un proceso ARMA(p,q), $\phi_p(B)Z_t = \theta_q(B)a_t$, la función de autocorrelación satisface

$$\phi_p(B)\rho_k = 0 \quad k > q$$

Además, si $\phi_p(B) = \prod_{i=1}^p (1 - G_i B)$, la solución de esta ecuación para la k -ésima autocorrelación es

$$\rho_k = A_1 G_1^k + A_2 G_2^k + \dots + A_p G_p^k \quad k > q - p \quad (3.1)$$

Bajo el supuesto de que el modelo es estacionario, los ceros de $\phi_p(B)$ deben estar fuera del círculo unitario. Esto implica que las raíces G_1, G_2, \dots, G_P son unitarias. Observando (3.1), tendremos que la función de autocorrelación decae rápidamente para valores grandes de k . Luego, si la función de autocorrelación no decae rápidamente, es signo de que es necesario cierto grado de diferenciación.

Una vez determinado el orden de diferenciación, se procederá a determinar los valores p y q del modelo.

Para determinar el modelo, utilizaremos principalmente las funciones de autocorrelación y de autocorrelación parcial. Los pasos a seguir serán los siguientes:

- 1) Graficar la serie y elegir las transformaciones apropiadas. Se suele aplicar la transformación de estabilización de varianza y usarla como serie original. Determinar si la serie necesita ser diferenciada para lograr estacionariedad.
- 2) Examinar las ACF y PACF. En el caso de que la ACF decaiga lento, estaría implicando que el proceso es no estacionario en Z_t pero posiblemente estacionario en alguna diferencia de Z_t . Así, deberíamos aplicar las diferencias correspondientes.
- 3) Examinar las ACF y PACF de la serie diferenciada con el objetivo de determinar los valores de p y q correspondientes. Para ello, utilizamos la siguiente tabla:

Proceso	ACF	PACF
AR(p)	Decae exponencialmente como una onda sinusoidal amortiguada	Se corta después de $k=p$
MA(q)	Se corta después de $k=q$	Decae exponencialmente como una onda sinusoidal amortiguada
ARMA(p,q)	Se corta después de $k=q-p$	Se corta después de $k=p-q$

Para el caso de las series estacionales, la identificación del modelo constará de dos partes. La primera, en la que identificaremos la componente periódica estacional de la series y la segunda donde se determinaran los parámetros no estacionales de ella. En efecto, los pasos a seguir serán los siguientes:

- 1) Encontrar d y D tal que

$$Y_t = (1 - B)^d (1 - B^s)^D Z_t$$

pueda considerarse como un proceso estacionario. Es decir, calculo las diferencias por ejemplo con lag=1 y lag=12

- 2) Examinar las ACF y PACF de Y_t en los lags $s, 2s, 3s, \dots$. Determinar P y Q para que el modelo $\text{ARMA}(P, Q)$ sea compatible con las funciones de autocorrelación y autocorrelación parcial observadas.
- 3) Examinar las ACF y PACF de Y_t en los lags $1, 2, \dots, s-1$. Determinar p y q para que el modelo $\text{ARMA}(p, q)$ sea compatible con las funciones de autocorrelación y autocorrelación parcial observadas.

3.2. Estimación de Parámetros

Luego de considerar el modelo a utilizar es necesario estimar los parámetros. Si elegimos un $\text{ARMA}(p, q)$:

$$\dot{Z}_t = \phi_1 \dot{Z}_{t-1} + \dots + \phi_p \dot{Z}_{t-p} + a_t - \theta_1 a_{t-1} - \dots - \theta_q a_{t-q}$$

Deberemos estimar ϕ, μ, θ y σ_a^2 . Utilizaremos dos posibles métodos de estimación; el de los momentos o el de máxima verosimilitud. Además comentaremos la aplicación del estimador de mínimos cuadrados en este contexto.

- Método de Momentos:

El estimador más simple de momentos para μ , resulta $\hat{\mu} = \bar{Z}$; para estimar ϕ usaremos que $\rho_k = \phi_1 \rho_{k-1} + \dots + \phi_p \rho_{k-p}$. Así, obtenemos las ecuaciones de Yule-Walker. Si reemplazamos ρ_k por

$$\hat{\rho}_k = \frac{\sum_{t=1}^{n-k} (Z_t - \bar{Z})(Z_{t+k} - \bar{Z})}{\sum_{t=1}^n (Z_t - \bar{Z})^2}$$

y resolviendo el sistema podemos obtener $\hat{\phi}_1, \dots, \hat{\phi}_p$.

Una vez obtenidos los estimadores de ϕ , calculamos

$$\gamma_0 = E(\dot{Z}_t \dot{Z}_t) = E((\dot{Z}_t)(\phi_1 \dot{Z}_{t-1} + \dots + \phi_p \dot{Z}_{t-p} + a_t - \theta_1 a_{t-1} - \dots - \theta_q a_{t-q})) = \phi_1 \gamma_1 + \dots + \phi_p \gamma_p + \sigma_a^2$$

y de esta forma estimamos σ_a^2

$$\hat{\sigma}_a^2 = \hat{\gamma}_0 (1 - \hat{\phi}_1 \hat{\rho}_1 - \dots - \hat{\phi}_p \hat{\rho}_p).$$

Para calcular los estimadores θ veamos como ejemplo un proceso $\text{MA}(1)$, calculamos

$$\rho_1 = \frac{\gamma_1}{\gamma_0} = \frac{-\theta_1}{1 + \theta_1^2}$$

Si reemplazamos ρ_k por $\hat{\rho}_k$, resolvemos y obtenemos $\hat{\theta}_1$

- Estimadores de Máxima Verosimilitud:
Consideremos un modelo ARMA(p, q)

$$\dot{Z}_t = \phi_1 \dot{Z}_{t-1} + \dots + \phi_p \dot{Z}_{t-p} + a_t - \theta_1 a_{t-1} - \dots - \theta_q a_{t-q}$$

como $a_t \sim N(0, \sigma_a^2)$, si $a = (a_1, \dots, a_n)$ tenemos

$$P(a|\theta, \mu, \phi, \sigma_a^2) = (2\pi\sigma_a^2)^{-n/2} e^{-1/\sigma_a^2 \sum_{t=1}^n (a_t)^2}.$$

Entonces, si escribimos a a_t de la siguiente forma

$$a_t = \theta_1 a_{t-1} + \dots + \theta_q a_{t-q} + \dot{Z}_t - \phi_1 \dot{Z}_{t-1} - \dots - \phi_p \dot{Z}_{t-p}$$

Sea entonces $Z = (Z_1, Z_2, \dots, Z_n)$ y $Z^* = (Z_{1-p}, \dots, Z_{-1}, Z_0)$ las condiciones iniciales conocidas, definimos la función condicional de máxima verosimilitud como

$$\ln L(\phi, \mu, \theta, \sigma_a^2) = -n/2 \ln 2\pi\sigma_a^2 - \frac{S(\phi, \mu, \theta)}{2\sigma_a^2},$$

donde

$$S(\phi, \mu, \theta) = \sum_{t=1}^n a_t^2(\phi, \mu, \theta|Z^*, a^*, Z). \quad (3.2)$$

Las cantidades que maximizan la función condicional de máxima verosimilitud en (ϕ, μ, θ) serán los estimadores de máxima verosimilitud condicional. Notar que basta minimizar (3.2) para encontrar estos estimadores.

Luego de estimar los parámetros, computamos $\hat{\sigma}_a^2 = \frac{S(\hat{\phi}, \hat{\mu}, \hat{\theta})}{d.f}$, donde *d.f* corresponde a la cantidad de términos usados en S menos el número de parámetros estimados.

- Estimación por mínimos cuadrados (OLS):

En general, dada una regresión lineal simple, $Z_t = \phi X_t + e_t$, con $E(e_t) = 0$, $E(e_t^2) = \sigma_e^2$, $E(e_t e_k) = 0$ para $t \neq k$ y $E(X_t e_t) = 0$, tenemos que el estimador de cuadrados mínimos ordinarios es $\hat{\phi} = \frac{\sum_{t=1}^n X_t Z_t}{\sum_{t=1}^n X_t^2}$ que es consistente e insesgado.

Si ahora considero $Z_t = \phi Z_{t-1} + e_t$, tenemos que:

$$\hat{\phi} = \frac{\sum_{t=2}^n Z_{t-1} Z_t}{\sum_{t=2}^n Z_{t-1}^2} = \frac{\sum_{t=2}^n Z_{t-1} (\phi Z_{t-1} + e_t)}{\sum_{t=2}^n Z_{t-1}^2} = \phi + \frac{\sum_{t=2}^n Z_{t-1} e_t}{\sum_{t=2}^n Z_{t-1}^2}$$

Para determinar si el estimador sigue siendo consistente e insesgado debemos analizar la naturaleza del error e_t . Veamos las siguientes situaciones:

- Si e_t es tal que $E(e_t) = 0$ y $V(e_t) = \sigma_a^2$, Z_t es un AR(1). Así, $\hat{\phi} = \hat{\rho}_1$ y , si además $|\phi| < 1$ tenemos que la función de autocorrelación es sumable por lo que $\hat{\phi}$ es consistente.
- Si $e_t = (1 - \theta B)a_t$, tenemos que e_t es un proceso MA(1). Luego, Z_t es un ARMA(1,1) con $Z_t = \phi Z_{t-1} + a_t - \theta a_{t-1}$. Notemos que si tomo esperanza, obtengo que $E(Z_{t-1}e_t) = E(Z_{t-1}(a_t - \theta a_{t-1})) = -\theta\sigma_a^2$; así estaría violando las hipótesis $E(e_t e_k) = 0$ y $E(X_t e_t) = 0$. Como $\hat{\phi} \simeq \hat{\rho}_1$ y , para un ARMA(1,1), se tiene que $\rho_1 = \frac{(\phi - \theta)(1 - \phi\theta)}{1 + \theta^2 - 2\phi\theta} \neq \phi$ esto indica que el estimador no es consistente.

3.3. Diagnóstico y Selección del Modelo

Luego de estimar los parámetros, será necesario realizar el chequeo del modelo. Esto consistirá principalmente en determinar si los a_t del modelo resultan realmente ruido blanco, esto se llevará acabo a través de los residuos \hat{a}_t .

Inicialmente podremos calcular $\frac{\hat{a}_t}{\hat{\sigma}_a}$ y analizar si tiene distribución $N(0,1)$ realizando un histograma o usando test de normalidad. Para determinar si tienen varianza constante, observaremos los ACF y PACF de los residuos para ver si forman un patrón y se encuentran o no correlacionados.

Supongamos ahora que al chequear el modelo, notamos por ejemplo que AR(1) produce un MA(1). Es decir:

$$(1 - \phi_1 B)(Z_t - \mu) = b_t$$

donde b_t corresponde a un MA(1).

En ese caso, propondremos un ARMA(1,1) y volveremos a estimar y luego chequear el modelo, de esta manera iteraremos este procedimiento hasta obtener ruido blanco.

Otra forma de determinar que modelo elegir es usando un criterio de selección de modelos. Existen muchos de ellos, sin embargo, el criterio de selección de Akaike (AIC) es el más utilizado.

Rigurosamente, definimos AIC(M) como

$$\text{AIC}(M) = -2 \log(L) + 2M \quad (3.3)$$

donde L corresponde a la verosimilitud y M es la cantidad de parámetros del modelo.

En particular, para un modelo ARMA, tenemos que

$$\log(L) = \frac{-n}{2} \log 2\pi\sigma_a^2 - \frac{S(\phi, \mu, \theta)}{2\sigma_a^2}$$

donde S es el definido en (3.2). Si maximizamos respecto a ϕ, μ, θ y σ_a^2 obtenemos

$$\log \hat{L} = \frac{-n}{2} \log \hat{\sigma}_a^2 - \frac{n}{2} (1 + \log 2\pi)^2.$$

Finalmente,

$$\text{AIC}(M) = n \log \hat{\sigma}_a^2 + 2M,$$

el criterio consiste en seleccionar aquel modelo que tenga mínima condición de Akaike.

Otro criterio a utilizar es el Criterio de Información Bayesiano (BIC). Este criterio penaliza también la cantidad de parámetros introducidos al modelo pero, a diferencia de Akaike, el BIC proviene de un marco bayesiano.

El BIC se define como

$$\text{BIC} = -2 \log(L) + M \log n,$$

notando que solo se diferencia del AIC en que el segundo término aparece n . Se elegirán los modelos que minimicen BIC. Desde una perspectiva bayesiana, BIC se utiliza para encontrar el modelo mas probable dados los datos.

3.3.1. Algunas herramientas de validación

En esta sección enumeraremos todos aquellos criterios o tests que utilizaremos para determinar y diagnosticar el modelo elegido

Test de Raiz Unitaria:

Este test se utiliza para determinar si la serie es o no estacionaria. Como vimos anteriormente, esto quedará determinado por el valor de las raíces.

Por ejemplo, para un modelo AR(1) de la forma $Z_t = \phi Z_{t-1} + a_t$, planteo el test

$$H_0 : \phi = 1 \quad \text{vs} \quad H_1 : \phi < 1$$

El test está basado en $\hat{\phi}$ un estimador de ϕ y se rechazará H_0 si $\hat{\phi} - 1 < 0$. Como el estimador de cuadrados mínimos de ϕ es

$$\hat{\phi} = \frac{\sum_{t=1}^n Z_{t-1} Z_t}{\sum_{t=1}^n Z_{t-1}^2}$$

se puede probar que bajo H_0

$$n(\hat{\phi} - 1) = \frac{n^{-1} \sum_{t=1}^n Z_{t-1} a_t}{n^{-2} \sum_{t=1}^n Z_{t-1}^2} \xrightarrow{\mathcal{D}} \frac{1/2[w(1)^2 - 1]}{\int_0^1 [w(x)]^2 dx}$$

donde $w(t)$ es un proceso browniano, es decir,

- $w(0)=0$
- $w(t) \sim N(0,t)$ para cada t
- $[w(t_2) - w(t_1)]$ y $[w(t_4) - w(t_3)]$ son independientes para cada (t_1, t_2) y (t_3, t_4)

En general, el estadístico a utilizar bajo H_0 es

$$T = \frac{\hat{\phi} - 1}{S_{\hat{\phi}}}$$

que converge en distribución a

$$\frac{1/2[(w(1))^2] - 1}{[\int_0^1 (w(x))^2 dx]^{1/2}}$$

y rechazaremos H_0 para valores pequeños de T .

Test de Shapiro-Wilk:

El test de Shapiro-Wilk testea la hipótesis nula de que X_1, \dots, X_n tienen una distribución normal. En este caso, se utilizaremos este test para observar la normalidad de los residuos. El estadístico utilizado es

$$W = \frac{(\sum_{i=1}^n a_i X_i)^2}{\sum_{i=1}^n (X_i - \bar{X}_i)^2}$$

donde $a = (a_1, \dots, a_n) = \frac{\mu V^{-1}}{(\mu V^{-1} V^{-1} \mu)^{1/2}}$, $E(X_i) = m_i$ y V es la matrix de covarianzas. Para este estadístico, rechazaremos la hipótesis nula para valores pequeños de W .

Test de Kolmogorov-Smirnov:

Este test sirve para comparar una muestra con una distribución dada. La hipótesis nula será que la muestra sigue esa distribución. El test está basado en la diferencia entre la distribución teórica y la hipotética. Dados n datos X_1, \dots, X_n , se define el test estadístico

$$T = \sup |F^*(x) - F_n(x)|, \quad (3.4)$$

donde F^* es la distribución hipotética y F_n es la distribución empírica basada en la muestra. La distribución del T , esta tabulada en tabla, luego si T excede el $1 - \alpha$ cuantil dado por la tabla de cuantiles para el Test de Kolmogorov, se rechaza la hipótesis nula con un nivel de confianza α .

Test de Dickey-Fuller

Este test chequea la hipótesis nula de que existe una raíz unitaria en una serie de tiempo. No entraremos en detalle sobre el estadístico que se construye en este caso. A modo de ejemplo si consideramos un proceso AR(1),

$$y_t = \rho y_{t-1} + \mu_t,$$

este tendrá una raíz unitaria si $\rho = 1$. Notemos que el proceso puede ser reescrito como

$$By_t = (\rho - 1)y_{t-1} + \mu_t = \delta y_{t-1} + \mu_t.$$

Luego el test se concentrará en testear si δ es 0 para comprobar si hay o no una raíz unitaria.

3.4. Forecasting

Luego de identificar el modelo y confirmar los supuestos asumidos, llega el momento de pronosticar. Tomaremos como ejemplo un modelo ARIMA(p, d, q) con $d = 0$ ó $d \neq 0$. Es decir, $\phi(B)(1 - B)^d Z_t = \theta(B)a_t$ donde $\phi(B)$ es el operador AR y $\theta(B)$ es el operador MA.

Reescribamos el modelo a tiempo $t + l$ en su representación AR.

$$\pi(B)Z_{t+l} = a_{t+l}$$

donde $\pi(B) = 1 - \sum_{j=1}^{\infty} \pi_j B^j = \frac{\phi(B)(1-B)^d}{\theta(B)}$ o equivalentemente

$$Z_{t+l} = \sum_{j \geq 1} \pi_j Z_{t+l-j} + a_{t+l}.$$

Si aplicamos a ambos lados de la ecuación el operador $1 + \psi_1 B + \dots + \psi_{l-1} B^{l-1}$:

$$0 = \sum_{j \geq 0} \sum_{k=0}^{l-1} \pi_j \psi_k Z_{t+l-j-k} + \sum_{k=0}^{l-1} \psi_k a_{t+l-k}$$

donde $\pi_0 = -1$ y $\psi_0 = 1$. Si elegimos ψ tal que $\sum_{i=0}^m \pi_{m-i} \psi_i = 0$ y llamamos $\pi_j^{(l)} = \sum_{i=0}^{l-j} \pi_{l-1+j-i} \psi_i$ es fácil ver

$$Z_{t+l} = \sum_{j \geq 1} \pi_j^{(l)} Z_{t-j+1} + \sum_{i=0}^{l-1} \psi_i a_{t+l-i}$$

Entonces tomando esperanza condicional

$$\widehat{Z}_n(l) = E(Z_{n+l}|Z_t, t \leq n) = \sum_{j \geq 1} \pi_j^{(l)} Z_{n-j+1}$$

pues $E(a_{n+j}|Z_t, t \leq n) = 0$. Así obtenemos que el error del forecast es:

$$e_n(l) = Z_{n+l} - \widehat{Z}_n(l) = \sum_{j=0}^{l-1} \psi_j a_{n+l-j} \quad (3.5)$$

Cuando actualicemos el forecast, obtendremos un nuevo error. Si el error obtenido viene dado por (3.5), tenemos que en particular $e_n(1) = a_{n+1}$. Entonces, $Z_n - \widehat{Z}_{n-1}(1) = a_n$. Usando (3.5) obtenemos que $e_{n-1}(l+1) = e_n(l) + \psi_l a_n$ y $e_n(l) = Z_{n+l} - \widehat{Z}_n(l)$. Luego,

$$\widehat{Z}_n(l) = \widehat{Z}_{n-1}(l+1) + \psi_l (Z_n - \widehat{Z}_{n-1}(1))$$

Lo que es lo mismo que:

$$Z_{n+1}(l) = \widehat{Z}_n(l+1) + \psi_l (Z_{n+1} - \widehat{Z}_n(1)). \quad (3.6)$$

En definitiva, el update del forecast se obtiene sumando al forecast anterior un múltiplo del error del pronóstico anterior. Además, se puede obtener un intervalo de confianza de nivel $1 - \alpha$ dado por

$$\widehat{Z}_t(l) \pm N_{\alpha/2} [1 + \sum_{j=1}^{i-1} \psi_j^2]^{1/2} \sigma_a^2,$$

donde $N_{\alpha/2}$ es tal que $P(N > N_{\alpha/2}) = \alpha/2$ con $N \sim N(0, 1)$.

Ejemplo: Sea un proceso AR(1) de la forma

$$(1 - \phi B)(Z_t - \mu) = a_t$$

donde $\phi = 0,6$, $\mu = 9$ y $\sigma_a^2 = 0,1$, supongamos que tenemos las siguientes observaciones $Z_{97} = 9,6$, $Z_{98} = 9$, $Z_{99} = 9$, $Z_{100} = 8,9$ y buscamos el forecast de Z_{101} y Z_{102} con un nivel de confianza del 95%. La forma en la que procederemos será la siguiente:

1) Si el modelo lo escribimos como

$$Z_t - \mu = \phi(Z_{t-1} - \mu) + a_t,$$

luego $\pi_1 = \phi$ y por lo tanto $\psi_1 = \phi$ entonces el forecast correspondiente vendrá dado por

$$\hat{Z}_t(l) = \mu + \phi(\hat{Z}_t(l-1) - \mu).$$

Así, obtenemos que

$$\begin{aligned}\hat{Z}_{100}(1) &= 9 + 0,6(8,9 - 9) = 8,94, \\ \hat{Z}_{100}(2) &= 9 + 0,6^2(8,9 - 9) = 8,964.\end{aligned}$$

2) Para computar los límites de confianza del forecast del 95 %, aplicamos (3.4). Así, obtenemos que para Z_{101} los valores son

$$8,94 \pm 1,96\sqrt{0,1},$$

y para Z_{102} son

$$8,964 \pm 1,96\sqrt{1 + (0,6)^2}\sqrt{0,1}.$$

3) Supongamos ahora que la observación a $t = 101$ resulta ser 8,8. Usando (3.6), actualizamos el forecast y obtenemos:

$$\hat{Z}_{101}(1) = \hat{Z}_{100}(2) + \phi[Z_{101} - \hat{Z}_{100}(1)] = 8,964 + 0,6(8,8 - 8,94) = 8,88.$$

En definitiva, estos serían los pasos a seguir a la hora de utilizar el modelo propuesto para pronosticar. Notemos que tomamos ϕ , μ y σ_a^2 como los valores poblacionales, sin embargo cuando trabajamos con datos reales el mismo procedimiento debería hacerse con los estimadores de estos.

3.5. Modelos RegARIMA

Con el objetivo de poder pronosticar, hemos estudiado modelos que explican la progresión de una variable en función de la variación de la misma. Esta forma de modelar no toma en cuenta los factores exógenos. Así, en el caso de pronosticar la demanda de una ruta, los modelos ARIMA no se ven afectados explícitamente por la variación de la oferta, la tarifa, el contexto económico, etc. Con este objetivo, introducimos los modelos RegARIMA. Estos consisten en modelar la variable dependiente

con una regresión lineal y, los residuos obtenidos, con un modelo ARIMA. En efecto, si y_t es la variable que queremos pronosticar,

$$y_t = \beta x_t + e_t,$$

donde x_t es la variable independiente y los errores e_t satisfacen

$$(1 - B)^d \phi(B)e_t = \theta(B)u_t,$$

donde d es el orden de diferenciación, ϕ es el operador AR de orden p y θ es el operador MA de orden q .

Una posible estimación para $(\beta, \phi_1, \dots, \phi_p, \theta_1, \dots, \theta_q)$ son los estimadores de máxima verosimilitud (ML) con errores normales. El cómputo de estos errores ya fue estudiado por Pesaran (1973), Pagan and Nicholls (1976), Harvey and Phillips (1979) y Otto, Bell y Burman (1987).

Otra opción consiste en un enfoque de cuadrados mínimos condicionados, el cual resulta en una aproximación a los estimadores de máxima verosimilitud exactos. En el caso de que los e_t sigan un proceso autoregresivo dado por $(1 - B)^d \phi(B)e_t = u_t$, este estimador será el de mínimos cuadrados.

Capítulo 4

Analisis de datos reales

El objetivo de este capítulo es aplicar la teoría presentada anteriormente utilizando datos reales. Las series a modelar corresponden a la demanda de pasajeros de ciertas rutas aéreas en el período enero de 2005 a mayo de 2012. Llamaremos demanda a los pasajeros que efectivamente volaron esa ruta en un determinado mes. El objetivo de generar estos modelos consiste en obtener funciones de demanda para poder realizar un planeamiento mas preciso de la red. Decisiones tales como inversión en equipos, aumento o disminución de frecuencias e ingreso a determinados mercados pueden ser sustentadas usando las predicciones obtenidas por estos modelos.

Como mencionamos anteriormente, utilizaremos los modelos ARIMA para la serie de datos correspondiente al período 01/05- 05/12 y ademas consideraremos las siguientes variables independientes ajustando los modelos RegARIMA

- 1) AKOS: Los akos son una métrica que se utiliza en la industria aeronáutica. Corresponde a Asientos por Kilómetro ofrecido; es decir, al producto de la oferta en asientos por los kilometros que tiene la ruta. Mide la oferta.
- 2) EMAE: El emae es el Estimador Mensual de Actividad Económica. Es un índice que realiza el INDEC con el objetivo de medir la actividad económica.
- 3) IPC: El ipc es el Índice de Precios al Consumidor. Es el índice que realiza el INDEC con el objetivo de medir la inflación.
- 4) ICC: El icc es el Índice de Confianza del Consumidor. Este índice es elaborado por la Universidad Torcuato Di Tella y es un indicador que *”...permite anticipar los cambios de tendencia de la actividad económica a lo largo del ciclo, a través de sus puntos de giro. El índice se compone de diez series seleccionadas en base a su conformidad, consistencia temporal, racionalidad económica, representatividad y disponibilidad”*.

A la hora de realizar los modelos utilizamos el software R. A su vez, los principales paquetes utilizados fueron *forecast* , *graphics* , *lmtest* , *stats* , *tseries* y *TSP*.

4.1. AEP-CNQ-AEP

El primer ejemplo a modelar consiste en la ruta Aeroparque-Corrientes-Aeroparque. En un principio ajustaremos un modelo ARMA o ARIMA según correspondiere. Los principales estadísticos de la serie son los siguientes:

Media	3786.24
Mediana	3415
Desviación Estandar	12228.54
Curtosis	-0.305
Mínimo	1282
Máximo	7254
Nro. Observaciones	89

Figura 4.1: Estadísticos Principales .

Para comenzar con el análisis de la serie observamos su gráfico en la figura 4.2. En un principio, analizando los principales estadísticos y el gráfico, podemos notar cierta no estacionariedad en la media, mas precisamente una esperable tendencia creciente. De todas maneras analizaremos la aplicación de una Transformación de Box-Cox a los datos para obtener más estacionariedad. La potencia elegida para transformar los datos lo podemos obtener del gráfico 4.2.

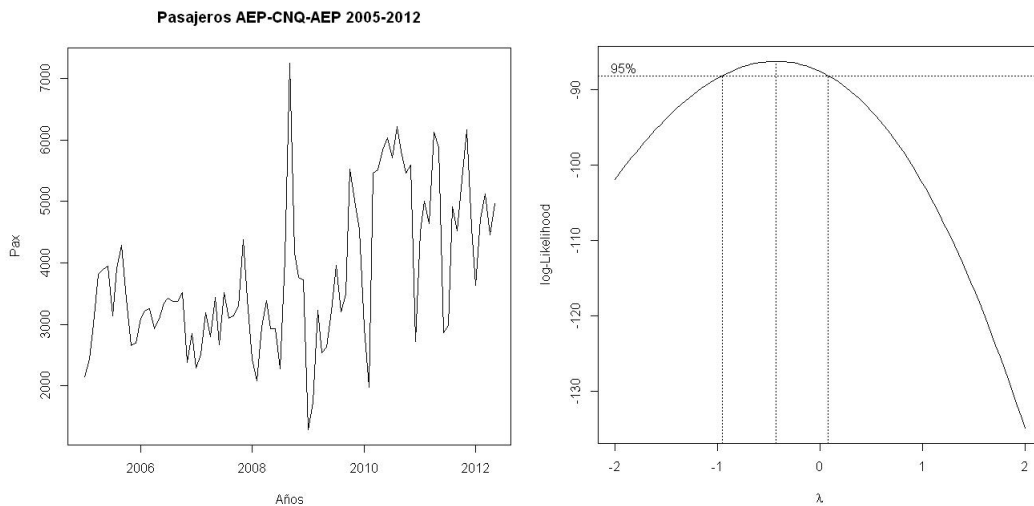


Figura 4.2: Gráfico Pax vs Mes y Gráfico Box-Cox

El gráfico sugiere aplicar dos posibles transformaciones a los datos con los valores $\lambda = 0$ y $\lambda = -0,5$. De esta manera, consideraremos dos posibles series para trabajar el modelo; $Y_t = \log(Z_t)$ y $W_t = \frac{Z_t^{-0,5}-1}{-0,5}$. Los gráficos de las nuevas series obtenidas vienen dados por la figura 4.3.

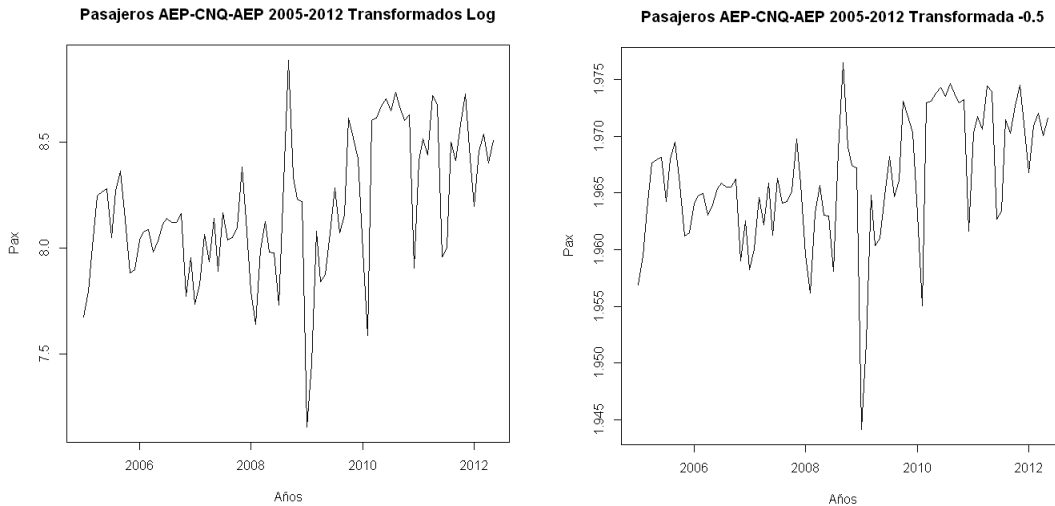


Figura 4.3: Series Transformadas

Una vez transformada la serie, utilizamos el Test de Dickey-Fuller para determinar si la serie es efectivamente estacionaria o no. En caso de ser necesario, aplicaremos diferencias hasta lograr la efectiva estacionariedad. Aplicamos el test a la serie sin diferenciar y a la serie diferenciada con $d = 1, D = 1$.

Luego tendremos en cuenta 5 series de datos:

- i) *mcnq1*: Es la serie sin diferenciar y sin transformación Box-Cox,
- ii) *mcnq2*: Es la serie sin diferenciar con transformación Box-Cox $\lambda = -0,5$,
- iii) *mcnq3*: Es la serie sin diferenciar con transformación Box-Cox $\lambda = 0$,
- iv) *mcnq2d1s1*: Es la serie diferenciada con transformación Box-Cox $\lambda = -0,5$,
- v) *mcnq3d1s1*: Es la serie diferenciada con transformación Box-Cox $\lambda = 0$,

Los estadísticos y p-valores obtenidos del Test de Dickey-Fuller se reportan en la figura 4.4.

Serie	Estadístico	p-valor
mcnq1	-3.30	0.0759
mcnq2	-3.58	0.0394
mcnq3	-3.46	0.04987
mcnq2d1s1	-5.82	0.01
mcnq3d1s1	-5.51	0.01

Figura 4.4: Test Dickey-Fuller

Observando los p-valores podemos rechazar con seguridad la hipótesis nula en las series con ambas diferencias; es decir, asegurar estacionariedad de los datos (notar además que la serie sin diferenciar tiene un p-valor mayor a 0.05). Las series sin diferencias también presentan p-valores menores a 0.05 por lo que también rechazaremos la hipótesis nula. En efecto, armaremos cuatro tipo de modelos; aquellos que combinan los dos posibles lambda con las dos posibles series diferenciadas. Al momento de seleccionar el modelo, notemos que en el caso de utilizar un criterio de información (por ejemplo el AIC), el criterio penalizará la sobrediferenciación. Así, tendremos una herramienta que nos permitirá eliminar los modelos innecesariamente complejos. A partir de ahora haremos los cuatro análisis por separado.

4.1.1. Caso $\lambda = -0,5$ y $d = D = 1$

Calculamos las funciones de autocorrelación y autocorrelación parcial, obteniendo la figura 4.5.

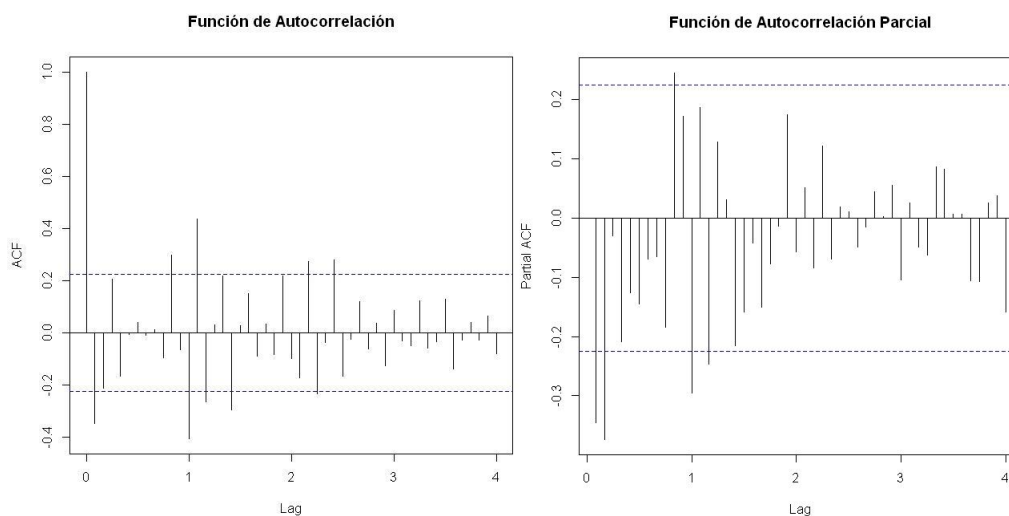


Figura 4.5: Función de Autocorrelación y Autocorrelación Parcial

A partir de estos gráficos, utilizando la teoría propuesta por Box-Jenkins, determinamos los valores p, q, P, Q del modelo ARIMA que vamos a elegir. Notemos que dado que conocemos la estacionalidad de los datos ($s=12$), tendremos efectivamente que encontrar los parámetros P, Q que modelan la estacionalidad. Una vez analizados los gráficos, los modelos propuestos vienen dados por la figura 4.6, donde la columna método se refiere al la estimación empleada para los coeficientes, nos referiremos a (ML) máxima verosimilitud o (CSS) máxima verosimilitud con utilizando como valores iniciales cuadrados mínimos condicionales. La columna “Modelos” les asigna un nombre a estos y el resto de las columnas corresponden a los parámetros de cada uno de los modelos. Aquellos que no presentan parámetros, son modelos generados por la función `auto.arima` del paquete `forecast`. Esta función genera una cantidad fija de modelos y elige según criterios de información. En particular, el modelo 3 usa el AIC y el modelo 4 el BIC.

Modelo	p	d	q	P	D	Q	Metodo
Modelo1.1	0	1	0	1	1	1	1 ML
Modelo2.1	2	1	0	1	1	1	1 ML
Modelo3							
Modelo4							
Modelo5.1	0	1	0	1	1	1	1 CSS
Modelo6.1	2	1	0	1	1	1	1 CSS

Figura 4.6: Parámetros de los modelos propuestos

Una vez ajustados los coeficientes mediante los métodos propuestos obtenemos finalmente los modelos mostrados en la figura 4.7.

```

Series: mcnq2
ARIMA(0,1,0) (1,1,1) [12]

Coefficients:
      sar1  sma1
      0.029 -1.00
s.e.  0.128  0.18

sigma^2 estimated as 0.0000258:  log likelihood=282
AIC=-558  AICc=-558  BIC=-551

Series: mcnq2
ARIMA(2,1,0) (1,1,1) [12]

Coefficients:
      ar1  ar2  sar1  sma1
      -0.39 -0.4  0.018 -1.00
s.e.  0.11  0.1  0.132  0.18

sigma^2 estimated as 0.0000196:  log likelihood=292
AIC=-574  AICc=-573  BIC=-562

Series: mcnq2
ARIMA(1,0,0) (1,0,0) [12] with non-zero mean

Coefficients:
      ar1  sar1  intercept
      0.494  0.23  1.966
s.e.  0.098  0.11  0.001

sigma^2 estimated as 0.0000211:  log likelihood=352
AIC=-697  AICc=-696  BIC=-687

Series: mcnq2
ARIMA(1,0,0) with non-zero mean

Coefficients:
      ar1  intercept
      0.55  1.966
s.e.  0.09  0.001

sigma^2 estimated as 0.0000222:  log likelihood=350
AIC=-695  AICc=-694  BIC=-687

Series: mcnq2
ARIMA(0,1,0) (1,1,1) [12]

Coefficients:
      sar1  sma1
      -0.083 -0.787
s.e.  0.127  0.099

sigma^2 estimated as 0.0000383:  part log likelihood=279

Series: mcnq2
ARIMA(2,1,0) (1,1,1) [12]

Coefficients:
      ar1  ar2  sar1  sma1
      -0.38 -0.38 -0.087 -0.76
s.e.  0.11  0.11  0.131  0.10

sigma^2 estimated as 0.0000298:  part log likelihood=288

```

Figura 4.7: Estimación de los coeficientes de los modelos.

Previa selección del modelo debemos ver que se cumplen los supuestos afirmados en la teoría de Box-Jenkins. Principalmente debemos ver si los residuos tienen o no una distribución normal. Para ello vamos a utilizar el Test de Kolmogorov-Smirnov. Este test compara cuan parecidos son los datos al de una distribución dada. En nuestro caso compararemos los residuos de los modelos con una distribución normal de media 0. Los estadísticos y p-valores vienen dados por la figura 4.8

Modelo	p-valor
Modelo1.1	0.1805
Modelo2.1	0.2545
Modelo3	0.3145
Modelo4	0.2901
Modelo5.1	0.0131
Modelo6.1	0.0028

Figura 4.8: Test Kolmogorov-Smirnov

Notemos que sacando los últimos dos modelos, no es posible rechazar la hipótesis nula; es decir, afirmar que los residuos no cuentan con distribución normal. Observe también que los últimos dos modelos son justamente los ajustados por CSS. Una vez observado que los supuestos no se violan podemos utilizar los criterios de información para seleccionar el modelo. La figura 4.9 muestra los valores de AIC y BIC de los modelos.

Modelo	AIC	BIC	Método
Modelo1.1	-558	-551	ML
Modelo2.1	-574	-562	ML
Modelo3	-697	-687	ML
Modelo4	-695	-687	ML

Figura 4.9: Valores de AIC y BIC

Efectivamente, la función `auto.arima` es la que encontró los modelos con AIC y BIC mas bajos. Estos dos modelos tienen a la serie sin diferenciar, por lo que no los tendremos en cuenta en este caso. Así, usando las condiciones, obtenemos que el modelo2.1 presenta el AIC y BIC mas bajo.

En la siguiente figura 4.10 presentaremos los gráficos de los pronósticos de cada modelo para los próximos 12 meses.

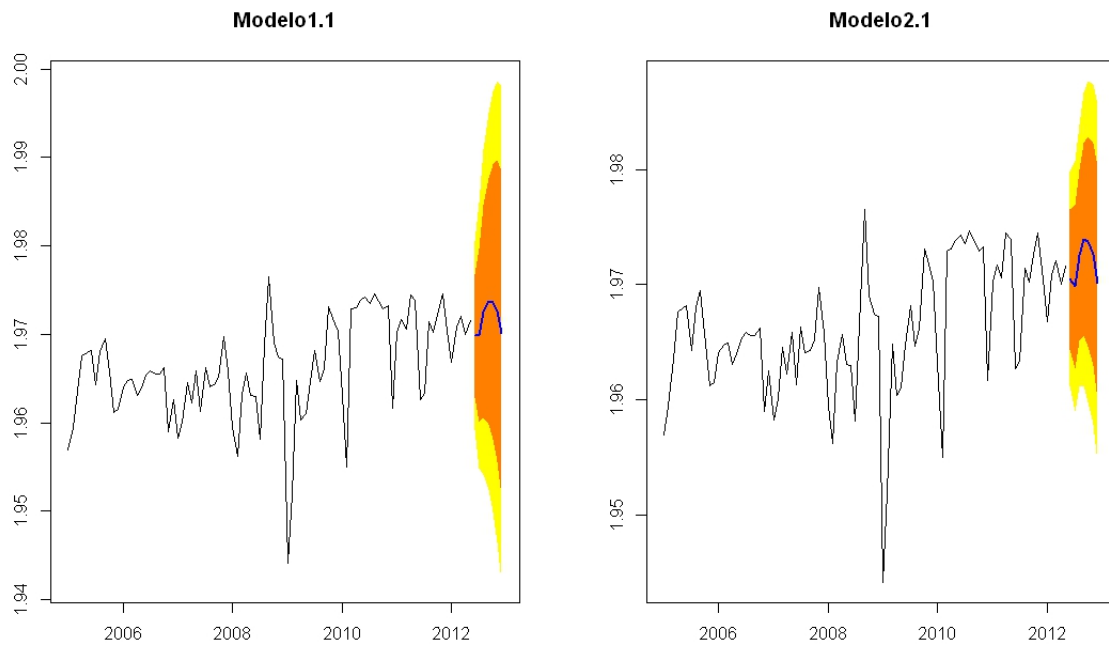


Figura 4.10: Forecast de los modelos

Finalmente, estos serán los dos modelos con los cuales nos quedaremos para este caso.

4.1.2. Caso $\lambda = -0,5$ y $d = D = 0$

La figura 4.11 muestra las funciones de autocorrelación y autocorrelación parcial de las series.

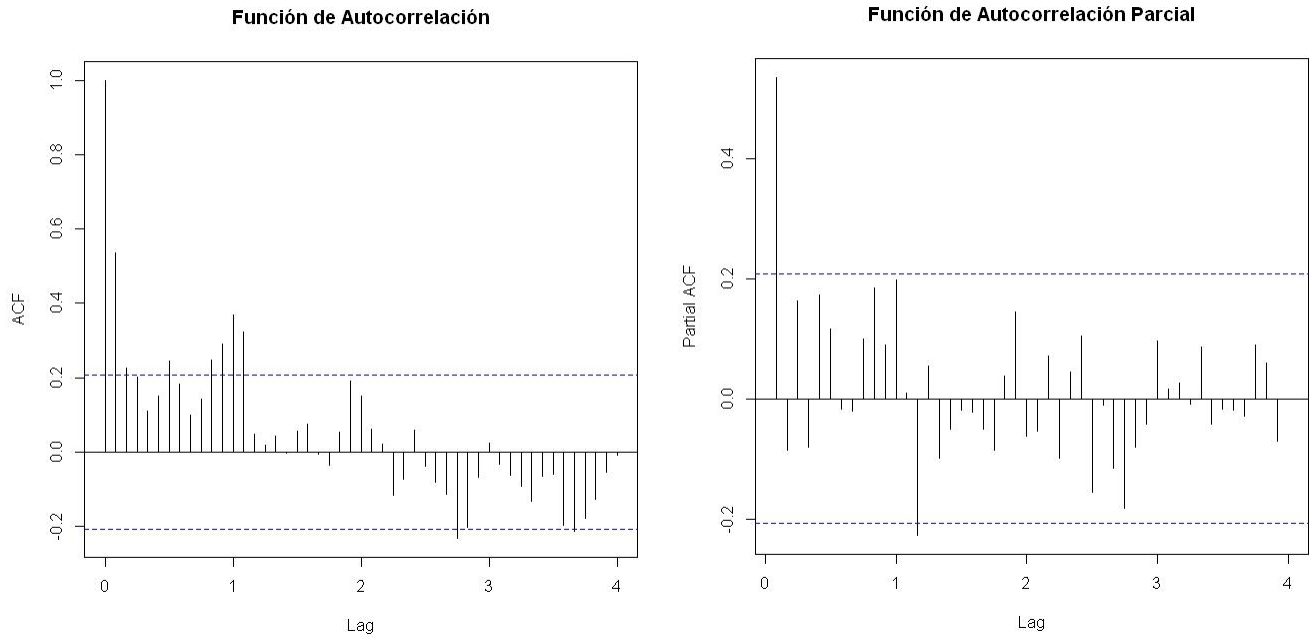


Figura 4.11: Función de Autocorrelación y Autocorrelación Parcial

Observando estos gráficos, determinamos los parámetros correspondientes de los modelos ARIMA, así proponemos un modelo que es estimado por los métodos de ML y CSS y denominamos modelos 7.1 y 8.1. Por otro lado los modelos 3 y 4 son sugeridos por la función auto.arima.

Modelo	p	d	q	P	D	Q	Método
Modelo7.1	1	0	0	0	0	1	ML
Modelo8.1	1	0	0	0	0	1	CSS
Modelo3	1	0	0	1	0	0	ML
Modelo4	1	0	0	0	0	0	ML

Figura: Parámetros de los modelos

Una vez determinados los parámetros, ajustamos los coeficientes usando los métodos propuestos. La figura 4.12 muestra la información referida a cada modelo.

```

Series: mcnq2
ARIMA(1,0,0) (0,0,1) [12] with zero mean

Coefficients:
      ar1      sma1
      1e+00  0.1284
s.e.  1e-04  0.1084

sigma^2 estimated as 2.764e-05:  log likelihood=340.19
AIC=-674.38  AICc=-674.1  BIC=-666.91

Series: mcnq2
ARIMA(1,0,0) (0,0,1) [12] with zero mean

Coefficients:
      ar1      sma1
      1.0001  0.1421
s.e.  0.0004  0.1149

sigma^2 estimated as 2.795e-05:  part log likelihood=340.3

Series: mcnq2
ARIMA(1,0,0) (1,0,0) [12] with non-zero mean

Coefficients:
      ar1      sar1  intercept
      0.4944  0.2272   1.9662
s.e.  0.0982  0.1094   0.0012

sigma^2 estimated as 2.109e-05:  log likelihood=352.38
AIC=-696.77  AICc=-696.29  BIC=-686.81

Series: mcnq2
ARIMA(1,0,0) with non-zero mean

Coefficients:
      ar1  intercept
      0.5526   1.9661
s.e.  0.0895   0.0011

sigma^2 estimated as 2.223e-05:  log likelihood=350.3
AIC=-694.6  AICc=-694.32  BIC=-687.14

```

Figura 4.12: Estimación de los modelos propuestos

A continuación testeamos si los residuos tienen una distribución normal o no. Para ello usamos el test de Kolmogorov-Smirnov. Los estadísticos y p-valores vienen se presentan en la figura 4.13.

Modelo	p-valor
Modelo7.1	0.4468
Modelo8.1	0.5422
Modelo3	0.3145
Modelo4	0.2901

Figura 4.13: Test Kolmogorov-Smirnov

Observando los p-valores correspondientes, notamos que no hay evidencia para rechazar la hipótesis nula. A continuación, procedemos a la selección del modelo. Ya sabemos que los modelos 3 y 4 presentan los AIC y BIC mas bajos (se puede ver en la figura 4.14).

Modelo	AIC	BIC	Método
Modelo7.1	-674	-674	ML
Modelo8.1			CSS
Modelo3	-696	-686	ML
Modelo4	-694	-687	ML

Figura 4.14: Valores de AIC y BIC

Los valores pronosticados para los 12 meses siguientes vienen dados por la figura 4.15.

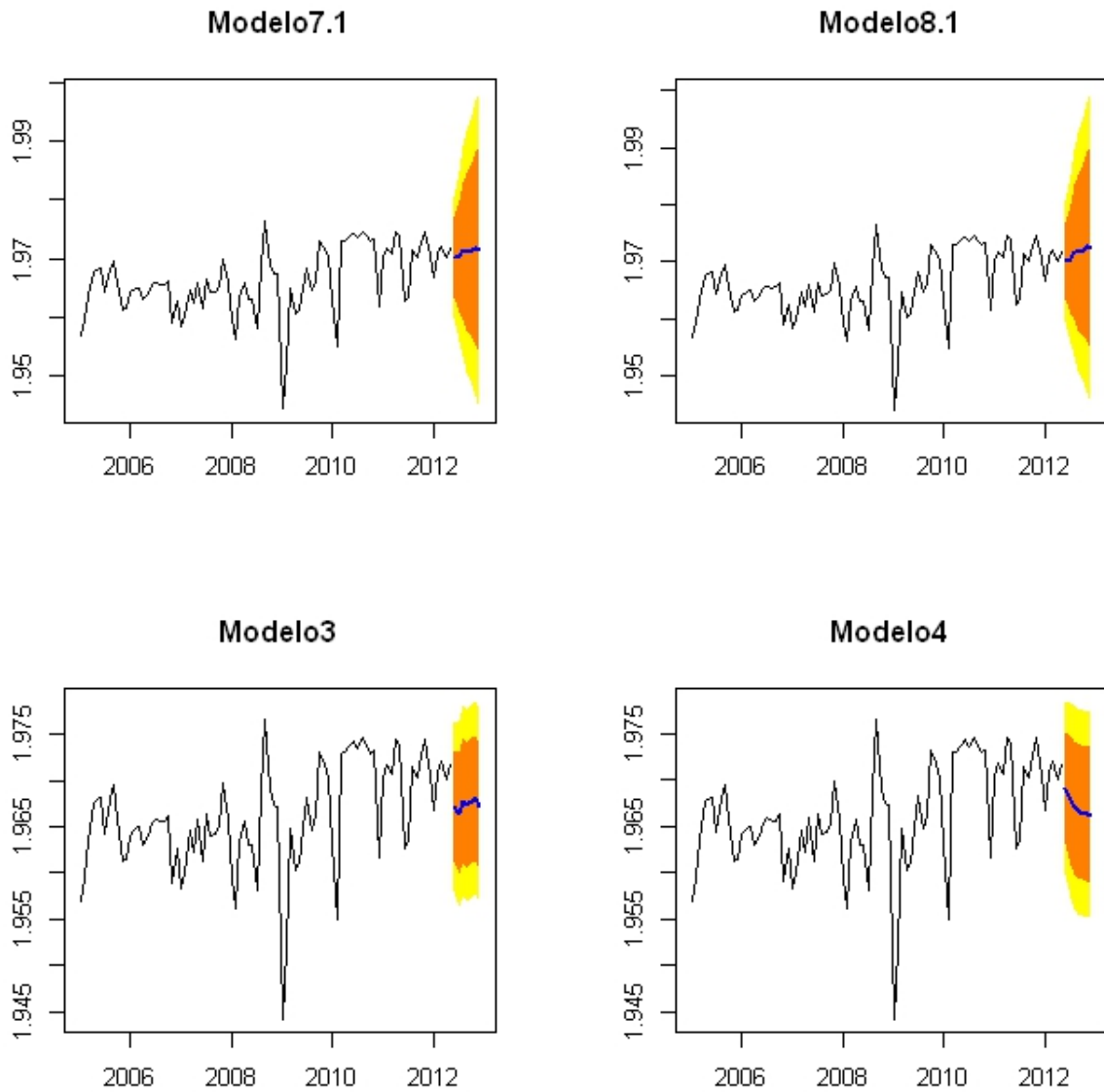


Figura 4.15: Forecast de los modelos

4.1.3. Caso $\lambda = 0$ y $d = D = 1$

Desde las figuras 4.16 hasta la figura 4.21 mostramos el análisis hecho para este caso, cuya procedimiento es análogo a lo realizado anteriormente

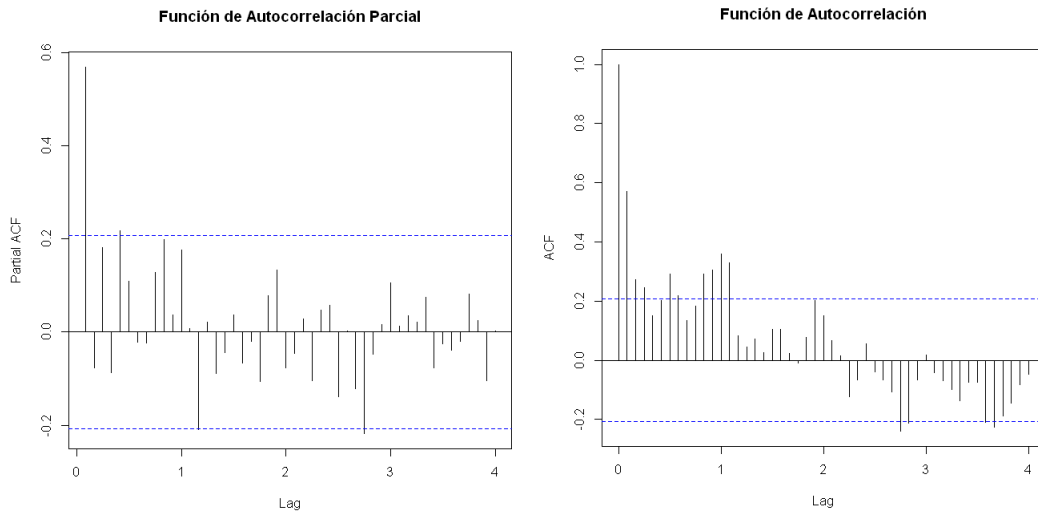


Figura 4.16: Funciones de Autocorrelación y Autocorrelación Parcial

Como procedimos anteriormente, del análisis de las ACF y PACF, proponemos los modelos 1.2, 2.2, 5.2 y 6.2, además la función auto.arima sugiere otros dos modelos que no tendremos en cuenta ya que se determinaron no incluyendo diferencias.

Modelo	p	d	q	P	D	Q	Método
Modelo1.2	0	1	0	1	1	1	ML
Modelo2.2	2	1	0	1	1	1	ML
Modelo5.2	0	1	0	1	1	1	CSS
Modelo6.2	2	1	0	1	1	1	CSS

Figura 4.17: Parámetros de los modelos

```

Series: mcnq3
ARIMA(0,1,0) (1,1,1) [12]

Coefficients:
      sar1      sma1
      -0.0012  -1.0000
s.e.   0.1289   0.2018

sigma^2 estimated as 0.0806:  log likelihood=-24.1
AIC=54.2  AICc=54.53  BIC=61.19

Series: mcnq3
ARIMA(0,1,0) (1,1,1) [12]

Coefficients:
      sar1      sma1
      -0.1233  -0.7718
s.e.   0.1301   0.1023

sigma^2 estimated as 0.12:  part log likelihood=-27.28

Series: mcnq3
ARIMA(2,1,0) (1,1,1) [12]

Coefficients:
      ar1      ar2      sar1      sma1
      -0.3762  -0.4156  -0.0231  -1.0000
s.e.   0.1042   0.1018   0.1301   0.2065

sigma^2 estimated as 0.06088:  log likelihood=-13.9
AIC=37.8  AICc=38.66  BIC=49.46

Series: mcnq3
ARIMA(2,1,0) (1,1,1) [12]

Coefficients:
      ar1      ar2      sar1      sma1
      -0.3726  -0.4013  -0.1399  -0.7320
s.e.   0.1043   0.1061   0.1333   0.1086

sigma^2 estimated as 0.09319:  part log likelihood=-17.66

```

Figura 4.18: Valores de los coeficientes de los modelos.

Continuamos ahora con el análisis de los residuos, en la figura 4.19 observamos que los modelos 5.2 y 6.2 (ambos ajustados por CCS) tienen residuos que no siguen una distribución normal. Así, por violar los supuestos, los eliminamos de la elección.

Modelo	p-valor
Modelo1.2	0.1012
Modelo2.2	0.1689
Modelo5.2	0.0099
Modelo6.2	0.0088

Figura 4.19: Test Kolmogorov-Smirnov

Finalmente, calculamos los criterios de selección para los dos modelos, resultando el modelo 2.2 más adecuado.

Modelo	AIC	BIC	Método
Modelo1.2	54,2	61,19	ML
Modelo2.2	37,8	49,46	ML

Figura 4.20: Valores de AIC y BIC

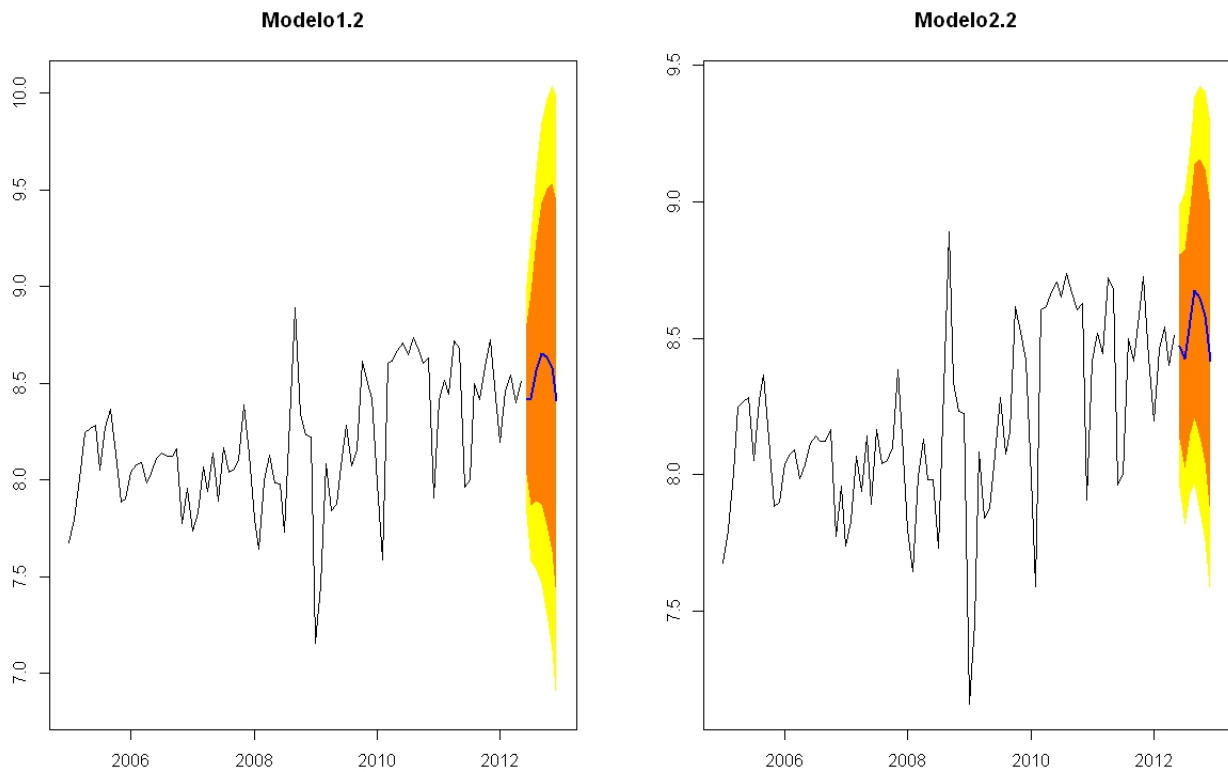


Figura 4.21: Forecast de los modelos propuestos

4.1.4. Caso $\lambda = 0$ y $d = D = 0$

Este caso comprende desde las figuras 4.22 a la 4.27. Observando las funciones ACF y PACF elegimos cuatro posibles modelos. Luego, verificamos que los residuos sean efectivamente normales usando el test de Kolmogorov-Smirnov. Notando que los cuatro modelos cumplen los supuestos, realizamos el forecast correspondiente.

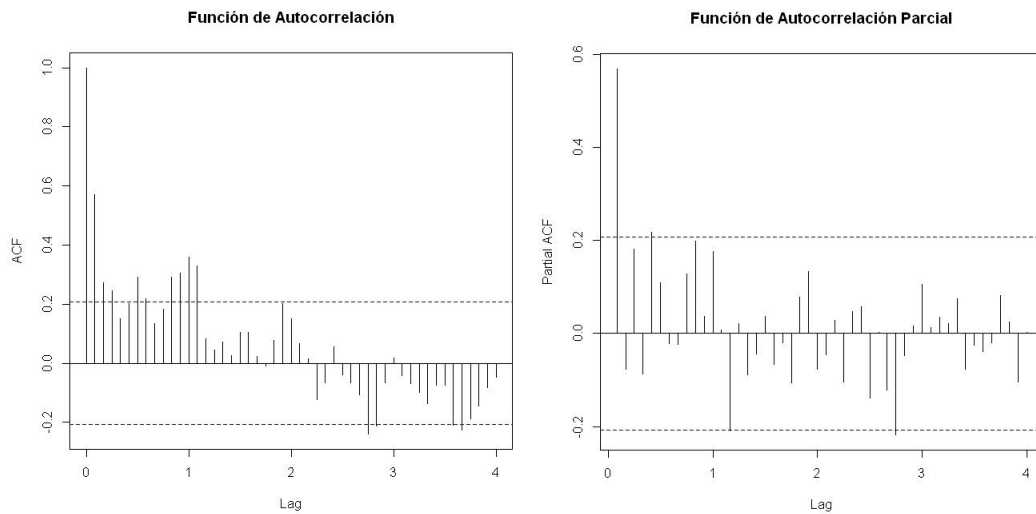


Figura 4.22: Funciones de Autocorrelación y Autocorrelación Parcial

Modelo	p	d	q	P	D	Q	Método
Modelo7.2	1	0	0	0	0	1	ML
Modelo8.2	1	0	0	0	0	1	CSS
Modelo9	1	0	2	0	0	0	ML
Modelo10	0	0	1	0	0	0	ML

Figura 4.23: Parámetros de los modelos

```

Series: mcnq3
ARIMA(1,0,0) (0,0,1) [12] with zero mean

Coefficients:
      ar1      sma1
      1e+00  0.0944
s.e.  1e-04  0.1103

sigma^2 estimated as 0.08666:  log likelihood=-18.01
AIC=42.02  AICc=42.3  BIC=49.48

Series: mcnq3
ARIMA(1,0,0) (0,0,1) [12] with zero mean

Coefficients:
      ar1      sma1
      1.0004  0.1009
s.e.  0.0042  0.1169

sigma^2 estimated as 0.08767:  part log likelihood=-17.97

Series: mcnq3
ARIMA(1,0,2) with non-zero mean

Coefficients:
      ar1      ma1      ma2  intercept
      0.9677  -0.3041  -0.5030  8.2023
s.e.  0.0404  0.1069  0.1004  0.1307

sigma^2 estimated as 0.0637:  log likelihood=-4.2
AIC=18.41  AICc=19.13  BIC=30.85

Series: mcnq3
ARIMA(0,0,1) with non-zero mean

Coefficients:
      ma1  intercept
      0.6452  8.1863
s.e.  0.0771  0.0460

sigma^2 estimated as 0.07026:  log likelihood=-8.38
AIC=22.76  AICc=23.05  BIC=30.23

```

Figura 4.24: Valores de los coeficientes de los modelos

Modelo	p-valor
Modelo7.2	0.3693
Modelo8.2	0.43
Modelo9	0.2082
Modelo10	0.832

Figura 4.25: Test Kolmogorov-Smirnov

Modelo	AIC	BIC	Método
Modelo7.2	42,02	49,48	ML
Modelo8.2			CSS
Modelo9	18,41	30,85	ML
Modelo10	22,76	30,23	ML

Figura 4.26: Valores de AIC y BIC

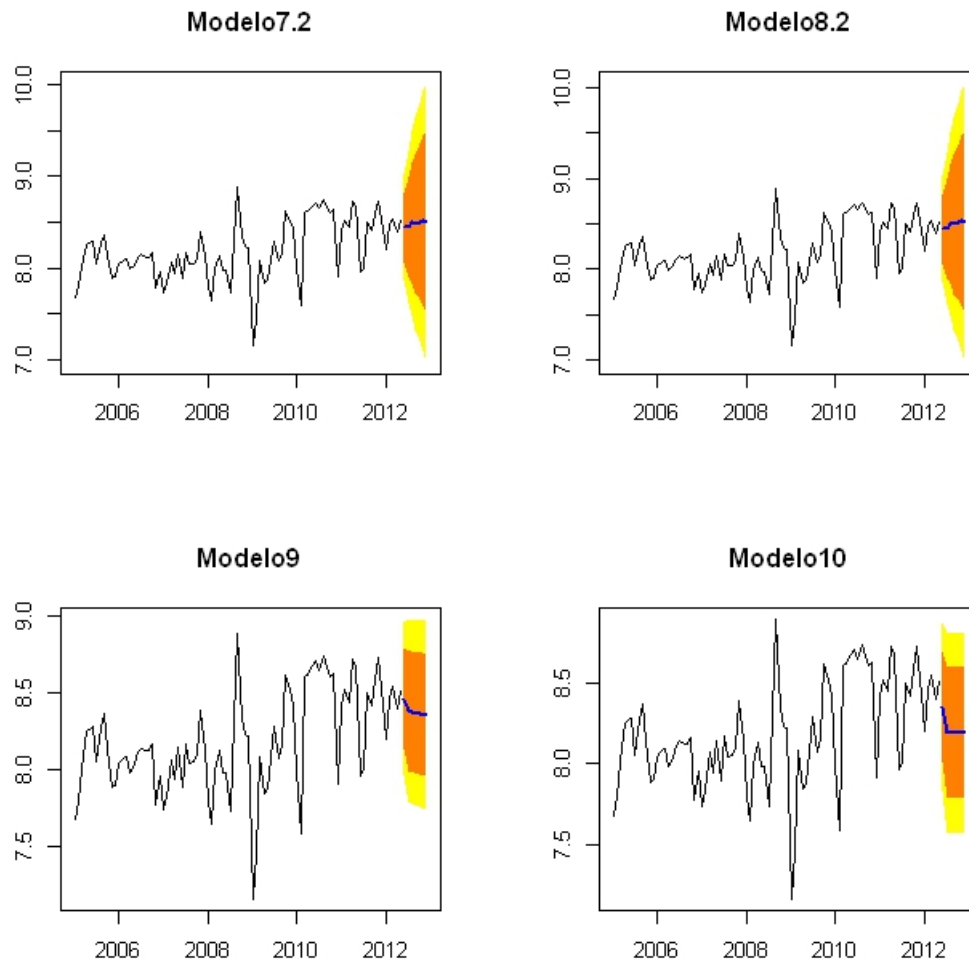


Figura 4.27: Forecast de los modelos propuestos

4.1.5. Modelos regARIMA

En la sección anterior solo buscamos explicar el comportamiento de la demanda en función de procesos propios de la ruta. Al aplicar modelos RegARIMA, veremos

como se relacionan la demanda con procesos exógenos a esta y si la inclusión de estas variables exogenas ayudan a explicar mejor la demanda. A su vez, cuando la regresión se hace en base a los logaritmos de los datos, economistas suelen usar el coeficiente que acompaña a la variable independiente (el cual se llama *elasticidad*, por ejemplo, elasticidad del precio con respecto a la demanda) para determinar cuan sensible es el cambio de la variable dependiente (la demanda) cuando varía la variable independiente (el precio, algún índice, etc).

En primer lugar ajustaremos un modelo regARIMA para ello consideramos la regresión

$$Y_t = \beta_1 + \beta_2 \log(akos) + \beta_3 \log(ema_e_{est}) + \beta_4 \log(ipc_gba) + \beta_5 \log(icc_utdt) + \epsilon_t$$

donde ϵ_t sigue un proceso ARIMA. El ajuste de la regresion se presenta en la siguiente figura.

```
Call:
lm(formula = log(Pax) ~ log(akos) + log(ema_e_est) + log(ipc_gba) +
    log(icc_utdt), data = mcng)

Residuals:
    Min       1Q   Median       3Q      Max
-0.37059 -0.09396  0.01337  0.08575  0.27889

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -7.16252     0.83251  -8.604 3.73e-13 ***
log(akos)     0.84870     0.05409  15.689 < 2e-16 ***
log(ema_e_est) 1.23607     0.25366   4.873 5.12e-06 ***
log(ipc_gba)  -1.14595     0.20556  -5.575 2.94e-07 ***
log(icc_utdt)  0.40841     0.10513   3.885 0.000203 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1328 on 84 degrees of freedom
Multiple R-squared:  0.8435,    Adjusted R-squared:  0.8361
F-statistic: 113.2 on 4 and 84 DF,  p-value: < 2.2e-16
```

Figura 4.28: Valores de la regresión

Observando esta figura, notamos que en un principio los coeficientes obtenidos son buenos. Al aplicarle el Test de normalidad de Shapiro-Wilk a los residuos, notamos que el p-valor obtenido es mayor que 0.05 (es 0.2272); por lo tanto no hay evidencia para rechazar la hipótesis nula y afirmar que los residuos no siguen una distribución normal.

Una vez analizada la información de la regresión pasaremos a modelar los residuos con un modelo ARIMA. Para eso, inicialmente le aplicaremos el Test Dickey-Fuller para determinar si hay raíces unitarias. EL p-value obtenido es de 0.02.

El test nos muestra que no es necesario diferenciar dado que la serie ya es estacionaria. Así, modelaremos los residuos de la regresión con un modelo ARMA.

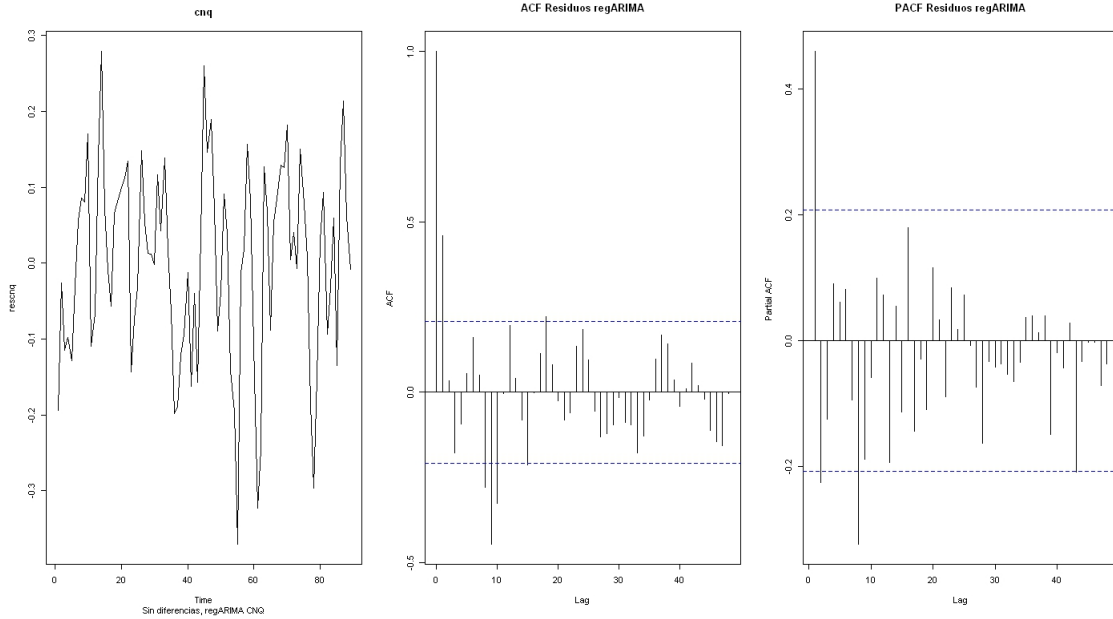


Figura 4.29: Serie de residuos, funciones de Autocorrelación y Autocorrelación Parcial

Al observar los gráficos de las correlaciones, elegimos los siguientes cuatro modelos, uno de los cuales fue elegido automáticamente por la función `auto.arima` (la cual selecciona el modelo que minimice el criterio de Akaike). (figura 4.30). Es decir, si denotamos con Y_t el logaritmo del proceso de demanda de pasajeros vamos a considerar los siguientes modelos:

$$\mathbf{regcnq1:} Y_t = \beta_1 + \beta_2 \log(akos) + \beta_3 \log(ema_e_{est}) + \beta_4 \log(ipc_gba) + \beta_5 \log(icc_utdt) + \epsilon_t \text{ donde } \epsilon_2(B)Y_t = \theta_1(B)a_t.$$

$$\mathbf{regcnq2:} Y_t = \beta_1 + \beta_2 \log(akos) + \beta_3 \log(ema_e_{est}) + \beta_4 \log(ipc_gba) + \beta_5 \log(icc_utdt) + \epsilon_t \text{ donde } \phi_2(B)\epsilon_t = \theta_{10}(B)a_t.$$

$$\mathbf{regcnq3:} Y_t = \beta_1 + \beta_2 \log(akos) + \beta_3 \log(ema_e_{est}) + \beta_4 \log(ipc_gba) + \beta_5 \log(icc_utdt) + \epsilon_t \text{ donde } \phi_8(B)\epsilon_t = \theta_{10}(B)a_t.$$

$$\mathbf{regcnq4:} Y_t = \beta_1 + \beta_2 \log(akos) + \beta_3 \log(ema_e_{est}) + \beta_4 \log(ipc_gba) + \beta_5 \log(icc_utdt) + \epsilon_t \text{ donde } \phi_3(B)\epsilon_t = \theta_2(B)a_t.$$

```

> regcnq1
Series: regcnq$residuals
ARIMA(2,0,1) with zero mean

Coefficients:
      ar1      ar2      ma1
      0.7961 -0.3429 -0.2306
s.e.  0.2790  0.1489  0.2871

sigma^2 estimated as 0.01222: log likelihood=69.54
AIC=-131.08  AICc=-130.6  BIC=-121.12

> regcnq2
Series: regcnq$residuals
ARIMA(2,0,10) with zero mean

Coefficients:
      ar1      ar2      ma1 ma2 ma3 ma4 ma5 ma6 ma7      ma8      ma9      ma10
      0.7182 -0.2686 -0.2558  0  0  0  0  0  0 -0.1094 -0.2458 -0.2507
s.e.  0.3758  0.2183  0.3787  0  0  0  0  0  0  0.1233  0.1033  0.1920

sigma^2 estimated as 0.009963: log likelihood=77.4
AIC=-140.8  AICc=-135.94  BIC=-108.44

> regcnq3
Series: regcnq$residuals
ARIMA(8,0,10) with zero mean

Coefficients:
      ar1      ar2 ar3 ar4 ar5 ar6 ar7      ar8      ma1 ma2 ma3 ma4 ma5 ma6 ma7      ma8      ma9
      0.4918 -0.1207  0  0  0  0  0 -0.3249 -0.0483  0  0  0  0  0  0  0 0.1799 -0.1738
s.e.  0.2702  0.1841  0  0  0  0  0  0.1566  0.2463  0  0  0  0  0  0  0 0.1584  0.1106
      ma10
      -0.3401
s.e.  0.1296

sigma^2 estimated as 0.009513: log likelihood=79.25
AIC=-142.49  AICc=-131.48  BIC=-95.21

> regcnq4
Series: regcnq$residuals
ARIMA(3,0,2) with zero mean

Coefficients:
      ar1      ar2      ar3      ma1      ma2
      1.4765 -1.4533  0.4642 -0.9633  0.8834
s.e.  0.1260  0.1296  0.1105  0.1008  0.1136

sigma^2 estimated as 0.01066: log likelihood=75
AIC=-138.01  AICc=-136.98  BIC=-123.08

```

Figura 4.30: Modelos ARMA para los residuos

Para estos cuatro modelos verificamos los supuestos de normalidad de sus residuos. Usando el Test de Kolmogorov-Smirnov, notamos que los residuos de los modelos no dan evidencia para rechazar la hipótesis nula (figura 4.31).

Modelo	p-valor
regcnq1	0.8469
regcnq2	0.7088
regcnq3	0.4601
regcnq4	0.7939

Figura 4.31: Test Kolmogorov-Smirnov

4.1.6. Pronósticos y conclusiones

En la figura siguiente mostramos el pronóstico obtenido para los siguientes 7 meses usando los modelos ARIMA y RegARIMA y comparamos con sus correspondientes valores reales. Antes de continuar, resumiremos los modelos que hemos analizado hasta aquí, además de los modelos de regresión. Llamaremos Z_t a la serie con $\lambda = -0,5$ e Y_t a la serie con $\lambda = 0$, de esta forma los modelos obtenidos son los siguientes:

Modelo1.1: $\Phi_1(B^{12})\phi_0(B)(1-B)(1-B^{12})Z_t = \Theta_1(B^{12})\theta_0(B)a_t$

Modelo1.2: $\Phi_1(B^{12})\phi_0(B)(1-B)(1-B^{12})Y_t = \Theta_1(B^{12})\theta_0(B)a_t$

Modelo2.1: $\Phi_1(B^{12})\phi_2(B)(1-B)(1-B^{12})Z_t = \Theta_1(B^{12})\theta_0(B)a_t$

Modelo2.2: $\Phi_1(B^{12})\phi_2(B)(1-B)(1-B^{12})Y_t = \Theta_1(B^{12})\theta_0(B)a_t$

Modelo3: $\Phi_1(B^{12})\phi_1(B)Z_t = \Theta_0(B^{12})\theta_0(B)a_t$

Modelo4: $\phi_1(B)Z_t = \theta_0(B)a_t$

Modelo7.1: $\Phi_0(B^{12})\phi_1(B)Z_t = \Theta_1(B^{12})\theta_0(B)a_t$

Modelo7.2: $\Phi_0(B^{12})\phi_1(B)Y_t = \Theta_1(B^{12})\theta_0(B)a_t$

Modelo8.1: $\phi_1(B)Z_t = \theta_0(B)a_t$

Modelo8.2: $\Phi_0(B^{12})\phi_1(B)Y_t = \Theta_1(B^{12})\theta_0(B)a_t$

Modelo9: $\phi_1(B)Y_t = \theta_2(B)a_t$

Modelo10: $\phi_0(B)Y_t = \theta_1(B)a_t$

En la figura 4.32 observaremos los resultados obtenidos. La tabla muestra los resultados por modelo para los siguientes 7 meses. Contiene los parámetros (valores p,d,q,P,D,Q), el lambda que se utilizó para transformar la serie y el método con el cual se ajustó la serie. Debajo aparecen los errores de estimación en valores absolutos. Los colores varían del verde (errores pequeños) al rojo (errores altos).

Nuestro objetivo final consiste en elegir el modelo con el cual explicar la serie de datos original. Notemos en un principio que las regresiones muestran los errores mas altos. Mas allá de presentar errores pequeños en el primer mes, se observa que en los meses posteriores los valores llegan al 30%. Una posible explicación es que las variables independientes elegidas pueden no ser las correctas, en el sentido que índices como el IPC o el ICC explican situaciones macro que pueden no influir directamente en la demanda de pasajeros. Es así que sería conveniente elegir variables que se relacionen

exclusivamente con la demanda.

Con respecto a los modelos ARIMA, los números 3,4,9 y 10 presentan altos errores. Sorprendente, estos son los modelos con condiciones AIC y BIC mas bajos para cada transformación ($\lambda = -0,5$ y $\lambda = 0$). Así, redujimos la cantidad de modelos elegibles a 8. Cabe notar que todos ellos presentar errores cercanos al 20% en los primeros dos meses de forecast. Sin embargo esto disminuye rotundamente para los siguientes meses, llegando a tener medias de 5% de error.

Finalmente, privilegiaremos los modelos con menor cantidad de parámetros. Es así que los modelos 8.1 (un proceso explicado como un AR(1)) y los 7.1,7.2 y 8.2 (procesos ARIMA(1,0,0)*(0,0,1), con diferentes valores de λ y métodos de ajuste) serían en ese orden los modelos a seleccionar para explicar el comportamiento de la ruta.

Nombre	Modelo	lambda	d	D	Metodo	1	2	3	4	5	6	7
Modelo 1.1	(0,1,0)*(1,1,1)[12]	-0,5	1	1	ML	4400	4408	5312	5800	5769	5334	4467
Modelo 1.2	(0,1,0)*(1,1,1)[12]	0	1	1	ML	4529	4537	5255	5735	5636	5330	4501
Modelo 2.1	(2,1,0)*(1,1,1)[12]	-0,5	1	1	ML	4584	4396	5290	5885	5800	5342	4468
Modelo 2.2	(2,1,0)*(1,1,1)[12]	0	1	1	ML	4765	4557	5243	5847	5686	5336	4517
Modelo 3	(1,0,0)*(1,0,0)[12]	-0,5	0	0	ML	3716	3551	3865	3746	3850	3939	3712
Modelo 4	(1,0,0)	-0,5	0	0	ML	4204	3859	3686	3595	3546	3520	3505
Modelo 7.1	(1,0,0)*(0,0,1)[12]	-0,5	0	0	CSS	4488	4520	4841	4790	4904	4981	4861
Modelo 7.2	(1,0,0)*(0,0,1)[12]	0	0	0	ML	4634	4654	4874	4838	4922	4986	4878
Modelo 8.1	(1,0,0)	-0,5	0	0	CSS	4488	4571	4986	4981	5173	5324	5243
Modelo 8.2	(1,0,0)*(0,0,1)[12]	0	0	0	CSS	4626	4661	4911	4887	4922	5076	4976
Modelo 9	(1,0,2)	0	0	0	ML	4737	4364	4339	4315	4292	4269	4247
Modelo 10	(0,0,1)	0	0	0	ML	4234	3591	3591	3591	3591	3591	3591
regcnq1	reg+(2,0,1)					5715	4842	3039	3572	4139	4042	2614
regcnq2	reg+(2,0,10)					5869	5126	3202	3590	4069	3940	2448
regcnq3	reg+(8,0,10)					5785	5064	3155	3524	4039	3927	2375
regcnq4	reg+(3,0,2)					5491	4719	3148	3852	4339	3938	2428
Real						5897	5534	5061	5180	5195	5520	4421
Modelo 1.1	Error Absoluto	25%	20%	5%	12%	11%	3%	1%				
Modelo 1.2		23%	18%	4%	11%	8%	3%	2%				
Modelo 2.1		22%	21%	5%	14%	12%	3%	1%				
Modelo 2.2		19%	18%	4%	13%	9%	3%	2%				
Modelo 3		37%	36%	24%	28%	26%	29%	16%				
Modelo 4		29%	30%	27%	31%	32%	36%	21%				
Modelo 7.1		24%	18%	4%	8%	6%	10%	10%				
Modelo 7.2		21%	16%	4%	7%	5%	10%	10%				
Modelo 8.1		24%	17%	1%	4%	0%	4%	19%				
Modelo 8.2		22%	16%	3%	6%	5%	8%	13%				
Modelo 9		20%	21%	14%	17%	17%	23%	4%				
Modelo 10		28%	35%	29%	31%	31%	35%	19%				
regcnq1		3%	13%	40%	31%	20%	27%	41%				
regcnq2		0%	7%	37%	31%	22%	29%	45%				
regcnq3	2%	8%	38%	32%	22%	29%	46%					
regcnq4	7%	15%	38%	26%	16%	29%	45%					

Figura 4.32: Modelos CNQ- Resultados - Error Absoluto

4.2. AEP-FMA-AEP

4.2.1. ARIMA

El segundo ejemplo a modelar consiste en la ruta Aeroparque-Formosa-Aeroparque. Para este caso los estadísticos son los siguientes:

Media	2760.56
Mediana	2559
Desviación Estandar	937.72
Curtosis	-0.67
Mínimo	1040
Máximo	4829
Nro. Observaciones	89

Figura 4.33: Estadísticos Principales .

El comportamiento de la ruta viene dado por el gráfico de la serie, mostrado en la figura 4.34. Al igual que en el análisis realizado para la otra ruta, el primer paso consiste en transformar la serie para lograr estacionariedad. Para ello utilizamos la Transformación Box-Cox. La elección del mejor λ viene dado por la figura 4.34.

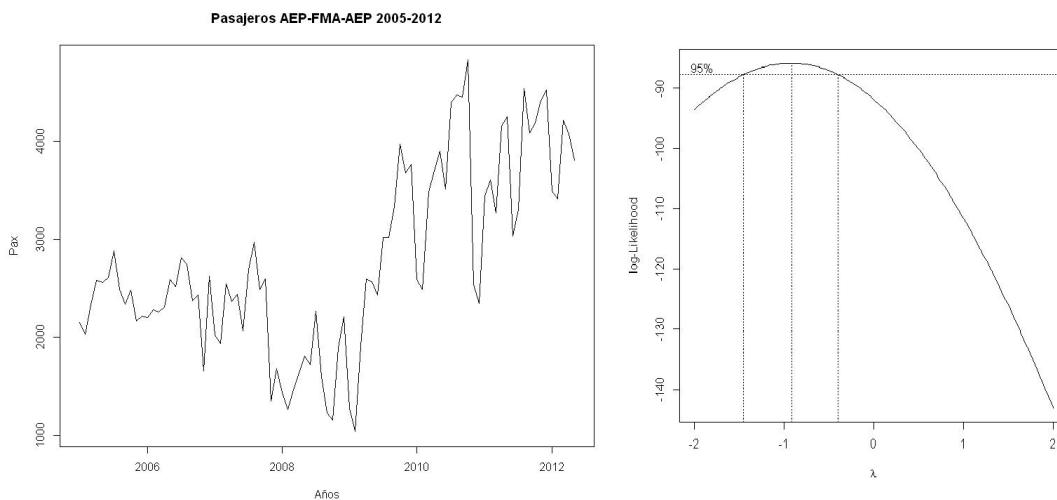


Figura 4.34: Serie de Datos y Gráfico Box-Cox

Observando el gráfico vamos a tomar $\lambda = -1$, es decir, tomaremos la serie $Z_t = 1/Y_t$. Una vez transformada la serie vamos a diferenciar hasta lograr la estacionariedad de la serie. Para determinar si se llegó a esta, utilizaremos el Test de Dickey-Fuller. Los valores obtenidos los encontramos en la figura 4.35.

Serie	Estadístico	p-valor
mfma2	-2.05	0.55
mfma2d1	-4.80	0.01
mfma2s1	-1.78	0.66
mfma2d1s1	-5.01	0.01

Figura 4.35: Test Dickey-Fuller

Observando la tabla, nos quedamos con la primera diferencia de lag 1. Una vez determinado el orden de diferenciación, vamos a calcular las funciones de autocorrelación y autocorrelación parcial. Estas funciones se presentan en la figura 4.36.

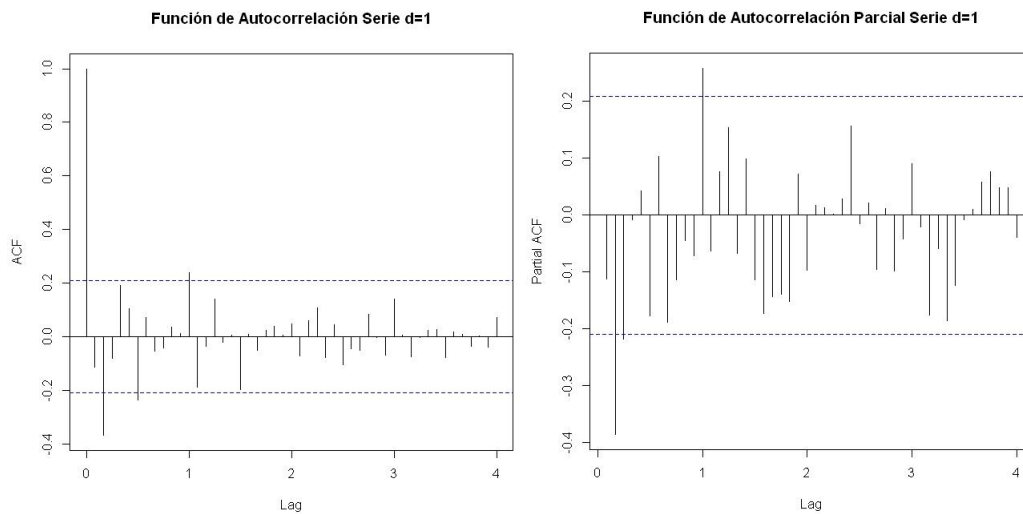


Figura 4.36: Gráficos ACF y PACF

El paso siguiente consiste en identificar el modelo analizando las dos funciones. Los modelos propuestos vienen dados por la figura 4.37

```

> modelo1
Series: mfma2
ARIMA(3,1,2) (1,0,1) [12]

Coefficients:
      ar1      ar2      ar3      ma1      ma2      sar1      sma1
    -0.2749 -0.4323 -0.2770  0.0492 -0.0400  0.9191 -0.7625
s.e.   0.4590  0.1898  0.2098  0.4720  0.2491  0.7302  1.2108

sigma^2 estimated as 8.083e-09:  log likelihood=692.82
AIC=-1369.63  AICc=-1367.81  BIC=-1349.81
> modelo2
Series: mfma2
ARIMA(2,1,2) (1,0,1) [12]

Coefficients:
      ar1      ar2      ma1      ma2      sar1      sma1
    0.3395 -0.5415 -0.5669  0.1898  0.9636 -0.8461
s.e.  0.2224  0.1678  0.2557  0.2323  0.4643  1.0190

sigma^2 estimated as 7.94e-09:  log likelihood=692.61
AIC=-1371.22  AICc=-1369.82  BIC=-1353.88
> modelo3
Series: mfma2
ARIMA(2,1,1) (1,0,0) [12]

Coefficients:
      ar1      ar2      ma1      sar1
    0.2180 -0.4120 -0.4172  0.2920
s.e.  0.1646  0.1047  0.1703  0.0977

sigma^2 estimated as 8.582e-09:  log likelihood=691.56
AIC=-1373.13  AICc=-1372.4  BIC=-1360.74
> modelo4
Series: mfma2
ARIMA(3,1,0) (1,0,0) [12]

Coefficients:
      ar1      ar2      ar3      sar1
    -0.2096 -0.4495 -0.2467  0.3010
s.e.  0.1033  0.0927  0.1029  0.0973

sigma^2 estimated as 8.469e-09:  log likelihood=692.1
AIC=-1374.19  AICc=-1373.46  BIC=-1361.8
> modelo5
Series: mfma2
ARIMA(3,1,2) (1,0,1) [12]

Coefficients:
      ar1      ar2      ar3      ma1      ma2      sar1      sma1
    -0.3236 -0.5167 -0.3184  0.0712 -0.0271  0.9561 -0.8741
s.e.  0.4387  0.1642  0.2154  0.4473  0.2064  0.0948  0.1550

sigma^2 estimated as 9.204e-09:  part log likelihood=689.29
> modelo6
Series: mfma2
ARIMA(2,1,2) (1,0,1) [12]

Coefficients:
      ar1      ar2      ma1      ma2      sar1      sma1
    0.2618 -0.5793 -0.5046  0.1424  0.9507 -0.8616
s.e.  0.1939  0.1415  0.2146  0.1808  0.1016  0.1655

sigma^2 estimated as 9.171e-09:  part log likelihood=689.45

```


Figura 4.37: Coeficientes de los modelos propuestos

Ya definidos los modelos vamos a determinar si el supuesto de normalidad de los residuos es violado o no. Para ello vamos a utilizar el Test de Kolmogorov-Smirnov. Los estadísticos y p-values encontrados aparecen en la figura

Modelo	p-valor
Modelo1	0.0561
Modelo2	0.05309
Modelo3	0.05742
Modelo4	0.05754
Modelo5	0.00676
Modelo6	0.00277

Figura 4.38: Test Kolmogorov-Smirnov

Observando los p-values dados por Kolmogorov-Smirnov, elegimos los modelos que no violen el supuesto de normalidad de los residuos. Notemos los modelos del 1 al 4 presentan p-values cercanos a 0.05, a diferencia de los modelos 5 y 6 que permiten rechazar la hipótesis nula. Así, seguimos con los modelos del 1 al 4. El forecast de se muestran en la 4.39.

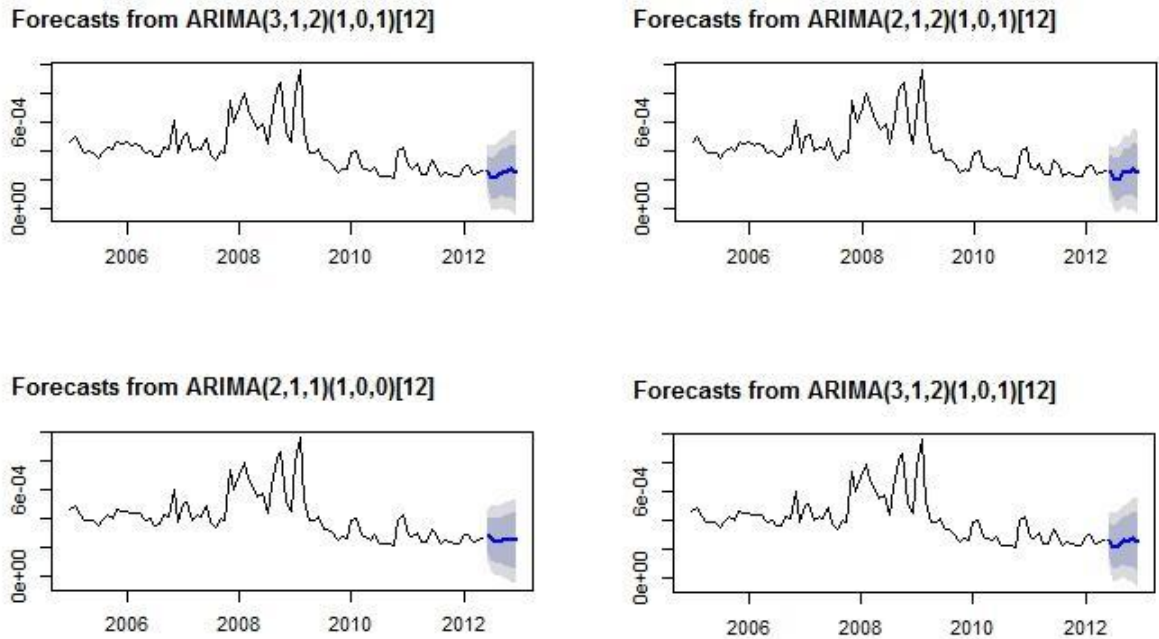


Figura 4.39: Forecast de los modelos

4.2.2. RegARIMA

A continuación consideraremos los modelos RegARIMA de manera similar a la empleada para el caso anterior.

Consideraremos la siguiente regresión

$$Y_t = \beta_1 + \beta_2 \log(akos) + \beta_3 \log(ema_e_est) + \beta_4 \log(ipc_gba) + \beta_5 \log(icc_utdt) + \epsilon_t$$

donde ϵ_t sigue un proceso ARIMA. Los estadísticos se presentan en la figura 4.41.

```
> summary(regfma)

Call:
lm(formula = log(mfma1) ~ log(akos) + log(ema_e_est) + log(ipc_gba) +
    log(icc_utdt), data = mfma)

Residuals:
    Min       1Q   Median       3Q      Max
-0.32677 -0.05274  0.00123  0.06677  0.21695

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -5.67607    0.57483   -9.874 1.04e-15 ***
log(akos)     0.77339    0.03247   23.818 < 2e-16 ***
log(ema_e_est) 0.84185    0.18920    4.450 2.62e-05 ***
log(ipc_gba)  -0.64263    0.16132   -3.984 0.000144 ***
log(icc_utdt)  0.16181    0.08555    1.891 0.062007 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1012 on 84 degrees of freedom
Multiple R-squared:  0.9233,    Adjusted R-squared:  0.9197
F-statistic: 252.8 on 4 and 84 DF,  p-value: < 2.2e-16
```

Figura 4.40: Coeficientes y estadísticos de la regresión

Para analizar la normalidad de los residuos vamos a aplicar nuevamente el Test de Kolmogorov-Smirnov. El p-valor es de 0.84 por lo que podremos afirmar que no hay evidencia para rechazar la hipótesis nula.

Para realizar el modelo ARIMA para sus residuos, aplicamos el Test de Dickey-Fuller para determinar si estos tienen raíces unitarias. El p-valor es de 0.01 por lo que podemos rechazar la hipótesis nula y suponer estacionariedad.

Para determinar como modelar la serie, observamos sus funciones de autocorrelación y autocorrelación parcial en la figura 4.41.

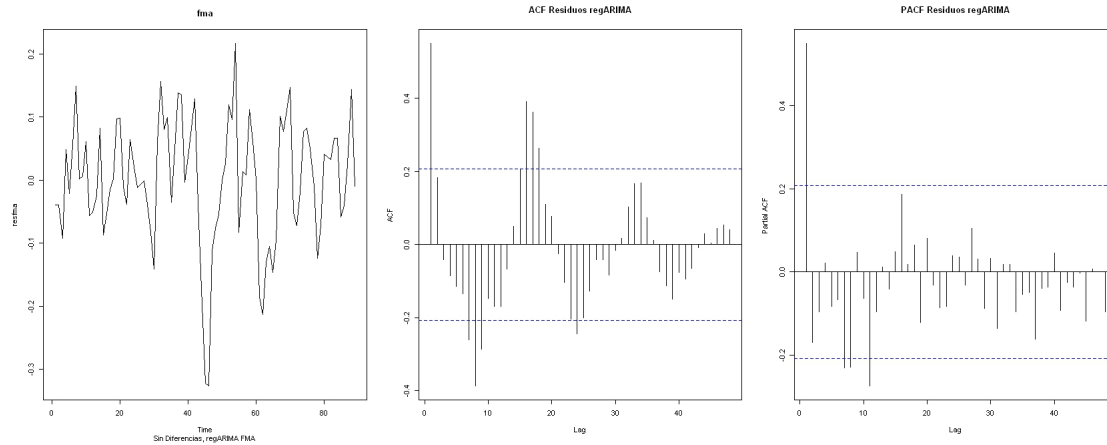


Figura 4.41: Gráficos ACF y PACF

Los modelos determinados a partir del análisis de las funciones vienen dados por la figura 4.42.

regfma1: $Y_t = \beta_1 + \beta_2 \log(akos) + \beta_3 \log(ema_e_{est}) + \beta_4 \log(ipc_gba) + \beta_5 \log(icc_utdt) + \epsilon_t$ donde $\epsilon_2(B)Y_t = \theta_1(B)a_t$.

regfma2: $Y_t = \beta_1 + \beta_2 \log(akos) + \beta_3 \log(ema_e_{est}) + \beta_4 \log(ipc_gba) + \beta_5 \log(icc_utdt) + \epsilon_t$ donde $\epsilon_1(B)Y_t = \theta_0(B)a_t$.

regfma3: $Y_t = \beta_1 + \beta_2 \log(akos) + \beta_3 \log(ema_e_{est}) + \beta_4 \log(ipc_gba) + \beta_5 \log(icc_utdt) + \epsilon_t$ donde $\epsilon_{11}(B)Y_t = \theta_0(B)a_t$.

```

> modeloreg1
Series: regfma$residuals
ARIMA(2,0,1) with zero mean

Coefficients:
      ar1      ar2      ma1
      1.4736 -0.5957 -0.9536
s.e.  0.0919  0.0837  0.0870

sigma^2 estimated as 0.006053:  log likelihood=100.38
AIC=-192.76  AICc=-192.28  BIC=-182.8

> modeloreg2
Series: regfma$residuals
ARIMA(1,0,0) with zero mean

Coefficients:
      ar1
      0.5450
s.e.  0.0876

sigma^2 estimated as 0.006741:  log likelihood=96.02
AIC=-190.04  AICc=-189.9  BIC=-185.06

> modeloreg3
Series: regfma$residuals
ARIMA(11,0,0) with zero mean

Coefficients:
      ar1  ar2  ar3  ar4  ar5  ar6      ar7      ar8  ar9  ar10      ar11
      0.4505  0  0  0  0  0 -0.079 -0.2244  0  0  0 -0.1094
s.e.  0.0863  0  0  0  0  0  0.102  0.1044  0  0  0  0.0822

sigma^2 estimated as 0.005894:  log likelihood=101.47
AIC=-194.94  AICc=-190.84  BIC=-165.08

```

Figura 4.42: Estimación de los coeficientes de los modelos

Una vez determinados los posibles modelos ARIMA vamos a utilizar el Test de Shapiro-Wilk para observar normalidad de sus residuos. En la figura 4.43 podemos observar que no hay evidencia para rechazar la hipótesis nula; es decir, no podemos rechazar que los residuos sean normales.

Modelo	p-valor
Modelo1	0.1805
Modelo2	0.2545
Modelo3	0.3145

Figura 4.43: Test Shapiro-Wilk

Dado que todos los modelos cumplen los supuestos, realizamos el pronóstico de la ruta que detallaremos en la sección siguiente.

4.2.3. Pronosticos y conclusiones.

Recordemos que notamos por Z_t a la serie transformada con $\lambda = -1$ y hemos encontrado 4 diferentes modelos además de los de regresión:

Modelo1: $\Phi_1(B^{12})\phi_3(B)Z_t = \Theta_1(B^{12})\theta_2(B)a_t$

Modelo2: $\Phi_1(B^{12})\phi_2(B)Z_t = \Theta_1(B^{12})\theta_2(B)a_t$

Modelo3: $\Phi_1(B^{12})\phi_2(B)Z_t = \Theta_0(B^{12})\theta_1(B)a_t$

Modelo4: $\Phi_1(B^{12})\phi_3(B)Z_t = \Theta_0(B^{12})\theta_0(B)a_t$

La tabla muestra los resultados por modelo para los siguientes 7 meses. Contiene los parámetros (valores p, d, q, P, D, Q), el λ que se utilizó para transformar la serie (que en este caso siempre fue -1) y el método con el cual se ajustó la serie. Debajo aparecen los errores de estimación en valores absolutos. Los colores varían del verde (errores pequeños) al rojo (errores altos).

Para el caso de esta ruta tenemos idéntica situación con las regresiones, tal es así que presentan los errores mas altos. El modelo 3 no solo presenta la menor cantidad de parámetros si no que también es el que menor error posee (promedio y media de 7%). Además tiene los segundos menores condiciones AIC y BIC (solo superado por el modelo 4). El modelo 4 presenta idéntica cantidad de parámetros, menor condiciones AIC y BIC y medias de error del 9%. Amparados por esta información, elegimos estos dos modelos como aquellos que explicarán la demanda de la ruta.

Nombre	Modelo	lambda	d	D	Metodo	1	2	3	4	5	6	7
Modelo1	(3,1,2)*(1,0,1)[12]	-1	1	0	ML	3819	4860	4630	3977	4131	3756	4125
Modelo2	(2,1,2)*(1,0,1)[12]	-1	1	0	ML	3845	5042	4653	3912	4038	3585	4098
Modelo3	(2,1,1)*(1,0,0)[12]	-1	1	0	ML	3545	3775	4129	3947	3969	4052	4087
Modelo4	(3,1,0)*(1,0,0)[12]	-1	1	0	ML	3443	3742	4144	3910	3896	4008	4071
regfma1	reg+(2,0,1)					3441	3521	3414	3265	3557	3418	3582
regfma2	reg+(1,0,0)					3485	3584	3477	3318	3602	3449	3602
regfma3	reg+(11,0,0)					3472	3492	3388	3322	3603	3367	3474
Real						3578	4553	3664	4035	4426	4184	4405
Modelo1	Error Absoluto	7%	7%	26%	1%	7%	10%	6%				
Modelo2		7%	11%	27%	3%	9%	14%	7%				
Modelo3		1%	17%	13%	2%	10%	3%	7%				
Modelo4		4%	18%	13%	3%	12%	4%	8%				
regfma1		4%	23%	7%	19%	20%	18%	19%				
regfma2		3%	21%	5%	18%	19%	18%	18%				
regfma3		3%	23%	8%	18%	19%	20%	21%				

Figura 4.44: Resultados Modelos - Error Absoluto

Bibliografía

- [1] Abraham, B. y Ledolter, J. (1983) *Statistical Methods for Forecasting*, Wiley.
- [2] Akaike, H. (1974) A new look at the statistical model identification, *IEEE Transactions on Automatic Control* 19 (6): 716–723.
- [3] Bianco A., Garcia Ben M., Martinez E. y Yohai V., *Robust Procedures for Regression Models with ARIMA errors*, Instituto del Calculo, Universidad de Buenos Aires.
- [4] Box, G. y Jenkins, G. (1984) *Time Series Analysis: Forecasting and Control*, 2da. Ed., Holden Day, San Francisco.
- [5] Brockwell, P. y Davis, R. (1996) *Introduction to Time Series and Forecasting*, Springer-Verlag.
- [6] Cochrane, J. (1997) *Time Series for Macroeconomics and Finance*, University of Chicago, Notes.
- [7] Cryer, J. y Chang, K. (2008) *Time Series Analysis*, Springer, Second Edition.
- [8] Dickey, Y. y Fuller, W. (1981) Likelihood Ratio Statistics for Autorregressive Time Series whith Unit Root, *Econometrica* 49 1063.
- [9] Enders, W. (1995) *Applied Econometric Time Series*, Wiley.
- [10] Farnsworth, G. (2008) *Econometrics in R*, Northwestern University.
- [11] Greene, W. (2008) *Econometric Analysis*, 6ta Ed., Prentice Hall.
- [12] Robinson, P. (2006) *Conditional-Sum-of-Squares Estimation of Models for Stationary Time Series with Long Memory*, London School of Economics.
- [13] Schwarz, G. (1978) Estimating the dimension of a model, *Annals of Statistics* 6 (2): 461–464.

- [14] Shapiro, S. y Wilk, M. (1965) An analysis of variance test for normality (complete samples), *Biometrika* 52 (3-4): 591–611
- [15] Wei, W. (2006) *Time Series Analysis, Second Edition*, Addison-Wesley.